

Software and Knowledge Engineering

Concept of Platforms as a Tool for Digital Transformation of Business Processes of the National Economy

Aigerim Bolshibayeva¹

¹ International Information Technology University, Almaty, Kazakhstan

Abstract

In the conditions of the emerging common information space, which unites a considerable number of firms, enterprises and government agencies around the world, fundamentally new conditions are created for business development. One of them is the digital economy, which is an integral part of the economy.

The digitalization of the economy is not a one-step process both in time and in space. It is a long-term and distributed process in space. Therefore, solving this problem requires both a time-consuming and a space-distributed solution - which is a digital platform.

The article discusses the concept of a digital platform given by various sources. A review was carried out and the classification of digital platforms that exist today is established. The theory of building a platform is stated: purposes, requirements, including the difference between platforms and applied systems, goals, a list of input parameters and output results, the architecture of the platform system and algorithms of its operation are defined. These data are necessary for the further development of the theory of digital platforms.

Keywords

Digitalization, digital platform, business process, platform classification, theory of platform

1. Introduction

Today the countries of the world are paying close attention to the development of the digital economy and information society. The understanding of the transition to the digital economy has developed throughout the world. So, in one of his speeches, the First Head of the Republic of Kazakhstan, Nursultan Nazarbayev, noted that digitalization is the future that can lead states to new leaders, the digital revolution is the time for decisive actions by ambitious states. In this regard, in 2017, the state program "Digital Kazakhstan" was adopted, which should become the basis for the rapid growth of technologies in the republic and reorientation to an electronic format for the provision of services, the basis for the development of which was the Address of the President of the Republic of Kazakhstan "The third modernization of Kazakhstan: global competitiveness" Dated January 31, 2017. The goal of the Program is to accelerate the development of the republic's economy and improve the quality of life of the population by digital technologies in the medium term, as well as create conditions for the transition of the economy of Kazakhstan to a fundamentally new development trajectory, ensuring the creation of the digital economy of the future in the long term [1]. The program, which is being implemented in the period from 2018 to 2022, should provide an additional impetus for the technological modernization of the country's flagship industries, as well as create conditions for large-scale and long-term growth in labor productivity.

Digital transformation ensures the fullest possible disclosure of the potential of digital technologies through their use in all aspects: business - processes, products and services, approaches to decision-making. It is important to emphasize that technology alone will never be enough for digital transformation. For the digital transformation process to be complete, clearly formulated business objectives and data are needed. Thus, digital transformation can be considered only at the intersection of three dimensions, business processes, data, infrastructure, which can be combined in the concept of a digital platform.

2. Platform as a tool for digital transformation

In world practice, T. Aisenman dealt with issues of the digital platform economy, who proposed that platforms include a single set of components (hardware, software and service modules with a given architecture) and rules (standards, protocols, policies and contracts with rights and responsibilities), used in interaction. The tools and building blocks of the platform provide ecosystem members with the ability to create powerful applications, which are then turned into benefits for end users [1].

D. Parker and S. Chaudary in their book offer the following definition of a platform: “A platform is an enterprise that provides mutually beneficial interactions between third-party manufacturers and consumers. It provides an open infrastructure for participants and sets the rules. The main task of the platform is to create connections between users and facilitate the exchange of goods or social currency, thereby contributing to the creation of value by all participants” [2].

The concept of a digital platform has evolved in recent decades in several areas of activity, which has led to many definitions of this concept. Thus, a digital platform is understood as a set of digital technologies, products or services that provide a technological basis on which external companies can create their own additional products, technologies, or services.

In turn, Intel experts define the concept of "platform" as "a complex set of components that provide implementation of intended use models, expand existing markets and create new ones, and also bring users much more benefits than the simple sum of the parts. The platform includes hardware, software, and services” [3].

According to I. Muti [4], platform technology should:

- perform one or more critical functions in a particular area;
- define some “standards” and influence the overall architecture of solutions / products;
- be open or semi-open to others to build on development opportunities through networking partnerships;
- allow both complementary companies (suppliers of complementary goods and services) and competitors to participate in the development of the platform.

Today, when the complexity of the latest technologies increases in direct proportion to the growth of their availability, more and more companies, regardless of the scale and areas of activity, are embarking on a new way of doing and developing business, based on "cloud" priorities ("Cloud First" is the main trend of the leading economies of the planet). Two key principles of the platform business - the service format of the product (everything-as-a-service) and the flexible format of payment (pay-as-you-go) - provide the unprecedented speed of new products launch to the markets and promise much richer, more positive, and productive experiences for millions of consumers. Thus, a digital platform is a completely high-tech business model that generates profit through exchange between two or more independent groups of participants. In the basic configuration, the platforms bring together manufacturers and end users directly, who get the opportunity to interact without intermediaries. They also enable different companies to share information and thus significantly improve collaboration and create innovative products and solutions [5].

The platform of the "Digital" economy is a digital environment (software and hardware complex) with a set of functions and services that meets the needs of consumers and manufacturers, as well as realizes the possibilities of direct interaction between them.

The value of the Platform lies in providing the very possibility of direct communication and facilitating the procedure for interaction between participants. Platforms reduce costs and provide additional functionality for both suppliers and consumers. They also involve the exchange of information between actors, which should significantly improve cooperation and contribute to the creation of innovative products and solutions. The "platform" as a business model has been around for a long time. A simple example is the classic market where sellers and buyers (producers and consumers) find each other. In the modern world, one can cite many actively growing companies based on the principles of the Platform Business Model, and the brightest ones are Uber and Airbnb [6].

Other approaches to defining a digital platform can be found, for example, in the monograph by Parker, Olstein or in the White Paper “Digital Platforms” prepared by the Ministry of Economic Development and Energy of the Federal Republic of Germany [7].

The interest of a general methodological nature for understanding the problem under study was the fundamental works of such domestic and foreign authors as D. Bell, J. Galbraith, and D. Tapscott, who for the first time considered the issues of the functioning of the information type of economy. Among the researchers of the role of scientific knowledge and information in economic development are F. Mahlup, M. Porat, T. Mesenburg. In the works of A. Shemet, I. Malik, A. Petrov, various approaches to the definition of the concept of "digital economy" are proposed, as well as ways of its implementation in the economic environment.

3. Classification of platforms as a method of studying the theory of platforms

The Center for Global Enterprise, based on a study of 176 platforms from different countries (The Rise of the Platform Enterprise: A Global Survey), identifies the following categories of digital platforms:

- Innovative platforms that allow platform leaders to attract a very large number of external innovators and serve as a technological foundation on which other companies develop additional products and services. Examples of such platforms are iOS from Apple Inc. and Google's Android, which have created very large innovative ecosystems of app developers for their mobile devices.
- Operating platforms that help individuals and organizations find each other, facilitating their various interactions and commercial transactions. The best examples of this type of platform are e-commerce platforms like Amazon and eBay. On-demand platforms such as Uber, Zipcar and Airbnb enable the exchange of goods and services between individuals.
- Integration platforms. It is mainly a few large companies like Apple and Google that offer both transactional and innovative platform capabilities. Both companies have created innovative platforms for their developers, which are then made available in their transactional paid forms. Likewise, Amazon and Alibaba are transactional platforms for their individual users and as innovative platforms for many vendors who also sell products on their e-commerce platforms.
- Investment platforms are holding companies that manage a portfolio of platform companies. For example, the Priceline Group is focused on online travel and related services, including Priceline, Kayak, and Open Table [8].

In turn, Deloitte University identifies four main types of platforms: aggregated platforms, social platforms, mobilization, and training platforms.

- Aggregation platforms bring together a wide range of relevant resources and help platform users connect to the most suitable resources. These platforms are typically transactional or task oriented — the essence of which is to express a need, get a response, complete a deal, and move on.

There are three subcategories in this category. First, there are platforms for collecting data or information, such as stock performance databases for investors or scientific databases. Second, there are marketplace and brokerage platforms such as eBay, Etsy, and the App Store, which have generated 85 billion app downloads as of October 2014. They provide an environment for suppliers to interact more effectively with their respective customers, wherever they are. In an increasing number of cases, these platforms are attracting resources that were previously not available to others. For example, Airbnb has created a platform that has grown more than tenfold, from 50,000 to 550,000 listings in less than four years, encouraging people to provide vacant rooms or portions of their homes to travelers and thus creating a market for these resources. And third, there are competition platforms like InnoCentive or Kaggle, where someone can post a problem or challenge and offer a reward or payment to the competitor who comes up with the best solution.

- Social platforms are like aggregation platforms in the sense that they bring many people together - think of all the broad-based social platforms we've come to know and love leading examples are Facebook and Twitter. They differ from aggregation platforms in some keyways. First, they end up building and strengthening long-term relationships between participants on the platform - it's not just about completing a transaction or task, but also getting to know people in areas of common interest. Many of these platforms are insurmountable, with US adults spending an average of 42.1 minutes per day on Facebook and 17.1 minutes on Twitter. Second, they tend to facilitate the creation of meshed networks of relationships, not interactions among themselves: people who connect

to enough platforms have been specially designed to be categorized into types. The three common types that exist today help their participants perform well on three different tasks. Over time, business ecosystems mature in one another in more varied ways that usually do not affect the organizer or owner of the platform.

- Mobilization platforms bring common interests to the level of action. These platforms aren't just conversations and interests; they focus on getting people to work together to achieve something that is beyond the capabilities of any individual participant. Due to the need for synergy over time, these platforms tend to foster longer-term relationships rather than focus on isolated and short-term transactions or tasks. But the key direction here is to connect and mobilize a specific set of people and resources to achieve a common goal. Participants are often viewed as “static resources” - they have a specific set of individual capabilities, and the challenge is to mobilize those fixed capabilities to achieve a long-term goal. There are many different forms of mobilization platforms. In a business context, the most common form of these platforms is “process networks” platforms that connect participants in extended business processes, such as supply chains or distribution operations, that help select and organize participants who need to collaborate flexibly over time. Li & Fung, a global sourcing company, offers a prime example of this kind of platform, although there are many other examples spanning a wide range of industries including motorcycles, financial services, diesel engines, and consumer electronics.

- Learning platforms. In a world of rising productivity pressures, we should also expect a fourth platform form to emerge. A dynamic and demanding environment favors those who can learn better and faster. Business leaders who understand this are likely to increasingly look for platforms that not only make things easier for their members, but also expand their knowledge, accelerate productivity gains, and hone their capabilities in the process. There are very few examples of learning platforms in business so far, but we can find very large-scale learning platforms in arenas as diverse as online war games (like World of Warcraft) and online platforms. Classification of platforms demonstrated on table1.

Table 1
 Classification of digital platforms

| | Instrumental digital platform | Infrastructure digital platform | Applied digital platform |
|---|---|--|---|
| The main activity based on the platform | Development of software and software hardware solutions | Providing IT services and information for decision making | Exchange of specific economic values in specified markets |
| The result of activities on the platform | A product (software or software and hardware) for processing information as a tool | IT service and the result of its work - the information necessary for making decisions in business activities | Transaction. A transaction that fixes the exchange of goods / services between participants in a given market |
| Information processing level | Technological operations of information processing | Generation of information for decision-making at the level of an economic entity | Processing information about the conclusion and execution of a transaction between several economic entities |
| The main beneficiary and its requirements | Developer of applied software or hardware- software solutions, technical requirements | Customer of IT service for the consumer (product engineer), functional requirements, requirements for the composition of information | The end consumer in the market solving a business problem, business requirements. Regulator (optional) - legal requirements |
| Example | Java, SAP HANA, Android OS, iOS, Intel x86, Bitrix, Amazon Web Services, Microsoft Azure, TensorFlow, Cloud Foundry | General Electric Predix, ESRI ArcGIS, ЕСИА, «CoBrain-Аналитика», «ЭРА-ГЛОУАСС» | Uber, AirBnB, Aliexpress, Booking.com, Avito, Boeing suppliers portal, Apple AppStore, «ПЛАТОН», AviaSales, FaceBook, Alibaba, Telegram, Yandex Taxi, Yandex Search, Facebook |

Digital platforms have a few advantages compared to traditional business conduct, but the issue of personal data privacy, information security, etc. is also acute in the development of platforms, in order

to demonstrate the strengths and weaknesses of the platform, a SWOT analysis of the use of digital platforms was carried out on Figure 1.

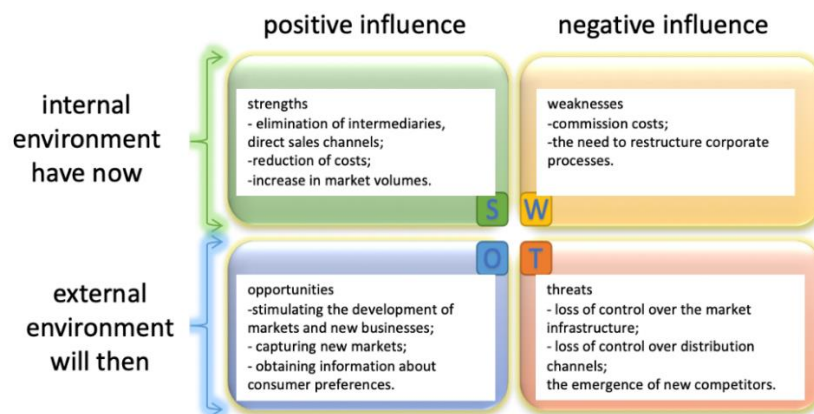


Figure 1: SWOT analysis of the use of digital platforms

A review of existing business process platforms shows that these platforms, although called platforms, are still a complete system and are used as a system.

A common disadvantage of all these platforms is that they are not intended to generate specific practical systems based on specific objects of labor and means of labor. Are not intended to generate new systems or the ability to generate new systems for new processes is problem domain low, i.e., limited, requires a lot of work to adapt. In this work, it is assumed that they are all systems.

Thus, since the work of the platform in all cases is based on the initial models of business processes, to build a platform that eliminates the shortcomings of existing platforms, it is necessary to find an adequate model. Further, based on this adequate model, build a platform.

The article proposes building a platform based on a new concept or model of a business process. This model should ensure the generation of a business process model, and then, based on this business process model, the generation of a business process automation system.

According to this model, a business process consists of a set of specialized processes and the relationship of business operations both among themselves within one specialized process, and between various specialized processes as part of one business process in different production situations is non-linear.

Currently, the following is relevant for the economy:

- conducting research and development of theories and methodology of platforms (and / or theoretical and methodological foundations, bases), based on a model that more adequately and realistically reflects the structure, architecture, methods, and procedures of real business processes.
- based on the new theory, create a platform that allows the effective generation of automation systems by business processes of a given local problem area of high functionality, i.e., conduct efficient and high-quality solutions to the automation problem.
- to increase the efficiency of interaction of the generated system:
 - with objects of the automated business process.
 - with objects of the space surrounding the business process.
 - harmonize with the rest of the components and elements of the infrastructure of the automated business process and form a single business process infrastructure.

The solution to the problems of digital transformation of an object is reduced to its or their automation. Therefore, the digital transformation of the business processes of the national economy is reduced to their automation, considering their interconnection.

4. The essence and features of the functioning of platforms for the tasks of digital transformation of business processes

Digital transformation can be seen as the third stage of digital embracing: from digital competence to the use of digital technologies. The transformation stage means that digital use inherently allows new types of innovation and creativity in a specific field, rather than simply reinforcing and supporting traditional methods.

For further research, let us turn to V. Mesropyan's definition: "These are revolutionary changes in business models based on the use of digital platforms, which lead to a radical increase in market volumes and the competitiveness of companies" [11].

Business process management is a complex procedure for analyzing business processes and automating them using workflow automation software.

Business Process Management (BPM) is an area where digital transformation can play a key role. However, in many organizations there are many misunderstandings about how to achieve digital transformation in business process management [12].

Within the framework of this study, theoretical developments determine the goals and criteria of platforms for business processes of the national economy. A business process model is established, which was given the name hierarchical semantic multidimensional non-linear model (or concept), on the basis of which there is a further superstructure of the business process and subsequently the platform. At the same time, the main requirements are always to accelerate and improve the quality of the process of creating these power supply automation systems.

Moreover, it is clear that it is impossible to build a universal platform suitable for creating control automation systems for all arbitrary business processes. Therefore, in this work, it is proposed to build such a problem-oriented platform that allows you to create automation systems for business processes from a specific problem area.

It should be noted that,

- firstly, the platforms themselves can be developed (created) at different levels of generality or universality;
- secondly, platforms can create (allow to create) automation systems at different levels of development (ie maturity and completeness) for the same business process.

Examples of platforms demonstrating different levels of versatility (or completeness: a list of functions and completeness of each function, i.e. functionality) can be the following

- more generalized and universal are platforms such as "operating systems", "MS Word" or "DBMS". These platforms allow you to build systems. But in the process of creating systems, their share of participation is limited, but the share of developer's labor is very significant. Or in the process of creating systems, these platforms allow and require the involvement of other specialized platforms. In this case, the platform consists of a collection of general, generic classes;
- platforms such as "Framework", "Pattern", "Microsoft, .Net Framework", etc. are more specialized, modified to some depth of the logic of the process of solving individual problems of a certain class of problems. In this case, the platform consists of a set of procedures and services, some of which are complete and ready to use;
- brought to the system implementation, ie, ready-to-use systems. Such systems can be used as platforms.

5. Building a new concept of platform theory based on the research

The developed platform for creating systems for automating business processes, i.e., BPMS-system consists of three levels: core, environment, and applications (see fig.2).

The core of the platform consists of the logic of the business process execution and its automation, which are an invariable part of the business processes of a certain or selected class of business processes. The logic or scenario for the execution of a business process and its automation is formally represented by a metamodel. The core of the platform is the first level of the transformed conceptual model.

The core of the platform consists of a set of reference or reference data and knowledge about business processes about the main characteristics, reliability, security, cost, as well as tools, models, and data metamodels for formalizing and building execution logic:

- business processes if the platform's mission is to build business processes.
- business process automation systems if the platform's mission is to automate business processes.

The core performs the following functions: surveys (pre-design) of the problem area, modeling the business process of the problem area as is "as-is", design based on the model (pattern) of the business process as it should be "to-be", design and modeling by decomposition a holistic problem into parts and the design of these parts, programming (individual modules), the creation or configuration of an automation system based on programs and their integration.

The generated systems for automating business processes can be both local and distributed, depending on the specifics of the business process.

Patterns, i.e., templates reflect the permanent parts of the system, inherent in the class of systems for the generation of which the platform is intended. The system consists of a changeable part (operations) that change or change from one instance to another instance of a given business process class

The platform environment consists of a changeable part (operations) from one instance to another instance of a given business process class. The content of the platform environment consists of services for automating business process operations, which are programmed based on WSDL / WS CDL technologies.

It is the second level of the model.

In the environment, a library is created from services, with the help of which the permanent part of the system is brought to a "combat" or practical option for automating a business process in a design and programmatic way.

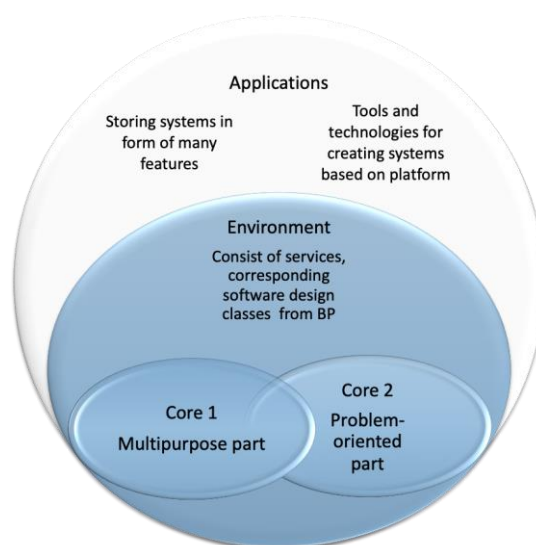


Figure 2: Content of the platform for automating business processes

The platform application consists of an archive of completed business processes of business process automation systems, i.e., BPMS systems.

With this composition of the platform, the procedure for servicing applications for the creation of automation systems, i.e., BPMS-systems depend on the completeness and "comprehensibility" of the content of the platform application.

The relative position of the components from the center of the platform is shown in Figure 2.

As you can see, the core of the platform plays an important role, since it is an invariable part of the platform, therefore, it is the optimal construction of business processes and the method of building an effective business process of an enterprise that is considered.

A business process includes many objects or items, many special processes, items and means of labor, and includes methodologies and technologies and those responsible for executing the business process. Thus, the business process has a complex structure and composition, i.e., architecture and complex components of this architecture.

Therefore, the presence of a model allows you to streamline and speed up the process of building both the components of the business process, and the whole business process itself, the creation of which is planned. The resulting model of a complex business process will allow:

- establish and disclose the composition, structure, and architecture of a complex business process of the selected class,
- build optimal models of the business process.
- automate a complex business process.
- to operate and manage a complex business process.

First, to determine an adequate model, it is necessary to note the main features and composition of business processes: the business process must ensure the achievement of a certain goal and the quality criterion of the output products of the business process. A business process includes many objects or objects, many processes, which we will call special processes, as well as objects and means of labor. In addition, the business process includes those responsible and executors for the implementation of the business process.

The analysis shows that a business process model is needed to build both a business process methodology and an automation system. In addition, it serves as the basis for all stages of the life cycle of a business process and automation system, i.e., the model should support project processes: from the pre-project stage to write-off (or inheritance) of both the business process and the automation system.

With such a concept, we can put the following statement of the problem:

Let there be "problem areas": $IP = \{IP_j\}$, $j=1,m$, where IP are sets of problem areas, which we will call classes; IP_j - problem area with identifier "j" or j-th class.

Let subclasses or instances $E_j = \{E_{ih}\}$, $h = 1,m_j$ of each problem domain IP_j be given.

If an instance E_{ih} of the problem area IP_j has an automation system AC_{ih} , then we will call this instance the base or main instance of the problem area IP_j .

Let m_j instances $E_j = \{E_{ih}\}$, $h = 1,m_i$ of the same "problem domain" class IP_j be given. Among them, an instance of E_{ih} , is the base or main, since for which the AC_{ih} automation system was created.

Task requirements: it is required to create an AC_{ik} for an instance E_{ik} of the problem domain IP_j .

Now, knowing the changeable parts of the platform we have proposed, we can apply various options for setting this task, as well as build universal platforms.

6. Conclusion

This article analyzes the development trends of the information services market, shows the relevance of the topic of dissertation research. The review of existing platforms is carried out, the classification of digital platforms is given.

Digital platforms have a number of advantages over traditional business conduct, but a review of existing platforms for business processes shows that these platforms, although called platforms, are still a complete system and are used as a system.

The common disadvantage of all these platforms is that they are not intended for generating specific practical systems based on specific objects of labor and means of labor, and also do not have the ability to generate new systems for new processes by the problem of the area, i.e. limited, requires a lot of work to adapt.

The platform can be designed to be overly versatile and therefore capable of creating systems for most business processes, but these systems will be too abstract and general. Therefore, in order to bring the systems created by the platform to perfection, i.e. to bring them to the required level requires significant development, in other cases, refinement.

If the development of automation of business process management processes has been brought to the system level (to system implementation), then on its basis it is not always possible to develop another, new system that provides automation of business process management processes. In contrast, platforms must provide the creation of systems for a wide class of business processes.

The versatility of the platform can be at different levels. Excessive versatility of platforms can

create automation systems for many business processes, but these systems may turn out to be functionally unsuitable in practice due to the excessive generalization and approximation of their functions, which are necessary for solving problems in practice.

The adjusted universality of the platform arises from taking into account the specific or individual properties of individual business processes, which are not common, inherent in all business processes of a given class. In turn, specializations can be at different levels.

Therefore, it is proposed to build a platform based on a new concept or model of a business process. This model should ensure the generation of a business process model, and then, on the basis of this business process model, the generation of a business process automation system so that the general requirements are met:

1. Creation of an automation system for a wide range of business processes (KS → MAX); The main purpose of the platform is that automation systems for a wide class of business processes should be created on its basis.

2. The list of implemented functions for each system must be wide enough to carry out missions (KF → MAX) e. a wide class of business processes (objects) and full functionality for each case of creating a system.

3. The level of completion of each function must be sufficient to fulfil the mission (ZF → MAX).

7. References

- [1] Gosudarstvennaya programma «Tsifrovoy KazakhstaN» ot 12 dekabrya 2017 goda.
- [2] Geoffrey G. Parker, Marshall W. Van Alstyne, Sangeet Paul Choudary, 2016. Revolyutsiya platform. Kak setevye rynki menyayut ehkonomiku – i kak zastavit' ikh rabotat' na vas //OOO «Mann, Ivanov i FerbeR», 2017.
- [3] Evans P.C., Gawer A. The Rise of the Platform Enterprise: A Global Survey. The Center for Global Enterprise, 2016.
- [4] Mootee, I. What's the difference between platform strategy vs. business strategy vs. product strategy? – Mode of access: <https://www.idr.is/do-you-know-the>.
- [5] Selin, A. Tsifrovye modeli biznesa: magistral'nyi trend sovremennogo rynka // Daidzhest novostei mira vysokikh tekhnologii – №5 – 2016. – 14 s.
- [6] Vvedenie v «TsifrovuYU» ehkonomiku/ A.V. Keshelava V.G. Budanov, V.YU. Rummyantsev i dr.; pod obshch. red. A.V. Keshelava; gl. «tsifr.» kons. I.A. Zimnenko. – VNIIGeosistem, 2017. – 28 s. (Na poroge «tsifrovogo budushchego»). Kniga pervaya), str.13.
- [7] Digital Platforms: Digital regulatory policy for growth, innovation, competition and participation. Berlin: Federal Ministry for Economic Affairs and Energy, 2017.
- [8] Kuprevich T.S. Tsifrovye platformy v mirovoi ehkonomike: sovremennye tendentsii i napravleniya razvitiya // Ehkonomicheskii vestnik universiteta. Sbornik nauchnykh trudov uchenykh i aspirantov. 2018. №37-1. URL: <https://cyberleninka.ru/article/n/tsifrovye-platformy-v-mirovoy-ekonomike-sovremennye-tendentsii-i-napravleniya-razvitiya>.
- [9] The power of platforms //Deloitte University https://www2.deloitte.com/content/dam/Deloitte/za/Documents/strategy/za_The_power_of_platforms.pdf.
- [10] Kuprevich T.S. Tsifrovye platformy v mirovoi ehkonomike: sovremennye tendentsii i napravleniya razvitiya // Ehkonomicheskii vestnik universiteta. Sbornik nauchnykh trudov uchenykh i aspirantov. 2018. №37-1. URL: <https://cyberleninka.ru/article/n/tsifrovye-platformy-v-mirovoy-ekonomike-sovremennye-tendentsii-i-napravleniya-razvitiya>.
- [11] Mesropyan V. Tsifrovye platformy – novaya rynochnaya vlast'. Moskva, 2018. URL: <https://www.econ.msu.ru/sys/raw.php?o=46781&p=attachment>.
- [12] Simonov N. Kazhdaya chetvertaya kompaniya poterpela neudachu v tsifrovoĭ transformatsii // Direktor informatsionnoĭ sluzhby. – 2017. – No9. – S. 6.

Application of Landsat-8/9 and Sentinel-2A/B Remote Sensing Data for Detecting Fire Zone in Kostanay Region, Kazakhstan, on September 3-4, 2022

Asset Akhmediya¹, Khuralay Moldamurat², Kazbek S. Baktybekov¹,
Guldana Kassymbayeva³ and Garyshbek Yechshanov²

¹ Kazakh Agro Technical University, Astana, Kazakhstan

² L.N. Gumilyov Eurasian National University, Astana, Kazakhstan

³ JSC NC Kazakhstan Garysh Sapary, Astana, Kazakhstan

Abstract

In this article, a study is conducted using available near- and infrared remote sensing data to detect the fire zone and its area. A forest fire in the Auliekol district, Kostanay region, Republic of Kazakhstan, was chosen as the event. The images were taken on September 3 and 4, 2022, the first by the Sentinel-2B remote sensing satellite and the second by Landsat-8. Supervised classification methods were applied to selected pixel samples belonging to fire and other classes. The Minimum distance, Mahalanobis distance and Support Vector Machine methods showed good repeatability, and the error in estimating the fire area was 1.5% in the case of Sentinel-2B for September 3, 2022.

Keywords

Fire zone, supervised classification, remote sensing, infrared band

1. Introduction

Fires cause massive damage to the ecosystem of forests, which are vital for producing oxygen. In addition, fires cause financial and material damage, leading to loss of housing and human casualties. If it is impossible to prevent the occurrence of fire face, the main task becomes early warning and prevention of its further spread. Local authorities and emergency services should promptly respond and assess the fire's damage and area. The necessary information about the fire area and its coordinates are received late, and there are not enough human resources to monitor a large area. Therefore, an alternative method may be space monitoring with the more frequent shooting (3-4 times a day) of one territory.

In Kazakhstan, fires occur mainly in summer and autumn, which affect the steppes and forests; they also cause substantial economic damage to agriculture. The territory of Kazakhstan has an area of more than 2 million 700 thousand square kilometers. Forests or forested areas occupy about 5% of this area. According to the state program «Zhasyl El 2022», an increase in the share and preservation of the existing forest fund has been announced. Recently, in September 2022, 43 thousand hectares of forest were affected in Kostanay region due to fires, this is the last significant emergency that occurred in Kazakhstan. Fire damage assessment is carried out, including with using remote sensing data.

Remote sensing satellites is able to detect fire zone; they can take images in different optical ranges, from ultraviolet to far infrared. High spatial resolution images are available for commercial use, and middle spatial resolution images are often freely available. Sentinel-2A/B and Landsat-8/9 remote sensing data are available to users with an average spatial resolution of 10 meters or more. Furthermore, the rest of the high-resolution less than 10 meters are purchased, the cost of which is higher, the higher the spatial resolution of the images obtained.

Fires are clearly visible from space by such signs as the presence of thick smoke trailing from the source of occurrence, the black colour of the burning where the fire took place, and an elevated temperature at the fire site. Visual interpretation using the visible spectrum is not enough; infrared spectra can significantly help detect fire foci. Small areas of ignition are visible with high-resolution images. However, cloud cover over the observation area substantially affects the quality of images,

where the ideal option is clear weather rarely happens. Therefore, the first step is a search for suitable images with minimal cloud cover, and then processing is carried out to identify the fire zone. These circumstances limit the use of satellite images in the operational mode.

Image processing of remote sensing data to determine the zone of fire damage is usually performed in specialized software or packages. Packages such as Erdas Imagine, ENVI and PCI Geomatics can be used to work with remote sensing data. They are licensed, the cost of some can reach up to \$ 10,000 with many modules, the user interface can be convenient. In addition, there are free programs (open-source), such as ESA SNAP developed by the European Space Agency and supported by the developer community. Methods for determining the fire zone were developed earlier, where remote sensing data from Landsat-8, Sentinel-2 and MODIS were used together with other high spatial resolution data images [1-6].

2. Study area and data

On September 2, 2022, a fire broke out in the Auliekolsky district of the Kostanay region near the village of Amankaragai on the northern side of the forest. The fire was extinguished entirely on September 10, when 85 houses had already burned. According to the media and emergency services, it was one of the most robust fires in the Kostanay region. Military and firefighters from all areas of Kazakhstan participated in extinguishing the fire. Google Earth data shows a dense forest in this research area, consisting of pine, poplar, birch and other trees. A steppe surrounds this forest area. The village of Amankaragai is located on the southern side of this forest area and is adjacent to it (Fig. 1).

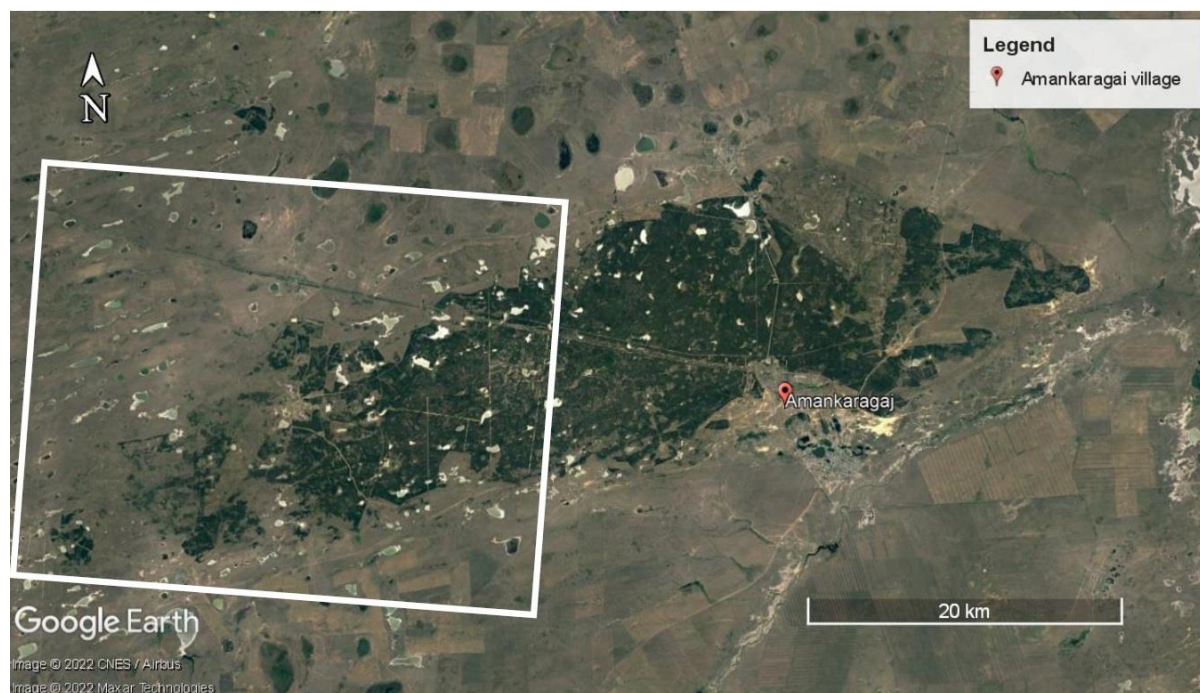


Figure 1: Study area from Google Earth, white rectangle outline – region of interest (ROI)

Landsat-8/9 and Sentinel-2B remote sensing satellite data for the study were obtained from the resources earthexplorer.usgs.gov, scihub.copernicus.eu. The list of satellite data is presented in Table 1. The scenes of satellite images that relate to the fire area completely or partially cover it (Fig.2). The remote sensing data were selected based on the minimum cloud cover of the study area with dates September 3 and 4, 2022.

Table 1
 Satellite data

| No | Satellite | Date | ID satellite data |
|----|-------------|------------|---|
| 1. | Sentinel-2B | 03.09.2022 | S2B_MSIL1C_20220903T065629_N0400_R063_T41 UNU_20220903T085606.SAFE |
| 2. | Landsat-8 | 04.09.2022 | LC08_L1TP_161023_20220904_20220913_02_T1 |

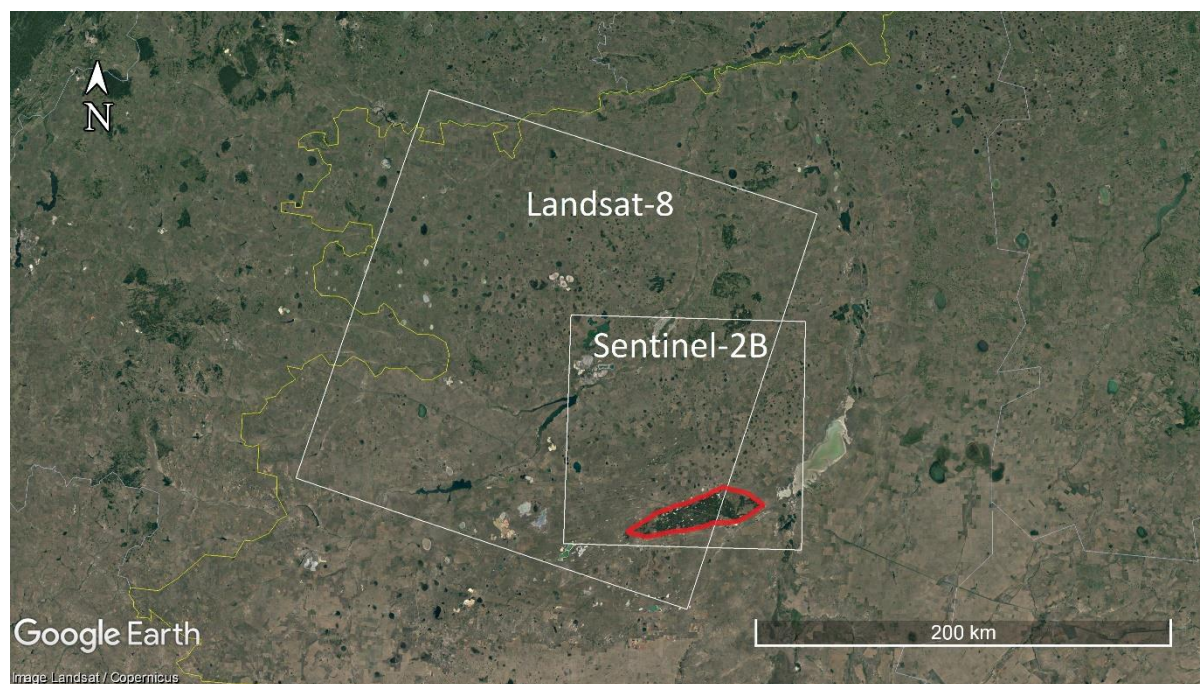


Figure 2: Scenes of satellite images of Landsat-8, Sentinel-2B (white square frames) and the study area (highlighted with a red line)

3. Methodology

The fire is a source of thermal radiation, it is detected through the infrared sensors, from which are formed satellite images at a wavelength from 0.85 to 2,29 μm . Therefore, the first process will be the extraction of the thermal component from the Landsat-8 and Sentinel-2B data. The number and spectrum of thermal components for Landsat-8 and Sentinel-2B are presented in Table 2.

Table 2
 Near and shortwave infrared bands, wavelength and pixel resolution

| Spectrum | Landsat-8 | Sentinel-2B |
|---------------------------|---|--------------------------------------|
| Near infrared (NIR) | Band 5 (0.85 – 0.88 μm), 30 m | Band 8a (0.865 μm), 20 m |
| Shortwave infrared (SWIR) | Band 6 (1.57 – 1.65 μm), 30 m | Band 9 (0.94 μm), 60 m |
| | Band 7 (2.11 – 2.29 μm), 30 m | Band 10 (1.375 μm), 60 m |
| | | Band 11 (1.61 μm), 20 m |
| | | Band 12 (2.19 μm), 20 m |

Let's visually assess the fire zone in true and false colors in RGB composite. Bands are chosen for the true color RGB composite:

1. Landsat-8: Red – Band 4; Green – Band 3; Blue – Band 2 (Figure 3);
2. Sentinel-2B: Red – Band 4; Green – Band 3; Blue – Band 2 (Figure 3).

Bands are chosen for the false color RGB composite:

1. Landsat-8: Red – Band 7; Green – Band 6; Blue – Band 5 (Figure 4);
2. Sentinel-2B: Red – Band 12; Green – Band 11; Blue – Band 8a (Figure 4).

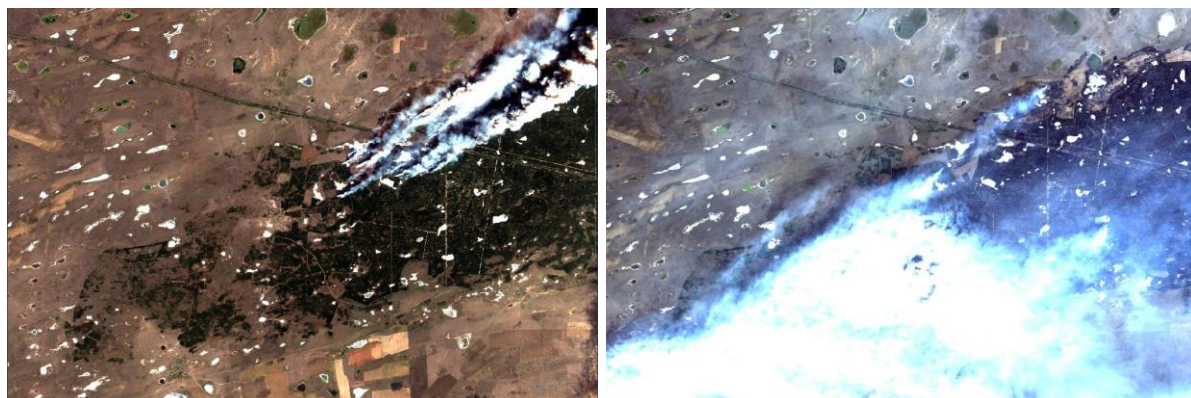


Figure 3: True color RGB composite for Sentinel-2B on September 3, 2022 (Left) and Landsat-8 on September 4, 2022 (Right)

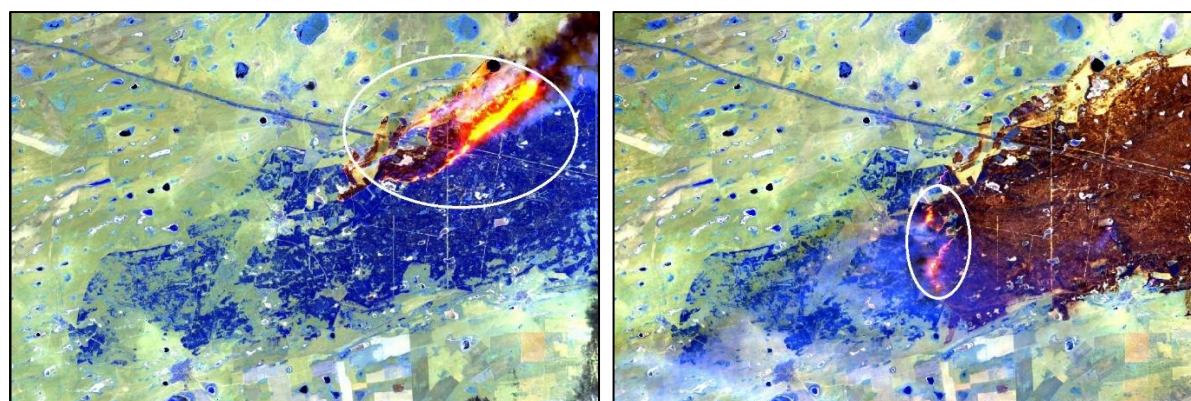


Figure 4: False color RGB composite for Sentinel-2B on September 3, 2022 (Left) and Landsat-8 on September 4, 2022 (Right). White oval contour shows the existence of a fire.

After a visual interpretation of the false color RGB composite, the supervised classification methods are conducted, and fire zone pixel samples are taken for it (Table 3). Several supervised classification methods are applied to identify the best one [7, 8]. A list of supervised classification methods is shown below:

1. Parallelipiped;
2. Minimum distance;
3. Mahalanobis distance;
4. Maximum Likelihood;
5. Spectral angle mapper;
6. Artificial Neural Network;
7. Support vector machine.

After choosing the best-supervised classification method, the number of pixels related to the fire is calculated, and thus the fire area is estimated at that time. The formula for calculating the fire area S_{fire} (in hectares):

$$S_{fire} = \frac{N_{fire} \times r^2}{10000} \quad (1)$$

where, N_{fire} – number of pixels, which belongs to fire zone; r – pixel resolution, meter.

Table 3

Training data, number of samples

| Class | Landsat-8 false color image | Sentinel-2B false color image |
|---|-----------------------------|-------------------------------|
| Fire zone | 364 pixels | 419 pixels |
| Others (water body, steppe, forest, agriculture field and etc.) | 8040 pixels | 21599 pixels |

The Landsat-8/9 and Sentinel-2A/B remote sensing data processing process is implemented in the ENVI software and has the following block scheme (Figure 5). The performance of the ESA SNAP software has not been tested. However, using the same processing steps, we can get the same supervised classification results if we use the same training samples and parameters. All processing processes can be combined into one processing that can be started from a single script or program module. ESA SNAP Graph Builder module can automate some processes and increase efficiency. It will be checked in future research.

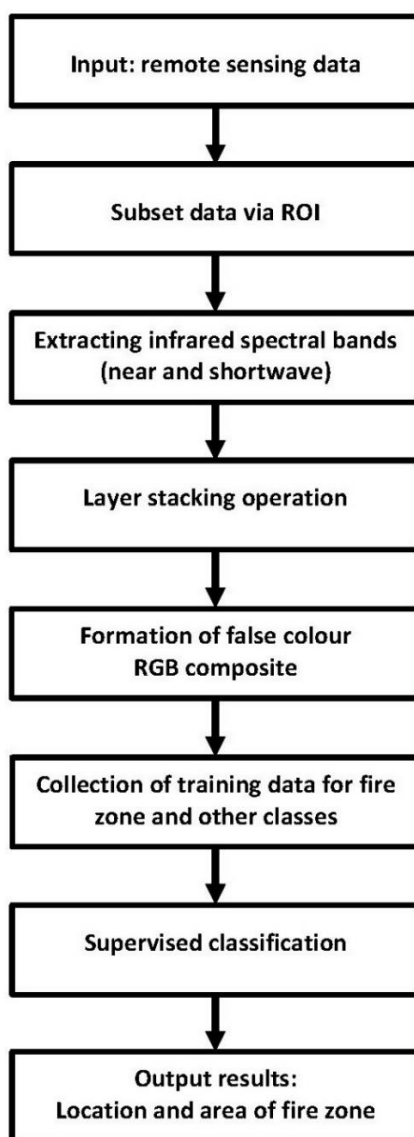


Figure 5: Flowchart of remote sensing data processing for detecting location fire zone and area

4. Results

The result of the fire zone detection consists of supervised classification methods using the infrared data of the Sentinel-2B and Landsat-8 remote sensing satellites. They are presented as supervised classification images (Figure 6-12). Here, the "fire zone" class is assigned a red color, and the remaining classes are given a white color in the image.

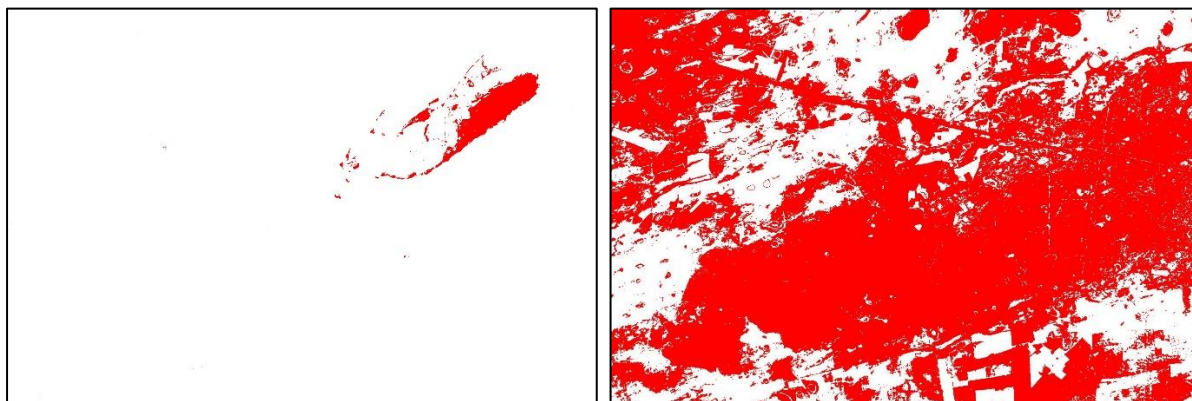


Figure 6: Classification image using the Parallelepiped method for Sentinel-2B on September 3, 2022 (Left) and Landsat-8 on September 4, 2022 (Right)

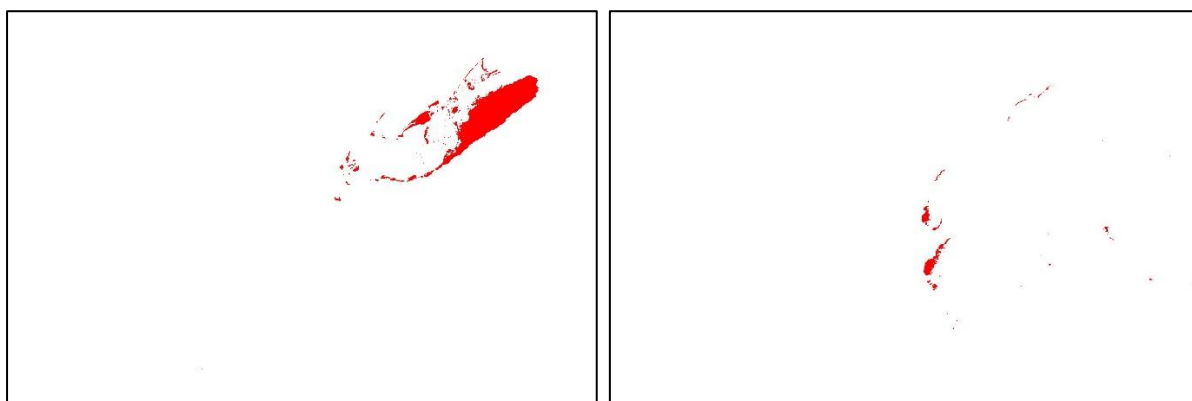


Figure 7: Classification image using the Minimum distance method for Sentinel-2B on September 3, 2022 (Left) and Landsat-8 on September 4, 2022 (Right)

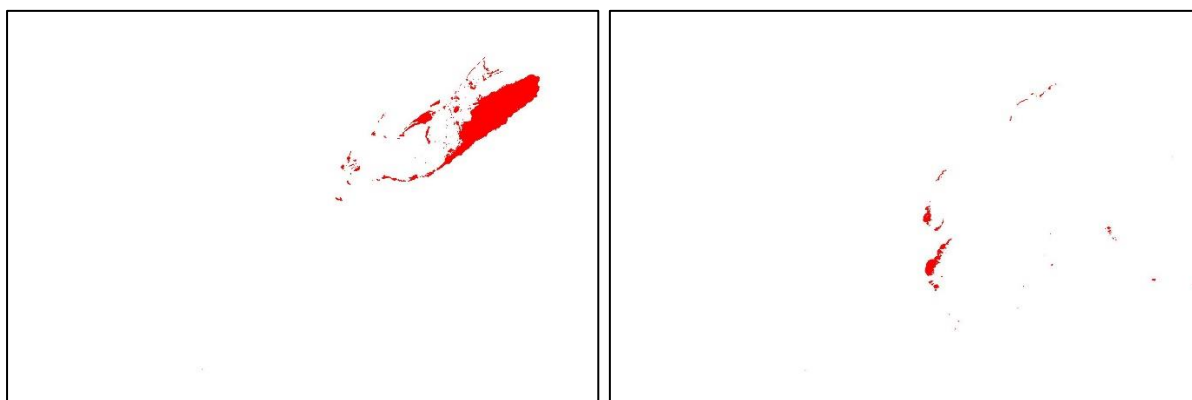


Figure 8: Classification image using the Mahalanobis distance method for Sentinel-2B on September 3, 2022 (Left) and Landsat-8 on September 4, 2022 (Right)

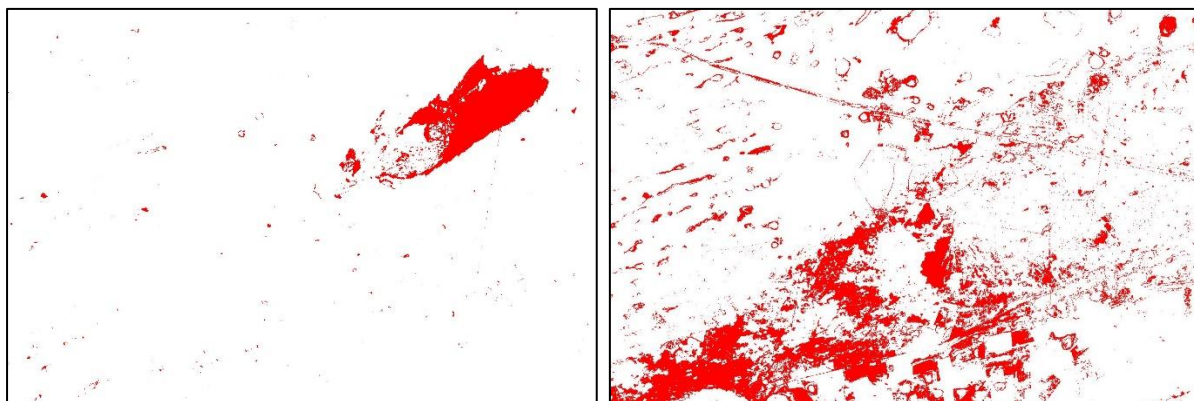


Figure 9: Classification image using the Maximum Likelihood method for Sentinel-2B on September 3, 2022 (Left) and Landsat-8 on September 4, 2022 (Right)

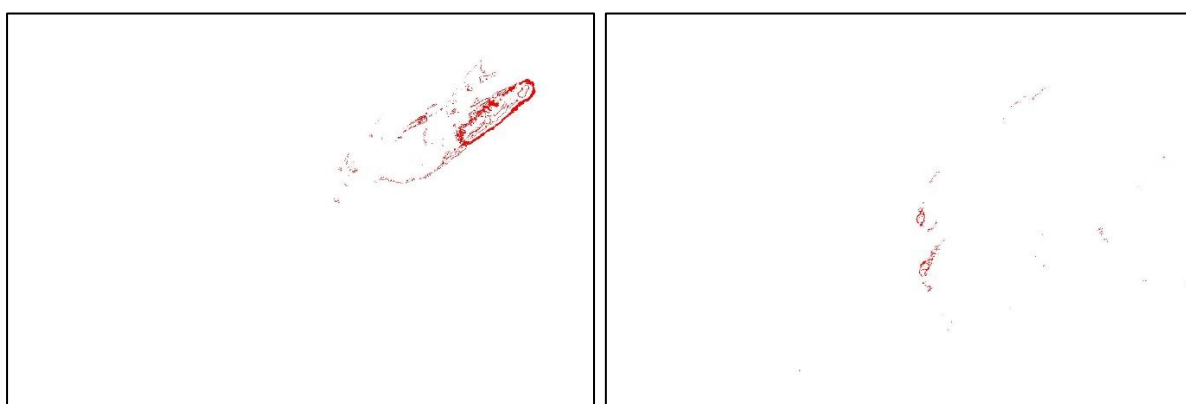


Figure 10: Classification image using the Spectral Angle Mapper method for Sentinel-2B on September 3, 2022 (Left) and Landsat-8 on September 4, 2022 (Right)

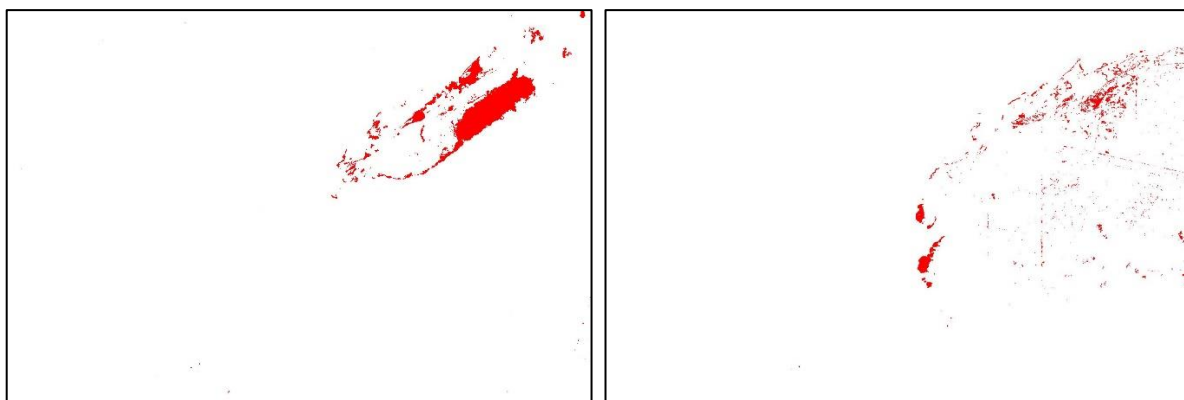


Figure 11: Classification image using the Artificial Neural Network method for Sentinel-2B on September 3, 2022 (Left) and Landsat-8 on September 4, 2022 (Right)

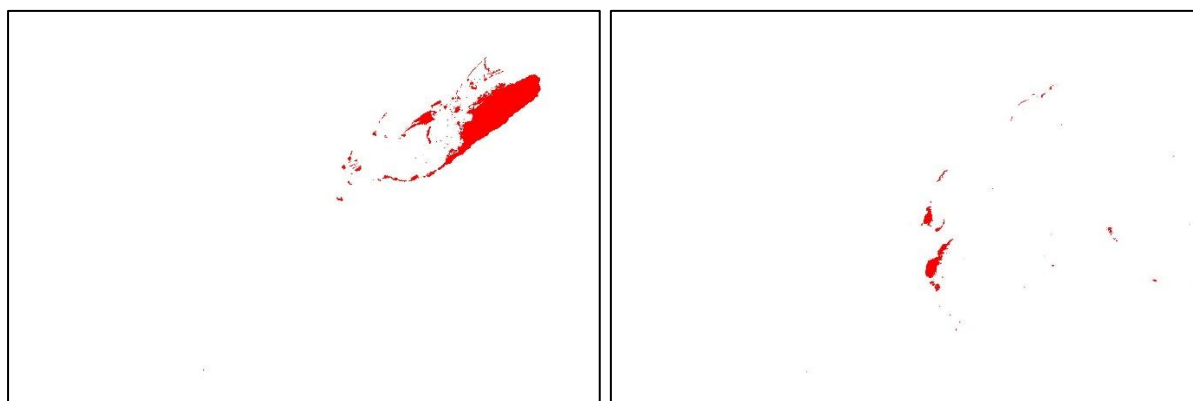


Figure 12: Classification image using the Support Vector Machine method for Sentinel-2B on September 3, 2022 (Left) and Landsat-8 on September 4, 2022 (Right)

The fire zone area for all supervised classification methods were calculated according to formula (1) and presented in Table 4. Here, in the case of three supervised classification methods (Minimum Distance, Mahalanobis Distance and Support Vector Machine), the results are close to each other with small deviations in area. The overall accuracy in these methods starts from 0.9962, and the kappa coefficient is starting 0.9488. This accuracy was achieved by merging fire zone classes together.

Table 4

Calculated fire zone area S_{fire}

| Supervised classification method | For Landsat-8 false color image, ha | For Sentinel-2B false color image, ha |
|----------------------------------|-------------------------------------|---------------------------------------|
| Parallelepiped | 81026,37 | 1837,28 |
| Minimum distance | 223,47 | 1846,48 |
| Mahalanobis distance | 243,72 | 1845,84 |
| Maximum Likelihood | 18312,48 | 4311,8 |
| Spectral Angle Mapper | 94,95 | 572,84 |
| Artificial Neural Network | 1077,84 | 2081,28 |
| Support Vector Machine | 294,21 | 1875,2 |

5. Conclusion

Remote sensing satellites must monitor one area more often, several times a day, to appear the occurrence of a fire. The cloud cover over the study area must be minimal to detect the fire during space monitoring. Thermal radiation can pass through thin clouds, so what is not visible in the visible range under clouds becomes visible in the infrared bands, including the fire zone. Thus, infrared bands such as near and shortwave infrared are necessary for finding the fire zone. It can be seen from the classification images that the Minimum Distance, Mahalanobis Distance and Support Vector Machine methods give the best and most repeatable results with minor deviations. As of September 3, with Sentinel-2B images, according to the above methods, the variation in the area of the fire zone was $(1875.2 - 1845.84) = 29.36$ ha, about 1.5%. As of September 4, with Landsat-8 images, the deviation was 31.6%. The last significant variations compared to the first are explained possibly with a small number of samples associated with a small area of the fire zone on September 4, 2022. It is generally possible to determine the fire zone during its active phase with satellite images Landsat-8/9 and Sentinel-2A/B in the infrared bands. However, they are not applicable after a fire in an already burned-out area. It is necessary to have at least two satellite images pre- and post-fire. Then, the burned-out site is founded by calculating the difference in the Normalized Burn Ratios [9, 10].

6. Acknowledgements

The authors would like to thank European Space Agency and NASA for free access to Sentinel-2A/B, Landsat-8/9 remote sensing data. Special thanks, warmth and appreciation to the following individuals for help and assistance: my co-supervisor, Dr. Sabyrzhan Atanov from Eurasian National University, for his vital support and assistance; to Dr. Sarnam Singh from Indian Institute of Remote Sensing, for practical knowledge about image processing.

7. References

- [1] A. D. Pacheco, J. A. D. Junior, A. M. Ruiz-Armenteros, and R. F. F. Henriques, "Assessment of k-Nearest Neighbor and Random Forest Classifiers for Mapping Forest Fire Areas in Central Portugal Using Landsat-8, Sentinel-2, and Terra Imagery," *Remote Sensing*, vol. 13, no. 7, Apr 2021, Art no. 1345, doi: 10.3390/rs13071345.
- [2] C. V. Angelino, L. Cicala, S. Parrilli, N. Fiscante, S. L. Ullo, and Ieee, "POST-FIRE ASSESSMENT OF BURNED AREAS WITH LANDSAT-8 AND SENTINEL-2 IMAGERY TOGETHER WITH MODIS AND VIIRS ACTIVE FIRE PRODUCTS," *Igarss 2020 - 2020 Ieee International Geoscience and Remote Sensing Symposium*, pp. 6770-6773, 2020, doi: 10.1109/igarss39084.2020.9324512.
- [3] P. Konkathi and A. Shetty, "Inter comparison of post-fire burn severity indices of Landsat-8 and Sentinel-2 imagery using Google Earth Engine," *Earth Science Informatics*, vol. 14, no. 2, pp. 645-653, Jun 2021, doi: 10.1007/s12145-020-00566-2.
- [4] J. Delegido *et al.*, "Fire severity estimation in southern of the Buenos Aires province, Argentina, using Sentinel-2 and its comparison with Landsat-8," *Revista De Teledeteccion*, no. 51, pp. 47-60, Jun 2018, doi: 10.4995/raet.2018.8934.
- [5] P. Garcia-Llamas *et al.*, "Evaluation and comparison of Landsat 8, Sentinel-2 and Deimos-1 remote sensing indices for assessing burn severity in Mediterranean fire-prone ecosystems," *International Journal of Applied Earth Observation and Geoinformation*, vol. 80, pp. 137-144, Aug 2019, doi: 10.1016/j.jag.2019.04.006.
- [6] S. Bar, B. R. Parida, and A. C. Pandey, "Landsat-8 and Sentinel-2 based Forest fire burn area mapping using machine learning algorithms on GEE cloud platform over Uttarakhand, Western Himalaya," *Remote Sensing Applications-Society and Environment*, vol. 18, Apr 2020, Art no. 100324, doi: 10.1016/j.rsase.2020.100324.
- [7] A. Akhmediya, N. Nabiyev, K. Moldamurat, K. Dyussekeyev, and S. Atanov, "Use of Sentinel-1 Dual Polarization Multi-Temporal Data with Gray Level Co-Occurrence Matrix Textural Parameters for Building Damage Assessment," *Pattern Recognition and Image Analysis*, vol. 31, no. 2, pp. 240-250, Apr 2021, doi: 10.1134/s1054661821020036.
- [8] A. Akhmediya, Q. M. Zeng, and Ieee, "USE OF SENTINEL-1 DATA FOR EARTHQUAKE DAMAGE ASSESSMENT IN CASES OF AMATRICE AND SARPOL-E ZAHAB," *Igarss 2018 - 2018 Ieee International Geoscience and Remote Sensing Symposium*, pp. 4877-4880, 2018.
- [9] E. Alcaras, D. Costantino, F. Guastaferrero, C. Parente, and M. Pepe, "Normalized Burn Ratio Plus (NBR plus): A New Index for Sentinel-2 Imagery," *Remote Sensing*, vol. 14, no. 7, Apr 2022, Art no. 1727, doi: 10.3390/rs14071727.
- [10] S. Veraverbeke, W. W. Verstraeten, S. Lhermitte, and R. Goossens, "Evaluating Landsat Thematic Mapper spectral indices for estimating burn severity of the 2007 Peloponnese wildfires in Greece," *International Journal of Wildland Fire*, vol. 19, no. 5, pp. 558-569, 2010, doi: 10.1071/wf09069.

Application of Data Mining to Calculate Power Loss

Bolatbek Amiyev¹, Gulnur Tyulepberdinova¹, Assem Baigara¹, Anelya Koilybekova¹,
Mazhit Orynbay¹, Askar Akhmedov¹, and Asset Baigara¹

¹ Al-Farabi Kazakh National University, Almaty, Kazakhstan

Abstract

The article presents the results of a study, the purpose of which was to develop a data retrieval system and determine non-technical energy losses in energy companies using Apache spark. To achieve this goal, the following tasks were performed: Analysis of the characteristics of data collected from computing systems that can be considered large; The problem of non-technical costs new technologies for big data analysis to analyze the advantages and disadvantages of their application in this area; Development of a system for detecting non-technical energy losses in energy companies using data retrieval methods and Apache spark. During the work, experiments were carried out with Apache Spark models: cluster computing and MLlib machine learning libraries to solve a specific problem of non-technical costs using a dataset of more than 1.6 million consumers showed that a new generation of energy data management can be effectively implemented in this technology.

Keywords

K-Medium, Clustering, Apache Spark, Data Retrieval, Non-Technical Losses

1. Introduction

In recent years, the problem of determining non-technical losses in energy distribution systems has been studied by electrical companies with the support of the Association for Academic Research. To study non-technical costs, energy metering systems require analysis of consumption data collected by consumers over several months or years. In addition, it is necessary to perform computational machine learning algorithms based on measurement data. Our modern society and daily activities depend on the availability of electricity. Electric grids allow the distribution and supply of electricity to consumers, such as residential buildings or factories, from generating infrastructures such as power plants or solar panels. Electric networks are the basis of modern society. Costs in the production and distribution of electricity, including financial costs for electricity suppliers, also reduce stability and reliability. One of the most frequently encountered problems is the costs in electrical networks, namely the difference between the energy produced or purchased and the money paid, which can be divided into two separate categories: technical and non-technical costs. In this article describes the various methods used to detect deviations or fraud. Also, according to, the first one is associated with system problems due to the physical characteristics of the equipment, that is, technical losses are energy lost in transport, conversion and measuring equipment, which becomes a very high cost for electric power companies [1]. Non-technical costs are costs associated with the commercialization of energy supplied to the user, and relate to supplied and unpaid energy, which leads to a loss of income. They are also defined as the difference between total costs and technical costs, which are closely related to illegal connections in the distribution system.

In recent years, the problem of determining non-technical losses in distribution systems has become very important. Theft and distortion of electricity meters in order to change data on energy consumed are the main reasons leading to non-technical losses in energy companies. Since then, conducting periodic checks to minimize such fraud can be very expensive, calculating or measuring the amount of expenses is a difficult task, and in most cases it is impossible to know where they will be. Several electric companies have come to the conclusion that illegal connections should be better classified in order to minimize fraud and theft of energy. Electricity supply companies will never be able to eliminate fraud, but reducing real costs guarantees investments in programs to improve the quality of energy, and also allows you to reduce its price for the consumer.

Currently, automatic determination of non-technical costs is being carried out using various

artificial intelligence methods. Despite the widespread use of machine learning methods to determine non-technical losses in energy systems, the problem of choosing the most popular characteristics has not been widely discussed in terms of non-technical costs.

Technical costs:

1. Copper losses depend on the I^2R losses characteristic of all inductors due to the finite resistance of the conductors;
 2. Dielectric losses are losses caused by the thermal effect on the dielectric material between the conductors;
 3. Induction and radiation losses caused by electromagnetic fields surrounding conductors.
- Technical costs are subject to calculation and control in the case when the considered power system consists of a certain number of loads. The following are the reasons for technical losses:

- Harmonic distortion;
- Correct grounding on the consumer side;
- Long single-phase networks;
- Unbalanced load;
- Losses associated with overload and low voltage;
- Costs associated with low quality equipment.

Non-technical costs:

1. Ensuring the fixation of low consumption meter readings
2. Interfere with the operation of counters for;
3. Errors in the calculation of technical costs;
4. Touch (connection) by LT lines;
5. Organization of false readings by purchasing meters;
6. Theft of the counter by circumvention or other illegal connection
7. Ignore only unpaid bills;
8. Faulty energy meters or unaccounted power;
9. Errors and delays in meter readings and invoices;
10. Non-payments on the part of consumers.

There are also two types of matching opponents. The first type is ordinary consumers using counters, let's call them internal opponents. Internal opponents may have some knowledge about smart meters, so they can fake these meters to reduce electricity consumption [2]. Or they may not know anything about smart meters, but they can get hacking tools for hacking counters for free [3]. The second type of opponents are external enemies. Let's call them external opponents. External attackers can remotely manage counters or manage payment messages in communication networks. They can increase electricity bills and decrease them (however, increasing electricity bills is another form of fraud beyond this thesis). In message-driven attacks, the counters remain unchanged. However, attackers must intercept communications between counters and the head system in order to obtain encryption keys. Thus, the utility company must find meters to replace its keys. According to the above opinion, a counter is called false when it is processed by messages or is false. Internal opponents can distort electricity consumption and attribute it to their neighbors. The room for this type of attack is the same as for message management. Power supply companies conduct research to assess the impact of technical losses in generating, sending and distribution networks, as well as the overall performance of electric networks [4,5]. Non-technical costs are one of the most important problems for electricity distribution companies worldwide. In 2004, Tenaga Nasional Berhad, the only electricity supplier on the Malaysian peninsula, recorded revenue expenses of US\$ 229 million per year due to electricity theft, incorrect accounting and calculation errors [6]. Non-technical expenses of energy supply companies in the United States were estimated from 0.5% to 3.5% of total annual income [7], which is relatively small compared to the expenses of electric power companies in developing countries such as Bangladesh [8], India [9] and Pakistan [10]. Considering that the revenue of utility companies in the United States in 1998 was about 280 billion US dollars, expenses range from 1 to 10 billion dollars [7]. Methods of effective management of non-technical costs in electric networks [11], income protection in the field of distribution [13], [14] and identification of fictitious electricity consumers [15] are proposed. The most effective way to reduce non-technical and

commercial costs is the use of intelligent electronic meters, which complicate fraudulent activities and facilitate their detection [14]. One of the methods for determining losses from the point of view of electrical engineering is the calculation of the energy balance given in [33], which requires topological information about the network. In countries with economies of particular interest due to the high proportion of non-technical costs, this is not possible for the following reasons: (i) the network topology is constantly changing to meet the rapidly growing demand for electricity, (ii) the infrastructure may fail and lead to an incorrect calculation of the energy balance, and (iii) it may damage transformers, feeders and connected meter readings must be read simultaneously. To determine non-technical costs, consumers are checked based on assumptions about possible non-technical costs. The test results are then used to improve predictions when studying algorithms. However, conducting inspections is expensive, as it requires the physical participation of technical specialists. Therefore, in order to reduce the number of false positives, it is necessary to make accurate predictions.

In recent years, several studies have been conducted in the field of electricity distribution on methods of searching and detecting and predicting data theft. These include statistical methods [16,17,18,19]; decision trees [20,21]; artificial neural networks [18,22,23]; knowledge determination in the database [23,24,25,26]; clustering methods [26,27,28]; support vector machine [29]; several classifiers using cross-identification and voting schemes [30]. Among these methods, load profiling is one of the most commonly used methods [31], which is defined as the structure of electricity consumption by a consumer or a group of consumers for a certain period [32]. Determining non-technical expenses are a difficult task due to many reasons for non-technical expenses, such as various types of theft from consumers.

The task of supervised learning to identify deviations. It should be noted that most non-technical methods of determining costs are controlled. Anomaly detection-a superclass of non-technical losses-is usually difficult to investigate in a controlled manner for the reasons described [34]: (i) the anomaly dataset contains very few positive examples and many negative examples, which leads to unbalanced classes, (ii) is used for this. Many kinds of anomalies, because learn from a few positive examples of which anomalies can be difficult for any algorithm, and (iii) there may be future anomalies that may look completely different than any other anomalous examples studied so far. In contrast, supervised learning (i) works well for many positive and negative examples, (ii) when there are enough positive examples for the algorithm to get an idea of what positive examples look like, and (iii) future positive examples may be similar to the ones they have. training kit. Literature review the definition of non-technical costs can be considered as a special case of theft detection, for which a general overview is shown [35].

2. Data set for measurement

The data set (data dictionary in Table 1) consists of 1,452,854 consumers with monthly records of total energy consumption for a 24-month period, i.e. from January 2019 to December 2020.

According to the data, the main energy management systems are active and reactive energy and power factor [12].

Table 1
 Dictionary of data in the dataset

| Field name | Type | Description |
|-------------------|--------|--|
| Id_customer | String | Alphanumeric national code that uniquely identifies the end user |
| year_month | String | Year and month that belong to the same record |
| Tot_active_energy | String | Total consumption upto a certain year-month |

Active energy is the energy that is converted into work and heat by electrical devices. Devices such as incandescent lamps consume only active energy. The unit of measurement is kWh (kilowatt-

hour). It is a unit of measurement of electricity; it refers to the energy consumed in 1 hour by a 1 kW device. The account shows electricity consumption in kWh.

Reactive energy is a part of the energy that is collected and released into the electrical network within a few seconds, instead of being consumed immediately by the consumer. The use of reactive energy refers to equipment that requires a magnetic field, such as electric motors, fluorescent lamps (neon), electronic devices (television, computers, etc.). The unit of measurement of reactive energy is var*h (volt-Ampere reactive hour). This energy is not commercialized; therefore, normal consumption of reactive energy should be considered physiological. Currently, the maximum amount of reactive energy sampling used only for power sources with a capacity of more than 16.5 kW is allowed, and if it is exceeded, a fine is imposed. The parameter that is usually taken into account to check whether the system consumes too much reactive energy is the power factor or $\cos \varphi$. This parameter evaluates the relationship between active energy and reactive energy, and in the case of an ideal load, only resistive, so reactive energy is not consumed, which is equal to 1. When the user's power factor ($\cos \varphi$) is higher than 0.9, reactive energy sampling is considered normal. Values below these limits indicate problems with the system and a simultaneous request for penalties from the electricity distributor with whom the contract has been agreed.

The deflection of the reactive power of the consumer device can be limited or even eliminated by some simple technological devices installed in the customer's electrical system, in which case it is necessary to talk about adjusting the power factor of the electrical system.

3. Methods for determining non-technical costs

This section describes various methods used to detect deviations or fraud. The strategies are briefly described below:

1. DNC clustering;
2. Clustering in GPC;
3. Interquartile interval method.

3.1. Interquartile interval method

The second method involves the quartile range. Abbreviated " IQR " is the width of a rectangle and a rectangle in a mustache graph. That is, $IQR = Q3 - Q1$. IQR can be used as an indicator of how scattered the values are. Statistics assume that the values are grouped into a specific central value. IQR shows how scattered the "average" values are; it can also be used to determine when some other values are "too far" from the central value. These "too far" points are called "outliers" because they are "outside" the expected range. IQR is the length of the rectangle in the rectangle and mustache graph. Output is any value from any end of the block that is one and a half times the length of the block. That is, if the data point is lower than $Q1 - 1.5 \times IQR$ or higher than $Q3 + 1.5 \times IQR$, it is considered too far from the central values to be reasonable. The values $Q1 - 1.5 \times IQR$ and $Q3 + 1.5 \times IQR$ are "fences" that separate "reasonable" values from emissions, and the emissions are located outside the fences. If this paragraph considers "extreme values" rather than emissions, then the values of $Q1 - 1.5 \times iqr$ and $Q3 + 1.5 \times iqr$ are "internal" fences, and the values of $Q1 - 3 \times iqr$ and $Q3 + 3 \times iqr$ are "external" fences. The IQR consumption was calculated using the DC data set for each client and thus compared with the current consumption value, which is reported in a new column called "check_filtering_IQR". The implemented code is presented in Figure 1.

```
#To automate the process of finding outliers by the IQR method,

Consumption_rePiv.registerTempTable("cs_pivot_selected")

df_quartile_1 = sqlContext.sql("select id_customer, percentile_approx(Consumption,0.25) as quartile_1 from cs_pivot_selected group by id_customer")
df_quartile_2 = sqlContext.sql("select id_customer, percentile_approx(Consumption,0.50) as quartile_2 from cs_pivot_selected group by id_customer")
df_quartile_3 = sqlContext.sql("select id_customer, percentile_approx(Consumption,0.75) as quartile_3 from cs_pivot_selected group by id_customer")

dataframe = Consumption_rePiv.join(df_quartile_1,['id_customer']).join(df_quartile_3,['id_customer']).join(df_quartile_2,['id_customer'])
dataframe = dataframe.withColumn('IQR', col('quartile_3') - col('quartile_1')).withColumn('lower_bound', col('quartile_1') - 3 * col('IQR'))\
    .withColumn('upper_bound', col('quartile_3') + 3 * col('IQR'))
dataframe = dataframe.withColumn('check_filtering_IQR', when((col('Consumption') < col('lower_bound')) | (col('Consumption') > col('upper_bound'))\
    ), lit(1)).otherwise(lit(0)))
```

Figure 1: Calculation of IQR consumption for each client using the DC data set

The display of results for a single client is shown in Figure 2.

| id_customer | year_month | Consumption | quartile_1 | quartile_3 | quartile_2 | IQR | lower_bound | upper_bound | check_filt |
|-------------|------------|-------------------|-------------------|--------------------|------------|--------------------|-------------------|--------------------|------------|
| 10096 | 2019-02 | 97.65517241379311 | 75.48387096774194 | 101.14285714285714 | 93.6 | 25.658986175115203 | 24.16589861751153 | 152.46082949308754 | 0 |
| 10096 | 2019-03 | 96 | 75.48387096774194 | 101.14285714285714 | 93.6 | 25.658986175115203 | 24.16589861751153 | 152.46082949308754 | 0 |
| 10096 | 2019-04 | 87.2 | 75.48387096774194 | 101.14285714285714 | 93.6 | 25.658986175115203 | 24.16589861751153 | 152.46082949308754 | 0 |
| 10096 | 2019-05 | 85.16129032258064 | 75.48387096774194 | 101.14285714285714 | 93.6 | 25.658986175115203 | 24.16589861751153 | 152.46082949308754 | 0 |

Figure 2: Representation of the IQR detection data frame

All customers with at least one month of consumption marked as extreme values were selected to create a different data set, and then shortened using "pivot_monthly_udf". This new data set contains only "Id_customer" and "check_filtering_IQR" as binary vectors for each month.

Then, using a user-defined function that calculates the serial number in the array, the longest number of serial numbers for each consumer, namely the month with the extreme value, was calculated. Consumers with at least 4 consecutive extreme values were marked as suspicious consumers. The code is shown in Figure 3.

```
dataframeToPiv = dataframe.filter(col("check_filtering_IQR")==1)
dataframe_1 = pivot_monthly_udf(dataframeToPiv,"id_customer","year_month","check_filtering_IQR")
vecAssembler = VectorAssembler(inputCols=dataframe_1.columns[1:], outputCol="features")
consAssembler = vecAssembler.transform(dataframe_1.na.fill(0)).select('id_customer', 'features')
dataset_count_ones = consAssembler.rdd.map(lambda x : [x[0],consecutive_one(x[1])])
dataset_count_ones.cache()
label_2_df = dataset_count_ones.filter(lambda x: int(x[1]) > 4)
label_2_Id = sqlContext.createDataFrame(label_2_df.map(lambda x : Row(x[0])),["id_customer"])
```

Figure 3: Using a user-defined function that calculates the serial number in the array

The following figure (Figure 4) shows the average daily consumption dynamics of these suspicious consumers.



Figure 4: Average number of suspected consumers detected by the IQR method by month

It is claimed that a sharp decrease in consumption can be caused by a real tendency of consumers,

such as a change in the type of contract or other use of energy consumed [19]. Such a slope can be caused by a failure of measuring equipment or a voluntary change of equipment, both situations lead to the emergence of non-technical costs for the company and the loss of money to it, such a determination gives 1.117 consumers.

3.2. Distance intra-cluster method

The first method that uses clustering in this work establishes the use of distances between points and the cluster center in the DCN dataset. In this method, the distance between points and cluster centers was calculated and a threshold level above each point was set, which could be set as common errors or potential customer fraud.

As mentioned above, the K-average algorithm should set the number of K clusters in advance, unfortunately, there is no clear answer to this problem. The optimal number of clusters is subjective to a certain extent and depends on the method used to measure similarity and the parameters used for distribution.

The main method for finding the required number of clusters is the elbow method and the average silhouette. The first method considers the total sum of squares inside the cluster, depending on the number of clusters: the number of clusters should be selected so that the addition of another cluster does not improve the total sum of squares inside the cluster. The second, average silhouette method calculates the average control silhouette for different k values. The optimal number of K clusters is the one that maximizes the average silhouette in the range of possible values for K. This second method is unlikely to be scalable, since it uses a pair distance, and it will always take $O(n^2)$ time to calculate it.

The optimal number of clusters in the Elbow method can be determined as follows:

1. Calculation of the clustering algorithm for different k values (for example, K-average clustering).
2. Calculate the total sum of squares inside the cluster (WSS) for each k.
3. Make a square curve of the sum inside the cluster depending on the number of K clusters.
4. The location of the bend (knee) on the graph is usually considered as an indicator of the corresponding number of clusters.

This section used a K-average clustering algorithm with a range of 5 to 30 for a number in a cluster figure 5 shows the trend of intra-cluster sum of squares for each k (each iteration used only 25% of the rows for time-related problems).

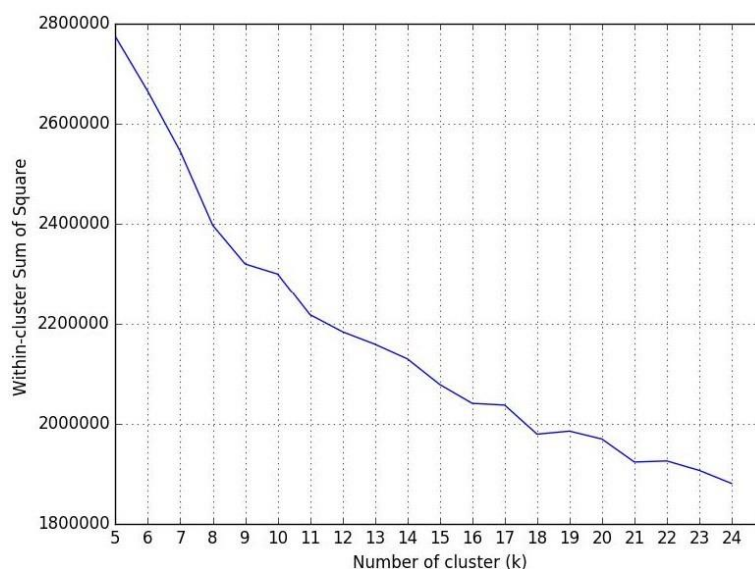


Figure 5: Elbow method with K-means

Therefore, for $k = 11$, the sum of squares inside the cluster changes slowly and changes less than for other k , so for this data set, 11 should be a good choice for the number of clusters.

Therefore, at the moment, a K-average value equal to $k-11$ was used, the configuration for the “initmode” parameter, which belongs to the initialization algorithm, was set to the “k-average” state in Step 4, the convergence tolerance and the maximum number of iterations were set to 0,001 and 100, respectively. The implemented code is shown in Figure 6.

```
vecAssembler = VectorAssembler(inputCols=DCN.columns[1:], outputCol="features")
consAssembler = vecAssembler.transform(DCN).select('id_customer', 'features')

kmeans_15 = KMeans(featuresCol="features",k=11, seed=1, initMode='k-means||', initSteps=4, tol=0.001, maxIter=100)
model_kmeans_15= kmeans_15.fit(consAssembler)

cons_transformed = model_kmeans_15.transform(consAssembler).select("id_customer","features", "prediction")
```

Figure 6: Application of the k-average value

“Vector Assembler” is a converter that combines a given list of columns into a single vector column. This is useful for combining raw objects and objects created by different object converters into a single object vector for learning ML models.

Figure 6 shows the normalized daily consumption trend for each cluster, for each individual item, we considered the average daily consumption for consumers belonging to this cluster this month, table 2 shows the cluster size.

Table 2

Number and size of clusters

| Cluster | Size |
|---------|---------|
| 0 | 111.471 |
| 1 | 63.989 |
| 2 | 121.858 |
| 3 | 66.582 |
| 4 | 123.425 |
| 5 | 107.384 |
| 6 | 81.341 |
| 7 | 58.931 |
| 8 | 50.816 |
| 9 | 116.965 |
| 10 | 47.222 |

This threshold is set for 2,484 consumers, and the figure below (Figure 8) shows the average daily consumption trend of these suspicious consumers.

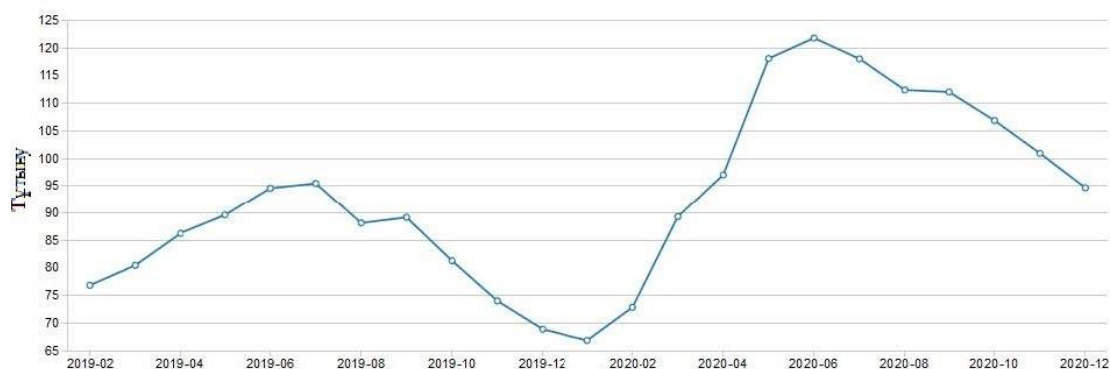


Figure 8: Average number of suspected consumers identified by the first clustering method

Consumption indicators may indicate that the algorithm marks all consumers who experiment with a sharp change in consumption as suspicious, in this case with an increase in consumption since January 2020.

The following images show all histograms of the distance between points with a limit separated by a vertical blue line and the center of the cluster. Table 3 shows how many abnormal clients are found in each cluster.

Table 3

Number and dimensions of clusters of abnormal identifiers

| Cluster | Size |
|---------|------|
| 0 | 316 |
| 1 | 140 |
| 2 | 252 |
| 3 | 520 |
| 4 | 130 |
| 5 | 225 |
| 6 | 93 |
| 7 | 200 |
| 8 | 114 |
| 9 | 302 |
| 10 | 192 |

3.3. Clustering method with a small number of clusters

The second clustering method was developed in a GCP dataset obtained by feature engineering, which is marked as errors or potential fraud by all clients belonging to the smallest clusters.

In addition, here the clustering algorithm k is the average, and the method of selecting the best is the elbow method shown in Figure 9.

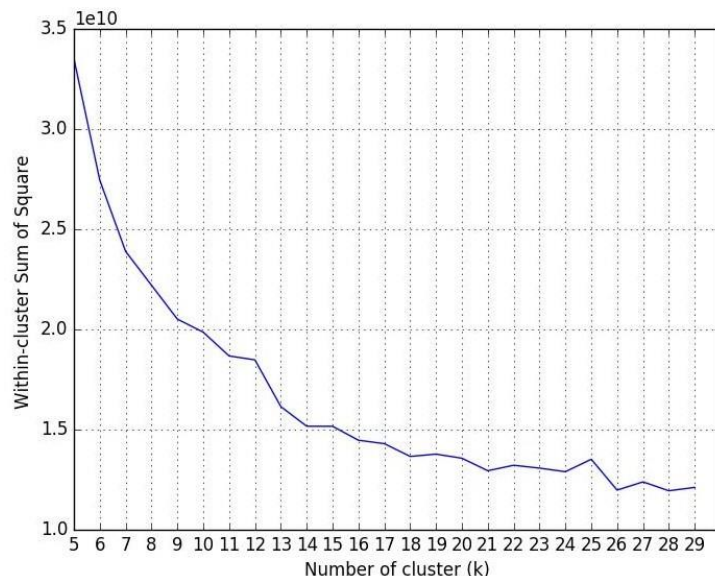


Figure 9: Elbow method for K-means

For k equal to 15, the sum of squares inside the cluster changes slowly and changes less than other k, but if k is larger in the next step along the slope, this may be the best choice for the approach, so 18 should be the best choice for this data.

In the same configuration as the first method, k-the average tolerance rate was 0.001 and the maximum number of iterations was 100.

The following figure (Figure 10) shows the normalized daily consumption trend for each cluster, as in the first method, table 4 shows the cluster size instead.

It is easy to understand at first glance that a cluster with a small number of consumers shows such a strange trend in consumption, against these arguments, it was decided to label consumers belonging to clusters #4,10,11,12,14 and 17 as suspicious consumers.

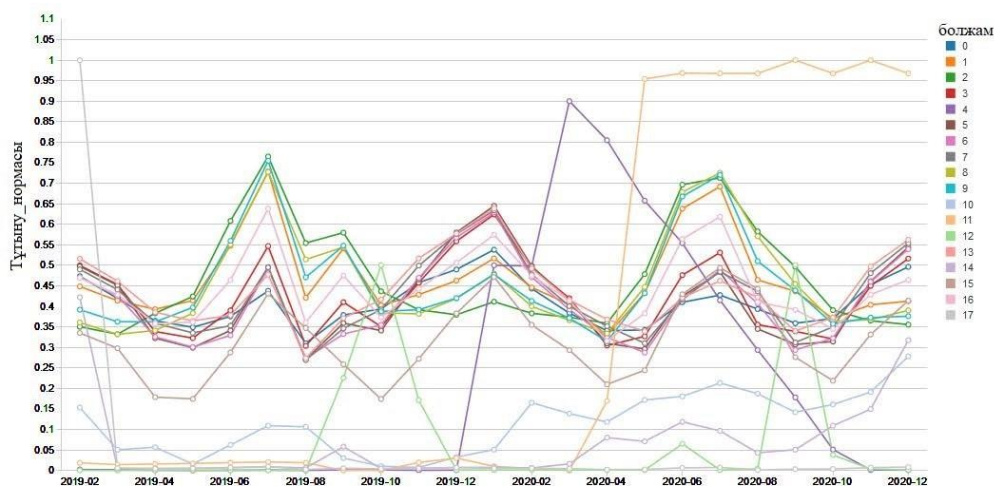


Figure 10: DCN on average by month of clusters

Table 4

| Cluster | Size |
|---------|---------|
| 0 | 259.358 |
| 1 | 3.604 |
| 2 | 204 |
| 3 | 12.622 |
| 4 | 2 |
| 5 | 29.259 |
| 6 | 84.664 |
| 7 | 218.853 |
| 8 | 766 |
| 9 | 1.831 |
| 10 | 44 |
| 11 | 1 |
| 12 | 2 |
| 13 | 337.557 |
| 14 | 19 |
| 15 | 632 |
| 16 | 6.362 |
| 17 | 2 |

Additional support for the method is shown in Figure 10 below, which means that the cluster for consumers is based only on daily consumption.

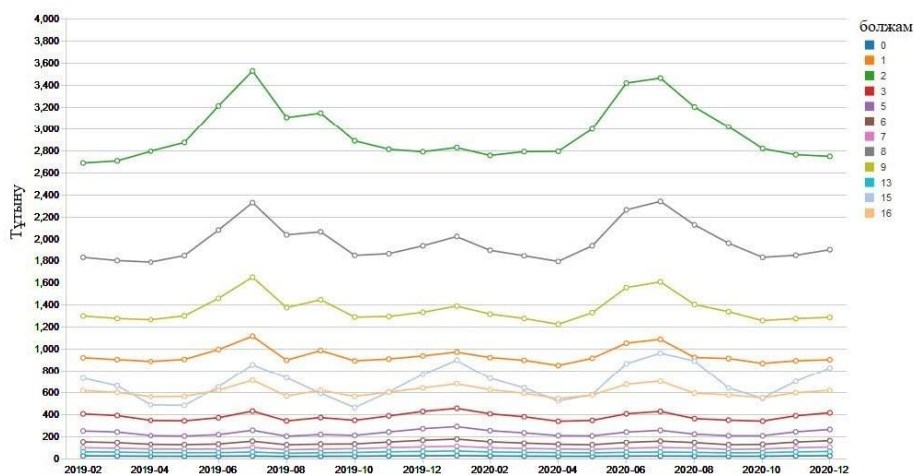


Figure 11: DCN on average by month of clusters removed the suspicious clusters

A total of 70 customers noted this threshold, and the figure below (figure 11) shows the average daily consumption trend of these suspicious consumers.

Moreover, in this case, the clustering algorithm revealed unusual behavior of buyers: a sharp decline in the first five months and a sudden increase since December 2019.

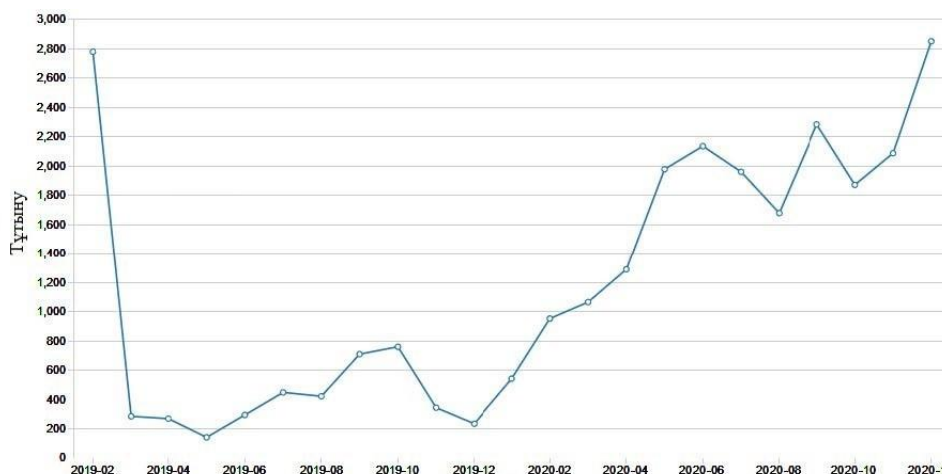


Figure 12: DC on average by month of suspicious customer found by Second Clustering Method

3.4. Result of methods for determining non-technical costs

Before sending consumers for verification, the results of all detection methods were checked to determine how many consumers were compared, and to make sure that different algorithms were not redundant. Thus, after combining the consumers found in each method, the results shown in Table 5 were obtained.

Table 5

Customer selected for verification

| Detection method | Size |
|---|-------|
| IQR method | 1.117 |
| Distance intra-cluster method | 2.484 |
| Clustering method with a small number of clusters | 70 |
| Total | 3.671 |

Thus, with the help of non-technical costs, a list of 3.671 consumers with a clear and suspicious consumption structure was obtained. Cases of these non-technical costs may be associated with a decrease in electricity demand for their businesses.

4. Conclusion

As a conclusion, it should be noted that non-technical costs are an important issue in the energy sector. Because it has a huge impact on the company's profits. However, the methodology for determining non-technical costs of companies is currently very limited, as these companies use detection methods that do not use data mining methods. Various methods for determining non-technical costs have been developed and tested in a specific database provided by the energy company. In particular, this article presents a direction of work based on 3 different methods for determining non-technical costs. The clustering method with a small number of clusters was the best because evaluating the performance of the clustering algorithm is not as trivial as counting the number of errors or the accuracy and recall of a controlled classification algorithm. In particular, any evaluation metric should not take into account the absolute values of cluster labels, but rather if this clustering defines data divisions similar to some basic set of true classes or satisfying some assumption, so that members belonging to the same class are more similar than members of different classes according to some similarity metric. Comparing the results of the used methods, we can conclude that although the IQR method is efficient in terms of execution time, the Distance intra-

cluster method has determined the largest number of customers with clear and suspicious consumption structure with the help of non-technical costs.

5. References

- [1] E. Sankari, R. Siva, R. Rajesh, P. Matheswaran. "Detection of Non-Technical Loss in Power Utilities using Data Mining Techniques.", *International Journal for Innovative Research in Science & Technology*, 1(9). 97-100.
- [2] W. Han, X. Yang. "Design a fast Non-Technical Loss fraud detector for smart grid.", *Security and Communication Networks*, 9(18). 5116-5132.
- [3] E. Protalinski. Smart meter hacking tool released, 2018, | ZDNet. [online] ZDNet. Available at: <https://www.zdnet.com/article/smart-meter-hacking-tool-released/>.
- [4] C.R. Paul. "System loss in a Metropolitan utility network.", *Power Engineering Journal*, 1(5). 305-307.
- [5] I.E. Davidson, A. Odubiyi, M.O. Kachienga, B. Manshire. "Technical loss computation and economic dispatch model for T&D systems in a deregulated ESI.", *Power Engineering Journal*, 16(2). 55-60.
- [6] Tenaga Nasional Berhad. Annual Report Tenaga Nasional Berhad 2018. Kuala Lumpur, KL: TNB, 2018.
- [7] T.B. Smith. "Electricity theft: a comparative analysis.", *Energy policy*, 32(18). 2067- 2076.
- [8] M.S. Alam, E. Kabir, M.M. Rahman, M.A.K. Chowdhury. "Power sector reform in Bangladesh: Electricity distribution system.", *Energy*, 29(11). 1773-1783.
- [9] A.K. Saxena. "Decision priorities and scenarios for minimizing electrical power loss in an Indian power system network.", *Electric Power components and systems*, 31(8). 717- 727.
- [10] R.M. Shrestha, M. Azhar. "Environmental and utility planning implications of electricity loss reduction in a developing country: A comparative study of technical options.", *International Journal of Energy Research*, 22(1). 47-59.
- [11] A.H. Nizar, Z.Y. Dong, Y. Wang. "Power utility nontechnical loss analysis with extreme learning machine method.", *IEEE Transactions on Power Systems*, 23(3). 946-955.
- [12] Zeuslog.com. (2018). Energia Attiva, Reattiva e Fattore di Potenza – ZeusLog. [online] Available at: http://www.zeuslog.com/?page_id=68&lang=it.
- [13] R. Mano, R. Cespedes, D. Maia, "Protecting revenue in the distribution industry: a new approach with the revenue assurance and audit process." in *Transmission and Distribution Conference and Exposition*, (San Paolo, Brazil, 2019), IEEE.
- [14] R.M. Krishna, S.H. Miller. "Revenue improvement from intelligent metering systems." in *Metering and Tariffs for Energy Supply*, (Birmingham, UK, 2016), IET, 218- 222.
- [15] A.J. Dick. "Theft of electricity-how UK electricity companies detect and deter." in *Security and Detection*, (Brighton, UK, 2017), IET, 90-95.
- [16] J.W. Fourie, J.E. Calmeyer. "A statistical method to minimize electrical energy losses in a local electricity distribution network." in *AFRICON, 2018. 7th AFRICON Conference in Africa*, (Gaborone, Botswana ,2018), IEEE, Vol. 2.
- [17] J.E. Cabral, E.M. Gontijo. "Fraud detection in electrical energy consumers using rough sets." in *Systems, Man and Cybernetics, 2018 IEEE International Conference on.*, (The Hague, Netherlands, 2018), IEEE, Vol. 4.
- [18] I. Monedero, C. Leon, J. Biscarri, R. Millan. "Midas: Detection of non-technical losses in electrical consumption using neural networks and statistical techniques." in *International Conference on Computational Science and Its Applications*. Springer, Berlin, 2019.
- [19] I. Monedero, C. Leon, J. Biscarri, R. Millan, J.I. Guerrero. "Detection of frauds and other non-technical losses in a power utility using Pearson coefficient, Bayesian networks and decision trees." *International Journal of Electrical Power & Energy Systems*, 34(1), 90-98.
- [20] E.M. Gontijo, J.R. Filho, A.C. Delaiba, J.E. Cabral, J.O. Pinto. "Fraud identification in electricity company customers using decision tree." In *Systems, Man and Cybernetics, 2018 IEEE International Conference on.*, (The Hague, Netherlands, 2018), IEEE, Vol. 4.

- [21] A.H. Nizar, Z.Y. Dong, J.H. Zhao, P. Zhang. "A data mining based NTL analysis method." in Power Engineering Society General Meeting, (Tampa, FL, USA, 2020), IEEE.
- [22] J.R. Galvan, A. Elices, A. Munoz, T. Czernichow, M.A. Sanz-Bobi. "System for detection of abnormalities and fraud in customer consumption." In Proc. of the 12th Conference on the Electric Power Supply Industry, (Pattaya, Thailand ,2016).
- [23] B.C. Costa, B.L. Alberto, A.M. Portela, W. Maduro, E.O. Eler. "Fraud detection in electric power distribution networks using an ANN-based knowledge-discovery process." International Journal of Artificial Intelligence & Applications, 4(6). 17.
- [24] A.H. Nizar, Z.Y. Dong, J.H. Zhao. "Load profiling and data mining techniques in electricity deregulated market." In Power Engineering Society General Meeting, (Montreal, Que., Canada, 2020), IEEE.
- [25] A.H. Nizar, Z.Y. Dong, M. Jalaluddin, M.J. Raffles. "Load profiling method in detecting non-technical loss activities in a power utility." in Power and Energy Conference PECon'06. IEEE International, (Putra Jaya, Malaysia, 2020). IEEE.
- [26] E.W. Angelos, O.R. Saavedra, O.A. Carmona Cortes, A.N. de Souza. "Detection and identification of abnormalities in customer consumptions in power distribution systems." IEEE Transactions on Power Delivery, 26(4). 2436-2442.
- [27] A.H. Nizar, Z.Y. Dong, P. Zhang. "Detection rules for non-technical losses analysis in power utilities." in Power and Energy Society General Meeting-Conversion and Delivery of Electrical Energy in the 21st Century, (Pittsburgh, PA, USA, 2018), IEEE.
- [28] L.A.P. Júnior, C.C.O. Ramos, D. Rodrigues, D.R. Pereira, A.N. de Souza, A.P. da Costa, J. P. Papa. "Unsupervised non-technical losses identification through optimum-path forest." Electric Power Systems Research, 140. 413-423.
- [29] J. Nagi, K.S. Yap, S.K. Tiong, S. K. Ahmed, M. Mohamad. "Nontechnical loss detection for metered customers in power utility using support vector machines.". IEEE transactions on Power Delivery, 25(2). 1162-1171.
- [30] R. Jiang, H. Tagaris, A. Lachs, M. Jeffrey. "Wavelet based feature extraction and multiple classifiers for electricity fraud detection." in Transmission and Distribution Conference and Exhibition 2022: Asia Pacific. IEEE/PES, (Yokohama, Japan, Japan, 2022), Vol. 3.
- [31] D. Gerbec, S. Gasperic, I. Smon, F. Gubina. "Allocation of the load profiles to consumers using probabilistic neural networks." IEEE Transactions on Power Systems, 20(2). 548-555.
- [32] S.V. Allera, A.G. Horsburgh. "Load profiling for the energy trading and settlements in the UK electricity markets." In Proc. DistribuTECH Europe DA/DSM Conference, (London, UK, 2020)
- [33] A. Méffe, C.C.B. Oliveira, N. Kagan, S. Jonathan, S. Caparroz, J.L. Cavaretti. "A new method for the computation of technical losses in electrical power distribution systems." in Electricity Distribution, (Amsterdam, Netherlands, 2019), IET.
- [34] M. Mansurova, M. Zubairova, N. Kadyrbek, G. Tyulepberdinova, T. Sarsembayeva. Data Analysis for The Student Health Digital Profile. 1-6. 2021. 10.1109/ICECCO53203.2021.9663804.
- [35] P. Glauner, J.A. Meira, P. Valtchev, R. State, F. Bettinger. "The challenge of non- technical loss detection using artificial intelligence: A survey.", International Journal of Computational Intelligence Systems (IJCIS), 10(1). 760-775.

Research and Development of Gamification Software for Teaching Technical Specialties

Aidar T. Jantikov¹, Laura M. Alimzhanova¹, Khudaibergen B. Nurlybekov¹, and
Rustem A. Malybayev¹

¹ International Information Technology University, Almaty, Kazakhstan

Abstract

The publication provides a thorough and detailed analysis of teaching methods and memorization of information, algorithms for gamification of education, as well as their implementation in teaching technical specialties. The article reveals in detail the topic of the effectiveness of using the technology of gamification of education, as well as increasing the motivation of students through video games and the introduction of game elements in teaching technical specialties, such as: fundamentals of programming, computer science and information systems. As a result of the research, software based on the use of educational gamification algorithms proposed and developed, with the help of which it is possible to significantly simplify the process of memorizing information in the study of subjects in the field of information technology.

Keywords

Gamification of education, game technologies, information technologies, teaching methods, ways of organizing training

1. Introduction

The value of information technology is growing exponentially. IT has become one of the most important industries of this century. Each organization and enterprise operate at the expense of basic and advanced programs built with the help of IT.

Now there is a tendency to increase the training of specialists in the IT field at the national level. Judging by the article published on the official information resource of the Prime Minister of the Republic of Kazakhstan dated May 20, 2022 [1], within the framework of the National Digitalization Project, the Head of State set the task of training at least one hundred thousand highly qualified IT specialists by 2025. To train one hundred thousand highly qualified IT specialists, the Ministry of Digital Development, Innovation and Aerospace Industry of the Republic of Kazakhstan has developed different tasks and activities. First, such as:

- Opening of twenty modern programming schools;
- Allocation of twenty thousand vouchers (grants for training) to IT programming schools;
- Patronage of specialized universities over regional universities;

On an ongoing basis, the Ministry is working to attract IT schools to mass training of IT specialists. In 2021, 6 such schools opened. It is planned to open 5 more schools in 2022. The IT schools themselves have been open in the regions since 2021.

Thus, the HackerU school launched in Almaty, the Da Vinci School in Aktau, the IT hub school in Uralsk, the Zhez-IT school in Zhezkazgan, and the Toraigyrov University in Pavlodar, sponsored by ERG and the Akimat of Pavlodar region, as well as the computer academy "Step".

To provide state support to students and IT schools, the Ministry has developed a mechanism for allocating vouchers (grants) to IT schools based on the corporate foundation "International Technopark of IT Startups "Astana Hub".

As a pilot in 2021, 100 grants given to 13 IT schools, and in 2022 it planned to provide 2000-3000 grants. The cost of one voucher is six hundred thousand tenge. The Ministry of Finance of the Republic of Kazakhstan has previously approved funding for the allocation of grants in a working manner.

This all shows the seriousness of the approach to training young specialists in the field of IT at the

state level. For a more simplified educational process, it proposed to create a mobile application based on gamification technology that teaches children aged 3 to 12 years the basics of programming. This is necessary to increase the demand for IT specialties in the country and facilitate training in IT areas for young professionals.

Gamification in education is a poorly studied method of teaching, which already brings satisfactory results in education. Using it, it can facilitate the understanding and application of any subject studied, regardless of the age, location, and capabilities of the student.

Mobile games are filling up increased segments of the younger generation. Modern children spend a lot of time playing mobile phone. As stated in the article "Children spend six hours or more a day on screens" [2], from 1995 to 2015, the amount of time spent by children aged 5 to 16 years in front of screens has increased significantly. At the time of 2015, children spend an average of 6.5 hours a day, instead of the 3 hours spent in 1995.

All this data taken from the annual reports called Connected Kids. They collect a general picture of the use of electronics and children's habits, which have analyzed since 1995. The data of the report for 2018 [3] and 2019 [4] years only give us more understanding that every year the interest of children in electronics with Internet access and the time spent behind screens per day is only growing. This also gives us the opportunity to introduce a new way of learning and memorizing new things with the help of information technologies and learning gamification algorithms.

Children memorize the material much faster if they apply it in practice. Playing a game that will teach them terms and basics of programming, they will have to memorize and apply everything they learned in the game to implement their own first project using knowledge of the programming language. This is one of the best and most interesting ways to learn.

2. Modern teaching methods and their classification

As shown in the article from Sadovskaya Irina Lvovna "Teaching methods: a new concept" [5], recently in the theory and practice of education, the term "teaching technology" has used as a "teaching methodology". This is because in the teaching methodology and teaching technology, the goals of education achieved using teaching methods. However, now there is no single formulation for the concept of "teaching method", each author strives to give his own definition and explains the absence of a recognized option by the versatility of the object defined. The analyses of the works on the topic give us three options for determining the methods of teaching:

- This is the way;
- This is the method;
- This is a set of techniques by which educational goals achieved.

Various teaching methods engage in learning processes that allow students to memorize information. The learning process called the purposeful active cognitive activity of the student, organized by the teacher, and implemented under his guidance, aimed at obtaining knowledge, skills, and abilities, forming interest in learning, and developing the creative potential of the child [6]. The learning process as a system consists of the subjects of learning (teacher and student), goals, methods, content, means and joint activities of the teacher and student. The learning process allows you to achieve a changed state for both subjects, where the student learns information, and the teacher gets experience in teaching and repeats the studied material with the student.

The teaching method is a way of transmitting and assimilating educational information. In the teaching method, it is important that the information presented in a certain way, recorded, transmitted, and perceived without any loss. Historically, there have been four main types of teaching methods:

- Auditory – information perceived by ear;
- Visual – information perceived with the help of vision;
- Kinesthetic – information perceived through muscular efforts;
- Polymodal – mixed channels of perception.

The polymodal type consists of:

- Audiovisual – designed for simultaneous auditory and visual fixation;
- Visual-kinesthetic – designed for simultaneous visual and kinesthetic fixation;

- Auditory-kinesthetic – designed for simultaneous auditory and kinesthetic fixation;
- Audiovisual-kinesthetic – conducting experiments.

With the correct presentation of the learning process, information should be perceived through all information channels. Currently, there is no doubt that any training should be based on a polymodal method of teaching, this pointed out by the great methodologists of the past V. A. Lai, V. Mrochek and F. Filippovich, F. Froebel, F. A. Ern, and others [5].

Now there is a slightly different classification of teaching methods, which Nina Afonina writes about in her article "Classification of teaching methods according to the source of knowledge" [7]. It divides teaching methods into the following classes:

- Practical teaching methods. These are the teaching methods that make up the practical direction of the student's activity. These include exercises, workouts, lab work, and errands.
- Visual teaching methods. They aimed at applying knowledge from authoritative sources, paper media, technical means, devices, and various equipment. These include illustrations, observations, display and demonstration.
- Verbal teaching methods. Based on the explanation of the educational material by the teacher. Verbal teaching methods include a story, a conversation, a lecture, an explanation.
- Methods of working with book manuals. Independent development of skills by studying the relevant literature. These are reading, taking notes, planning for the text, writing abstracts, and quoting.
- Video methods. These methods include the use of modern information technologies, which already include components of practical, visual, and verbal teaching methods.

With the help of computer technologies and video methods, it is possible to achieve the implementation of all the above methods of teaching, through distance learning. Thus, the student can both view information on ready-made video tutorials, perform exercises on special test sites, learn new things from electronic textbooks and articles.

Video methods can be attributed to the polymodal type, since with the use of computer technologies, it is possible to obtain an audiovisual perception of new material, as well as kinesthetic, based on the performance of independent work by students. Thus, the use of video methods becomes one of the leading options for assimilation of information.

3. Features of methods of teaching technical specialties

Training in technical specialties, including programming, is slightly different from basic training in any other subjects. Programming training requires a student to have a set of knowledge in mathematics and an understanding of algorithms and welcomes an analytical mindset that can easily understand problem solving. This field is not as simple as teaching humanities or solving math problems. Everything is more complicated in programming. In addition to understanding the problem and the logic of its solution, the student should be able to build a structure in the programming language used and develop a software product that will produce the expected results when solving the problem.

Unfortunately, a lot of students have difficulties in studying this technical specialty. Since the methods of teaching programming are different from those used in the study of other subjects at school. The study of technical specialties requires a different approach to the student to involve him more deeply in the learning process and further understanding of this subject.

In the article by I. V. Bazhenova "Visualization of knowledge as a method of cognitive approach to programming training" [8], it proposed to use such means of knowledge visualization as conceptual and mental maps and visualization of algorithms for programming training. The methodology is based on a cognitive approach to learning, which allows considering the cognitive characteristics of the trainees.

Modern neurophysiological and psychophysiological studies of the human brain show that 80-90% of information received by many people through the organs of vision, i.e., they are "visual". In addition, in the era of the Internet, there is a need to consider the peculiarities of sensory perception and thinking of the modern generation of schoolchildren and students, which called the "network" or

"digital generation" [8].

It follows from this study that the visual, audiovisual, and video methods prescribed earlier in the article are among the most effective ways of teaching programming.

I. V. Bazhenova believes that the most important value for the learning process are the methods of visualization of knowledge. In the programming training course, the following scheme should use:

- declarative knowledge → visualization of programming concepts and data structures;
- procedural knowledge → visualization of algorithms.

The most convenient means of visualizing programming concepts are conceptual and mental maps. They have the following key characteristics:

- hierarchical data display;
- the presence of links between the concepts, in the form of arrows;
- permissibility of cross-links;
- availability of examples of events and objects that contribute to the understanding of the displayed concepts.

Mental maps allow a person to immediately grasp the entire task and intuitively solve it. In addition, in the future he will be able to use it as a hint, which will simplify the process of memorizing information and its further use in the field of programming.

Figure 1 shows an example of the implementation of a mental map on the topic "Pointers in C++". The object of study visualized in the center of the map, then the main ideas associated with the object of study diverge from the center as branches, and the branches themselves form the structure of the tree, which allows the student to visualize the process of solving the problem and the key characteristics of the concept studied.

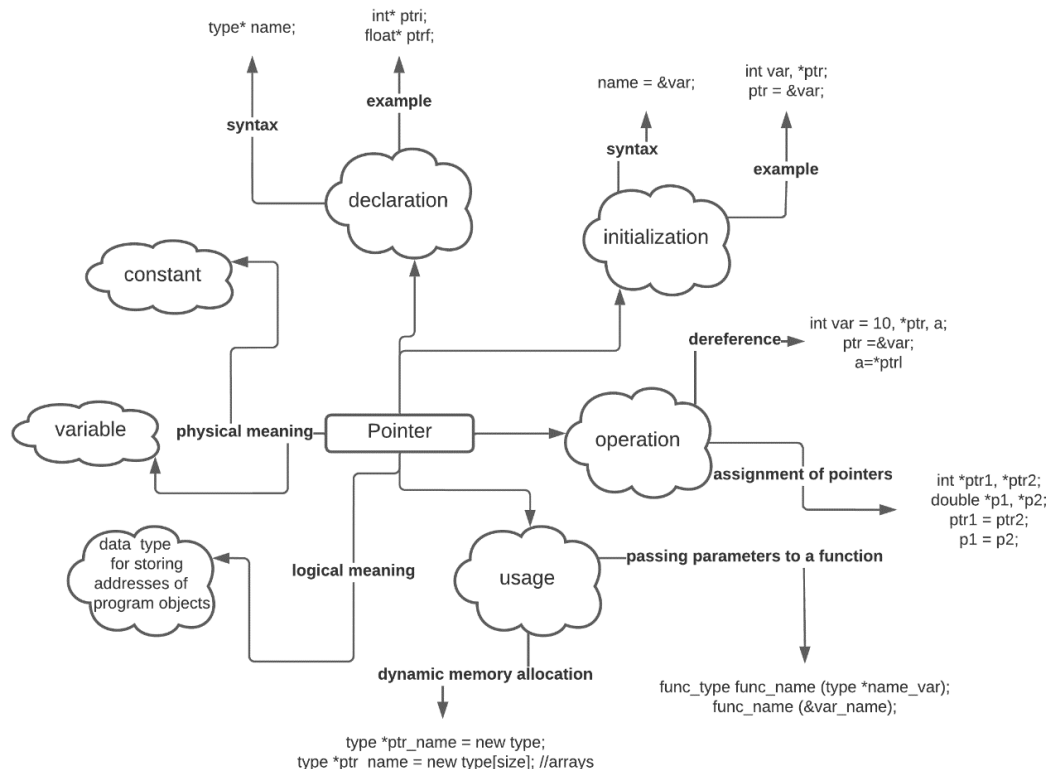


Figure 1: Mental map on the topic "Pointers in C++"

Thus, the method of visualization of knowledge is a powerful tool for cognition and study of technical specialties. The use of various visualization tools in teaching programming allows to significantly activate the cognitive activity of students, motivate them to independently acquire

knowledge, arouse interest in obtaining new knowledge.

4. Gamification as a method of teaching technical specialties

The gamification method chosen for teaching technical specialties. However, first it is necessary to understand more deeply how gamification works and why it solves the problems of education so effectively.

Gamification is the appearance of game elements in non-game processes - for example, in education. The elements of the game create constant feedback, which, in turn, allows you to adjust the behavior of the "player", helps to optimize the assimilation of the material, increases engagement, and allows you to gradually complicate and complicate tasks due to increased involvement – just like in a normal game we move from simpler levels to more complex. The game helps to increase motivation, and high motivation helps not to be afraid of difficult material. The implementation of game elements can be different; good and bad grades or an increased scholarship can also regard as part of the game. But usually, gamification still understood as the use of special digital technologies characteristic of computer games.

Gamification itself shows itself in all its glory in the process of online learning. Self-education lends itself best to gamification. This method of teaching has its pros and cons. From the positive sides, one can consider the fact of self-discipline, setting goals, achieving results along a single trajectory, etc. Of the disadvantages, we can consider the lack of control, which encourages the student to postpone his studies for an indefinite period.

Now, there is an increasing trend of gamification in training. According to a study by foreign scientists in a paper titled "Does educational gamification improve students' motivation? If so, which game elements work best? " [9], the game form of learning increases the motivation to learn by two-thirds. The introduction of game elements into the learning process itself considered especially important, especially in the distance learning format. This makes it possible to increase the involvement of students in the educational process.

However, there are not only pros, but also cons. In gamification, the issue of motivation of students is acute, and judging by the study [10] conducted by foreign scientists, the introduction of external motivation factors affects the internal motivation of a person. This may lead to the fact that the student will strive to achieve the goal only in the conditions of the game and will not be able to use the knowledge in real life. That is why gamification algorithms should design in such a way that they can adapt to each user individually. This approach will allow many students to be interested in the learning process, and in the future, they will be able to apply knowledge in real life.

Thus, there are certain models that often used in gamification. The most common of them is the model of the BPL (badges, points, leaderboards). It reserves the use of badges, points, and leaderboards. For completing tasks, the student given points, badges given for a certain number of points and levels assigned. This is how the leaderboard formed, which motivates students to compete in the amount of material studied.

The simplest example of gamification is learning English with Leo the lion cub in a game called *LinguaLeo*. In this game, you play as a lion cub who, as the game progresses, earns achievements and learns English on the fly. Leo allows you to control the entire learning process (Fig. 4), without letting you get bored. Together with him, you can play various games that train your memory, speed, and more (Fig. 5). He will also not let you forget about education, constantly reminding you of the importance of continuing education and giving various bonuses for continuous learning (Fig. 2). This approach gives the user motivation to continue learning until they achieve certain results. In addition, the game provides not only one bare theory, but also allows you to immediately practice and learn what you have learned (Fig. 3). After that, the user learns English perfectly.

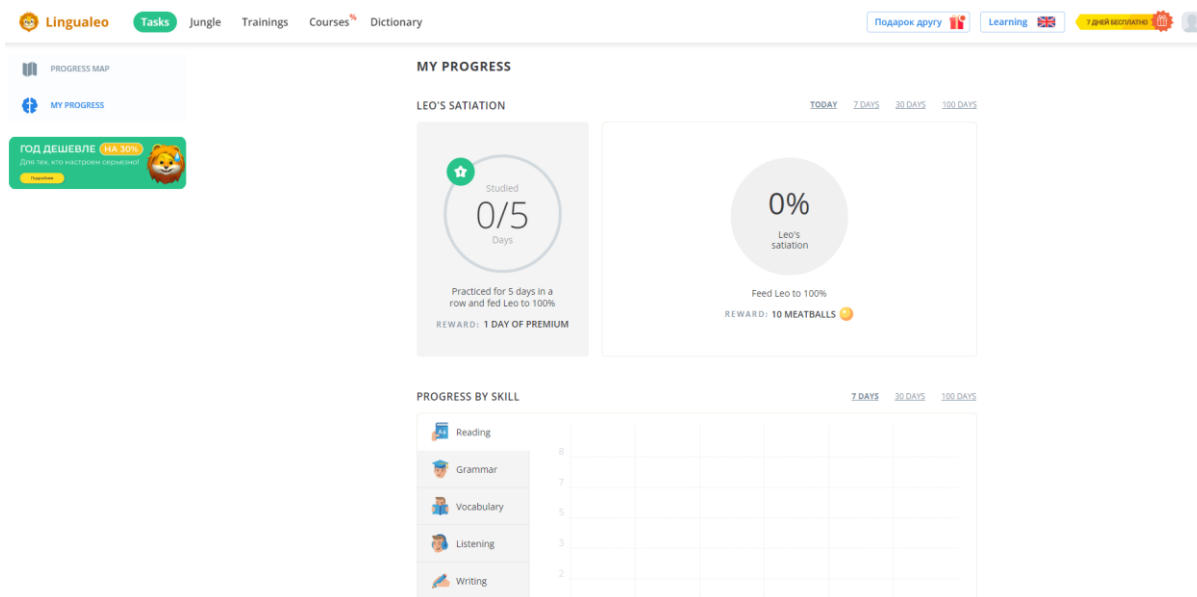


Figure 2: Bonuses for continuing education

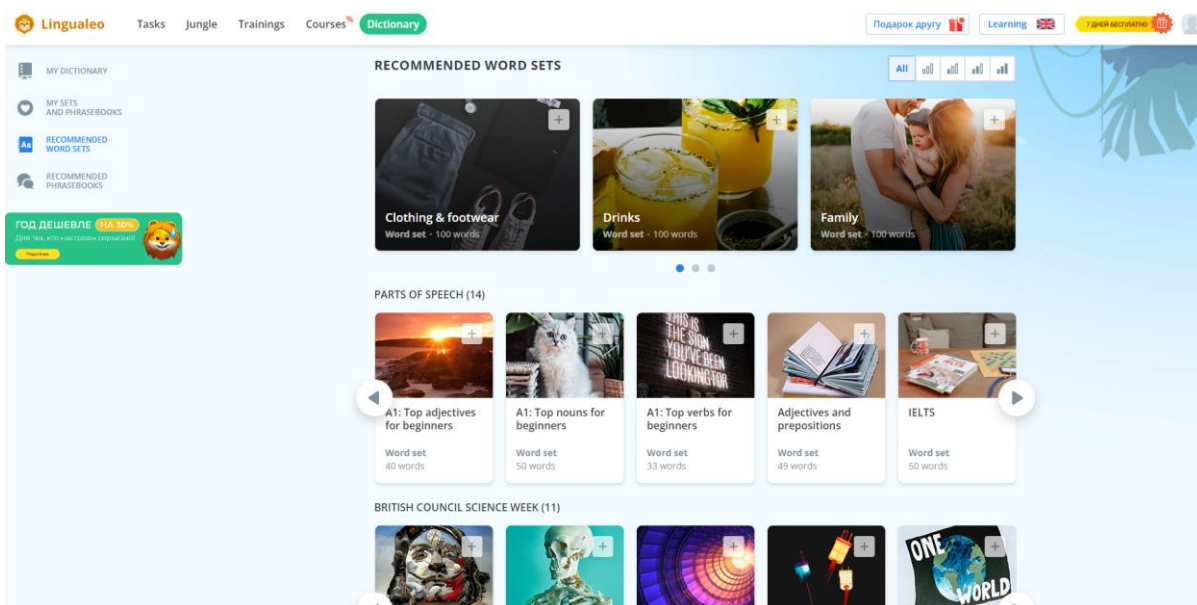


Figure 3: Recommendation on theory and practice

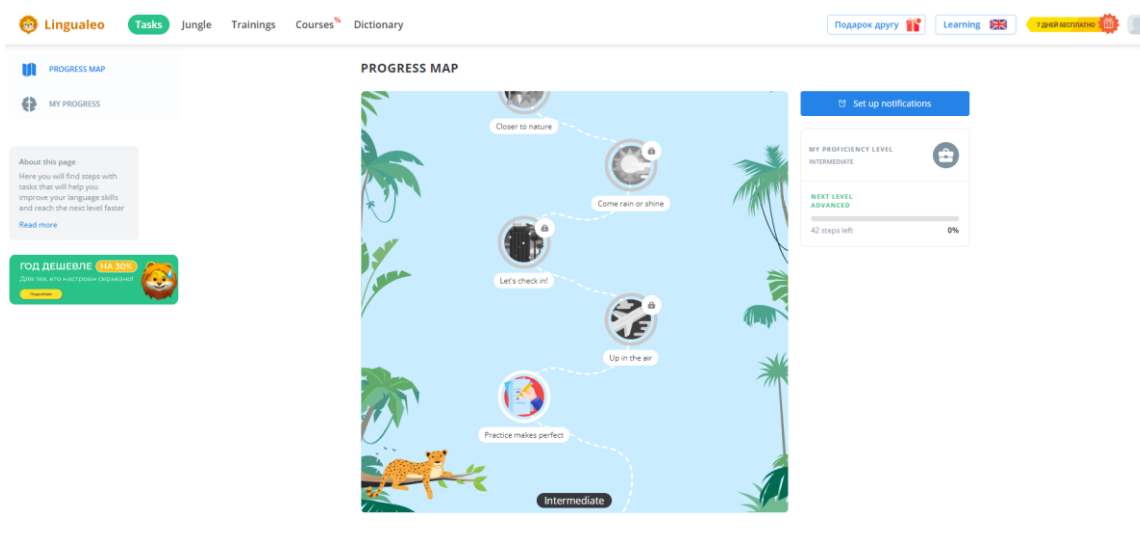


Figure 4: Control of the learning process

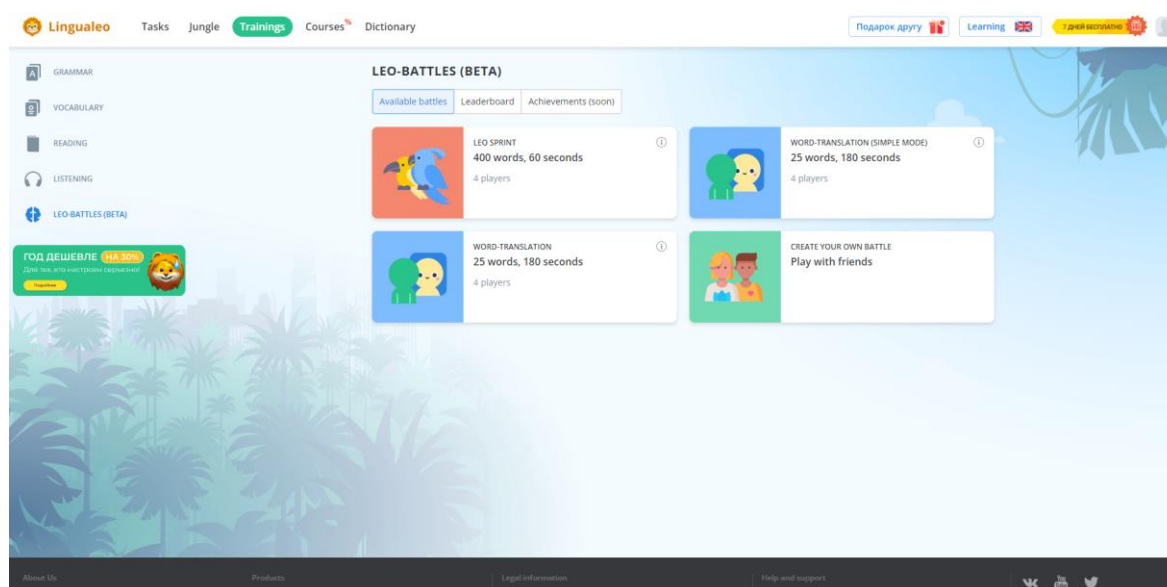


Figure 5: Learning games

Based on all the above, this game has an incredible popularity among users, as it is one of the best for learning English. Thus, the game proves the effectiveness of using gamification in education.

This study shows that gamification will have a positive effect in teaching information technology. It proposed to develop software that teaches the basics of programming based on the technology of gamification of education.

Gamification of the process of learning the basics of programming will have a positive impact on the overall picture of the material studied by students. Gamification will increase interest in the educational material, as well as teach you to delve deeper into the subject and learn until you completely memorize information.

To train in technical specialties, the gamification process will need to optimize. Using the previously studied features of teaching technical specialties, it can be determined that the gamification process will build with the help of information technology, based on audiovisual and video teaching methods.

To implement the learning process, visual learning with the help of mental maps will take as an

example. The gamification algorithm will adapt to the player's success, starting with the most important link of training and its gradual transition to the rest of its branches. For example, for the most basic training, let us take parts of a personal computer. To begin with, the following list of words will enter, which the player will learn as the game progresses:

- Computer science;
- Algorithm;
- Calculating machine;
- System unit;
- Screen;
- Keyboard;
- Personal computer;
- Hardware;
- Software;
- Operating system;
- Memory;
- RAM (Random access memory).

These words are the most used words in the IT field. Before studying them, the user will give the opportunity to see these objects in pictures, listen to their names, as well as learn about the scope of these objects and their definition. Such an audiovisual method will allow you to memorize them faster.

5. Development of gamification software for teaching technical specialties

A game, an application for Android-based mobile devices, was chosen for the development of gamification. The game was developed on a special Unity game engine, using the C# programming language.

The game is to complete as many levels as possible and learn as many words as possible, which means that a platformer will be a more suitable genre of the game. This is a genre of computer games in which the main feature of the gameplay is jumping on platforms, climbing ladders, collecting items usually needed to complete a level. The number of words studied by the user also influenced the decision. For the convenience of studying, it was necessary to have many levels that would allow covering a larger amount of information from the terminology of information technology.

Now there was a choice of the player for whom the user would play. Using an open library of assets and because of the training of terminology related to information technology, the choice fell on a small robot. The search criteria were as simple as possible: animation, attitude to the IT sphere, open-source code. For this reason, this character, called RobotBoy, chosen (Fig. 6).



Figure 6: RobotBoy character

After choosing the main character, work began on the implementation of the terminology training functionality. A system created to collect oil drops (bonuses, stars), without which he would not have been able to pass the level. The goal of the game was not to let the little work rust and try to collect all the drops of oil so that he could lubricate all his mechanisms.

At each level, three drops of oil created which the robot had to collect. They performed the function of bonuses or "stars", which took roots for all standard platformer games. In addition, the user had three lives that he could waste by bumping into various obstacles that scattered in different corners of the game and looked completely different. Lights, abysses, saws, etc. functioned as

enemies for little Rusty.

The name of the game was composed from the main character of the game and its purpose, collecting oil. The game called "Rusty Rusty", which implies a little wordplay. The first word, translated from English, means "rusty", and the second already means the name of this little robot. That is, the game called "Rusty Rusty". And then the goal of the game is immediately visible – not to let Rusty rust by collecting oil droplets.

Regarding the study of terms, three terms in English hidden at each level. They were right under the bonuses (oil drops), and the catch of each level was based on collecting at least one bonus, which means that the user will not be able to go to the next level without collecting at least one bonus (oil drops, words). After the user collects all three bonuses, a window opens for a few seconds, in which there is a translation of all the words that he collected into Russian and Kazakh. And only after collecting these words, an additional item opens in the menu, in which the user can find out the translations of all the collected words and their brief description.

This approach to learning terminology motivates the player to pass all levels collecting all bonuses, despite all the difficulties. As the levels progress, the player offered more complex words. This allows you not to lose interest in the game and try to pass as many levels as possible and learn as many words as possible.

The game has a favorable interface that will appeal to both adults and children. It does not scare away some sharp moments of graphics, but on the contrary, the sound and graphic design makes you walk through the world of the game and find out more words, thereby fulfilling the main task of the developer - to train the user.

The very first level of the game (Fig. 7) shows the user the main points of the game control, the left-right arrows mean moving along the X axis, the up arrow means jumping. The first level consists of learning how to use, it will not be difficult to pass it even for the ineptest player.



Figure 7: Level 1

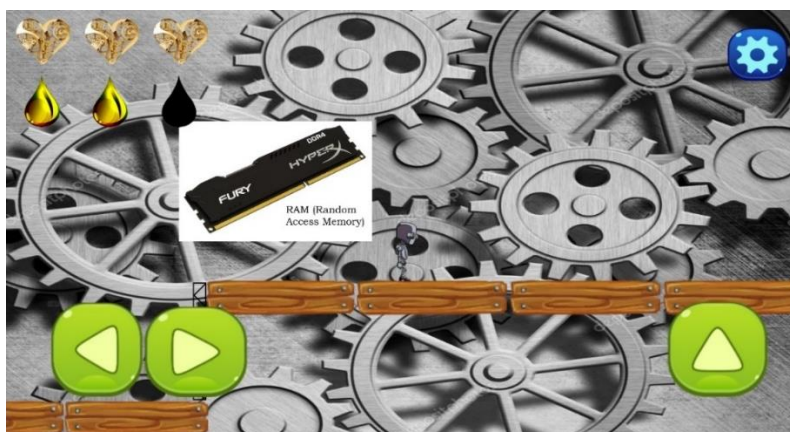


Figure 8: Collecting bonuses

After collecting a drop of oil, the user shown a word in English, taken from information technology, and its graphical representation, so that it is easier to connect the word and its representation in his head. (fig. 8).

After collecting all three bonuses, the player shown a window with translations of words into Russian and Kazakh (Latin) languages (Fig. 9). As well as words with their full explanation opened in the Dictionary menu section.

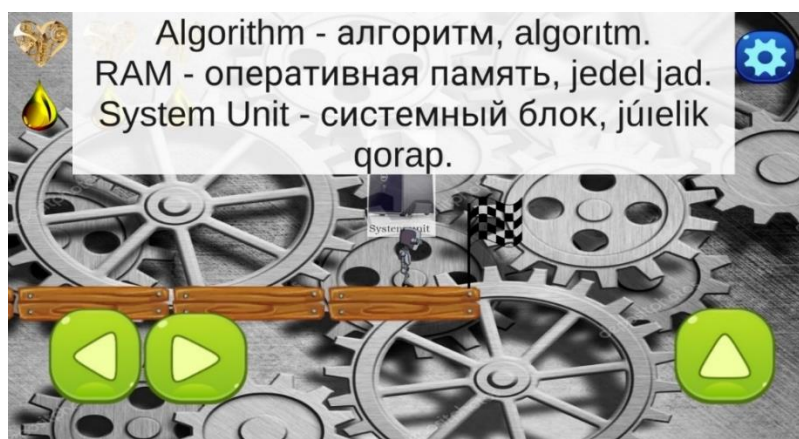


Figure 9: Word translation

As you can see in the photos, for people to understand how they look when collecting terms, photos added next to the names. This simplified the process of memorizing unfamiliar words, which had a positive effect on the overall picture of learning.

The entire game code written on a game engine called Unity (Fig. 10). This engine is extremely popular in the development of 2D games. Unity also equipped with a free store of various pieces of code, a training service, and ready-made projects, which noticeably accelerated the development process of the Rusty Rusty game.

Unity is a multifunctional platform that has full functionality for working with two-dimensional graphics and creating games for any platform, including mobile phones. This allows you to develop full-fledged gaming applications of any scale for any mobile device, as well as immediately export ready-made installation files to the Google Play app store.

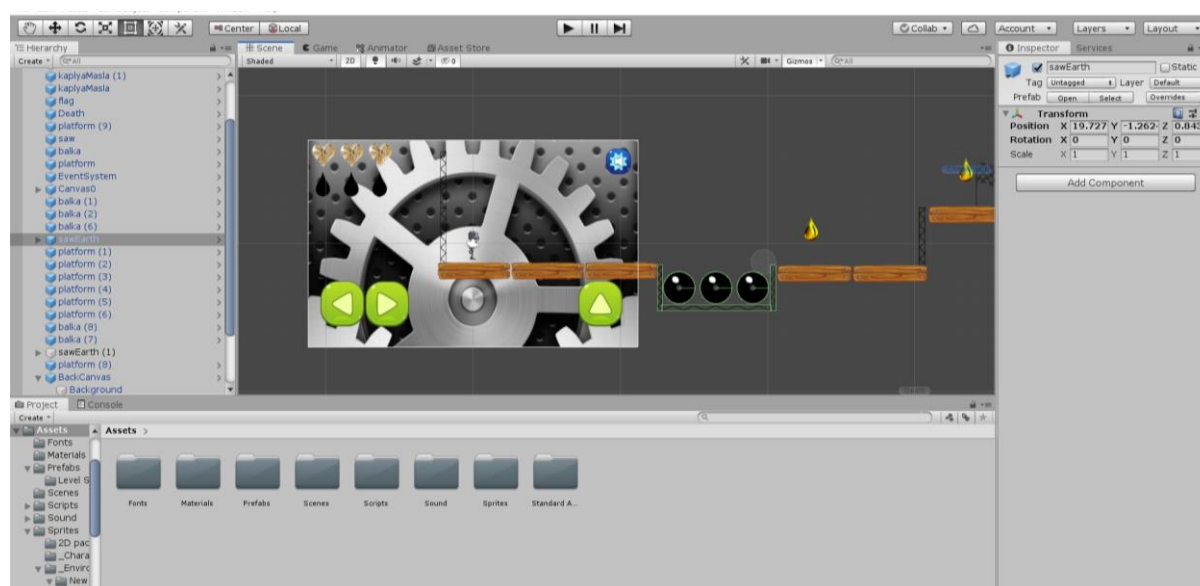
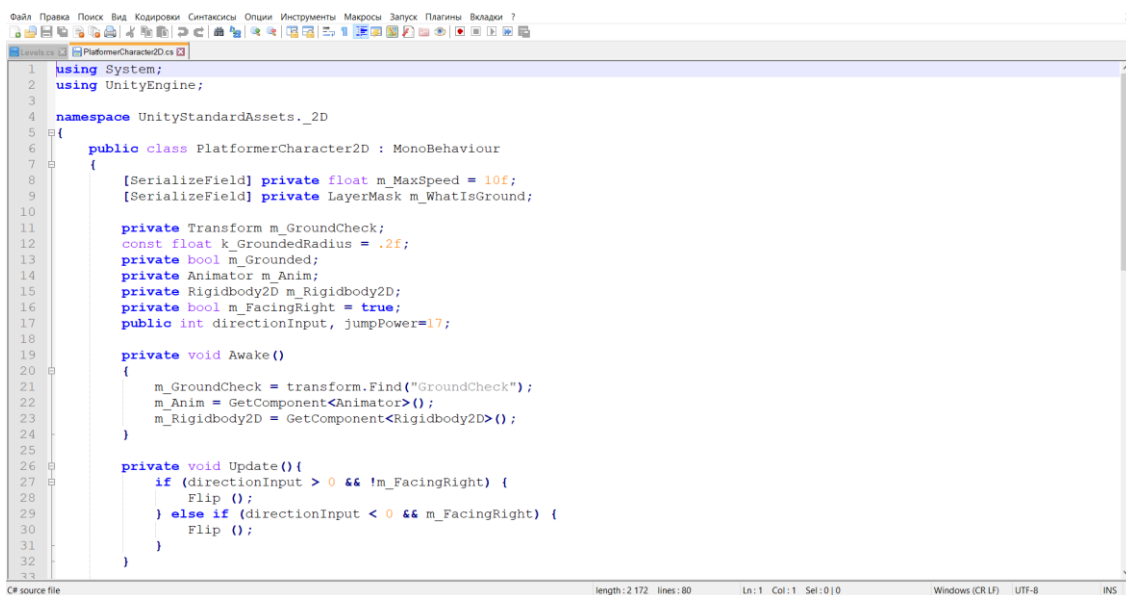


Figure 10: Unity Game Engine

The entire player movement code and bonus collection written in C# (Fig. 11). A multifunctional programming language that made it possible to implement any idea that required when creating a game application.



```
1 using System;
2 using UnityEngine;
3
4 namespace UnityStandardAssets._2D
5 {
6     public class PlatformerCharacter2D : MonoBehaviour
7     {
8         [SerializeField] private float m_MaxSpeed = 10f;
9         [SerializeField] private LayerMask m_WhatIsGround;
10
11         private Transform m_GroundCheck;
12         const float k_GroundedRadius = .2f;
13         private bool m_Grounded;
14         private Animator m_Anim;
15         private Rigidbody2D m_Rigidbody2D;
16         private bool m_FacingRight = true;
17         public int directionInput, jumpPower=17;
18
19         private void Awake()
20         {
21             m_GroundCheck = transform.Find("GroundCheck");
22             m_Anim = GetComponent<Animator>();
23             m_Rigidbody2D = GetComponent<Rigidbody2D>();
24         }
25
26         private void Update() {
27             if (directionInput > 0 && !m_FacingRight) {
28                 Flip ();
29             } else if (directionInput < 0 && m_FacingRight) {
30                 Flip ();
31             }
32         }
33     }
34 }
```

Figure 11: C# programming language

6. Testing the game, releasing the game for general use

After the development of the game, it was fully tested for errors and improved every time. After all the requirements of the testers considered, a beta test launched, a test of the game by a small group of real users. Minor design flaws identified, and soon everything considered and corrected. Thus, the necessary data obtained from the first users of the game. After a week of using the game, positive feedback received about the system of teaching terms through unobtrusive study during the game.

As soon as the game was ready to launch, it sent to the Android games and applications platform - Play Market.

The game released under the name Rusty Rusty, by the developer of Binary Approach. The game can be found in the app store, learn about reviews, system requirements and description. Now, the game has gained more than one hundred downloads and an overall rating of 4.7.

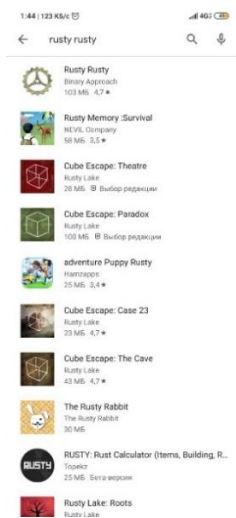


Figure 12: Search for a game in the app store

To search for a game in the app store, enter the name of the game "Rusty Rusty" in the search (Fig. 12) and click on the first link to go to the screen shown in Fig. 13.

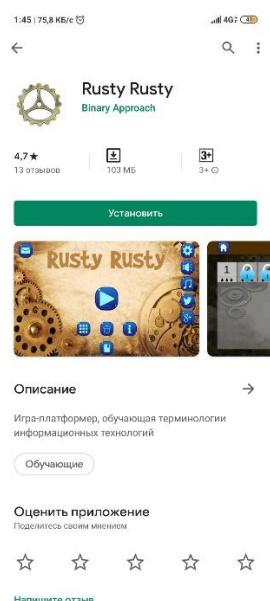


Figure 13: Rusty Rusty Game

The game has already received the first reviews and is freely available in the application marketplace.

7. Conclusion

Kazakhstan is one of the most developed countries in terms of information technology. However, with the growth of technology, there is an acute lack of knowledge in this area. This can be seen by the significant growth in the capabilities of organizations in the field of IT and the weak level of training in the country's schools. This is one of the most acute problems of our century. To solve this problem, it is necessary to modernize the system of teaching technical specialties, as well as to optimize teaching methods with the realities of our time, using the most modern technologies of teaching and memorizing information.

According to the study, the most effective teaching method is polymodal and video teaching methods, since they directly affect different senses at once, allowing the student to understand the subject more deeply and remember information. These methods work well in conjunction with information technology. It is much easier to use audiovisual teaching methods in the application than to try to show this topic of the board. It is for this reason that the most effective way to teach technical specialties was to introduce them into software with learning gamification algorithms.

In this article, the peculiarity of distance learning through gamification of the educational programs familiar to us revealed. Gamification allows you to remotely get an education in the right direction, while maintaining motivation to study and not needing the personal presence of a student in an educational institution, and not needing a teacher, since the whole gamification process often takes place in the example of self-education.

The project offers a remote solution to the problem, based on the gamification of learning available anywhere in the country, regardless of the availability of high-quality educational institutions nearby.

In our world, it is extremely important to gain knowledge that will help us get an excellent job. Unfortunately, not everyone has access to all the latest technologies that appear daily. Children in remote corners of the country do not even know about such a science as "Computer Science". To raise children's awareness and involve the younger generation in the science of information technology, it was necessary to create something incredible.

During the research, the game “Rusty Rusty” developed, which allows children to get the necessary knowledge from the field of information technology through the game. The words that the game taught taken from the most used terms from the Computer Science section.

At the end of the game's development, alpha and beta tests conducted to identify the presence of errors. After completing all the tests, the game put on the Play Market application market. Judging by the feedback from the players, it was an interesting project that allowed you to learn something new and at the same time without spending a second on training.

The effectiveness of using gaming applications as teaching any sciences, in particular computer science, has proven. According to the study, it is easier for people to perceive and remember all information through audiovisual and kinesthetic channels, and entertainment applications are more suitable for this.

Based on the results of the project, it proposed to study and propose a completely new and highly effective method of teaching technical specialties through the introduction of gaming applications, which will serve to improve the quality of education in the country.

8. References

[1] Official information resource of the Prime Minister of the Republic of Kazakhstan. Grants in IT and de—bureaucratization - how digitalization is developing in Kazakhstan. 20 May 2022. URL: <https://primeminister.kz/ru/news/reviews/granty-v-it-i-debyurokratizaciya-kak-razvivaetsya-cifrovizaciya-v-kazahstane-2344022>.

[2] Jane Wakefield. Children spend six hours or more a day on screen. 27 Mart 2015. URL: <https://www.bbc.com/news/technology-32067158>.

[3] Mediacom. Connected Kids Report. Trend Watch 2018. URL: <https://groupmp15170118135410.blob.core.windows.net/cmscontent/2018/09/Connected-Kids-Trends-Watch-2018.pdf>.

[4] Mediacom. Connected Kids Report. September 2019. URL: https://groupmp15170118135410.blob.core.windows.net/cmscontent/2019/10/Connected-Kids-Report-2019-V2_compressed.pdf.

[5] Sadovskaya Irina Lvovna, Teaching methods: a new concept // Bulletin of the KSPU named after V.P. Astafyev. 2007. №1. URL: <https://cyberleninka.ru/article/n/metody-obucheniya-novaya-kontseptsiya>.

[6] Julia Alexandrovna Herzen. The learning process // Educational portal "Guide". — Date of the last update of the article: 28.03.2022. URL: https://spravochnick.ru/pedagogika/process_obucheniya/.

[7] Nina Afonina. Classification of teaching methods by the source of knowledge // Educational portal "Guide". — Date of the last update of the article: 30.08.2022. URL: https://spravochnick.ru/pedagogika/klassifikaciya_metodov_obucheniya_po_istochniku_znaniy.

[8] I. V. Bazhenova Visualization of knowledge as a method of cognitive approach to teaching programming // Reshetnev readings. 2014. №18. URL: <https://cyberleninka.ru/article/n/vizualizatsiya-znaniya-kak-metod-kognitivnogo-podhoda-k-obuchenyu-programirovaniyu>.

[9] Jared R. Chapman & Peter J. Rich (2018) Does educational gamification improve students' motivation? If so, which game elements work best? // Journal of Education for Business, 93:7, 315-322, DOI: 10.1080/08832323.2018.1490687. URL: <https://www.tandfonline.com/doi/abs/10.1080/08832323.2018.1490687>.

[10] Ji-Won Oak1 and Jae-Hwan Bae. Smart Multiplatform-Based CPR Game App Design. Advanced Science and Technology Letters Vol.39 (Games and Graphics 2013), pp.20-23 <http://dx.doi.org/10.14257/astl.2013.39.04> URL: https://web.archive.org/web/20180601222239id_/http://onlinepresent.org/proceedings/vol39_2013/4.pdf.

Development of the Multispectral Microcontroller System for Analyzing Air Quality for the Presence of the Hazardous Gas Mixtures

Zhanna Mukanova¹, Sabyrzhan Atanov², and Marat Baydeldinov²

¹ Turan University, Almaty, Kazakhstan

² L.N. Gumilyov Eurasian National University, Astana, Kazakhstan

Abstract

Air pollution is one of the important problems in the world, especially in the urban areas of developing countries. The use of the gas analyzers allows timely determination of the quantitative or qualitative composition of the analyte based on the measurement of parameters characterizing its physical or physicochemical properties. The paper presents a diagram of a simple and budgetary device that allows you to quickly conduct an express test of air mixtures for the hazardous gas presence. To test the scheme, a number of experiments were carried out, where a sample of air in the laboratory room, carbon dioxide (CO₂) and a mixture consisting of pure oxygen (O₂) with nitrogen (N₂) in a ratio of 9:1 were supplied in turn to the container for the circulation of the analyzed air. The results of the experiments are presented in the graphs.

Keywords

Gas analyzer, infrared radiation, sensor, security, high-frequency scanning

1. Introduction

With the growth of the technological process in the modern world, the number of industrial enterprises is increasing, where the level of safety of must meet high standards [1, 2]. In order to comply with safety and fire regulations, factories and enterprises must be equipped with automatically operating gas analyzers that ring an alarm (light or sound) in advance, that is, before reaching the gas content corresponding to the lower concentration limit of ignition.

The use of the gas analyzers allows timely determination of the quantitative or qualitative composition of the analyte based on the measurement of parameters characterizing its physical or physicochemical properties.

Most molecules can absorb infrared radiation. This process is possible when the radiation wavelength coincides with the natural vibration frequency of the molecules and, as a result, the energy state of the molecule changes. Thus, when radiation is absorbed at a certain wavelength, the amplitude of the atomic vibrations increases. It follows, when IR radiation is absorbed, the temperature of the gas rises. The amount of the absorbed infrared radiation is proportional to the concentration of the gas.

It should be noted that IR radiation interacts with dipole gas molecules. A molecular dipole is formed when the vibrations of a molecule are not symmetrical with respect to different atoms, or the atoms are arranged in an asymmetric manner.

As they vibrate, the atoms deform the interatomic bonds, which form a dipole or multipole moment, as a result of which they bend, stretch, or twist.

From the foregoing, it follows that all gases whose molecules consist of two or more different atoms have the ability to absorb infrared waves. For example, methane (CH₄), carbon monoxide (CO), carbon dioxide (CO₂), etc. The listed gases are dangerous and toxic to humans [3].

There are specialized sensors for detecting one particular type of gas. For example, in [4], when conducting experimental tests of the developed software application, MQ-135 and ME2-O2-Φ20 sensors were used. The MQ-135 gas sensor module is used to detect carbon dioxide. It has high sensitivity and short response time. The ME2-O2-F20 sensor is designed to test the oxygen concentration in the air, which is based on the principle of an electrochemical cell. However, the

development or adjustment of sensors to detect several types of gas mixtures can increase the scope of the device.

The purpose of this work is to create a simple and low-cost device that allows you to quickly conduct an express test of air mixtures for the presence of hazardous gases.

2. Related work

Currently, there are a large number of sensors that are used in various fields of human activity [5, 6, 7, 8]. Gas sensors play a vital role in daily life, especially in environmental monitoring and industry. The development of automobile traffic and high industrialization led to the production of many gases. The concentration of gases must be within acceptable limits when released into the environment. Therefore, the gas must be able to detect, even if its concentration is very small.

There are many different types of sensors for measuring. Sensors are widely used in various fields. For example, in scientific research, testing, quality control, automated control systems and other areas.

Nimmala Harathi and Argha Sarkar [9] presented and analyzed the operation of a gaseous surface acoustic wave sensor based on a nanostructure, designed to detect various concentrations of hydrogen gas with the high sensitivity. I.V. Sherstov, D.B. Kolker et al. [10] studied a photoacoustic methane gas analyzer based on a quantum-cascade laser ($\sim 7.7 \mu\text{m}$), a resonant differential PAR detector, and a sealed gas-filled cuvette. Mohamed A.R.A., Alnaqbi, Muna S. Bufaroosha et al. [11] described the design, optimization and application of a portable gas analyzer based on an amperometric detector, which allows measuring the sulfur dioxide content in a gas stream in real time. In [12], Maosen Xu, Yan Xu et al. described the design of an integrated non-dispersive infrared (NDIR) gas sensor for determining the respiratory coefficient ($q\text{CO}_2$) with ultra-compact dimensions and low power consumption.

The optical-acoustic effect was discovered in 1880 by Bell, Tyndel and Roentgen. Optical methods of gas analysis are among the most selective and sensitive. One of the first places among them in this regard is occupied by the optical-acoustic method, the selectivity of which, unlike conventional spectroscopic methods, is achieved without spectral decomposition, using selective optical-acoustic receivers that use the specificity of infrared absorption spectra of gaseous, vaporous and liquid substances.

The paper presents a diagram of an optical-acoustic gas analyzer. Increasing the accuracy of measuring the concentration of gas compounds occurs using spectral sensors and temperature and pressure sensors, combined with visible, ultraviolet and infrared LEDs. This scheme greatly simplifies the design of the gas analyzer and improves the measurement accuracy.

3. Methods

As part of the development of a prototype gas analyzer designed to determine the concentration of multicomponent gas mixtures in air in laboratory and industrial conditions, AS7265x and BMP180 sensors were used.

The SparkFun Triad Spectroscopic Sensor is a powerful optical inspection sensor, also known as a spectrophotometer [13]. Three AS7265x spectral sensors are combined together with UV and IR LEDs to illuminate and test various surfaces for light spectroscopy (Figure 1). The triad consists of three sensors: AS72651, AS72652 and AS72653 can detect light from 410nm (UV) to 940nm (IR). Each sensor contains a 16-bit resolution A/D converter that integrates the current from each channel's photodiode. Upon completion of the conversion cycle, the integral result is transferred to the appropriate data registers. Data transfer is carried out with double buffering to ensure their integrity [14].

Due to the use of spectral sensors combined with ultraviolet and infrared LEDs, it greatly simplifies the design of the gas analyzer and improves measurement accuracy.

The sensors are focused on applications in areas such as:

- assessment of the metal temperature during smelting, forging, heat treatment;
- sorting seeds by color in agriculture;
- non-contact tissue analysis in medicine;
- analysis of the composition of substances in chemistry;
- control of freshness of products;
- water quality control;
- assessment of the state of vegetation cover in agriculture and ecology;
- analysis of the spectral composition of light sources.

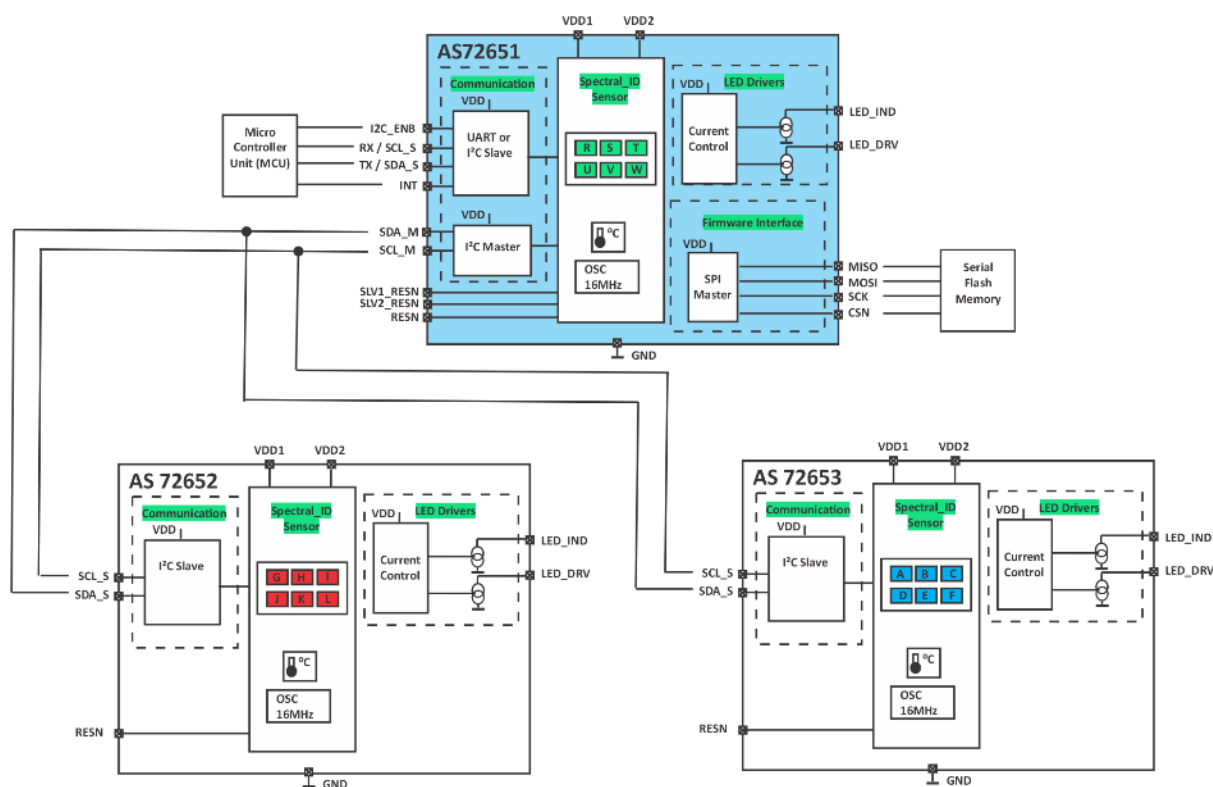


Figure 1: AS7265x Functional Block Diagram

The BMP180 sensor allows you to measure absolute atmospheric pressure in the range of 300...1100hPa (+9000....-500 meters above sea level). The module can be used in home weather stations, aircraft, as an altimeter, etc. The GY-68 module on the BMP180 chip combines an atmospheric pressure sensor and a thermometer.

To create a test device, an experimental stand was built, which is a small closed gas dynamic system for collecting and quantifying the concentration of gases in air mixtures. The system consists of the following main components: an Arduino microprocessor board, AS7265x and BMP180 sensors, a sample air circulation container with a parabolic diffuser, a gas generation tank and a circulation pump. The layout of the experimental stand is shown in Figure 2.

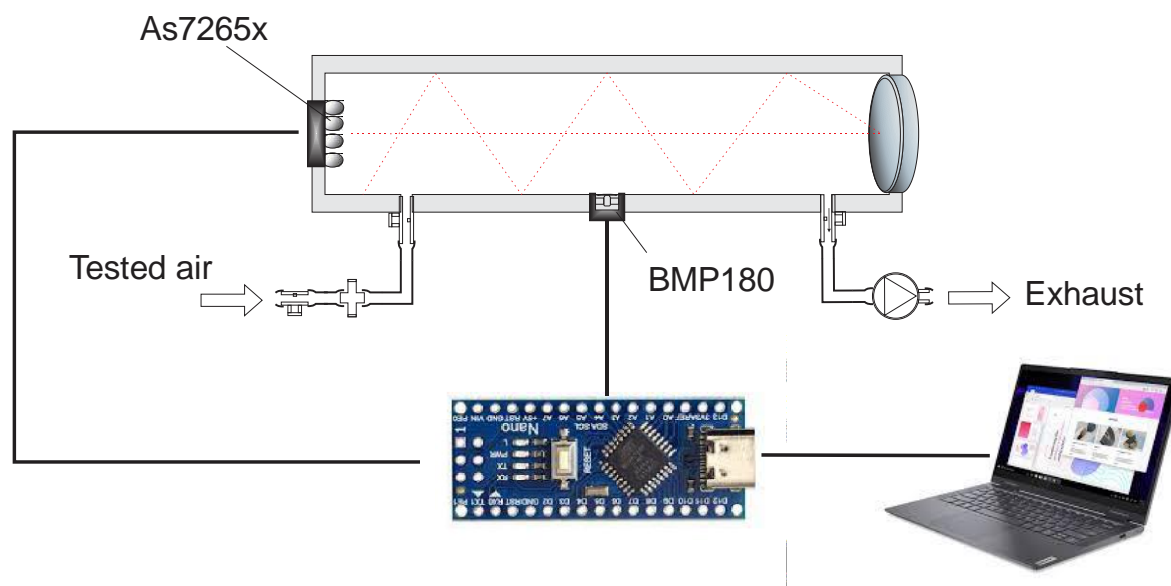


Figure 2: The experimental stand scheme

The use of the parabolic emitter in this design will make it possible to provide multi-path passage of rays through the gas chamber, thereby increasing the gas pressure in the chamber. The optical-acoustic method is very versatile: it allows analyzing all gaseous substances, with the exception of single-element ones, in the presence of appropriate IR emitters.

The amplitude P of pressure fluctuations in an optical-acoustic chamber can be approximately expressed by formula (1) (for an infinitely large acoustic resistance of the chamber and microphone walls):

$$P = \frac{P_0 Q}{TG \sqrt{1 + \omega^2 C_V^2 / G^2}}, \quad (1)$$

where P_0 is the static gas pressure in the chamber;

T is thermodynamic temperature;

G is thermal conductivity of the gas-chamber system;

C_V^2 is the heat capacity of the gas;

ω is circular frequency of modulation;

Q is the amplitude of the heat flux absorbed in the chamber.

The Arduino board was used as a microcontroller to receive and transmit digitized sensor signal data to a laptop via a USB port. The output of digital voltage data obtained from the sensor with a frequency of 200 Hz was recorded in an ASCII text file.

4. Results

To carry out experimental tests of the stand, a sample of air in the laboratory room, carbon dioxide (CO_2) and a mixture consisting of pure oxygen (O_2) with nitrogen (N_2) in a ratio of 9:1 were supplied in turn to the container for the circulation of the analyzed air. The experiments were carried out in a well-ventilated room 20 m² in size. Figure 3 shows a general view of a laboratory stand for the collection and analysis of gases.



Figure 3: General view of a laboratory bench for collecting and analyzing gases

The tested gas mixture, which is released during the reaction under the pressure of the circulation pump, gradually moves into the container for collecting the analyzed gas. The AS7265x and BMP180 sensors are polled at a frequency of 200 Hz and the data, after digital processing in the Arduino, is sent to the laptop. Figure 4-6 shows graphs of the data obtained from the experiment.

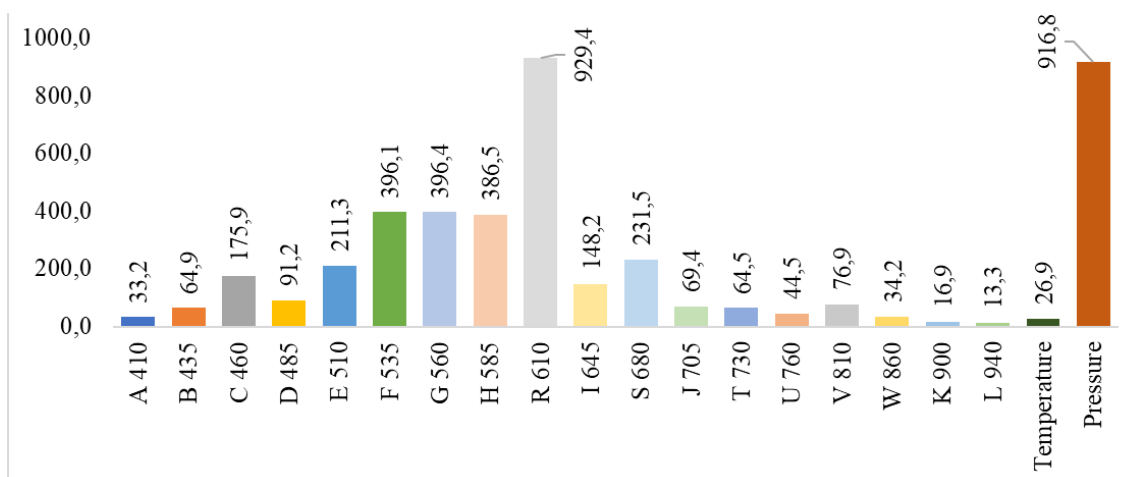


Figure 4: Measurement results of the air sample in the laboratory room

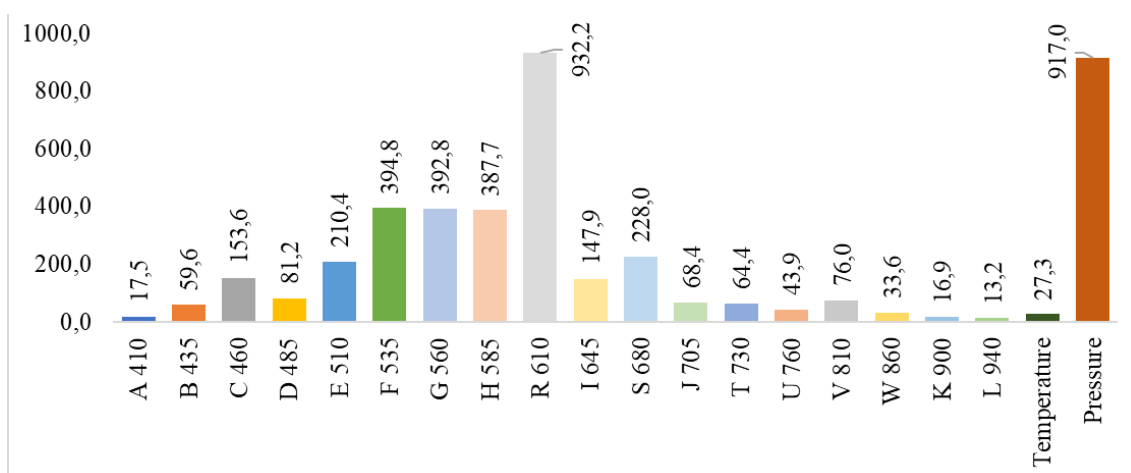


Figure 5: Carbon Dioxide (CO₂) Measurement Results

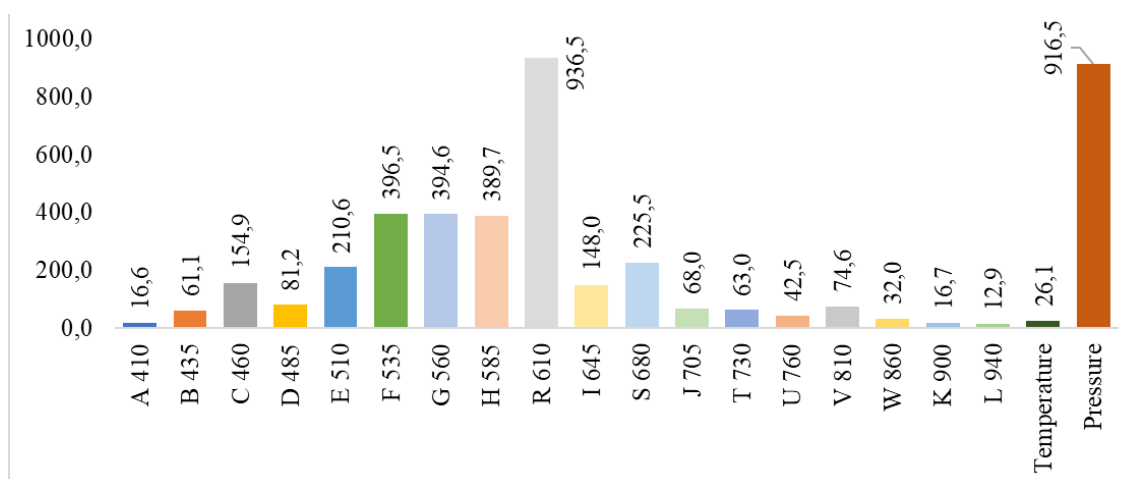


Figure 6: Measurement results of a mixture of pure oxygen (O₂) with nitrogen (N₂)

It can be seen from the graphs that the wavelength of 610 nm has the highest peak response for all the studied samples of gas mixtures.

5. Conclusions

For the experiments, an experimental bench was built, which is a small closed gas dynamic system for collecting and quantifying the concentration of gases in air mixtures. The system includes an Arduino microprocessor board, AS7265x and BMP180 sensors, a container for the circulation of the analyzed air with a parabolic diffuser, a container for generating gases and a circulation pump. The scheme of the gas analyzer proposed in this research was patented in the Kazakhstan patent bureau [15].

To collect data, a sample of the air in the laboratory room, carbon dioxide (CO₂) and a mixture consisting of pure oxygen (O₂) with nitrogen (N₂) in a ratio of 9:1 were supplied in turn to the container for the circulation of the analyzed air. The experiments were carried out in a well-ventilated room 20 m² in size.

As a result of the experiments, it can be concluded that the sensors used respond to changes in the composition of air mixtures and the proposed scheme can be used to assemble a simple and low-cost device that allows you to quickly conduct an express test of air mixtures for the hazardous gases presence.

The technical result of the presented scheme of the gas analyzer is to increase the accuracy of measuring the concentration of gas compounds through the use of a combination of spectral sensors and pressure and temperature sensors. For a more detailed study, it is planned to increase the number of tested gases. It is also planned to create and train an artificial neural network to automatically determine the composition of air mixtures and increase the amount of detected gases.

6. References

[1] Z. Wang, W. Wei. "Effects of modifying industrial plant configuration on reducing air pollution-induced agricultural loss. " *Journal of Cleaner Production* 277 (2020): 124046. <https://doi.org/10.1016/j.jclepro.2020.124046>.

[2] X. Jiang, P. Zhang, J. Huang. "Prediction Method of Environmental Pollution in Smart City Based on Neural Network Technology." *Sustainable Computing: Informatics and Systems*, 2022: 100799. <https://doi.org/10.1016/j.suscom.2022.100799>.

[3] A. I. Kitajgorodskij. "Vvedenie v fiziku" *Izd-vo «Nauka», Glavnaya redakciya fiziko-matematicheskoy literatury*, 197: 684.

[4] Z. Mukanova, S. Atanov and M. Jamshidi, Features of Hardware and Software Smoothing of Experimental Data of Gas Sensors, 2021 IEEE International Conference on Smart Information Systems and Technologies (SIST), 2021, pp. 1-6, doi: 10.1109/SIST50301.2021.9465981.

[5] Zh. Nurlan, T. Zhukabayeva, M. Othman, A. Adamova, N. Zhakiyev. "Wireless Sensor Network as a Mesh: Vision and Challenges." IEEE Access (2021): 1-1. doi: 10.1109/ACCESS.2021.3137341.

[6] A.K. Kereyev, S.K. Atanov, K.P. Aman, Z. Kulmagambetova, B. Kulzhagarova. "Navigation system based on Bluetooth beacons: implementation and experimental estimation." Journal of Theoretical and Applied Information Technology 98 (2020). <http://www.jatit.org/volumes/Vol98No8/6Vol98No8.pdf>.

[7] K.K. Adilzhan, S. Atanov, T.Z. Timur. The Usage of Extended Kalman Filter to Increase Navigation Accuracy of Mobile Units in Closed Spaces, SIST 2021 - 2021 IEEE International Conference on Smart Information Systems and Technologies, 2021, 9465903.

[8] S.S. Brimzhanova, S.K. Atanov, K. Moldamurat, D.M. Kalmanova, T. Tabys. Problems of detecting fuzzy duplicates. In Proceedings of the 5th International Conference on Engineering and MIS (ICEMIS '19). Association for Computing Machinery, New York, NY, USA, Article 23 (2019), pp. 1–5. <https://doi.org/10.1145/3330431.3330455>.

[9] H. Nimmala, A. Sarkar. "TiO₂ based surface acoustic wave gas sensor with modified electrode dimensions for enhanced H₂ sensing application." International Journal of nano dimension 12 (2021): 83-89.

[10] I.V. Sherstov, D.B. Kolker, A.A. Boyko, V.A. Vasiliev, R.V. Pustovalova. Methane photo-acoustic gas analyzer based on 7.7- μ m quantum cascade laser. Infrared Physics & Technology 117 (2021) 103858. <https://doi.org/10.1016/j.infrared.2021.103858>.

[11] A.R. Mohamed, A. Alnaqbi, M.S. Bufaroosha, M.H. Al-Marzouqi, S.A.M. Marzouk. "Portable analyzer for continuous monitoring of sulfur dioxide in gas stream based on amperometric detection and stabilized gravity-driven flow." Sensors and Actuators B: Chemical 225 (2016): 24-33. <https://doi.org/10.1016/j.snb.2015.11.008>.

[12] M. Xu, Y. Xu, J. Tao, Y. Li, Q. Kang, D. Shu, T. Li, Y. Liu. "A design of an ultra-compact infrared gas sensor for respiratory quotient (qCO₂) detection. " Sensors and Actuators A: Physical 331 (2021): 112953. <https://doi.org/10.1016/j.sna.2021.112953>.

[13] <https://www.sparkfun.com/>.

[14] <https://ams.com/>.

[15] Zh.A. Mukanov, S.K. Atanov. "Gazoanalizator" Patent na poleznuyu model'. №5141 from 10.07.2020.

Development of a Mobile Application "ZhEDUniver" Integrated with the Smart Zhetysu Platform for Students of Zhetysu University Named After I. Zhansugurov

Shynar Assylbekova¹, Bagdat Serikov¹, Yerasyl Zhailau¹, Almas Suleimenov¹ and Daulet Batyrbekov¹

¹ Zhetysu university named after I. Zhansugurov, Taldykorgan, Kazakhstan

Abstract

An important problem in the educational process in higher education is the task of automating the educational process. Currently, the educational process is widely provided with information systems. Automation systems in higher education institutions are being implemented primarily to improve the quality of education. In particular, the majority of students are leaders in the effective use of mobile technologies and mobile devices in the educational process.

The article is devoted to the development of a mobile application of Zhetysu University named after I. Zhansugurov, integrated with the Smart Zhetysu platform. To achieve the goal, the subject area and the object area were analyzed, sketches and a product project were prepared, and a mobile application filled with data from the official website of Zhetysu University was implemented.

The article discusses the results of the developed mobile app and presents the results of a study that surveyed students on the implementation of the mobile app.

A step-by-step algorithm for creating an application is presented. The application is tested on different devices running the Android operating system, with different display diagonal.

Keywords

Development, mobile application, mobile device, Android operating system, electronic journal

1. Introduction

In the era of Industrial Revolution 4.0, marked by ongoing automation of traditional manufacturing and industrial practices and using modern smart technology, smartphones and their applications (apps) are an inseparable part of our lives [1]. The number of mobile app users increases every year. This has led to an extreme demand for developing software that runs on mobile devices [2].

A mobile application is software designed to work on smartphones, tablets and other mobile devices [3].

The development of mobile applications for higher education institutions will modernize and intensify the educational process. Therefore, it is not surprising at all that mobile apps are now commonly used in higher education, especially thanks to the ubiquity of smartphones, i.e. their use anywhere and at any time, their interactivity and multimodal character of mobile apps [4]. Educational researchers have started exploiting the potential and scope of smart mobile technologies in education [5]. Many studies indicated that mobile devices applications, when designed appropriately, may impact students' academic achievement, which motivates them towards learning [6], [7], [8], [9]. In addition, the creation of mobile applications not only makes the learning process more convenient, but also simplifies the solution of organizational problems. In particular, it will make it possible to instantly exchange information about changes in the schedule, the dates of exams or students' independent work.

Each mobile application for a higher education institution is created taking into account all the peculiarities of the educational process and the specifics of the subjects studied at the university.

The mobile application "ZhEDUniver" is designed to provide effective information support for the management processes of the education system, as well as the management of the educational process of the higher education institution. In other words, it is defined as, "With the use of mobile devices, learners can learn anywhere and at any time" [10].

The mobile application "ZhEDUniver" allows you to provide a set of tasks in the following areas:

- Improving the quality of educational services through the improvement of information and technical support activities for students;
- Increasing students' awareness of the educational process.

"ZhEDUniver" is an application for quick and convenient access to university information resources. Students can view the latest news related to the educational process, view the latest scores in the progress log with the schedule of classes, information on the current assessment, as well as download educational materials. If there are questions from students, an appeal is provided in the form of applications in the ready form of its registration.

The article describes the ways of creating and implementing a mobile application for a university in the educational process.

2. Purpose and objectives of the research

The purpose of this work is to develop a mobile application "ZhEDUniver", integrated with the Smart Zhetysu platform. To achieve this goal, we need to perform the following main tasks:

- Research of the subject area;
- Designing the mobile application;
- Designing the mobile application;
- Modeling the information system;
- Mobile application development using Node JS, MS SQL, HTML, JavaScript, CSS, Java technologies;
- Testing and debugging of the Android mobile application;
- Testing the software on various devices (cell phone, tablet, smartphone).

In addition to the main task, the developer's duties included creating a user-friendly application. It should work continuously, be intuitive and functional.

3. Research methods

Currently, university students use the web version of the Smart Zhetysu platform personal account. When logging in to the web application, learners use different browsers, which provide their own inconveniences. In particular: the browsers use extensions and modules that prevent the execution of some scripts of the web application, files, pop up commercials, news, you need to remember the address of the site, or search for it in a search engine and other inconvenient and uncontrollable functions.

In this regard, a survey of students of 1-4 courses of the specialty "6B06102-Information Systems" for the introduction of a mobile application of the electronic journal in the learning process of the I. Zhansugurov Zhetysu University was conducted. About 100 students took part in the survey. The result of the survey allowed to make changes and additions necessary for the e-journal and check the technical and psychological training of students.

According to the survey, most students use mobile operating systems iPhone OS/iOS (Apple) 26.8% and Android OS 73.2%.

What mobile operating system do you use?

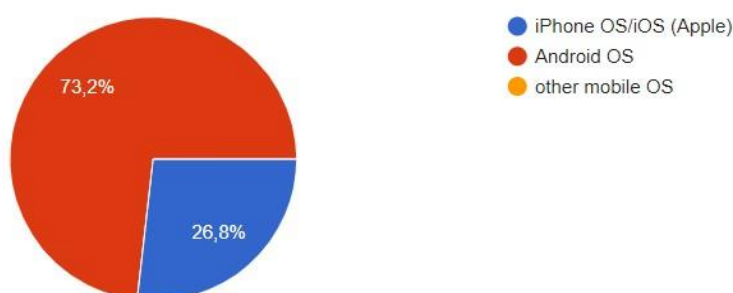


Figure 1: Using mobile operating systems

To the question "Do you want the University to have a mobile application of an electronic journal?", the answers were as follows 99.1% of students answered "yes" and 0.9% "no".

Do you want the university to have a mobile application of an electronic journal?

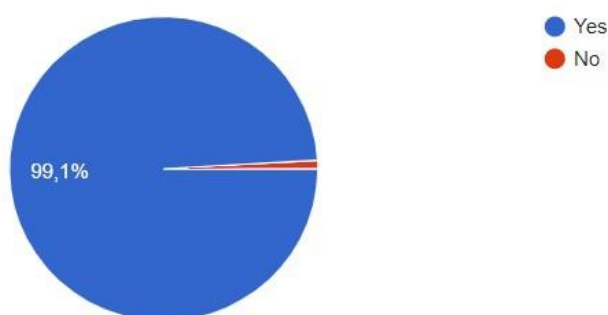


Figure 2: Response to the availability of an electronic journal mobile application

As shown in Figure 3 "How do you feel about the possibility of viewing information related to the educational process (class schedule, exam schedule, academic performance, etc.) through a mobile application?" 87.5% of students answered the question "Yes, it is necessary", 1.8% answered "I do not think it is necessary" and 10.7% answered "Yes, if the students' requests are satisfied".

How do you feel about the possibility of viewing information related to the educational process (class schedule, exam schedule, academic performance, etc.) through a mobile application

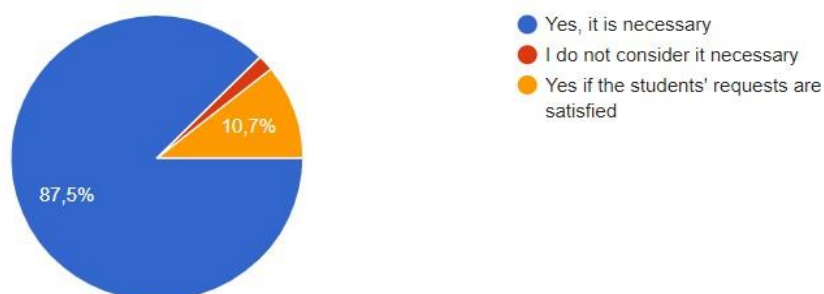


Figure 3: Responses to opportunities to view information related to the educational process

"Do you think a mobile e-journal app is necessary?" to the question, 96.4% said "necessary" and 3.6% said "not necessary."

Do you think a mobile e-journal app is necessary?

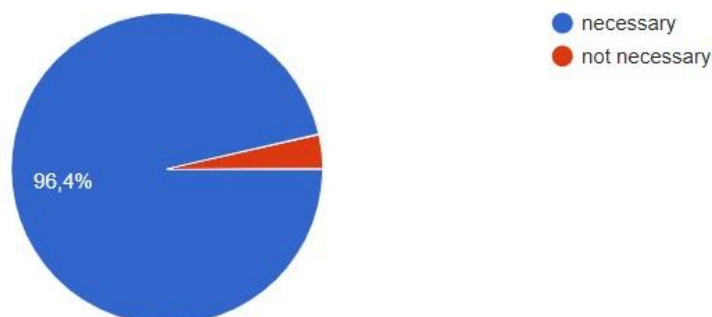


Figure 4: Answers to the question "Do we need a mobile app for the e-journal?"

According to the results of the survey, it was found that the majority of students use the mobile operating system Android and it is necessary to create a mobile application of the electronic journal.

In this regard, it was decided to develop a mobile application of e-journal on the operating system Android, fully incorporating information with the platform Smart Zhetysu. When comparing and analyzing the data, it was found that the vast majority of students are technically and psychologically ready to use the mobile application in learning.

4. Research results

During the implementation of the research, the participants developed a work plan for the development of a mobile application:

1. Subject area research;
2. Designing the mobile application;
3. Modeling of the information system;
4. Mobile application development;
5. Testing a mobile application on Android;
6. Alpha - software testing;
7. Debugging and verification of the software;
8. Testing of the software product;
9. Beta testing of the mobile application.

Collection and analysis of the work of the structural subdivisions of teaching and learning department, project office, registrar's office, distance learning faculty, digitalization and process automation department were conducted. Functionality of Smart Zhetysu, Moodle, Platonus, Univer 2.0 platforms were studied. The analysis of advantages and disadvantages in the work of information systems was carried out. The object of the study is to develop applications for mobile devices on the Android OS.

The subject of the research is the development of a mobile application "ZhEDUniver" for mobile devices and university students on the Android OS.

The data structure in the database platform Smart Zhetysu is constructed.

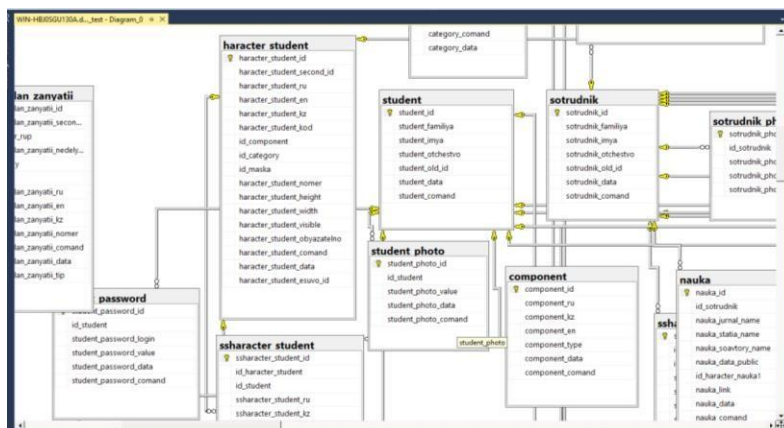


Figure 5: Data structure in the Smart Zhetysu platform database

The information system of the mobile application was integrated with the Smart Zhetysu platform. The research participants created a three-link architecture "User Interface→Application→DMS".

This architecture has the following characteristics:

- tolerability;
- modularity;
- centralization of management;
- reliability;
- easy integration;
- productivity.

Developed modules, as well as the interface of the mobile application. All modules and units are assembled into one project and compiled into a single mobile application. A unique combination of login and password is generated for users. Single login for different categories of users (student, lecturer).

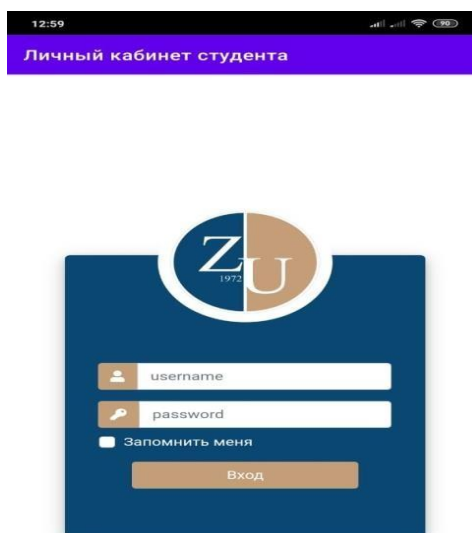


Figure 6: Log in to the system

After identification, the main page "Student's personal account" opens. The center of the page contains a list of available modules. An example of the appearance of the page is shown in Figure 7. The mobile application includes 15 modules necessary for the student's learning process. With the help of the modules listed below, students can get the necessary help or leave a request for any questions related to the educational process.

Modules of the mobile application "ZhEDUniver":

- Electronic journal;
- Examination schedule;
- Applications;
- Online testing;
- Attestation committee;
- Video lectures and electronic textbooks;
- Questionnaire;
- Scholarships/Grants transfers;
- How to protect yourself;
- Directory guide;
- Medical info;
- Questions/Answers;
- Mass discovery online courses;
- Information for learners with the use of distance learning technologies;
- Digital library.



Figure 7: Student's personal account

A module that students use daily and frequently is the electronic journal. The electronic journal includes the following menus:

- Current Control;
- Rating 1;
- Rating 2;
- Summation;
- Schedule;
- Individual curriculum;
- Transcript;
- Registration;
- Academic debt.

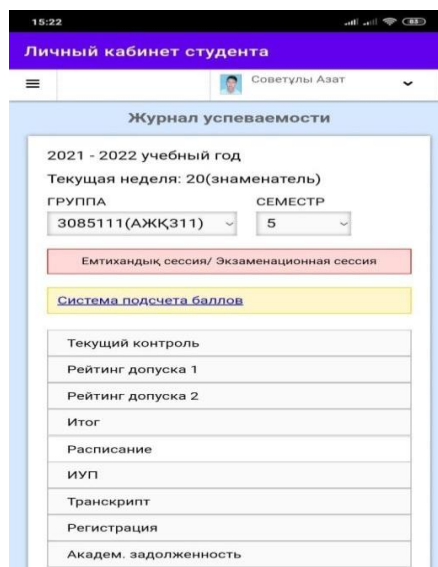


Figure 8: Journal of academic performance

If we briefly talk about the functions that each menu performs:

- The current control provides a display of the student's progress over the entire period of study with the possibility of choosing a semester to view the progress, as well as grouping the results by type of certification. After clicking the current control, the window of electronic educational and methodological support will open. Electronic educational and methodological support - lecturers upload a lecture, practice, laboratory work, student's independent work for 15 weeks and additional materials to this module. Accordingly, students can view and download assignments electronically.

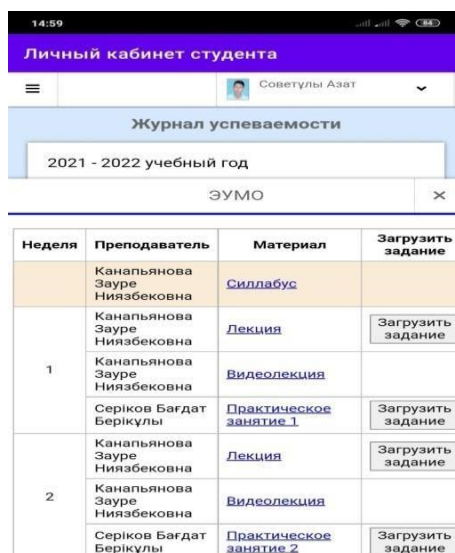


Figure 9: Electronic educational and methodological support

- Rating of admission 1,2 - depending on the semester of training in the studied disciplines in the I, II ratings indicate the points gained by students in each discipline and the total score.
- The final menu indicates the score scored by the student in the I and II ratings for each discipline and the score obtained on the exam. The scores obtained by rating and exam are summed up, and the student's total score for each discipline is indicated.
- The schedules menu contains information about the name of the discipline, the type of classes, the surname, the name of the lecturer, the audience, the time of the lesson.

- "Individual curriculum" is a student's document that contains information about the list and terms of study of academic disciplines selected for mastering.
- "Transcript" is a document of a prescribed form containing a list of completed disciplines for the corresponding period of study, indicating credits and grades in alphabetic and numeric terms. The transcript is an official document of the student.
- In the "Registration for disciplines" menu, students choose a discipline online in accordance with the registration schedule approved by the Dean of the faculty in the "ZhEDUniver" application within the time limits set by the academic calendar.
- Academic debt is an unsatisfactory result of intermediate certification (deuce, non-appearance and non-payment) in one or more disciplines, or not passing the intermediate certification in the absence of valid reasons.

All procedures and functions were tested, as well as the operation of the mobile application interface. During testing and analysis of the algorithms, technical errors were revealed in the codes of MS SQL stored procedures, as well as inefficient algorithms in the project units. As a result of the work performed, all technical errors were eliminated and debugged. Below is a code fragment of 1 error out of 120 technical errors.

```
declare
    @gruppa_id bigint,
    @kontingent_kurs_id bigint,
    @ssharcater_student_zayavlenie_nomer int=(select max(ssharcater_student_zayavlenie_nomer)+1 from
ssharcater_student_zayavlenie where id_harcater_student_zayavlenie=@harcater_student_zayavlenie_id)
    set @ssharcater_student_zayavlenie_nomer=isnull(@ssharcater_student_zayavlenie_nomer,1)
    select top(1) @gruppa_id=gruppa_id,@kontingent_kurs_id=kontingent_kurs_id from
    ucheb_god.kontingent.kontingent_kurs,kurs.gruppa,gruppa_kontingent_kurs
        where id_ucheb_god=ucheb_god_id and
        gruppa.id_kontingent=kontingent_id and
        kontingent_kurs.id_kontingent=kontingent_id and
        gruppa.id_kontingent=kontingent_kurs.id_kontingent and id_kurs=kurs_id and id_gruppa=gruppa_id and
        id_kontingent_kurs=kontingent_kurs_id and gruppa_kontingent_kurs_status=1 and id_student=@student_id and
        kurs_nomer=(2021-ucheb_god_nachalo+1)
        if @gruppa_id is not null and @kontingent_kurs_id is not null and (exists (select * from
        ssharcater_student_zayavlenie where id_student=@student_id
        and id_harcater_student_zayavlenie=@harcater_student_zayavlenie_id and DATEDIFF ( MONTH ,
        ssharcater_student_zayavlenie_user_data, getdate())>=6) or not exists(select * from ssharcater_student_zayavlenie
        where id_student=@student_id and id_harcater_student_zayavlenie=@harcater_student_zayavlenie_id))
begin
insert into
ssharcater_student_zayavlenie(id_harcater_student_zayavlenie,id_gruppa,id_kontingent_kurs,id_st
udent,ssharcater_student_zayavlenie_nomer,ssharcater_student_zayavlenie_status,ssharcater_stud
ent_zayavlenie_src_file,ssharcater_student_zayavlenie_user_data,ssharcater_student_zayavlenie_d
ata,ssharcater_student_zayavlenie_comand)
values(@harcater_student_zayavlenie_id,@gruppa_id,@kontingent_kurs_id,@student_id,
@ssharcater_student_zayavlenie_nomer,0,replace(@ssharcater_student_zayavlenie_src_file,'D:','..\\
'),getdate(),getdate(),1)
end
```

Beta testing and approbation of the software product has been carried out. Changes and additions to the mobile application are made on the recommendation of specialists responsible for the educational process of the university. Technical features:

- Supports operation with Android operating systems;
- "ZhEDUniver" is integrated by Smart Zhetysu platform, displays real-time information to the user;
- Data for mobile application is transferred through dedicated channel via single communication server;
- User password combinations are stored on the university's server.

5. Conclusion

As part of the study, a mobile application was created for the Android operating system, which

allows students to get acquainted with their schedule, attendance and academic performance. The mobile application "ZhEDUniver" is an application that provides students with quick access to digital services and educational content. The availability of the mobile version is widespread among trainees. In the analysis of existing technologies for the development of mobile applications, as well as on the basis of analysis of design specifications, the operating system of the mobile device for the application was selected. Mobile application "ZhEdUniver" is developed in Russian language.

In the course of the research work the set tasks were solved:

- Analyzed the subject area;
- A product project was developed;
- A mobile application running on the Android operating system was created;
- Filling the application with data taken from the university database.

Based on the above, the objectives were achieved, the result of the effective use of the created application among students, the research work has achieved its goal and implemented at the university.

6. Acknowledgments

The study was carried out within the framework of the competition of Zhetysu University named after I. Zhansugurov for grant funding of projects of commercialization of the results of scientific and scientific-technical activity of young scientists "Jas ǵalym". The authors would like to express their gratitude to the team of project participants for their help in preparing and processing the experimental data.

7. References

- [1] B. Klimova. "Evaluating Impact of Mobile Applications on EFL University Learners' Vocabulary Learning – A Review Study", The 11th International Symposium on Frontiers in Ambient and Mobile Systems (FAMS-2021), Procedia Computer Science 184 (2021) 859–864, Warsaw, Poland, March 23-26, 2021, pp.1-6, <https://doi.org/10.1016/j.procs.2021.03.108>.
- [2] A. Kaur, K. Kaur. "Systematic literature review of mobile application development and testing effort estimation", Journal of King Saud University - Computer and Information Sciences (2018), <https://doi.org/10.1016/j.jksuci.2018.11.002>.
- [3] Developed by the Autonomous Nonprofit Organization Russian Quality System "Comparative testing of mobile applications for smartphones" PNST №277-2018, June 26, 2018. (in Russian).
- [4] R. Gafni, D.B. Achituv, G.J. Rachmani. "Learning foreign languages using mobile applications." Journal of Information Technology Education Research, 16, pp. 301-317, 2017.
- [5] P. Punithavathi, S. Geetha. "Disruptive smart mobile pedagogies for engineering education", 9th World Engineering Education Forum 2019, WEEF 2019. Procedia Computer Science 172 (2020) pp.784–790.
- [6] M. Masood, M. Thigambaram. "The usability of mobile applications for pre-schoolers", Procedia-Social and Behavioral Sciences, Volume 197, 25 July 2015, pp.1818-1826.
- [7] H.B. Miller, J.A. Cuevas "Mobile learning and its effects on academic achievement and student motivation in middle grades students", International Journal for the Scholarship of Technology Enhanced Learning, 2017; 1(2): pp.91-110.
- [8] A. Baker, C. Dede, J. Evans "The 8 essentials for mobile learning success in education", Qualcomm Wireless Research, 2014.
- [9] C.K. Joergensen. "Quality panel project" 2010, Social and Health School Greve, Denmark.
- [10] M.L. Crescente, D. Lee. "Critical issues of m-learning: design models, adoption processes, and future trends." Journal of the Chinese institute of industrial engineers, vol. 28, №2, pp. 111- 123, 2011.

Overview of the Current State of Research, Devoted to the Problems of Information Retrieval and Natural Language Processing

Askar Bekishev¹, Kuanysh Nursakitov¹, Saule Kumargazhanova¹ and Aliya Urkumbaeva¹

¹ D. Serikbayev East Kazakhstan Technical University, Ust-Kamenogorsk, Kazakhstan

Abstract

This article analyzes the current state of research on the problems of information retrieval and processing of natural language texts. Particular attention is paid to the natural language collection, the source of which is the Internet. In order to understand the nature, characteristic and distinctive properties of the information contained in such collections, the concept of big data is introduced into the conceptual field of research. The content of this concept is revealed in relation to the use of the collection, on the basis of which it is concluded that it is necessary to create fundamentally new models and methods for searching, processing and analyzing information in such a collection.

Keywords

Information system, information retrieval, artificial intelligence, document classification, machine learning, neural networks, natural language processing

1. Introduction

The post-industrial society of our days is characterized by the total penetration of global information technologies into the social reality of everyday life and the virtualization of social space. The individual plunges deeper and deeper into cyberspace. This process, on the one hand, enables the individual to organically integrate into the global space of social interaction with qualitatively changed and complicated requirements for the nature of participation in social communication and effectively use its capabilities. On the other hand, cybersocialization makes a person more vulnerable to threats and destructive challenges coming from the Global Network. Among the most serious of them is the threat of the impact of suicidal content and cyberbullying on children and young people. Therefore, the search and identification of such content in the cyberspace of Kazakhstan is a topical and important issue, given the very high level of suicides in our country. According to WHO, Kazakhstan ranks third in the world in terms of the number of suicides and is the leader among the countries of Central Asia [4].

Google Chief Economist H. Varian formulates this need as follows: "Data is becoming so widely available that there is not enough ability to extract knowledge from it" [1]. So simply and concisely, Google defines one of the main scientific and technological challenges of our time - the phenomenon of big data.

IBM in its concept of "4V" identifies 4 problematic aspects that are inherent in the phenomenon of big data: large volumes (volume), high speeds (velocity), high diversity (variety) and high unreliability (veracity) [2] (Figure 1). According to this concept, big data refers to huge collections (aspect of volume) of heterogeneous (aspect of diversity), frequently changing (aspect of speed) and containing distortions (aspect of reliability) data.

Due to these aspects, big data is difficult to process by classical methods and on classical architectures, and require the use of fundamentally new approaches and processing methods.

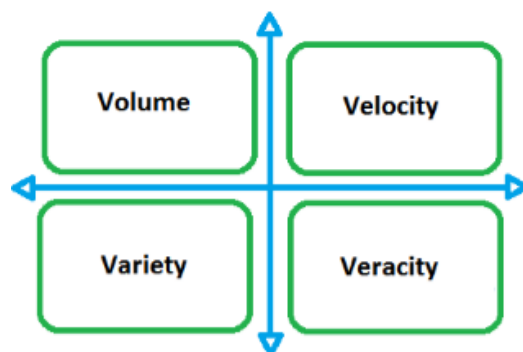


Figure 1: Problematic aspects of big data

Due to these aspects, big data is difficult to process by classical methods and on classical architectures, and require the use of fundamentally new approaches and processing methods.

Initially, big data was considered as a technological phenomenon, which, on one hand, demonstrated the inconsistency and inefficiency of computing technologies, and, on another hand, stimulated their development. Over time, the focus of research has shifted to the scientific field. It became clear that depending on the level of complexity of the data, the level of computational problems also changes. In particular, when it was just about large data arrays, then with an increase in their volumes, the complexity of calculations grew linearly, and they were amenable to simple parallelization [3]. If the data arrays were not only large, but also heterogeneous, then with an increase in their volumes, the complexity of calculations began to grow non-linearly, and additional analysis methods were required [3]. Thus, technological problems moved to a more complex level of scientific problems.

A typical example of big data is large collections of natural language texts presented on the World Wide Web. At present, studies devoted to the problems of processing and analyzing such text collections are relevant. In particular, these tasks include the tasks of automatic search and extraction of texts contained in distributed heterogeneous sources; tasks of automatic classification and clustering of texts; tasks of spam recognition, unwanted content filtering, etc.

At the heart of all these tasks is the complex scientific problem of natural language modeling, which generates a whole range of tasks related to the development and development of the corresponding mathematical apparatus, methods and algorithms for processing texts. The problem of language modeling is central to all research on machine processing and machine understanding of natural language, starting from the earliest works of Chomsky, Tenier, D. Miller [9] and ending with research conducted in the laboratories of the largest computer corporations such as IBM, Microsoft, Google [1].

For a long time, this problem was solved on the basis of linguistic and statistical approaches. The linguistic approach is based on the idea of creating a formal model of a language capable of describing all its constructions in a language understandable to machines. It is very inefficient in terms of labor costs, since it requires manual marking of a large number of auxiliary resources (dictionaries, corpora), but even so, linguistic models do not completely cover all the diversity and richness of living, continuously developing forms of natural language. The statistical approach is based on the use of frequency and probabilistic models of the occurrence of words to identify concepts and relationships contained in the texts under study. Its advantage is algorithmic simplicity, which makes it possible to obtain a practical result in a short time, and its disadvantage is the lack of knowledge about the world and the language, which limits the possibility of pragmatic and semantic analysis.

Therefore, over time, the prospect of applying a new approach based on statistical models and methods of processing and analyzing texts, but enriched with knowledge of the subject area, became obvious. If we accept as true the statements of analysts that up to 85% of knowledge is contained in texts [91], then the more texts of the subject area at the disposal of an expert, the more complete and accurate it is possible to build a knowledge model. The question of choosing a method for extracting and presenting the knowledge necessary to enrich the statistical models of text processing and

analysis remains open.

In the context of the information explosion, the traditional problems of processing and analyzing textual information have only intensified. The volumes, update rates and format diversity of textual information have sharply increased, which actualized the need for research and development of more efficient approaches to the processing and analysis of natural language texts.

2. Problems of information retrieval and text processing

A natural language is understood as a language is used for human communication (unlike formal languages) and not purposefully created (unlike artificial languages) [6]. Natural language texts are the most common form of knowledge representation. These texts are easily perceived and interpreted by humans, but difficult for machine understanding.

Machine understanding of natural language is a broad term that is used to refer to various levels of human-machine interaction in natural language [7, 8]. The simplest level of interaction is provided by systems that understand a limited set of human commands formulated in natural language. The medium and complex level of interaction is demonstrated by systems that replace a human expert when processing information or communicating with another person. The conceptual breadth of the term "machine understanding" allows us to interpret it as a designation of a simple act of processing information in a certain communicative environment (natural language processing), and as a cognitive process of thinking (meaning understanding). In the first case, the text is not understood, but processed with the installation of what the user expects to receive from it.

It is noted in [9] that such a "weak" interpretation of machine understanding goes back to the "machine" metaphor of N. Chomsky and D. Miller, who, by understanding operator, mean a certain information converter that produces a result in accordance with the user's expectations (Figure 2).

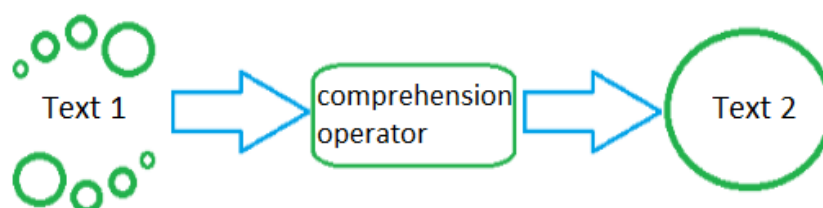


Figure 2: Machine understanding as an act of information processing

Proponents of the second, more "strong" interpretation consider the problem of machine understanding of language in close connection with the study of the underlying mechanisms of the functioning of thinking and speech [10,11,12]. The founders of the school of a strong semantic approach in Russian-speaking science are Igor Melchuk [13] and Yuri Apresyan [14], whose work is related to the creation of complex language models. The Meaning-Text model, created by I. Melchuk, gave rise to the theory of the integral description of the language and systemic lexicography. The model is a description of a natural language, understood as a device that provides a person with a transition from meaning to text (text construction) and from text to meaning (text understanding) [13].

A strong approach to machine understanding of a language implies the creation of some kind of semantic metalanguage, consisting of basic functions and concepts that can describe all other concepts and language constructs. As a result, the strong approach is forced to claim complete coverage and accuracy of the created metalanguages, which causes serious problems in their implementation.

Obviously, unlike the strong approach, the weak approach is pragmatic and more effective. When using a weak approach, the status of understanding models is acquired by any natural language processing systems, even those whose goal is not to understand the meaning, but to interpret it according to the needs of the user [10]. Such systems, as noted earlier, include systems for information retrieval, summarizing and categorizing documents, and machine translators.

However, even a weak interpretation of machine understanding requires the ability of a computer to overcome the informality and ambiguity of natural language. Computer processors do well with formal grammars and languages that have strict and unambiguous rules for constructing and

interpreting expressions. Natural languages, unlike formal ones, do not have the accuracy and rigor of constructing and understanding statements.

Machine understanding of natural language is a branch of knowledge that is at the intersection of computational linguistics and artificial intelligence, and whose purpose is to create computers capable of understanding or interpreting the meaning of natural language texts in accordance with user expectations.

The problems of constructing formal models of a language are dealt with by such a scientific discipline as computational linguistics [15]. More precisely, the subject of computational linguistics is the development of a formal apparatus for describing the structure of natural and some artificial languages [16]. The methods of computational linguistics are based on the apparatus of mathematical logic and its sections such as propositional calculus, automata theory and algorithm theory.

The earliest research in the field of computational linguistics is related to the practical problems of machine translation [17]. In January 1954, the first public demonstration of a machine translation system from Russian into English, performed on an IBM-701 machine, was held in the USA. Although the translation algorithm used in the system did not have scientific value, the very appearance of such a system gave a powerful impetus to research in the field of machine translation, both in the United States and beyond. In particular, in the USSR, the first experience of translating from English into Russian using the BESM machine was obtained by the end of the next year [18].

The first attempts to describe natural languages using formal grammars were proposed at about the same time in the works of the American linguist Noam Chomsky and the French linguist Lucien Tenier. Chomsky and Tenier are the founders of two opposing approaches to the description of natural language: the grammar of constituents and the grammar of dependencies [19]. Both of these approaches are used in modern computational linguistics, but the grammar of constituents is used, as a rule, for languages with a fixed word order (as in English), and the grammar of dependencies is used for languages with a free word order (as in Russian).

N. Chomsky defined the main task of computational linguistics as the search for “simple and explanatory grammars for natural languages” [20]. He believed that the basis of such grammars are sets of generative rules, with the help of which speakers construct sentences. In general, these rules are pairs of chains (left and right) and determine the possibility of replacing the left chain with the right one to generate sentences. The ideas of Chomsky's generative grammars are successfully used in the development of compilers for programming languages [21], but there are also their implementations for natural language texts [22].

Unlike Chomsky, Tenier proposed to consider the structure of a sentence not as a chain of syntactic units (components), but as a hierarchy of syntactic links (dependencies) [23]. The top of the sentence, according to Tenier's concept, was the verb-predicate, all other syntactic units included in the sentence were subordinate to it.

In a certain sense, it can be argued that Tenier's dependency grammar is still closer to understanding the meaning than Chomsky's grammar of constituents, since it shows explicit semantic rather than formal connections between sentence elements.

Despite the differences, both approaches have the same global drawback, which is inherent to a greater or lesser extent in all formal grammars. The essence of this drawback lies in the fact that none of the formal grammars is capable, precisely because of its formalism, of processing deviations of live speech from grammatical rules. In other words, no formal natural language model is able to cover the structure of the language as a whole. It is unlikely that this problem will be solved in the future. Therefore, today the problem of natural language processing is more of an engineering than a theoretical problem.

The engineering approach involves breaking down the natural language processing process into several well-defined stages (procedures) [24]:

- Graphematic analysis;
- Morphological analysis;
- Syntactic analysis;
- Semantic analysis.

It is not necessary that all procedures must be performed. For example, to implement simple

natural language processing systems, it is sufficient to implement the first two procedures. Each subsequent procedure is more complicated than the previous one and, to a certain extent, allows you to eliminate the shortcomings (remove the ambiguity) that arose during the implementation of the previous procedure.

Graphematic analysis. The purpose of the procedure is to divide the text into structural elements. The input is a text in natural language, the output is formed by a set of elements of which it consists (paragraphs, paragraphs, sentences, words and other tokens). Tokens are such indivisible units of text as punctuation marks, numbers, dates, currency signs, abbreviations, etc.

Despite its apparent simplicity, graphematic analysis is not a trivial task, since a living language is constantly updated, and new graphemes are constantly generated in it.

Morphological analysis. The purpose of the procedure is the normalization of words and pseudowords (tokens) identified at the stage of graphematic analysis. Words are given as input, and their normal (dictionary) forms and morphological characteristics are formed at the output. Normalization allows you to simplify and formalize the dictionary, which is important from the point of view of the final goal of processing - text structuring. Like the problem of the previous stage, this problem is non-trivial. Normalization is inherently ambiguous, which is expressed in the fact that several normal forms can correspond to one word. This ambiguity can be removed at the parsing stage.

For example, morphological ambiguity arises when parsing the sentences "Children ate porridge" and "I was asked." In the first sentence, the ambiguity is generated when parsing the word "ate", for which in Russian there are 2 normal forms "is" and "spruce". Parsing helps to determine that the word "is" is the normal form because it is preceded by a subject. In the second sentence, ambiguity arises when parsing the word "Me", for which there are also 2 normal forms "I" and "Men" (surname). In this case, parsing is powerless, so semantic analysis is needed.

For example, morphological ambiguity arises when parsing the sentences "Children ate porridge" and "I was asked." In the first sentence, the ambiguity is generated when parsing the word "ate", for which in Russian there are 2 normal forms "is" and "spruce". Parsing helps to determine that the word "is" is the normal form because it is preceded by a subject. In the second sentence, ambiguity arises when parsing the word "Me", for which there are also 2 normal forms "I" and "Men" (surname). In this case, parsing is powerless, so semantic analysis is needed.

Syntactic analysis. The purpose of the procedure is to parse the sentence and highlight its syntactic structure. The input is a sentence and the words included in the sentence that have passed morphological analysis (i.e., words in normal form and with established morphological features). At the output, a parse tree is formed - a structure of relationships between individual words and parts of a sentence.

The need for parsing arises in more complex systems, such as automatic fact extraction systems. Fact extraction is impossible without highlighting the syntactic structure, because the lion's share of meaning is conveyed not by the words themselves, but by the relationship between them. To extract a fact, its reference element (for example, the name of a person) is first determined, then, based on the parse tree, the relations of this element with other words are determined. Thus, a certain surface scheme of the fact is formed, which can then be enriched at the stage of semantic analysis.

Semantic analysis. The purpose of this procedure is the transition from the structure of surface syntactic links to semantic interpretation. The syntax tree of parsing the sentence is given as an input to the procedure. At the output, a set of semantic structures is formed, built in accordance with the accepted formal notation (semantic model).

It was noted above that natural language processing consists in the ability of a machine to perform two tasks: to recognize the structure (syntax) of a text and to extract its meaning (semantics). The creation of machines with such abilities is a difficult but achievable goal. Unlike natural language processing, these abilities are not enough to understand the language. The machine must be able to understand the context (pragmatics) of the text. The creation of such machines capable of understanding the context is still an unattainable goal.

At first glance, it may seem that the ability to understand the semantics (meaning) of the text is the same as the ability to understand the pragmatics (context). In fact, there is a fundamental difference between these abilities. Semantic analysis is the process of extracting the meaning of a text based on

some given knowledge model (for example, a domain ontology). Pragmatic analysis goes beyond static models of knowledge and relies on extralinguistic factors such as the intentions of the author, the social context of the utterance, etc. In other words, semantic analysis is performed within the framework of an explicitly given knowledge model, while pragmatic analysis requires the machine to be able to analyze implicit factors and draw conclusions based on them.

In the early period of the development of the discipline of artificial intelligence, researchers did not distinguish between natural language processing and machine understanding [25]. Currently, most researchers are of the opinion that natural language understanding, although it is the highest level of natural language processing, is an unattainable goal [25].

In the textbook work on information retrieval [26], the following definition is given: “Information retrieval is the process of searching in a large collection (usually stored in computer memory) of some unstructured material (usually a document) that satisfies information needs.” This definition is classic and links together three key concepts of information retrieval (see Figure 3):

- Purpose of the search - information need, expressed as a search query;
- Search area - a static collection of documents;
- Search result - a list of documents whose content matches the search query from the point of view of the search engine.

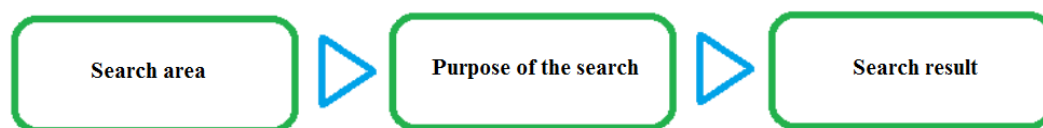


Figure 3: Key concepts of information retrieval

Based on the above definition, there are 5 stages in the process of information retrieval of natural language texts [26].

- Determining the purpose of the search (clarification of the information need or search topic, formulation of the request).
- Determining the scope of the search (determining the totality of text sources, repositories or holders of documents).
- Search and extraction of documents from the specified area and in accordance with the specified purpose.
- Ranking of search results according to their relevance.
- Issuing results to the user and evaluating the quality of the search.

Relevant documents are documents whose content corresponds to the search query from the point of view of the search engine. Relevant documents corresponding to the search query do not always correspond to the information needs of the user, since the search query itself does not always fully and accurately express this need.

Documents, the contents of which correspond to the information needs of the user, are called pertinent. Obviously, ideally, the sets of relevant and pertinent documents should coincide, but in practice this happens very rarely.

The method for determining the relevance of documents, as well as the formats for presenting queries and documents, are set by the information retrieval model [26].

The efficiency of the search engine is evaluated using the criteria for the completeness and accuracy of the search [26,27]. The completeness of the search is the proportion of pertinent documents found by the machine in the total number of pertinent documents in the collection. Search accuracy is the proportion of pertinent documents found by the machine in the total number of relevant documents found by the machine.

Information retrieval of natural language texts is based on four main models, and all modern search engines, one way or another, use these models or their modifications [26] (Figure 4).

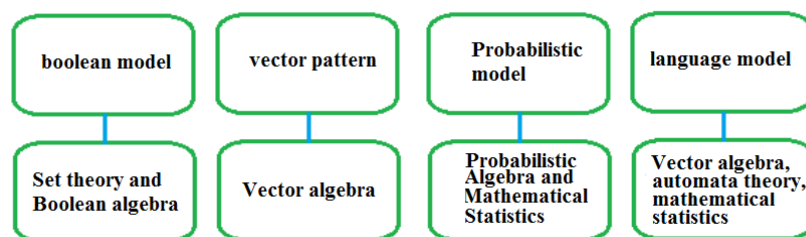


Figure 4: Basic models of information retrieval and their mathematical apparatus

Boolean information retrieval model. This is the simplest information retrieval model, in which both the document and the query are described using a set of keywords (terms). The model got its name because the query in it is formed as a Boolean expression, in which the terms are interconnected by Boolean operations AND (logical "and"), OR (logical "or"), NOT (logical negation) [26]. Relevance assessment in the Boolean model is performed using a logical comparison operation: the query expression is compared with the search image of the document, which is a set of terms. The Boolean model is rigidly tied to the dictionary of terms, so it can only be used in systems where the accounting of documents by terms (keywords) is strictly regulated. These are mainly library systems, corporate knowledge bases, electronic document management systems, etc.

The vector information retrieval model, like the Boolean one, is based on the description of the document and the query using terms. It was proposed by Gerard Salton in the early 1970s and is now the most popular information retrieval model [27,28,29,30,31]. Its advantage over the Boolean model is that the document is described not as a set of terms, but as a vector whose coordinates are the weights of the use of terms in the document. Thus, documents and queries are defined as multidimensional vectors, and the proximity between them is estimated as a vector distance. Currently, there are a huge number of modifications of the vector search model, which differ in the methods of calculating the weights of terms, methods of measuring the distance between vectors, etc.

Probabilistic model of information retrieval. The probabilistic model is based on an elegant mathematical idea of estimating the probability of a document matching a query [26,32,33]. The search engine learns to distinguish between relevant and irrelevant documents based on a training collection of documents. Training collections are generated by the user or retrieved automatically with some initial guess. Due to this, this model is inferior to the vector model in performance. After training, the probability is estimated. The probability that an incoming document is relevant is determined based on the ratio of the occurrence of the terms of this document in the relevant and irrelevant subsets of the training collection.

Language model of information retrieval. The language model goes back its basic idea to the theory of finite automata. The essence of the idea is that the probability of the appearance of the next word in speech depends on the use of previous words. This idea was first considered in [34]. In particular, it noted that this idea can be used in information retrieval if each document has its own language model. Moreover, if a certain query is submitted to the input of the search engine, then it is possible to rank documents based on the probability that their language models will generate the text of this query, i.e. that the request can be obtained randomly from the language model of the document [26]. Thus, the relevance of a document to a query is based on the credibility score of the query. Language models are built on the very natural idea of generating speech, but their application is limited due to the sparseness of the training collections. The prospects for the use of such models are concentrated around the study of ways to smooth the language models of documents.

The main issue of classical information retrieval is the search itself. However, over the half century of the existence of the discipline of information retrieval, the list of issues addressed within the framework of this discipline has expanded significantly [26,37,38]. It currently includes issues such as:

- Modeling the presentation of documents and queries; Search area - a static collection of documents;
- Designing the architecture of search engines and agents, including for the Web;

- Search and presentation of knowledge in the Semantic Web;
- Classification (categorization) of texts;
- Clustering of texts;
- Extracting facts (knowledge) from texts;
- Annotation of texts;
- Detection of texts containing unwanted content;
- Search for similar texts (in particular, plagiarism);
- Parallel search of texts in different languages;
- Improving the performance of information retrieval through the use of new distributed technologies and methods.

Classification (categorization) of texts. The task of classifying documents (natural language texts) is to distribute the initial set of documents into several specified classes (categories) based on their content (semantics, meaning) [39,40,41]. Natural language texts are difficult objects for classification, since they are unstructured data arrays. This necessitates preliminary processing of documents in order to obtain their feature descriptions suitable for classification algorithms [26]. As a rule, all the same vectors formed from a set of keywords are used as feature descriptions of documents. Classical classification algorithms such as support vector machine, naive bayes classifier, Rocchio method are based on machine learning technology. To build a classifier, these algorithms use a training sample, i.e. a set of documents already divided into classes. Based on the analysis of the training sample, these algorithms try to restore the relationship between documents and classes, and then, based on the found relationship, they build a classifier that can determine the class for any new document.

Text clustering. Unlike classification, which involves the distribution of documents into known classes, clustering is the process of automatically splitting documents into a predetermined number of classes (clusters) based on the degree of their similarity. Clustering makes it possible to identify natural groups of documents in a collection of documents, characterized by internal homogeneity and external isolation [42,43,44]. Most clustering algorithms involve comparing documents with each other based on some measure of similarity (similarity) [43]. Proximity measures are established using special metrics, very often the Euclidean distance is used as such a metric. There are two types of document clustering: hierarchical, which allows for subclusters, and flat, which involves direct division into clusters [45].

Annotation of texts. Annotation is a process of analytical and synthetic processing of information, the purpose of which is to convey a concise description of the document, revealing its logical structure and summary. The most important problems of automatic annotation of documents are to ensure completeness, reduce repetition in the presentation of information, and ensure coherence [46]. There are two groups of methods for automatic annotation of documents: extracting and generating [47]. Extracting methods use the extraction of the most significant information from the document in the form in which it is present there, generating methods create new text that briefly conveys the content of the original document. Extraction methods are based on machine learning algorithms, while the document is represented as a semantic graph of concepts, and concepts are combined into semantic clusters. For each block of text, its weight is determined based on the semantic graph, the most significant blocks are included in the annotation [48]. Generating methods are based on lexical-syntactic analysis of text sentences. Typically, these methods use a tagged dictionary of predicates (for scientific texts, for example, the dictionary may contain such predicates as "represents", "mentions", "contains", etc.). At each step, the generating algorithm analyzes the next fragment of the text and combines it with the most suitable predicate [48,49].

Extracting information from texts. Information extraction is the process of automatically extracting structured data (facts, knowledge, relationships) from unstructured text documents. To extract information, modifications of the common pattern specification language (CPSL) are used, which make it possible to recognize fact patterns in the text [50,51]. Also, when extracting information, ontologies and thesauri of subject areas are actively used to identify the objects of the subject area mentioned in the text of the document [50].

Typical subtasks of information extraction are: recognition of named elements (entities), for

example: names of persons, names of organizations, geographical names, events [52]; resolution of anaphoras and coreferences, i.e. search for links related to the same object [53,54]; terminology selection: finding keywords and phrases (collocations) for a given text [55,56,57]; sentiment analysis and Opinion Mining: extracting semantic, emotive, evaluative, and other information from the text [58,59].

The problem of extracting knowledge from unstructured information arrays is currently so relevant and complex that it is proposed to look for its solution in the creation of a semantic Web [60]. The Semantic Web is a kind of add-on to the regular Web that allows you to standardize the presentation of information in a form suitable for machine processing and automatic extraction [60,61]. The traditional Web uses the HTML hypertext markup language, which describes the presentation format of information, but not the content of the data. As a result, it is very difficult to semantic analyze, and even such simple fact extraction tasks as searching for persons, upcoming events, analyzing political statements require the use of intellectual methods. The use of Semantic Web technology will allow computers to interpret the information presented on the Web at the level of human understanding, for which the resources presented on the Web will have to be annotated using the RDF (Resource Description Framework) resource description model [62,63].

Detection of texts containing unwanted content is a particular task of content analysis. Content analysis refers to the process of quantitative and qualitative analysis of the content of documents in order to identify or measure various facts and trends reflected in these documents [64]. The difference between the task of content analysis and the task of automatically determining the topic of a document (topic detection) is that content analysis is aimed at studying documents in their social context. Thus, the problem of automatic detection of documents by their content is closer to pragmatic text analysis than to semantic one. The need for pragmatic analysis dictates certain requirements for methods for analyzing documents with prohibited or unwanted content. These methods should be aimed at finding not only linguistic features, but also pronounced social and psycholinguistic features (meta-information, emotions, images, etc.) that can indicate that a document belongs to a prohibited category. The selection of such features for machine learning of the classifier is the most difficult (intelligent) stage of content analysis. Here it is very important to have a good training collection of documents, training on which will allow the classifier to better cope with the assignment of new documents to a prohibited or allowed category.

3. Conclusion

The problems of searching, processing and analyzing information in large collections of natural language texts on the Internet can be divided into two large classes: traditional problems associated with natural language processing and new problems associated with processing big data.

The traditional problems of information search associated with the processing of the natural language are due to the phenomenon of the ambiguity that arises at all levels of the organization of the natural-language text: morpho-syntactic (structural), semantic and pragmatic (contextual). To eliminate ambiguity, the analysis of the text at each of the levels is necessary.

The new problems of information search associated with the processing of large data are due to constantly growing volumes, high update speed, high variety, the presence of errors and inaccuracies.

Currently, there are several most promising approaches to information search, which are able to overcome these problematic aspects: parallelization of data processing, approximation (data smoothing), randomization, transfer of calculations closer to data and deep learning.

Along with problems that carry out big data, two positive aspects allow us to talk about their important role in solving information search problems. Firstly, big data increase the effectiveness of machine learning. The more data is covered by training, the better the results of the computer. Secondly, big data increase the value of information search. The higher the completeness of the search, the more valuable the results for the user.

Thus, the review of the current state of the problems shows that this direction is relevant and multifaceted. Models, methods and algorithms for searching, processing and analyzing texts in the conditions of big data, models, methods and algorithms for machine learning on big data, methods for

designing x architectures of interconnections for distributed data processing require additional research.

Based on the above, in order to level most of the problems, when developing intelligent information retrieval and text processing systems, we suggest using the following development algorithm:

1. Develop a method for automatically creating a thesaurus of a subject area used to improve the efficiency of information retrieval of natural language texts. The method of creating a thesaurus is based on a statistical approach to natural language processing.
2. For the developed method, implement a distributed algorithm for its operation.

4. References

- [1] Cukier, K. Data, data everywhere: A special report on managing information. // *The Economist*, 394(8671), 2010.
- [2] Schroeck M. et al. Analytics: The real-world use of big data //IBM Institute for Business Value—executive report, IBM Institute for Business Value. – 2012.
- [3] Кузнецов С. Большие хлопоты с большими объемами данных // *Открытые системы. СУБД*. – 2008. – №. 4. – С. 64-69.
- [4] Marca D. A., McGowan C. L. SADT: structured analysis and design technique. // McGraw-Hill, Inc., 1987.
- [5] Tan W. et al. Social-network-sourced big data analytics // *Internet Computing, IEEE*. – 2013. – Т. 17. – №. 5. – С. 62-69.
- [6] Stenlund S. Language and Philosophical Problems. // Psychology Press. – 1990.
- [7] Waldrop M. Natural Language Understanding // *Science*. – 1984. – Т. 224.– №. 4647. – С. 372-74.
- [8] Pantic M., Pentland, A., Nijholt, A., Huang, T.S. Human computing and machine understanding of human behavior: a survey // *Artificial Intelligence for Human Computing*. – Springer Berlin Heidelberg, 2007. – С. 47-71.
- [9] Попович М.В., Крымский С.Б. и др. Доказательство и понимание. // Киев, Наукова думка. – 1986.
- [10] Гаврилова Т. А., Хорошевский В. Ф. Базы знаний интеллектуальных систем. // СПб: Питер. – 2000. – С. 384.
- [11] Пospelov Д.А. Моделирование рассуждений: Опыт анализа мыслительных актов. – Радио и связь, 1989.
- [12] Баранович А.Е. О систематизации аксиоматического аппарата предметной области Искусственный интеллект // *Интеллектуальные системы*. – 2010. – Т. 14. – №. 1-4. – С. 5-34.
- [13] Мельчук, И.А. Опыт теории лингвистических моделей «Смысл↔Текст». // М.: Наука. – 1974.
- [14] Апресян Ю.Д. Лексическая семантика. // М.: Наука, 1974.
- [15] Chomsky N. Computational Linguistics and Cognitive Science // *IEEE Intelligent Systems*. – 2011. – Т. 26. – №. 4. – С. 12.
- [16] Grishman R. Computational linguistics: an introduction. // Cambridge University Press, 1986.
- [17] Hutchins J. Machine translation: A concise history // *Computer aided translation: Theory and practice*. – 2007.
- [18] Панов Д.Ю., Ляпунов А.А., Мухин И.С. Автоматизация перевода с одного языка на другой. В сб. Сессия по научным проблемам автоматизации производства. М. Изд. АН СССР, 1956.
- [19] Schneider G. A linguistic comparison of constituency, dependency and link grammar // Master's thesis, University of Zurich. – 1998.
- [20] Chomsky N. Three models for the description of language // *Information Theory, IRE Transactions on*. – 1956. – Т. 2. – №. 3. – С. 113-124.
- [21] Aho A. V., Ullman J. D. The theory of parsing, translation, and compiling. – Prentice-Hall, Inc., 1972.

[22] Charniak E. Statistical parsing with a context-free grammar and word statistics // Proceedings of the 14th National Conference on Artificial intelligence and 9th Conference on Innovative applications of artificial intelligence. – AAAI Press, 1997. – С. 598-603.

[23] Теньер Л. Основы структурного синтаксиса. // М.: Прогресс, 1988. – 656 с.

[24] Дунаев А.А. Исследовательская система для анализа текстов на естественном языке // Проблемы интеллектуализации и качества систем информатики. Н.: Ин-т систем информатики имени А.П. Ершова СО РАН –2006– Вып. – С. 55-66.

[25] Liddy E.D. Natural Language Processing. // Encyclopedia of Library and Information Science. – NY. Marcel Decker. – 2001.

[26] Маннинг К., Рагхаван П., Шютце Х. Введение в информационный поиск // М.: Вильямс. – 2011.

[27] Нугуманова А.Б., Ораккызы М. К проблеме выбора информативных признаков в задачах автоматической классификации текстов. // Информационные и телекоммуникационные технологии: образование, наука, практика: Труды Международной научно-практической конференции. – Алматы, 5-6 декабря 2012. – Алматы, КазНТУ им. Сатпаева, 2012. – С. 379-383.

[28] Salton G., Wong A., Yang C. S. A vector space model for automatic indexing // Communications of the ACM. – 1975. – Т. 18. – №. 11. – С. 613-620.

[29] Ning B., Ji Z. Research on web information retrieval based on Vector Space Model // Journal of Networks. – 2013. – Т. 8. – №. 3. – С. 688-695.

[30] Jing L., Ng M. K., Huang J. Z. Knowledge-based vector space model for text clustering // Knowledge and information systems. – 2010. – Т. 25. – №. 1. – С. 35-55.

[31] Kiela D., Clark S. A Systematic Study of Semantic Vector Space Model Parameters // Proceedings of the 2nd Workshop on Continuous Vector Space Models and their Compositionality (CVSC)@ EACL. – 2014. – С. 21-30.

[32] Sparck Jones K., Walker S., Robertson S. E. A probabilistic model of information retrieval: development and comparative experiments: Part 1 // Information Processing & Management. – 2000. – Т. 36. – №. 6. – С. 779-808.

[33] Sparck Jones K., Walker S., Robertson S. E. A probabilistic model of information retrieval: development and comparative experiments: Part 2 // Information Processing & Management. – 2000. – Т. 36. – №. 6. – С. 809-840.

[34] Ponte J.M., Croft W.B. A language modeling approach to information retrieval // Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval. – ACM, 1998. – С. 275-281.

[35] Voorhees E.M., Buckland L.P. The Twenty-First Text Retrieval Conference Proceedings (TREC 2012) // NIST Special Publication. – 2012. – С. 500- 298.

[36] Smeaton A.F. et al. Analysis of papers from twenty-five years of SIGIR conferences: what have we been doing for the last quarter of a century? // ACM SIGIR Forum. – ACM, 2003. – Т. 37. – №. 1. – С. 49-53.

[37] Salton G., Harman D. Information retrieval. // John Wiley and Sons Ltd., 2003. – с. 858-863.

[38] Baeza-Yates R. et al. Modern information retrieval. // New York: ACM press, 1999. – Т. 463.

[39] Joachims T. Transductive inference for text classification using support vector machines // ICML. – 1999. – Т. 99. – С. 200-209.

[40] Forman G. An extensive empirical study of feature selection metrics for text classification // The Journal of machine learning research. – 2003. – Т. 3. – С. 1289- 1305.

[41] Sebastiani F. Machine learning in automated text categorization // ACM computing surveys (CSUR). – 2002. – Т. 34. – №. 1. – С. 1-47.

[42] Aggarwal C. C., Zhai C. X. A survey of text clustering algorithms // Mining Text Data. – Springer US, 2012. – С. 77-128.

[43] Aggarwal C.C., Zhai C.X. A survey of text clustering algorithms // Mining Text Data. – Springer US, 2012. – С. 77-128.

[44] Jing L., Ng M. K., Huang J. Z. Knowledge-based vector space model for text clustering // Knowledge and information systems. – 2010. – Т. 25. – №. 1. – С. 35- 55.

[45] Shehata S., Karray F., Kamel M. S. An efficient concept-based mining model for enhancing text clustering // Knowledge and Data Engineering, IEEE Transactions on. – 2010. – Т. 22. – №. 10. – С. 1360-1371.

[46] Лукашевич Н.В., Добров Б.В. Автоматическое аннотирование новостных кластеров на основе тематического представления // Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной международной конференции «Диалог». – 2009. – №. 8. – С. 15.

[47] Осминин П.Г. Современные подходы к автоматическому реферированию и аннотированию // Вестник Южно-уральского государственного университета. Серия: лингвистика. – 2012. – №. 25.

[48] Morales L. P., Esteban A. D., Gervás P. Concept-graph based biomedical automatic summarization using ontologies // Proceedings of the 3rd Textgraphs Workshop on Graph-Based Algorithms for Natural Language Processing. – Association for Computational Linguistics, 2008. – С. 53-56.

[49] Saggion H. A classification algorithm for predicting the structure of summaries // Proceedings of the 2009 Workshop on Language Generation and Summarisation. – Association for Computational Linguistics, 2009. – С. 31-38.

[50] Appelt D.E., Onyshkevych B. The common pattern specification language // Proceedings of a workshop on held at Baltimore, Maryland: October 13-15, 1998. – Association for Computational Linguistics, 1998. – С. 23-30.

[51] Reiss F. et al. An algebraic approach to rule-based information extraction // Data Engineering, 2008. ICDE 2008. IEEE 24th International Conference on. – IEEE, 2008. – С. 933-942.

[52] Etzioni O. et al. Unsupervised named-entity extraction from the web: An experimental study // Artificial Intelligence. – 2005. – Т. 165. – №. 1. – С. 91-134.

[53] Толпегин П. В., Ветров Д. П., Кропотов Д. А. Алгоритм автоматизированного разрешения анафоры местоимений третьего лица на основе методов машинного обучения // Компьютерная лингвистика и интеллектуальные технологии: Труды международной конференции «Диалог». – 2006. – С. 504-507.

[54] Поцепня В. Н. Разрешение местоименной анафоры в многоязычных информационных системах // Искусственный интеллект. – 2006. – №. 4. – С. 619- 626.

[55] Pecina P., Schlesinger P. Combining association measures for collocation extraction // Proceedings of the COLING/ACL on Main conference poster sessions. – Association for Computational Linguistics, 2006. – С. 651-658.

[56] Pecina P. Lexical association measures and collocation extraction // Language resources and evaluation. – 2010. – Т. 44. – №. 1-2. – С. 137-158.

[57] Liu J. et al. Advertising keywords extraction from web pages // Web Information Systems and Mining. – Springer Berlin Heidelberg, 2010. – С. 336-343.

[58] Baccianella S., Esuli A., Sebastiani F. SentiWordNet 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion mining // LREC. – 2010. – Т. 10. – С. 2200-2204.

[59] Liu B. Sentiment analysis and opinion mining // Synthesis Lectures on Human Language Technologies. – 2012. – Т. 5. – №. 1. – С. 1-167.

[60] Berners-Lee T. et al. The semantic web // Scientific American. – 2001. – Т. 284. – №. 5. – С. 28-37.

[61] Horrocks I. et al. SWRL: A semantic web rule language combining OWL and RuleML // W3C Member submission. – 2004. – Т. 21. – С. 79.

[62] Decker S. et al. The semantic web: The roles of XML and RDF // Internet Computing, IEEE. – 2000. – Т. 4. – №. 5. – С. 63-73.

[63] Allemang D., Hendler J. Semantic web for the working ontologist: effective modeling in RDFS and OWL. – Elsevier, 2011.

[64] О.Т. Манайев. Kontent-analiz – opisaniye metoda (in Russian) URL: <http://psyfactor.org/lib/kontent.htm> (access date 25.07.2014).

[65] Barroso L. A., Dean J., Holzle U. Web search for a planet: The Google cluster architecture // Micro, IEEE. – 2003. – Т. 23. – №. 2. – С. 22-28.