ISSN 2663-3051
(ISSN 0555-2656 до 2019 р.)

БІОНІКА ІНТЕЛЕКТУ

ІНФОРМАЦІЯ, МОВА, ІНТЕЛЕКТ

№ 1 (99) **2023**

НАУКОВО-ТЕХНІЧНИЙ ЖУРНАЛ

Заснований у жовтні 1967 р.

Засновник та видавець
Харківський національний університет радіоелектроніки

Періодичність видання – *2 рази на рік*

Харків • ХНУРЕ • 2023

Науково-технічний журнал

«БІОНІКА ІНТЕЛЕКТУ»

ISSN 2663-3051

Заснований Харківським національним університетом
радіоелектроніки у 1967 році

Реферування та індексування:

Google Scholar

BASE

INDEX COPERNICUS
INTERNATIONAL

Національна бібліотека України
імені В. І. Вернадського

Журнал включено до списку наукових спеціалізованих видань України
з технічних та фізико-математичних наук
згідно з наказом Міністерства освіти і науки України № 820 від 11.07.2016

УДК 004.4: 004.4

DOi 10.30837/ bi.2023.1(99).01

Mazurova Oksana[1], Ramazanov Rasul[2]

[1] Kharkiv National University of Radio Electronics, Kharkiv, Ukraine,
oksana.mazurova@nure.ua; ORCID ID: https://orcid.org/0000-0003-3715-3476

[2] Kharkiv National University of Radio Electronics, Kharkiv, Ukraine,
rasul.ramazanov@nure.ua; ORCID ID: https://orcid.org/0009-0000-8656-1869

RESEARCH ON TECHNOLOGIES FOR ACCESSING RELATIONAL DATABASES USING MS SQL SERVER

In the modern world, hundreds of large projects are developed every day, and thousands of startups with a wide variety of topics appear, developers start creating their projects, companies modernize old projects. Each of these activities has something in common - the way they store and manipulate data. Almost all modern projects have a database, because it becomes simply impossible to interact with the client without it. Now, no matter what your application, service, game or anything else is, it is necessary for the user to be able to have his account page, save progress, anything that needs to be saved locally or in the cloud or on the server. And a modern developer can use the database directly or through special database access technologies, including through ORM or Micro ORM. The subject is the study of access technologies for working with relational databases. The goal is to compare the efficiency, flexibility and use of various access technologies when working with the MS SQL SERVER database. Task: to investigate access technologies to MS SQL SERVER. Collect the technical characteristics of queries when using different database access technologies. Compare the results. Methods: analysis of MS SQL SERVER access technologies, experimental research, statistical analysis of results. Results: it is shown how access technologies such as ORM Entity Framework, Micro ORM Dapper and ADO.NET differ in use, it is shown that the performance is most effective in ADO.NET, followed by Dapper and in third place among the access technologies used in terms of efficiency is Entity Framework . But it is shown how the type of request contributes to the efficiency of execution by various technologies. Conclusions: ORMs are used in cases where it is necessary to work with a database using an object-oriented approach. ORM transforms data from the database into objects in the application, which facilitates interaction with the database and reduces the amount of code that must be written to work with the database. Micro ORMs are used when the speed of execution of requests to the database is required, or if the project is small in scope and does not need the full functionality of ORM. Micro ORM is smaller, faster and easier to use than ORM. ADO.NET is recommended when direct interaction with databases is needed. ADO.NET allows you to create multithreaded and distributed applications, interact with databases using transactions, and manage data security. It is more extensible and scalable than ORM and Micro ORM, but requires more code to interact with the database.

ACCESS TECHNOLOGY; DATABASE; MS SQL SERVER; ENTITY FRAMEWORK; DAPPER; ADO.NET.

О.О. Мазурова, Р.Ш. Рамазанов. Дослідження технологій доступу до реляційних баз даних під керуванням MS SQL SERVER. У сучасному світі кожного дня розробляються сотні великих проектів, з'являються тисячі стартапів із найрізноманітнішими тематиками, розробляються нові проекти та модернезують старі. Майже всі подібні проекти мають базу даних, бо зберігати та обробляти дані без неї стає просто неможливо. Сучасний розробник має змогу використовувати базу даних через спеціальні технології доступу до них, в тому числі через ORM або Micro ORM. Предметом дослідження є технологій доступу до реляційних баз даних, які займають значимі позиції в програмних рішеннях. Мета роботи – порівняти продуктивність використання найбільш ефективних на сьогодні технологій доступу до бази даних під керуванням MS SQL SERVER та розробити рекомендації, що дозволять розробникам зробити більш ґрунтовний їх вибір. Методи: аналіз технологій доступу до баз даних MS SQL SERVER, експериментальне дослідження, статистичний аналіз результатів. Результати: зібрано вагомі метрики під час експериментального дослідження продуктивності таких технології доступу як ORM Entity Framework, Micro ORM Dapper та ADO.NET; розроблено рекомендації стосовно використання цих технологій та вибору найбільш ефективного варіанту в залежності від видів запитів до баз даних. Висновки: Micro ORM більш рекомендована до використання, коли критичною є швидкість виконання запитів до бази даних, або якщо проект має невеликий обсяг та не потребує повного функціоналу ORM; ADO.NET рекомендовано в разі, коли потрібно пряма взаємодія з базами даних; ADO.NET дозволяє створювати багатопоточні та розподілені додатки, використовувати механізм транзакцій та керувати безпекою даних. Він є більш розширеним та масштабованим, ніж ORM та Micro ORM, але вимагає більше коду для взаємодії з базою даних.

ТЕХНОЛОГІЯ ДОСТУПУ; БАЗА ДАНИХ; MS SQL SERVER; ENTITY FRAMEWORK; DAPPER; ADO.NET.

Introduction

In today's world, the development of almost any software application is closely related to the use of databases (DB) [1, 2]. Regardless of the type of application, service, or game, there is always a need for users to have their own accounts, store progress, or other information locally, in the cloud, or on a server. Therefore, a modern developer must choose and utilize a specific database access technology, including ORM or Micro ORM, when working with databases in their projects.

For developers, the question arises as to which database access technologies are best to use and which ones are most suitable for their system. Traditional analysis of documentation on various database access technologies

and general recommendations provided by developers often do not provide a detailed description of the technical characteristics regarding the usage of these technologies within a specific stack of software tools.

One of the widely adopted solutions for working with relational databases is developing on the .NET platform using the Microsoft SQL Server database management system (DBMS) [3-4]. Currently, the most popular database access technologies in such an environment are Object-Relational Mapping (ORM) tools like Entity Framework, Micro ORM solutions such as Dapper, and the well-established ADO.NET technology.

When designing a system that interacts with SQL databases, it is crucial for the developer to have a clear understanding of the database's logic and the tools offered by the chosen database access technology [1, 5]. Due to limited practical information available about specific software connections, many commercial projects hesitate to adopt new database access technologies, as implementing such a transition requires significant time for performance modeling and data migration [6].

Therefore, a relevant direction for research is to establish clearer, practically validated recommendations for using database access technologies based on experimental measurements and comparisons of significant database performance metrics [7].

1. Analysis of the problem and existing methods

Relational databases have been widely used until now for developing applications that require strong support for ACID transaction properties, such as in the fields of banking, healthcare, and so on. With the emergence of the first technologies for accessing relational databases, many new trends and improvements to existing approaches have appeared [8, 9].

On the .NET platform, there are several ORM frameworks (Object-Relational Mapping) [10-11] that allow for easy and convenient interaction with relational databases. A thorough analysis has been conducted on several of them, namely:

– Entity Framework (EF) is an ORM framework developed by Microsoft for the .NET platform. EF allows developers to work with databases using an object-oriented approach rather than writing SQL queries. It supports various relational databases, including Microsoft SQL Server [12-13], Oracle, and MySQL;

– NHibernate is a popular ORM framework for the .NET platform. It allows developers to work with databases by mapping objects to database tables, making the interaction much simpler;

– Dapper is a lightweight ORM framework developed by StackExchange. It falls under the category of Micro ORM and enables developers to have more precise control over the database interaction process by using simple SQL queries;

– DevExpress XPO is an ORM framework that provides a wide range of functionality for interacting with relational databases, including code generation based on the database schema.

These frameworks enable developers to interact easily and conveniently with relational databases on the .NET platform, while also providing support for various functional capabilities such as data caching, lazy loading, data migrations, and more.

The rapid and widespread adoption of ORM and Micro ORM is driven by their simplicity. They allow programmers to work with databases using familiar objects and programming languages, instead of complex SQL code. ORM enables developers to interact with databases through an object-oriented interface, making the process more intuitive. Additionally, ORM automatically generates SQL code for interacting with the database, making the development process faster and less error-prone.

Micro ORM is a simplified version of an ORM that provides only basic database interaction functionalities. This makes it even simpler and easier to use, resulting in faster execution of database queries and reduced memory consumption.

As a result, the simplicity of working with ORM and Micro ORM [14] allows developers to focus on developing program functionality rather than spending time writing and testing complex SQL queries to the database.

Additionally, the well-established object-oriented technology ADO.NET [15-16], which is part of the .NET Framework, is quite commonly used on the .NET platform. ADO.NET enables developers to establish connections with a database, execute SQL queries, and retrieve query results in the form of a DataSet or DataReader.

During the selection of an access technology, one can rely on research dedicated to the characteristics of access technologies and their architectural features. For instance, in a three-tier architecture, ORM such as Entity Framework resides in the Data Access Layer and serves as a wrapper that communicates with the database and maps data from the database to the data layer model used by the developer. This accelerates development and data manipulation processes.

In addition to the architectural considerations, developers can also rely on analyzing current trends in database development [1, 17]. When considering the .NET platform, the two most popular ORM technologies are Entity Framework and the micro ORM Dapper. Until recently, Dapper and ADO.NET were comparable in terms of performance, but with each new version of .NET developed by Microsoft, the efficiency of development using Entity Framework has been increasing and, in some cases, may surpass Dapper and ADO.NET. The availability of open-source products has fostered a large community of developers dedicated to the advancement of not only Entity Framework but also other ORM and micro

ORM solutions. However, it is unfortunate that some access technologies are not open-source. ORM has become so prevalent that some developers may not be familiar with SQL and rely solely on ORM for writing queries. This can be a significant problem because such developers become heavily dependent on a single ORM and may not recognize alternative options or understand the criteria for selecting more efficient technologies for specific software solutions.

2. Objective of the Work

The purpose of this article is to conduct an experimental investigation of database access technologies on the .NET platform, specifically focusing on a relational database managed by the MS SQL Server DBMS. The goal is to evaluate their productivity and develop practical recommendations for their effective usage in various software projects.

For the experimental research, the most popular representatives of their respective classes have been selected: the ORM Entity Framework, the Micro ORM Dapper, and the object technology ADO.NET.

This research requires the following steps to be conducted: Analysis of the domain-specific application area and designing a database based on it for further experimental research.

– Development of software solutions based on Entity Framework, Dapper, and ADO.NET database access technologies.

– Conducting experimental research on the performance of implemented database access technologies and providing recommendations on the suitability of using these technologies.

The evaluation of the effectiveness of using database access technologies should be conducted considering the following metrics: query execution speed (in milliseconds), query execution speed (in ticks), and consumed resources of the working memory (in bytes).

3. Materials and Methods

The chosen subject area for designing the database for the research is the field of e-commerce. E-commerce, or electronic commerce, refers to the process of buying and selling goods and services over the Internet. It can include online stores, internet auctions, digital goods (such as music and videos), online booking and payment services (such as hotels, airline tickets, etc.), electronic marketplaces, and more. For conducting the experiments, a simplified database [18] for an online clothing store was designed.

A database containing the following basic concepts and their interrelationships has been developed for conducting experiments:

Season: Can be described by attributes such as "name" and "start date" (modeled by the "Seasons" table);

– Catalog: Within a season, there can be multiple catalogs of different categories (the "Catalogs" table);

– Category: Has an attribute "name" (the "Categories" table);

– Product: Can be described by attributes such as name, price, color, description, and category (the "Products" table);

– Good: Represents the relationship between a product and a catalog (the "Goods" table);

– Order: References the user who placed the order and the products included (the "Orders" table);

– User: Can be described by attributes such as name, email, and phone number (the "Users" table).

Classes were designed [16] for the use of Entity Framework and Dapper technologies based on the developed database model. The Code First approach was employed during the creation of software solutions using these technologies. The diagram of the developed classes can be seen in Fig. 1.

The solutions for the research were developed as web applications using ASP.NET Core Web API. In general, the architecture of ASP.NET Core Web API allows for the development of fast, scalable, and reliable web services that can be integrated into various applications and platforms.

Experimental research planning was conducted, and queries were developed as the basis for measuring performance metrics and investigating the productivity of access technologies.

The following queries were developed and used for conducting series of experiments:

– GetUsers query: retrieves all fields from the Users table.

– GetUserWithOrders query: retrieves data that requires joining the Users, Orders, Products, Goods, Catalogs, and Seasons tables (see Fig. 2).

– GetSeasonsQuery: retrieves the count of products present in a season; this query utilizes join operations between tables and the aggregate function Count() (see Fig. 3).

– ultiple aggregate functions.

– CreateCategory query: creates a new category with a specified name.

– DeleteCategory query: deletes a category with a specified name.

4. Research results and their discussion

Let's consider the results of executing queries for the investigated technologies: Entity Framework, Dapper, and ADO.NET. MS SQL Server was used as the database management system, and the tables in the database contained 1000 and 10,000 records, respectively.

For the purpose of conducting a pure research study, caching was disabled for Entity Framework. Caching allows Entity Framework to execute the same query in a very short time, which is one of the advantages of an ORM. By disabling caching, we can observe the actual performance of Entity Framework without the influence of cached results.

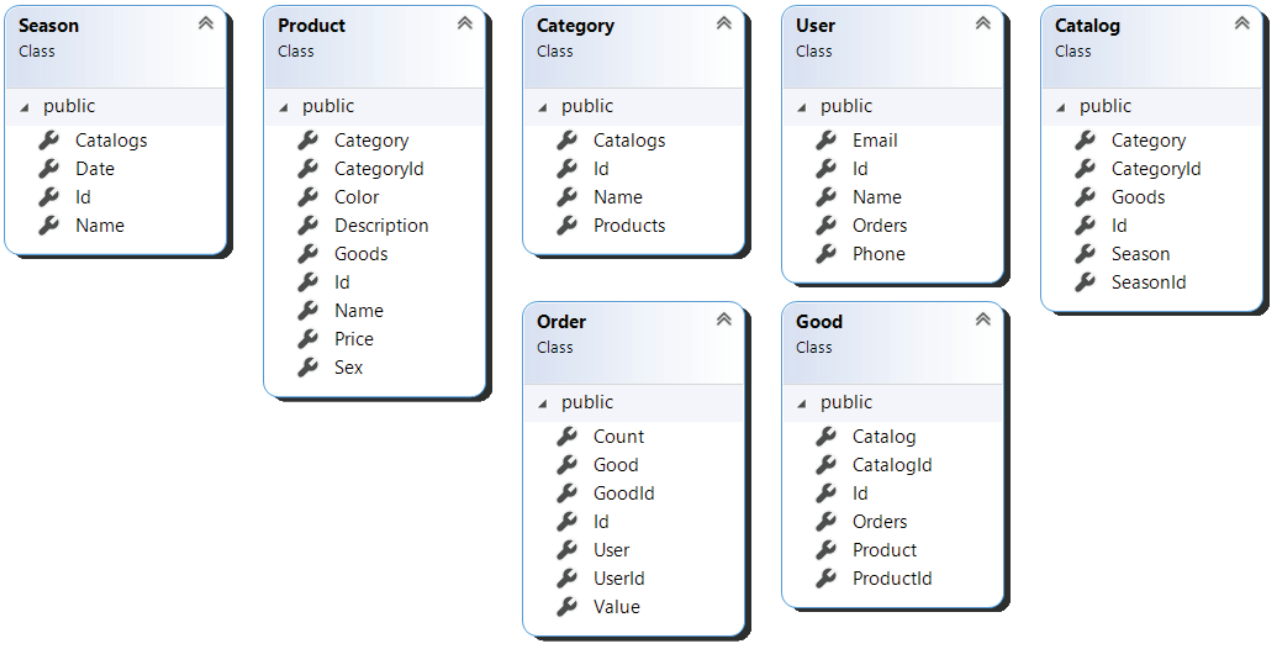

Fig. 1. Class diagra

```
SELECT users.Name as UserName, products.Id as ProductId,
               products.Name as ProductName, orders.Count, seasons.Date

               FROM Users as users,
               Orders as orders,
               Products as products,
               Goods as goods,
               Catalogs as catalogs,
               Seasons as seasons

               WHERE orders.UserId = users.Id
               AND orders.GoodId = goods.Id
               AND goods.ProductId = products.Id
               AND goods.CatalogId = catalogs.Id
               AND catalogs.SeasonId = seasons.Id
               AND seasons.Date < GETDATE()

               GROUP BY users.Name, products.Id, products.Name, orders.Count, seasons.Date
```

Fig. 2. Query GetUserWithOrders

```
SELECT seasons.Name AS SeasonName, seasons.Date, catalogs.Id AS CatalogId, COUNT(goods.CatalogId) AS GoodsCount
          FROM Seasons AS seasons
          JOIN Catalogs AS catalogs ON catalogs.SeasonId = seasons.Id
          JOIN Goods AS goods ON goods.CatalogId = catalogs.Id
          JOIN Products AS products ON products.Id = goods.ProductId
          GROUP BY seasons.Name, seasons.Date, catalogs.Id
```

Fig. 3. Query GetSeasonsQuerry

In the research, it was also decided to measure the performance metrics of queries for Dapper and ADO. NET, taking into account the resources spent on database connection and excluding the connection time. This is reasonable because, in the case of Dapper and ADO. NET, the database connection is established explicitly by the developer when executing queries. Therefore, the measurement of query performance includes the time required for establishing the database connection.

Let's also note that the first query to the database takes

longer to execute compared to subsequent queries because there may be additional tasks to perform before executing the query, such as query plan preparation and query optimization. Query plan preparation is the process of creating an execution plan that outlines the necessary steps to execute the query. It involves selecting the access method for database tables, determining the join method for table joins, and deciding the order of operations in the query. Query optimization is the process of improving query performance by modifying the query or utilizing additional indexes to speed up data retrieval and selection. The execution speed of the first query depends on the MS SQL Server and the time it takes to establish a connection to the database. The average query execution time for the server and database residing in the same physical space depends on the hardware and was computed based on 1000 runs in a series of experiments. Considering the consumed resources is important to understand how much memory is required for executing a specific query and to properly configure the database server.

The performance evaluation of the queries was conducted taking into account the following metrics:

– speed of execution of the first launch (ticks);
– speed of execution of the first launch (ms);
– average request execution time (ticks);
– average request execution time (ms);
– spent RAM resources (bytes) when executing the request.

Fig. 4 shows the results of a series of experiments on the GetUsers request. We see that the average execution time is equal to one for all access technologies, taking into account the execution of the request with and without connection to the database server. This means that the request is running too fast to be measured in ms, which shows us how beneficial it is to use more precise units of measurement like ticks.

Request	GetUsers				
	Entity Framework	Dapper	Dapper W/O connection	ADO.NET	ADO.NET W/O connection
First run (ticks)	2005217	3851348	3392629	3216040	206956
First run (ms)	182	400	352	321	20
Average execution time (ticks)	5623	5806	6181	4702	3209
Average execution time (ms)	1	1	1	1	1
Spent memory (bytes)	8512	4571	4384	4544	704

Fig. 4. Results of experiments with the request GetUsers

The first query run takes significantly longer than all other queries. If we take the ratio of the speed of the first request in ticks, then in ADO.NET w/o connection the execution speed is the best. This means that the server took less time to process the first request in ADO.NET w/o connection than for other access technologies. Also, the experiment made it possible to observe significantly lower costs in ADO.NET w/o connection for the first start in ms.

If we look at the average execution time in ticks, we will see that it will be less for ADO.NET and ADO.NET w/o connection than for Entity Framework. Dapper by this metric remains the longest running.

In terms of memory consumption, ADO.NET w/o connection consumed significantly less resources, namely 704 bytes. Then comes Dapper w/o connection, followed by ADO.NET and Dapper, and Entity Framework required the most resources, almost twice as much as ADO.NET and Dapper.

Fig. 5 shows the results of experiments with GetUsersWithOrders queries. Execution of this request is also faster than can be measured in ms. Let's look at the measurements in ticks. The first run shows the same ratios as the GetUsers query, namely ADO.NET w/o connection is the fastest, followed by Entity Framework and ADO.NET, followed by Dapper w/o connection and Dapper. The same behavior for the first run can be seen in the execution speed in ms. If you look at the average query execution time in ticks, then ADO.NET w/o connection is in first place, followed by ADO.NET, Dapper w/o connection and Dapper, and Entity Framework took the longest time to execute this query. According to the consumed memory resources, it can be found that ADO.NET with and without connection requires much less memory than Dapper, and Entity Framework consumes the most of this resource.

Request	GetUserWithOrders				
	Entity Framework	Dapper	Dapper W/O connection	ADO.NET	ADO.NET W/O connection
First run (ticks)	2555442	4138181	4002818	3283229	244381
First run (ms)	260	391	352	346	27
Average execution time (ticks)	15413	12814	11782	11867	8633
Average execution time (ms)	1	1	1	1	1
Spent memory (bytes)	12496	10552	10329	1731	1324

Fig. 5. Results of experiments with the request GetUsersWithOrders

Fig. 6 shows the results of experiments with the GetSeasonsQuerry query. The speed of execution of the first run is similar to the GetUsers and GetUsersWithOrders queries. But according to the average request execution time, it can be emphasized that ADO.NET w/o connection and Dapper w/o connection are executed the fastest, followed by ADO.NET and Dapper, and Entity Framework, which executes in almost 9 times longer than ADO.NET w/o connection. According to the used resources, we can also say that ADO.NET and Dapper use 3 times less memory resources than Entity Framework.

Request	GetSeasonsQuerry				
	Entity Framework	Dapper	Dapper W/O connection	ADO.NET	ADO.NET W/O connection
First run (ticks)	2510343	4273887	3499499	3447487	246980
First run (ms)	248	414	355	351	40
Average execution time (ticks)	44839	8590	6888	7092	5319
Average execution time (ms)	1	1	1	1	1
Spent memory (bytes)	5048	1720	1688	1688	896

Fig. 6. Results of experiments with the request GetSeasonsQuerry

Fig. 7 shows the results of experiments with the GetSeasonsTotalPrice query. In terms of the speed of the first run, the situation is similar to the previous requests, judging by the metrics in ticks and ms. In terms of average execution time, ADO.NET w/o connection and ADO.NET lead, followed by Dapper w/o connection and Dapper, and the Entity Framework query took the longest, which lagged behind ADO.NET w/o connection by almost 2 times.

Request	GetSeasonTotalPrice				
	Entity Framework	Dapper	Dapper W/O connection	ADO.NET	ADO.NET W/O connection
First run (ticks)	2983407	4065444	3811651	4184960	199574
First run (ms)	281	441	359	349	27
Average execution time (ticks)	13621	9117	9328	8586	7400
Average execution time (ms)	1	1	1	1	1
Spent memory (bytes)	9992	3022	2978	2764	1586

Fig. 7. Results of experiments with the request GetSeasonsTotalPrice

Fig. 8 shows the results of experiments with the CreateCategory request. Technologies can be divided according to the execution time of the first launch as follows: ADO.NET w/o connection, Entity Framework, ADO.NET, Dapper w/o connection and Dapper. According to the average execution time in ms, you can see that Entity Framework lags behind significantly. A more revealing ratio can be obtained by comparing the average execution time in ticks, where it can be seen that ADO.NET w/o connection is 2.6 times faster than Entity Framework. According to the used memory, similar to previous requests, Entity Framework uses 3 times more resources than other access technologies.

Request	CreateCategory				
	Entity Framework	Dapper	Dapper W/O connection	ADO.NET	ADO.NET W/O connection
First run (ticks)	1592580	3900297	3601395	3294704	214774
First run (ms)	161	421	329	313	16
Average execution time (ticks)	26597	11690	11098	10710	9817
Average execution time (ms)	2.5	1.5	1	1	1
Spent memory (bytes)	7664	2554	2474	2468	1238

Fig. 8. Results of experiments with the request CreateCategory

Fig. 9 shows the results of experiments with DeleteCategory queries. By the time of execution of the first run, the same sequence of technologies is preserved, as in the previous requests. But in terms of average execution time, Entity Framework is now the fastest: 5 times faster than Dapper and 4 times faster than ADO.NET. This can be seen by the average execution time in ticks and ms. In terms of memory usage, everything is similar to the previous queries, where Entity Framework uses many times more memory than the other investigated technologies.

Request	DeleteCategory				
	Entity Framework	Dapper	Dapper W/O connection	ADO.NET	ADO.NET W/O connection
First run (ticks)	2949541	3714834	3741236	3655405	196057
First run (ms)	273	375	318	329	17
Average execution time (ticks)	15092	77176	73560	71484	63049
Average execution time (ms)	1.3	7	6.9	6.4	6.7
Spent memory (bytes)	9064	2288	1934	936	772

Fig. 9. Results of experiments with the request DeleteCategory

Based on the analysis of the results of the experiments and taking into account the provided functionality, separate recommendations can be formulated for each of the technologies.

ADO.NET technology has shown itself to be very resource-efficient compared to the ORM and Micro ORM studied. The execution of the request is almost the fastest for almost all types of requests. Certain problems can arise only due to the incorrect writing of queries in the SQL language. That is, the use of this technology requires the developer to have some experience with SQL. Other complications may arise when it is necessary to take the data returned from the database, because there is no internal automapping in ADO.NET.

Dapper technology showed itself very well in the speed of execution of requests, where there was almost no lag behind ADO.NET. This Micro ORM outperformed Entity Framework by several times. You can see from the memory consumption why Dapper is considered a lightweight Micro ORM. It consumes resources in almost the same way as ADO.NET. As with ADO.NET, Dapper requires SQL knowledge to use, but it has no problem with automapping extracted data. This is a very useful tool that even allows you to populate fields in custom classes from Dapper extracted data belonging to other classes.

Entity Framework is a very popular ORM and it is fully confirmed by experiments why this is so. Thanks to the internal functionality, you can use different development approaches, such as Code First, DataBase and Model First, perform database migrations, easily write queries that the ORM will automatically send to the DB server. But there are disadvantages to this. As you can see, the more complex the query, the longer it took to use Entity Framework compared to other access technologies. But when the queries are simple, Entity Framework took less time than Dapper.

5. Conclusions

In the work were investigated such database access technologies as ORM Entity Framework, Micro ORM Dapper and ADO.NET from the point of view of performance when working with the popular RDBMS MS SQL Server. A series of experiments was conducted to measure the performance of database queries.

To conduct the research, a software solution was developed using the .NET platform, C# 7, ASP.NET Core Web API, Swagger. To conduct experiments, a relational database in the field of e-commerce and a set of requests for performing CRUD operations were designed, the performance of which was investigated.

During the experiments, metrics were used regarding the speed of the first request and the average speed in milliseconds and ticks, the amount of memory spent on the request (byte).

The research showed that none of the technologies used can be called unequivocally the best. Based on the results of the experiments and taking into account the features of the functionality, we can conclude that if the development of a simple application is planned or it is necessary to speed up the execution of requests to the database as much as possible, it is better to use Micro ORM Dapper. If a large and complex program is being created, and at the same time it is planned to use an object-oriented approach, then ORM Entity Framework may be the best choice. ADO.NET is more efficient for executing complex queries, especially if they require optimization or use special database functions, ADO.NET works at a lower level of abstraction compared to Dapper and Entity Framework and provides direct control over transaction management, i.e. the ability to manually start, commit or cancel transactions, which gives you more flexibility and control over this process.

REFERENCES

[1] Filatov, V., & Semenets, V. (2018). Methods for Synthesis of Relational Data Model in Information Systems Reengineering Problems. In 2018 International Scientific-Practical Conference Problems of Infocommunications. Science and Technology (PIC S&T). IEEE.

[2] Maran, V., Machado, A., Machado, G. M., Augustin, I., de Oliveira, J. P. M. (2018), "Domain content querying using ontology-based context-awareness in information systems", Data and Knowledge Engineering, No. 115, P. 152–173. DOI: 10.1016/j.datak.2018.03.003.

[3] Michael Lee, Gentry Bieker: SQL Server 2008. DOI: https://doi.org/10.1002/9781118257388.ch17

[4] Christian Nagel Professional C# 7 and .NET Core 2.0 DOI: https://doi.org/10.1002/9781119549147.ch31

[5] Pérez-Castillo, R., De Guzmán, I. G. R., Caivano, D., Piattini, M. (2012), "Database schema elicitation to modernize relational databases", ICEIS 2012 - Proceedings of the 14th International Conference on Enterprise Information Systems, P. 126–132.

[6] Maran M. M., Paniavin N. A., Poliushkin I. A. Alternative Approaches to Data Storing and Processing. V International Conference on Information Technologies in Engineering Education (Inforino). 2020. P. 1–4, DOI: https://doi.org/10.1109/inforino48376.2020.9111708

[7] Renée M. P. Teate SQL for Data Scientists: A Beginner's Guide for Building Datasets for Analysis. DOI: https://doi.org/10.1002/9781119669388.ch1

[8] Filatov, V., Radchenko, V. (2015), "Reengineering relational database on analysis functional dependent attribute", Proceedings of the X Intern. Scient. and Techn. Conf. "Computer Science & Information Technologies" (CSIT'2015), 14-17 sept. 2015, Lviv, Ukraine, P. 85–88.

[9] Sahatqija, K., Ajdari, J., Zenuni, X., Raufi, B., Ismaili, F., (2018), "Comparison between relational and NOSQL databases", 41st International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO), P. 216-221. DOI: https://doi.org/10.23919/mipro.2018.8400041

[10] Ying Bai SQL Server Database Programming with Visual Basic.NET: Concepts, Designs and Implementations. DOI: https://doi.org/10.1002/9781119608493.ch3

[11] Ying Bai Oracle Database Programming with Visual Basic. NET: Concepts, Designs, and Implementations. DOI: https://doi.org/10.1002/9781119734529.ch3

[12] Itzik Ben-Gan. Microsoft SQL Server 2012 T-SQL Fundamentals - Microsoft Press, 1st edition July 15, 2012.- 442 c.

[13] Christian Nagel Entity Framework Core. DOI: https://doi.org/10.1002/9781119549147.ch26

[14] Riadh Ghlala Analytic SQL in SQL Server 2014/2016. DOI: https://doi.org/10.1002/9781119649540.ch1

[15] Ying Bai Practical Database Programming with Visual C# .NET DOI: https://doi.org/10.1002/9780470567845.ch5

[16] Jonathan Eckstein, Bonnie R. Schultz Introductory Relational Database Design for Business, with Microsoft Access. DOI: https://doi.org/10.1002/9781119430087.ch4

[17] Paulraj Ponniah Ph.D. Database Design and Development: An Essential Guide for IT Professionals. DOI: https://doi.org/10.1002/0471728993.ch1

[18] Bagui, S., Earp, R. (2011), Database Design Using Entity-Relationship Diagrams (Foundations of Database Design), Auerbach Publications, 371 P., ISBN 978-143-986-177-6. DOI: https://doi.org/10.1201/9781439861776

The article was delivered to editorial stuff on the 15.02.2023

УДК 004.4: 004.4 DOi 10.30837/ bi.2023.1(99).02

К.С. Смеляков[1] , І.В. Кириченко[2], Г.Ю. Терещенко[3], Д.П. Панасенко[4]

[1]ХНУРЕ, м. Харків, Україна, kyrylo.smelyakov@nure.ua, ORCID iD: 0000-0001-9938-5489
[2]ХНУРЕ, м. Харків, Україна, iryna.kyrycheno@nure.ua, ORCID iD: 0000-0002-7686-6439
[3]ХНУРЕ, м. Харків, Україна, hlib.tereshchenko@nure.ua, ORCID iD: 0000-0001-8731-2135
[4]ХНУРЕ, м. Харків, Україна, daniil.panasenko@nure.ua

ВИКОРИСТАННЯ МАШИННОГО НАВЧАННЯ ДЛЯ ОПТИМІЗАЦІЇ ДОСТУПУ ДО ДАНИХ В ГІБРИДНОМУ СХОВИЩІ ЗОБРАЖЕНЬ

Робота присвячена дослідженню можливостей інтеграції машинного навчання для оптимізації доступу до даних в гібридному сховищі зображень. Основна задача полягає у пошуку схожих зображень серед великої кількості візуальних даних, що зберігаються у гібридному сховищі. Було розроблено систему, яка використовує методи глибинного навчання для вилучення ознак зображень, зокрема модель ResNet50, яка забезпечує високу точність вилучення ознак завдяки своїй глибинній архітектурі. Для ефективного пошуку схожих зображень застосовувалися MongoDB для зберігання зображень та метаданих, а також ElasticSearch для швидкої індексації та пошуку за векторами ознак. Проведено експериментальні дослідження з використанням датасету зображень різних тварин для оцінки продуктивності запропонованого підходу. Результати дослідження показали, що обраний підхід забезпечує високу швидкість та точність пошуку схожих зображень, підтверджуючи доцільність використання гібридних сховищ з використанням методів машинного навчання для ефективного управління великими обсягами візуальних даних. Використання попередньо навчених моделей значно знижує витрати на обчислювальні ресурси та час, необхідний для навчання, забезпечуючи при цьому високу точність і ефективність результатів.

ГІБРИДНЕ СХОВИЩЕ ЗОБРАЖЕНЬ, ДОСТУП ДО ДАНИХ, МАШИННЕ НАВЧАННЯ, НЕЙРОННА МЕРЕЖА, ELASTICSEARCH, MONGO DB, PYTHON

K.S. Smelyakov, I.V. Kyrychenko, G.Yu. Tereshchenko, D.P. Panasenko. Using Machine Learning to Optimize Data Access in Hybrid Image Storage. This work is dedicated to exploring the integration of machine learning to optimize data access in a hybrid image storage system. The primary task is to search for similar images among a large volume of visual data stored in the hybrid repository. A system has been developed that employs deep learning methods for feature extraction from images, specifically using the ResNet50 model, which provides high accuracy in feature extraction due to its deep architecture. For efficient image search, MongoDB was used for storing images and metadata, while ElasticSearch was utilized for fast indexing and vector feature search. Experimental studies were conducted using a dataset of various animal images to assess the performance of the proposed approach. The research results demonstrated that the chosen approach ensures high speed and accuracy in searching for similar images, confirming the feasibility of using hybrid storage with machine learning methods for effective management of large volumes of visual data. The use of pre-trained models significantly reduces computational resource costs and the time required for training, while still providing high accuracy and efficiency in the results.

HYBRID IMAGE STORAGE, DATA ACCESS, MACHINE LEARNING, NEURAL NETWORK, ELASTICSEARCH, MONGO DB, PYTHON

Вступ

В сучасному світі обсяг даних постійно зростає і значну частину цих даних становлять зображення. Важливість ефективного зберігання, управління та доступу до візуальних даних важко переоцінити, оскільки ці задачі виникають у різних галузях, включаючи медицину, безпеку, комерцію, наукові дослідження та багато інших.

Зберігання та управління зображеннями є важливим завданням, яке потребує ефективних рішень. Звичайні реляційні бази даних часто не забезпечують достатньої продуктивності для зберігання та обробки великих обсягів візуальних даних. Для таких задач підходять нові підходи, такі як NoSQL бази даних, об'єктні сховища та гібридні сховища.

Гібридні сховища поєднують різні типи сховищ, надаючи можливість зберігати дані в оптимальному форматі в залежності від їх характеристик та вимог до доступу. Наприклад, метадані зображень можуть зберігатися в реляційних або документних базах даних, тоді як самі зображення можуть зберігатися в об'єктних сховищах. Це забезпечує ефективне управління та доступ до даних.

Машинне навчання значно змінює підходи до обробки візуальних даних. ML моделі можуть використовуватися для автоматичного витягання ознак зображень, класифікації, розпізнавання об'єктів, покращення якості зображень та багато іншого. Використання попередньо навчених моделей дозволяє значно знизити час та ресурси, необхідні для навчання, забезпечуючи при цьому високу точність і ефективність.

Гібридні сховища даних у поєднанні з методами машинного навчання надають потужний інструмент для ефективного управління великими обсягами візуальних даних.

Таким чином, сучасні підходи до зберігання та обробки зображень базуються на використанні гібридних сховищ та методів машинного навчання, що забезпечує

ефективне управління візуальними даними, підвищує швидкість та точність доступу до них, а також дозволяє вирішувати складні завдання в різних галузях.

1. Гібридні сховища зображень

Гібридні сховища даних представляють собою комбінацію різних типів сховищ, які працюють разом для забезпечення оптимальної продуктивності, масштабованості та гнучкості [1]. Вони поєднують переваги реляційних баз даних (SQL) та нереляційних баз даних (NoSQL) для задоволення різних потреб у зберіганні та обробці даних.

Основна ідея гібридного сховища полягає в тому, щоб використовувати найкращі властивості кожного типу сховища для вирішення конкретних завдань. Наприклад, реляційні бази даних відмінно підходять для транзакційних систем, де важлива цілісність та консистентність даних, тоді як NoSQL бази даних, такі як MongoDB, забезпечують високу продуктивність та масштабованість для зберігання великих обсягів даних, особливо неструктурованих або напівструктурованих.

Основні особливості використання гібридних сховищ зображення.

Гнучкість у виборі технологій: Гібридні сховища дозволяють використовувати різні бази даних для різних типів даних. Це означає, що можна вибрати найкращий інструмент для кожної конкретної задачі.

Масштабованість: Використання NoSQL баз даних, таких як MongoDB, дозволяє легко масштабувати сховище для обробки великих обсягів даних. Це особливо важливо для додатків, які мають справу з великими обсягами неструктурованих даних.

Висока продуктивність: Гібридні сховища забезпечують високу продуктивність за рахунок використання NoSQL баз даних для швидкого доступу до великих обсягів даних та реляційних баз даних для забезпечення цілісності транзакцій.

Забезпечення цілісності даних: Реляційні бази даних, такі як PostgreSQL, або MySQL, забезпечують цілісність даних та підтримку складних транзакцій, що є важливим для критично важливих додатків.

Зниження витрат: Використання гібридних сховищ дозволяє знизити витрати за рахунок оптимізації використання ресурсів. Наприклад, часто використовувані дані можуть зберігатися у швидкодоступних NoSQL базах даних, тоді як менш часто використовувані дані можуть зберігатися у дешевших реляційних базах даних.

Проблеми, які вирішують гібридні сховища.

Управління різноманітними типами даних: Гібридні сховища дозволяють ефективно управляти як структурованими, так і неструктурованими даними. Це особливо корисно для додатків, які обробляють різноманітні типи даних, такі як текстові документи, зображення, відео та інші мультимедійні дані.

Швидкий доступ до даних: Використання NoSQL баз даних дозволяє забезпечити швидкий доступ до великих обсягів даних, що є критично важливим для додатків з високими вимогами до продуктивності.

Масштабованість: Гібридні сховища легко масштабуються, що дозволяє забезпечити безперебійний доступ до даних навіть при збільшенні обсягів даних.

Забезпечення безпеки даних: Реляційні бази даних забезпечують високий рівень безпеки даних, що є важливим для додатків, які обробляють конфіденційну інформацію.

Загалом, гібридні сховища забезпечують гнучкість, продуктивність та масштабованість, що робить їх ідеальним рішенням для багатьох сучасних завдань, пов'язаних з обробкою та зберіганням даних. Гібридні сховища дозволяють ефективно управляти різноманітними типами даних та забезпечувати швидкий доступ до них, що є критично важливим у сучасному світі великих даних.

2. Застосування машинного навчання

Машинне навчання стає дедалі важливішим інструментом у різних сферах, де зберігаються та обробляються великі обсяги зображень. Його застосування в поєднанні з гібридними сховищами зображень дозволяє вирішувати низку складних задач, що вимагають автоматизації та підвищеної точності.

Ось кілька прикладів задач, які можуть бути ефективно вирішені за допомогою машинного навчання в контексті гібридних сховищ зображень:

– автоматична класифікація зображень: машинне навчання може бути використане для автоматичної класифікації зображень у великій базі даних [2];

– аналіз і обробка зображень: алгоритми машинного навчання здатні автоматично аналізувати зображення для виявлення певних об'єктів або особливостей;

– сегментація зображень: у сфері обробки зображень важливо вміти сегментувати зображення, тобто розділяти його на значущі частини, машинне навчання дозволяє автоматично виділяти об'єкти на зображеннях;

– пошук схожих зображень: це одна з найпоширеніших задач, яка може бути вирішена за допомогою машинного навчання – система, що використовує алгоритми машинного навчання, здатна швидко знайти схожі зображення в базі даних на основі їхніх ознак;

– анотація та розпізнавання зображень: машинне навчання може автоматично додавати анотації до зображень, розпізнаючи об'єкти або сцени і це спрощує процес організації та пошуку зображень у великих базах даних.

Використання машинного навчання в контексті гібридних сховищ зображень дозволяє значно покращити ефективність та точність обробки візуальних даних.

3. Постановка задачі

Розібравшись в особливостях доступу до даних в гібридних сховищах зображень і проблемами які вони вирішують, для експериментальної інтеграції машинного навчання було вирішено взяти показову задачу, яка вирішується завдяки машинному навчанню. Однією з найактуальніших задач в цій області є пошук схожих зображень. Ця задача полягає в тому, щоб знайти велику кількість зображень, схожих на задане, серед інших зображень в гібридному сховищі зображень. Пошук схожих зображень є дуже практичною задачею, вирішення якої може принести значну користь у різних сферах.

Проблема, яка вирішується в даній роботі, полягає в оптимізації процесу пошуку схожих зображень у гібридному сховищі.

Основна мета цього дослідження – визначити, які методи машинного навчання можуть бути інтегровані в процес пошуку для підвищення його ефективності.

Важливо також оцінити, як саме використання гібридного сховища впливає на ефективність рішення даної задачі, а також які переваги застосування такого підходу.

4. Вибір методу машинного навчання

Вибір підходу до інтеграції машинного навчання є критично важливим етапом данного дослідження, оскільки саме методи машинного навчання визначають ефективність і точність пошуку схожих зображень у гібридному сховищі. Існує кілька популярних методів і моделей, які можуть вирішувати цю задачу [3].

Серед найбільш поширених підходів для пошуку схожих зображень можна виділити наступні методи машинного навчання.

Автокодувальники (Autoencoders) [4]: ці нейронні мережі можуть навчитися стискати дані до латентного простору меншої розмірності, зберігаючи важливу інформацію. Вони часто використовуються для вилучення ознак і можуть бути застосовані для порівняння схожості зображень. Автокодувальники можуть працювати з неструктурованими даними і вилучати значущі ознаки, але вони можуть бути менш точними у порівнянні з більш складними моделями, такими як CNN або ResNet.

Сверточні нейронні мережі (Convolutional Neural Networks, CNNs) [5]: CNN широко використовуються для аналізу зображень. Вони можуть вилучати складні ознаки з зображень, що робить їх ідеальними для задач класифікації та пошуку схожих зображень. CNN здатні вилучати складні та значущі ознаки зображень, що робить їх високоефективними для задач класифікації, але навчання CNN може бути ресурсомістким та вимагати великої кількості даних.

Мережі глибинного навчання для вилучення ознак [6] (Deep Feature Extraction): Цей підхід полягає у використанні попередньо навчених моделей, таких як VGG, Inception або ResNet, для вилучення ознак зображень, які потім використовуються для порівняння схожості [7]. Використання попередньо навчених моделей дозволяє знизити витрати на обчислювальні ресурси та час на навчання, а також забезпечує високу точність результатів, але можуть виникати проблеми з адаптацією до специфічних завдань або даних.

Зібрані дані про методи машинного навчання для пошуку схожих зображень, було зведено в порівняльну таблицю.

Таблиця 1

Порівняння методів машинного навчання

Метод	Переваги	Недоліки
Автокодувальники	Працюють з неструктурованими даними, вилучають значущі ознаки	Менш точні
CNN	Вилучають складні та значущі ознаки, висока ефективність	Ресурсомісткі, потребують великої кількості даних
Глибинне вилучення ознак	Знижені витрати на обчислення та навчання, використання попередньо навчених моделей, висока точність результатів	Можливі проблеми з адаптацією до специфічних завдань чи даних

Виходячи з проведеного аналізу було вирішено для інтеграції машинного навчання в систему пошуку схожих зображень обрати метод глибинного вилучення ознак, так як він найбільш підходить для проведення даного дослідження, не вимагає витрат на навчання, через використання попередньо навчених моделей, а також надає високу точність результатів.

Серед цих методів ми обрали модель ResNet50 (Residual Networks) [8]. ResNet50 є однією з найпопулярніших моделей для вилучення ознак завдяки своїй високій точності. Модель ResNet50 вже попередньо навчена на великому датасеті зображень, що дозволяє використовувати її для вилучення ознак без додаткового навчання, знижуючи витрати на обчислювальні ресурси. Крім того, ResNet50 забезпечує високу точність результатів, що робить її ідеальним вибором для нашого дослідження.

ResNet50 – це глибока нейронна мережа, яка складається з 50 шарів і використовує концепцію "залишкових блоків" (residual blocks) для подолання проблеми затухання градієнтів у глибоких мережах. Основна ідея залишкових блоків полягає в тому, що кожен блок не намагається вивчити точну карту відображення вхідних даних до вихідних, а замість цього вивчає різницю (залишок) між вхідними та вихідними даними.

Коли зображення подається на вхід ResNet50, воно проходить через кілька згорткових шарів, шарів підвибірки (pooling layers) і повнозв'язаних шарів.

На кожному етапі мережа вилучає все більш абстрактні ознаки, що дозволяє моделі будувати багатошарове представлення вхідного зображення. На виході ResNet50 ми отримуємо вектор ознак (feature vector), який можна використовувати для порівняння схожості з іншими зображеннями.

Для порівняння схожості між векторами ознак ми використовуємо метрику косинусної подібності [9]. Косинусна подібність вимірює кут між двома векторами у векторному просторі і визначає, наскільки ці вектори схожі між собою. Воно обчислюється за формулою 1:

$$similarity = \cos(\theta) = \frac{A \cdot B}{\|A\| \|B\|}, \quad (1)$$

де A і B – два вектори.

Косинусна подібність набуває значень від –1 до 1, де 1 означає, що вектори ідеально вирівняні (максимальна схожість), 0 означає, що вони ортогональні (немає схожості), а –1 означає, що вони ідеально протилежні. Для задачі пошуку схожих зображень ми зазвичай розглядаємо значення косинусної подібності від 0 до 1.

5. Вибір гібридного сховища зображень

Виходячи з проведеного аналізу, визначеної задачі і обраного методу машинного навчання, визначимо що функціональністю гібридного сховища зображень має бути: збереження зображень та пошук за вектором ознак. Вибір БД для виконання поставленої задачі проведемо за даною необхідною функціональністю.

Виділимо головні вимоги до збереження зображень:

– підтримка великих обсягів даних;

– гнучка схема зберігання для різних форматів даних (зображення, метадані);

– висока продуктивність при записі і читанні даних;

– масштабованість і надійність.

Виходячи з вимог до збереження зображень, БД, які відповідають цим вимогам, можуть бути такі БД, як: MongoDB, Cassandra та Couchbase [10]. В якості БД для збереження зображень для виконання дослідження було обрано MongoDB, так як вона являється найбільш популярною, з великою спільнотою користувачів.

Виділимо головні вимоги до пошуку за векторами ознак:

– швидка індексація та пошук за векторними ознаками;

– підтримка складних запитів і агрегацій;

– висока продуктивність при обробці великих обсягів даних;

– масштабованість і надійність.

Виходячи з вимог до пошуку за векторами ознак, БД, які відповідають цим вимогам можуть бути, такі БД, як: Elasticsearch та Faiss [11]. В якості БД для здійснення пошуку за векторами ознак для виконання дослідження було обрано Elasticsearch. Elasticsearch є оптимальним вибором для пошуку за векторами ознак завдяки своїй високій швидкості пошуку, швидкій індексації, підтримці складних запитів та агрегацій, а також розподіленій архітектурі. Крім того, Elasticsearch має велику та активну спільноту, що забезпечує підтримку та розвиток технології.

6. Опис програмного рішення

Програмне рішення базується на використанні двох головних компонентів. Перша компонента – це розгорнутий Docker Compose, який відповідає за контейнеризацію гібридного сховища зображень. Друга компонента це Python – додаток у JupyterLab, який використовується для виконання основного коду даного дослідження.

В якості гібридного сховища зображень було обрано використання комбінації двох баз даних – MongoDB та ElasticSearch.

MongoDB – це документо-орієнтована база даних, яка забезпечує ефективне зберігання та доступ до великих обсягів даних. Це найбільш ефективний спосіб зберігання зображень у базі даних. Кожне зображення зберігається у форматі BSON разом з інформацією про його метадані, такі як назва файлу, розмір та формат.

Так як задача даного дослідження вимагає пошук схожих зображень за допомогою векторних подібностей, було обрано ElasticSearch. ElasticSearch – це найкращий інструмент для швидкого пошуку векторів ознак зображень, так як в цій БД присутня індексація та Elasticsearch підтримує потужні функції, що робить його ідеальним вибором для виконання даного завдання [12].

Docker Compose використовується для управління контейнерами, що забезпечують роботу MongoDB та Elasticsearch. Це дозволяє легко розгортати та масштабувати систему, забезпечуючи ізоляцію середовища виконання для кожного компонента.

Python-додаток складається з кількох функціональних модулів, які забезпечують виконання різних функцій.

Модуль завантаження зображень: відповідає за завантаження зображень у гібридне сховище даних. Він обробляє завантажені зображення та зберігає їх у MongoDB і ElasticSearch.

MongoDB-клієнт: забезпечує взаємодію з базою даних MongoDB для зберігання та отримання зображень та їх метаданих.

ElasticSearch-клієнт: Відповідає за індексацію та пошук векторів ознак зображень у Elasticsearch.

ML–модель: Використовується для генерації векторів ознак зображень. Було обрано модель ResNet50, яка завантажується з попередньо навченими вагами ImageNet [13]. За допомогою даної моделі відбувається порівняння схожості зображень.

Дана архітектура представлена у вигляді діаграми компонентів:

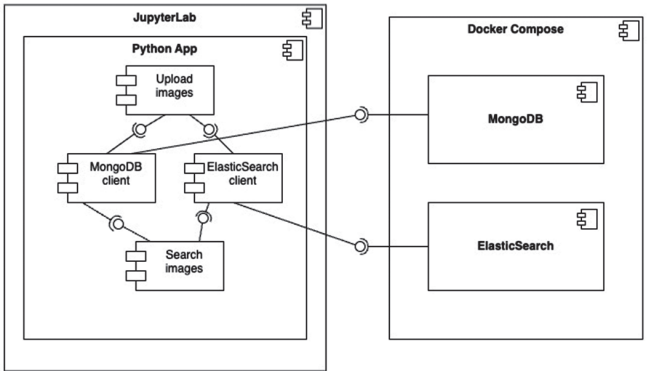

Рис. 1. Діаграма компонентів програмної реалізації дослідження

Ця архітектура забезпечить ефективну обробку, зберігання та пошук зображень, використовуючи переваги обраних нами сучасних технологій та інструментів.

Для успішної реалізації проекту було обрано широкий спектр технологій та інструментів, які забезпечують ефективну розробку, тестування та проведення якісного дослідження. У цьому розділі буде детально розглянуто кожну з обраних технологій та пояснені причини їх вибору.

Основною мовою програмування для проведення дослідження, було обрано мову програмування Python. Python відомий своєю простотою та читабельністю коду, що робить його дуже зручним інструментом. Крім того, Python має велику кількість бібліотек для роботи з даними та машинного навчання, що допоможе виконати поставлену задачу.

Для розробки були використані наступні бібліотеки.

Tensorflow Keras [14]. Для реалізації та тренування моделей машинного навчання використовується бібліотека Keras з бекендом TensorFlow. Це дозволило використати модель нейронної мережі, таку як ResNet50, яка використовується для визначення векторів ознак зображень. Використання попередньо натренованої моделі ResNet50 значно спростило процес машинного навчання для обробки зображень та підвищило точність результатів даного дослідження.

Pandas та NumPy [15]. Використання даних бібліотек допомагає дуже зручно обробляти і маніпулювати даними, багатовимірними масивами і числовими обчисленнями.

PyMongo та elasticsearch-py. Дані бібліотеки використовувались у якості клієнтів для забезпечення доступу до даних у сховища даних.

Matplotlib [15]. Для візуалізації результатів та аналізу даних обрана бібліотека Matplotlib. Вона дозволяє створювати графіки для візуального представлення результатів експериментів, а також виводити зображення у вікно виводу IDE.

Sklearn. З даної бібліотеки використовувалась функція cosine_similarity для визначення косинусної подібності векторів.

Розробка виконувалась в веб IDE – JupyterLab, що є інтерактивним середовищем для роботи з Python. JupyterLab дозволяє виконувати код, переглядати результати та візуалізувати дані у реальному часі, що значно спрощує процес дослідження та тестування моделей машинного навчання. Це середовище є особливо корисним для роботи з даними та побудови експериментів.

Для реалізації поставленої задачі також необхідно створити гібридне сховище зображень. В якості гібридного сховища зображень використовується комбінація MongoDB та ElasticSearch.

MongoDB – це документо-орієнтована база даних, яка забезпечує високу гнучкість та масштабованість. MongoDB дозволяє зберігати великі обсяги даних у форматі JSON та легко їх обробляти. MongoDB в даному дослідженні використовується для зберігання зображень та пов'язаних з ними метаданих.

ElasticSearch – це пошуковий рушій з відкритим вихідним кодом, який спеціалізується на повнотекстовому пошуку та аналізі великих обсягів даних у реальному часі. ElasticSearch дозволяє швидко проводити пошук схожих зображень за допомогою методів машинного навчання та підтримує індексацію та пошук векторних даних.

Kibana використовувалась в якості веб GUI для роботи з ElasticSearch.

MongoDB Compass використовувався в якості десктоп GUI для роботи з MongoDB.

Для забезпечення зручного розгортання та управління інфраструктурою проекту був обран Docker. Docker дозволяє створювати контейнери, які ізолюють середовище виконання додатків, забезпечуючи їхню портативність та легкість розгортання. Docker Compose був використаний для спрощення управління багатосервісною архітектурою, що включає MongoDB, Elasticsearch та Kibana.

7. Проведення експериментів

Результатами проведених експериментів мають бути дані для аналізу і формування висновків щодо доцільності використання обраного методу машинного навчання ResNet50 з гібридним підходом до зберігання зображень у вирішенні проблеми – швидкого пошуку схожих зображень в сховищі даних.

Для цього визначено провести порівняння трьох різних підходів для вирішення цієї проблеми.

Перший підхід. Збереження зображення в єдиному сховищі MongoDB. При пошуку схожих зображень перебрати усі зображення зі сховища, розрахувати вектор ознак і вирахувати коефіцієнт їхньої схожості.

Другий підхід. Збереження зображення в єдиному сховищі MongoDB з попереднім прорахунком вектору ознак зображення і збережені їх в цьому ж сховищі. При пошуку схожих зображень перебрати усі зображення зі сховища, взяти вектор ознак зі сховища і вирахувати коефіцієнт їхньої схожості.

Третій підхід. Збереження зображення в сховищі MongoDB. Попередній прорахунок вектора ознак зберігається в ElasticSearch. Для пошуку схожих зображень робимо запит в ElasticSearch, пошук відбувається на стороні сховища.

Проведення експерименту має відбуватись з використанням цих трьох підходів для визначення їх ефективності.

Для якісного проведення експерименту необхідно виконати ці три тести на різних обсягах даних.

Необхідно зробити заміри часу виконання кожного експерименту. Це і буде головним критерієм для оцінки

Експерименти будуть проводитись на локальному комп'ютері з 8 ядрами процесора M1 Apple Silicon, 8 ГБ оперативної пам'яті, використовуючи Python 3.8, MongoDB версії 4.4, Elasticsearch версії 7.10 та TensorFlow версії 2.4.

В якості тестових зображень, необхідно було обрати великий дата-сет зображень, гарної якості, одного формату, для якісного і зручного проведення експериментів. Зображення мають бути, як різноманітні, так і мати певну схожість (деякі з них), для наочного бачення коректності роботи програми.

Було обрано дата-сет з зображеннями різних тварин з Kaggle [16]. Даний дата-сет налічує 5400 зображень 90 видів тварин розподілених за окремими директоріями. Зображення якісні в форматі .jpg.

Під час виконання експериментів даний дата-сет буде завантажено в сховище і на основі нього буде відбуватися пошук схожих зображень.

Правильним результатом виконання тестових програм буде, те що результуючі схожі зображення відносяться до одного типу тварин.

Експеримент буде проведено на різних обсягах даних. Визначимо наступні обсяги даних – 10, 20, 50, 100, 500, 1000, 5400. Для цього буде взято необхідну кількість випадкових зображень з цього дата-сету.

Проведено визначені експерименти.

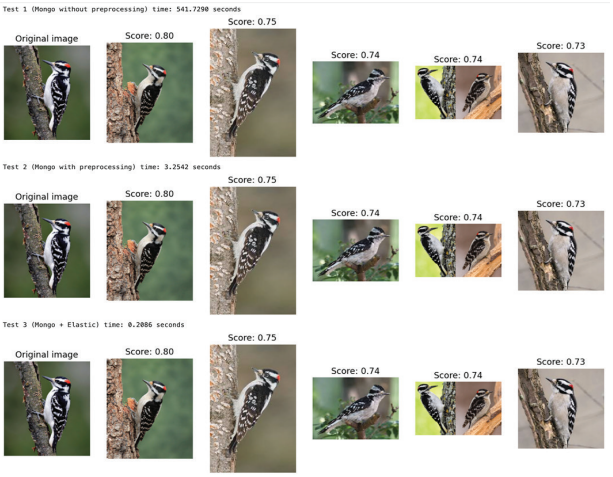

Рис. 2. Результати дослідження

Порівнявши час виконання кожного експерименту, варіант з використанням гібридного сховища зображень виявився найбільш ефективним. Усі зафіксовані результати трьох експериментів представлено в таблиці. Результати представлені у секундах.

Таблиця 2

Результати виконання експериментів

Експеримент	10	20	50	100	500	1000	5400
MongoDB	0.881	1.591	4.008	7.54	38.126	81.656	541.729
MongoDB з попереднім прорахунком	0.08	0.095	0.113	0.143	0.442	0.687	3.254
MongoDB з ElasticSearch	0.156	0.120	0.173	0.138	0.391	0.289	0.208

Результати проведених експериментів показують значні відмінності у швидкості пошуку схожих зображень між трьома підходами: використання тільки MongoDB, використання MongoDB з попереднім прорахунком векторів ознак, та використання гібридного сховища у якості комбінації MongoDB з Elasticsearch.

Для розуміння залежності часу виконання пошуку від обсягів даних побудуємо графік (див. рис. 3).

На графіку наглядно можна побачити лінійну залежність часу виконання від кількості зображень в методі використання тільки MongoDB без попереднього прорахунку. Причому даний підхід досить затратний. Це пояснюється тим, що кожного разу під час пошуку схожих зображень система повинна перебирати всі зображення та обчислювати вектори ознак «на льоту». Це призводить до суттєвих витрат часу при великих обсягах даних. Оцінка складності алгоритму – O(n).

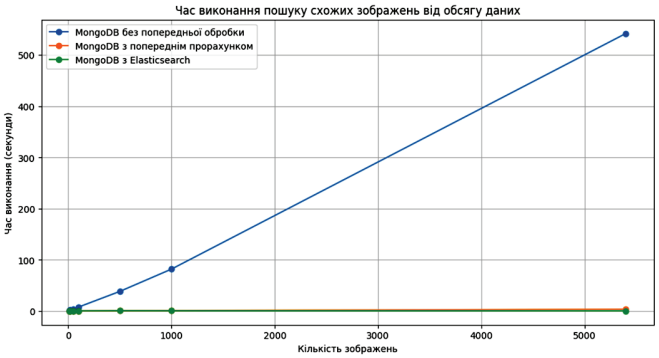

Рис. 3. Графік залежності часу виконання від обсягів даних трьох підходів

Для більш наглядної оцінки наступних двох підходів, було прийнято рішення виключити цей підхід з графіку (див. рис. 4).

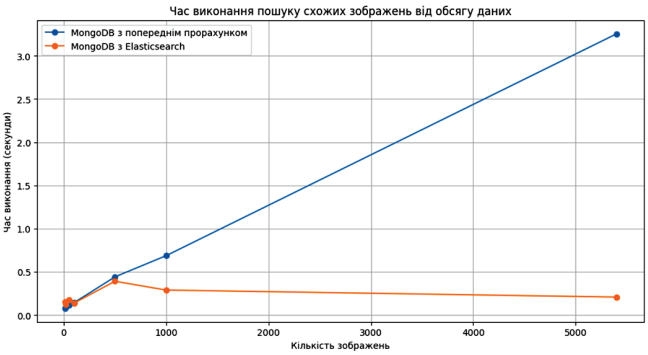

Рис. 4. Графік залежності часу виконання від обсягів даних двох підходів

Використання MongoDB з попереднім прорахунком векторів ознак значно покращує час виконання пошуку. Це пояснюється тим, що вектори ознак вже збережені у базі даних і система не витрачає час на їх обчислення під час пошуку. Однак все одно ми бачимо лінійне збільшення часу виконання від обсягів даних. Це пояснюється тим що виконується повний перебір в сховищі для пошуку найбільш схожих зображень. Оцінка складності алгоритму – O(n).

Зовсім інша залежність з використанням гібридного сховища зображень у якості MongoDB з ElasticSearch. На результатах не видно залежності часу від обсягів даних. Це обумовлюється тим що, використання ElasticSearch для пошуку максимально ефективне і за рахунок індексації має логарифмічну складність –O(log(n)), для якої обрані дані об'єми даних значно малі.

Використання гібридного сховища у якості комбінації MongoDB та Elasticsearch дозволяє максимально використати сильні сторони кожної технології. MongoDB забезпечує гнучке та надійне зберігання даних, тоді як Elasticsearch надає потужні можливості ефективного пошуку.

Висновки

В ході проведення дослідження було виконано аналіз предметної області та розглянуто можливості інтеграції машинного навчання для доступу до даних в гібридних сховищах зображення. Описано їх принципи роботи, виділено особливості, основні відмінності, переваги та недоліки.

У процесі роботи було поставлено задачу реалізувати алгоритм пошуку схожих зображень в сховищі, було розроблено та протестовано програмну частину, яка реалізує кілька підходів до зберігання та пошуку зображень, що дозволило зробити висновки щодо їх ефективності та доцільності використання в реальних умовах.

Перш за все, ми визначили, що традиційний підхід, який полягає у зберіганні зображень у єдиному сховищі MongoDB без попереднього розрахунку векторів ознак, є найменш ефективним з точки зору швидкості пошуку. Результати експериментів показали, що цей метод має значні затримки при обробці великих обсягів даних, що робить його непридатним для задач, де потрібна висока швидкість доступу до зображень. А при збільшені обсягів даних час виконання пошуку збільшується, що унеможливлює пошук.

Водночас, використання підходу з попереднім розрахунком векторів ознак та їх зберіганням у MongoDB дозволяє значно скоротити час пошуку схожих зображень. Такий підхід забезпечує швидший доступ до даних за рахунок попередньої обробки зображень і збереження результатів обчислень, що робить його більш ефективним рішенням на невеликих обсягах даних. Однак даний підхід все одно має лінійну залежність часу виконання від обсягів даних, тому не може використовуватись на великих обсягах даних.

Найбільш ефективним виявився гібридний підхід, який поєднує зберігання зображень у MongoDB та векторів ознак у Elasticsearch. Завядки індексації і потужним можливостям пошуку ми досягли логарифмічної залежності часу виконання від обсягів даних, що робить його Цей метод дозволяє не тільки швидко знайти схожі зображення, але й забезпечує гнучкість у пошукових запитах та високу точність результатів завдяки можливостям Elasticsearch. Експерименти підтвердили, що цей підхід має найнижчий час виконання запитів, особливо при обробці великих обсягів даних.

Таким чином, результати дослідження підтверджують доцільність використання гібридних сховищ зображень з інтеграцією методів машинного навчання для оптимізації процесу пошуку. Використання MongoDB та Elasticsearch у поєднанні з попереднім розрахунком векторів ознак дозволяє досягти високої ефективності та точності при роботі з великими обсягами візуальних даних.

Важливим висновком є також те, що правильний вибір методів машинного навчання та технологій зберігання даних значно впливає на результати роботи системи. Використання моделі ResNet50 для екстракції ознак зображень та алгоритму косинусної подібності для порівняння векторів ознак виявилося ефективним рішенням для задачі пошуку схожих зображень.

Загалом, проведене дослідження показало, що інтеграція машинного навчання в процеси управління та обробки даних у гібридних сховищах зображень відкриває нові можливості для покращення ефективності та точності роботи з великими обсягами візуальної інформації. Це дозволяє забезпечити швидкий та надійний доступ до необхідних даних у різних сферах застосування, включаючи медицину, безпеку та електронну комерцію.

Мета завдання досягнута за рахунок визначення у роботі ефективності використання методів машинного навчання для доступу до даних в гібридних сховищах зображень. У результаті роботи було підібрано найефективнішу модель машинного навчання і гібридне сховище зображень для вирішення поставленої задачі.

Результати цього дослідження можуть бути використані для прийняття рішення про те, який підхід використовувати для збереження і організації зображень у сховищі і який алгоритм та модель машинного навчання використовувати для пошуку схожих зображень.

Список літератури:

[1] James, Blessing & Asagba, P. (2017). Hybrid Database System for Big Data Storage and Management. International Journal of Computer Science, Engineering and Applications. 7. 15-27. 10.5121/ijcsea.2017.7402.

[2] Kyrychenko, I., Tereshchenko, G. Proniuk, G., Geseleva, N. "Predicate Clustering Method and its Application in the System of Artificial Intelligence", 2023 7th International Conference on Computational Linguistics and Intelligent Systems (COLINS-2023), 2023. – CEUR-WS, 2023, ISSN 16130073. - Volume 3396, PP. 395 - 406.

[3] Kyrychenko, I., Nazarov, O., Huliiev, N., Avdieiev, O. "Selection of Artificial Neural Networks for Disease Prediction", 2023 7th International Conference on Computational Linguistics and Intelligent Systems (COLINS-2023), 2023. – CEUR-WS, 2023, ISSN 16130073. - Volume 3387, PP. 236-248.

[4] Berahmand, Kamal & Daneshfar, Fatemeh & Salehi, Elaheh & Li, Yuefeng & Xu, Yue. (2024). Autoencoders and their applications in machine learning: a survey. Artificial Intelligence Review. 57. 10.1007/s10462-023-10662-6.

[5] Tabian, I., Fu, H., & Khodaei, Z.S. (2019). A Convolutional Neural Network for Impact Detection and Characterization of Complex Composite Structures. Sensors, 19(22), 4933. DOI: 10.3390/s19224933.

[6] Asokan, Anju & Jude, Anitha & Patrut, Bogdan & Danciulescu, Dana & D, Jude. (2020). Deep Feature Extraction and Feature Fusion for Bi-temporal Satellite Image Classification. Computers, Materials & Continua. 66. 373-388. 10.32604/cmc.2020.012364.

[7] He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 770-778. doi:10.1109/CVPR.2016.90.

[8] Meghana, Avuthu Sai. "Age and Gender prediction using Convolution, ResNet50 and Inception ResNetV2." International Journal of Advanced Trends in Computer Science and Engineering 9, no. 2 (April 25, 2020): 1328–34. http://dx.doi.org/10.30534/ijatcse/2020/65922020.

[9] Understanding Cosine Similarity and Cosine Distance in Depth. URL: https://pub.aimind.so/understanding-cosine-similarity-and-cosine-distance-in-depth-cc91eac3ef2 (дата звернення: 29.04.2024).

[10] Matallah, Houcine & Belalem, Ghalem & Bouamrane, K.. (2020). Evaluation of NoSQL Databases: MongoDB, Cassandra, HBase, Redis, Couchbase, OrientDB. International Journal of Software Science and Computational Intelligence. 12. 71-91. 10.4018/IJSSCI.2020100105.

[11] Znakhur, Serhii & Znakhur, Liudmyla. (2022). Similar goods search based on FAISS. Bulletin of Kharkov National Automobile and Highway University. 40. 10.30977/BUL.2219-5548.2022.96.0.40.

[12] Wong, Wai-Tak. (2019). Advanced Elasticsearch 7.0: A practical guide to designing, indexing, and querying advanced distributed search engines.

[13] Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A. C., & Fei-Fei, L. (2015). ImageNet Large Scale Visual Recognition Challenge. International Journal of Computer Vision, 115(3), 211-252. doi:10.1007/s11263-015-0816-y.

[14] Keras Applications: ResNet. URL: https://keras.io/api/applications/resnet/ (дата звернення: 20.04.2024).

[15] Nelli, Fabio. (2018). The NumPy Library: With Pandas, NumPy, and Matplotlib. 10.1007/978-1-4842-3913-1_3.

[16] Animal Image Dataset (90 Different Animals) on Kaggle. URL: https://www.kaggle.com/datasets/iamsouravbanerjee/animal-image-dataset-90-different-animals (дата звернення: 04.05.2024).

Надійшла до редколегії 1.06.2023

УДК 004.8:004.9 DOi 10.30837/ bi.2023.1(99).03

О.В. Чала[1], Є.О. Богатов[2]

[1]професор кафедри інформаційних управляющих систем,
Харківський національний університет радіоелектроніки, Україна, oksana.chala@nure.u

[2]асистент кафедри інформаційних управляющих систем, Харківський національний університет радіоелектроніки, Україна, ievgen.bogatov@nure.ua

ПОБУДОВА МОДЕЛІ БІЗНЕС-ПРОЦЕСУ З ВИКОРИСТАННЯМ ТЕМПОРАЛЬНИХ ЗНАНЬ ПРИ ВПРОВАДЖЕННІ ПРОЦЕСНОГО УПРАВЛІННЯ

Предметом дослідження є бізнес-процеси, що представляють собою послідовність робіт, яка забезпечується ресурсами та створює продукти та послуги, що мають цінність для клієнтів таких процесів. Метою є розробка підходу до вирішення задач впровадження процесного управління на основі автоматизованого виявлення темпоральних знань, що дозволяє встановити умови та обмеження на виконання дій процесу з урахуванням їх фактичної упорядкованості в часі і, тим самим, дає можливість врахувати персональні знання виконавців процесу як при побудові, так і при уточненні його моделі. Завдання: структуризація задач процесного управління з урахуванням взаємодії задач впровадження процесного управління та задач управління бізнес-процесами; розробка методу побудови прототипу бізнес-процесу з використанням темпоральних знань. Використовуваними підходами є: методи процесного управління, методи process mining, методи формування темпоральних знань. Наукова новизна отриманих результатів полягає в наступному. Розроблено метод побудови моделі бізнес-процесу «як є», що фактично виконується на підприємстві, з використанням темпоральних знань. Метод містить етапи формування темпоральних знань у вигляді правил, що визначають послідовність дій процесу у часі, виділення підмножин правил, що задають послідовне та паралельне або альтернативне виконання дій процесу, а також етапи моделі бізнес-процесу у вигляді графу потоків робіт, що враховує виявлені темпоральні правила. В практичному аспекті метод створює умови для реалізації постійного удосконалення бізнес-процесу на основі ітеративного виявлення та подальшого використання неявних знань виконавців, які знайшли відображення у темпоральних правилах.

ІНФОРМАЦІЙНА СИСТЕМА, БІЗНЕС-ПРОЦЕС, ПРОЦЕСНЕ УПРАВЛІННЯ, ЗНАННЯ, ТЕМПОРАЛЬНЕ ПРАВИЛО, ЖУРНАЛ ПОДІЙ, ПРИЧИННО-НАСЛІДКОВИЙ ЗВ'ЯЗОК

O.V. Chala, E.O. Bogatov Building a business process model using temporal knowledge when implementing process management. The article's subject matter are business processes, which are a sequence of work, which are provided with resources and create products and services that are valuable for the customers of such processes. The aim is to develop an approach to solving the problems of implementing process management based on the automated detection of temporal knowledge, which makes it possible to establish conditions and restrictions on the execution of process actions, considering their actual temporal order, and thus makes it possible to consider the personal knowledge of process performers both in building and refining their model. Objectives: To structure the process management tasks, considering the interaction between the subtasks of process management implementation and business process management tasks; to develop a method for building a business process prototype using temporal knowledge. The approaches used are process management methods, process mining methods, temporal knowledge methods. The scientific novelty of the obtained results is as follows. The article develops a method for building a model of the business process "as it is", performed in an enterprise, using temporal knowledge. The method includes stages of forming temporal knowledge in the form of rules that determine the sequence of process actions in time, selection of subsets of rules that specify sequential and parallel or alternative execution of process actions, and stages of building the business process model in the form of a workflow graph that considers the identified temporal rules. In practice, the method creates conditions for the implementation of continuous improvement of the business process based on the iterative identification and further use of the tacit knowledge of the performers, which is reflected in the temporal rules.

INFORMATION SYSTEM, BUSINESS PROCESS, PROCESS MANAGEMENT, KNOWLEDGE, TEMPORAL RULE, EVENT LOG, CAUSE AND EFFECT RELATIONSHIP

Вступ

Процесне управління підприємством орієнтовано на побудову та подальше використання моделей бізнес-процесів (БП) [1]. Такі моделі визначають послідовність виконання робіт, а також необхідні для цих робіт ресурси [2]. При впровадженні процесного управління виділяють п'ять рівнів зрілості – від початкового до рівня постійного покращення бізнес-процесів [3]. На початковому рівні зрілості процеси є неформалізованими. Їх виконання повністю базується на знаннях та ініціативі виконавців. На п'ятому рівні використовуються моделі бізнес-процесів, які повністю відображають особливості фактичних процесів, що виконуються на підприємстві. Перехід між рівнями зрілості передбачає постійне удосконалення процесних моделей. Вказане удосконалення потребує виявлення знань щодо особливостей виконання бізнес-процесу на конкретному підприємстві. Такі знання представлені в явній та неявній формі. Явні знання зазвичай задокументовані, а неявні – є персональними знаннями виконавців. Використання останніх приводить до змін порядку робіт процесу у часі, які не відображені в процесній моделі [4].

Існуючі підходи до побудови моделей бізнес-процесів базуються на використанні явних знань, а також додатковому опитуванню експертів з метою виявити неявні знання [5]. Проте не приділяється достатньо уваги виявленню неявних поведінкових знань виконавців процесу з метою подальшого використання цих знань для удосконалення процесної моделі, оскільки представлення таких знань у вербальній формі пов'язано із суттєвими труднощами. Проте неявні поведінкові знання можуть бути виявлені на основі аналізу записів про виконання бізнес-процесів [6, 7] та виявлення темпоральних залежностей, що задають упорядкованість пар подій бізнес-процесу у часі. Дослідження підходів до побудови й представлення темпоральних знань представлені в роботах [8-10]. Однак в наведених дослідження не розглядаються питання використання темпоральних знань для підтримки впровадження процесного управління. Зазначене свідчить про актуальність розробки підходу до впровадження процесного управління на основі побудови моделей бізнес-процесів з використанням темпоральних знань.

1. Постановка задачі

Метою статті є розробка підходу до вирішення задач впровадження процесного управління на основі автоматизованого виявлення темпоральних знань, що дозволяє встановити умови та обмеження на виконання дій процесу з урахуванням їх фактичної упорядкованості в часі і, тим самим, дає можливість врахувати персональні знання виконавців процесу як при побудові, так і при уточненні його моделі.

Для досягнення мети дослідження вирішуються такі задачі: структуризація задач процесного управління з урахуванням взаємодії задач впровадження процесного управління та задач управління бізнес-процесами; розробка методу побудови прототипу бізнес-процесу з використанням темпоральних знань.

2. Структуризація задач процесного управління

Процесний підхід до управління передбачає організацію управління підприємством з використанням процесного опису його діяльності.

Головна ідея процесного управління полягає у зменшенні витрат на організацію діяльності підприємства шляхом формування горизонтально-орієнтованих послідовностей робіт, що забезпечують ефективну взаємодію виконавців з різних підрозділів при створенні товарів та послуг для клієнтів підприємства. Ці послідовності робіт є ключовим елементом відповідних бізнес-процесів.

Безпосередня взаємодія виконавців робіт дає можливість суттєво знизити витрати на досягнення цілей підприємства із забезпеченням заданого рівня якості продукції. Якість продукції згідно процесного підходу визначається тим, наскільки результати бізнес-процесів становлять цінність для користувача.

Процесний опис діяльності підприємства містить моделі взаємопов’язаних бізнес-процесів. Моделі бізнес-процесів інтегрують послідовності робіт, що дають можливість досягти мети діяльності підприємства, необхідні для цих робіт ресурси, а також враховують організаційні аспекти управління бізнес-процесами.

Процесне управління передбачає вирішення двох груп задач: управління бізнес-процесами; впровадження процесного управління.

Управління бізнес-процесами включає вирішення таких задач: управління підприємством з використанням моделей бізнес-процесів; моніторинг виконання бізнес-процесів.

Властивості задач процесного управління наведено в табл. 1. Зв’язок між задачам процесного управління представлено на рис. 1.

Таблиця 1

Задачі процесного управління

Група задач	Задача	Властивості
Управління бізнес-процесами	Управління підприємством з використанням моделей бізнес-процесів	Реалізується управління за відхиленням. Використовуються показники результативності та ефективності для оцінки бізнес-процесу.
	Моніторинг виконання бізнес-процесів	Послідовність дій виконаного бізнес-процесу фіксується у журналі подій. Журнал подій містить окремі записи для кожної реалізації бізнес-процесу.
Впровадження процесного управління	Побудова й удосконалення моделей бізнес-процесів	Формується апріорно задана модель бізнес-процесу. Дана модель враховує явні знання про діяльність підприємства і може не враховувати неявні знання виконавців щодо дій в рамках процесу.
	Вилучення знань для удосконалення моделей бізнес-процесів	Формується модель бізнес-процесу «як є» на основі аналізу журналу подій. Дана модель відображає як явні, так і неявні знання щодо процесу. Модель порівнюється з апріорно заданою моделлю БП. На основі порівняння виявляються неявні знання, які дають можливість удосконалити бізнес-процес.

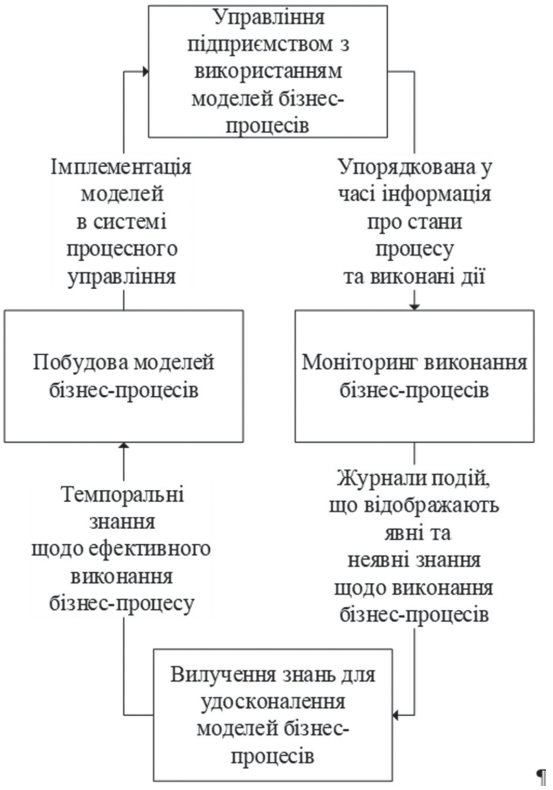

Рис. 1. Взаємозв'язок задач процесного управління

При імплементації першої задачі управління бізнес-процесами реалізується управління за відхиленням згідно послідовності робіт, визначеної у моделі бізнес-процесу. Оцінка виконання проводиться за показниками результативності та ефективності.

Показник результативності визначає ступінь досягнення результатів бізнес-процесу. Показник ефективності визначається з урахуванням вартості ресурсів, що були витрачені для досягнення вказаного результату.

Задача моніторингу бізнес-процесів передбачає фіксацію послідовностей робіт, які були виконані при реалізації цих процесів. Запис послідовностей робіт зазвичай виконується у журналах подій (логах). Журнал подій містить множину записів кожної реалізації БП.

Порівняння послідовності дій в журналі та в моделі БП дає можливість перевірити відповідність реального процесу та його модельного опису, що створює умови для постійного удосконалення процесних моделей.

При впровадженні процесного управління виконуються такі задачі: побудова моделей бізнес-процесів; вилучення знань для удосконалення моделей бізнес-процесів.

При вирішенні першої задачі зазвичай використовуються документальний опис діяльності підприємства, а також знання, що можуть бути вилучені в результаті бесід із експертами у відповідній предметній області.

На основі цих знань формується апріорна «ідеальна» модель виконання робіт бізнес-процесу. Проте дана модель не завжди враховує неявні знання виконавців процесу. Ці знання в більшості випадків не можуть бути вилучені в результаті бесід з виконавцями, тому що вони є неявними поведінковими (tacit) знаннями. Останні можуть бути вилучені за результатами виконання бізнес-процесів, що створює умови для подальшого удосконалення процесної моделі.

Задача удосконалення моделі бізнес-процесу полягає у виявленні невідповідностей між моделлю та реальним процесом, що виконується на підприємстві, і подальшій адаптації процесу з тим, щоб останній забезпечував досягнення заданих показників результативності та ефективності.

Цільова модель бізнес-процесу в даній задачі є апріорно заданою. Модель бізнес-процесу, що фактично виконується, може бути сформована методам process mining [11] на основі аналізу журналу подій, що містить упорядковані у часі записи про послідовність виконання дій бізнес-процесу.

Порівняння моделей дає можливість виявити неявні знання, що відображають рішення виконавців процесу, та не враховані в цільовій моделі БП. При удосконаленні моделі БП враховуються ці неявні знання, оскільки останні відображають практичні прийоми та навички, що відповідають корпоративній культурі підприємства і можуть спрощувати імплементацію бізнес-процесу для виконавців.

3. Метод побудови прототипу бізнес-процесу з використанням темпоральних знань при впровадженні процесного управління

Запропонований метод автоматизованої побудови процесної моделі орієнтований на вирішення задачі впровадження процесного управління на основі порівняння знань у апріорно заданій моделі бізнес-процесу та отриманій на основі аналізу журналу подій моделі бізнес-процесу.

Для обґрунтування методу розглянемо вербальну постановку комплексної задачі впровадження процесного управління з вирішенням задач побудови процесних моделей та вилучення знань для удосконалення бізнес-процесів.

Вхідними даними для побудови процесної моделі є інформація про фактично виконані бізнес-процеси, яка відображає явні та неявні знання про допустимі послідовності робіт для всіх реалізованих екземплярів бізнес-процесу. Під екземпляром бізнес-процесу будемо розуміти виконану на підприємстві послідовність робіт згідно цільової моделі БП. Тобто процес, описаний в одній моделі, може виконуватись багаторазово, в різні періоди часу. Кожна імплементація такого БП розглядається як екземпляр процесу.

Вхідні дані представлені у форматі журналу подій E який складається з множини записів про виконання кожного екземпляру бізнес-процесу E_i $E=\{E_i\}$. Кожна послідовність E_i кладається із подій з мітками часу $t_{i,j}$:

$$E_i = \langle e_{i,1},...e_{i,j},...,e_{i,J} : (\forall i \forall j)\exists t_{i,j} \in e_{i,j} \rangle. \quad (1)$$

Наявність міток часу для подій дає можливість визначити їх темпоральну упорядкованість.

Семантика такого подієвого опису полягає в тому, що кожна подія характеризує стан бізнес-процесу після виконання однієї з дій у послідовності робіт. Події мають властивості, які відображають властивості елементів бізнес-процесу, що були пов'язані із виконанням дій. Такі властивості, наприклад, включають: назву дії; рівень важливість дії; назву елементу організаційної структури, де було виконано дію; виконавця дії, а також роль виконавця, стан виконання дії тощо. Склад властивостей події визначається для кожного бізнес-процесу. Єдиною спільною властивістю, яка використовується для опису всіх бізнес-процесів в журналах подій, є мітка часу.

Слід зазначити, що на початку впровадження процесного управління на підприємстві бізнес-процеси зазвичай є неформалізованими. Вони базуються на усталених в рамках корпоративної культури послідовностях робіт з виробництва продукції та послуг, а також усталених правилах взаємодії між співробітниками. Такі процеси зазвичай виконуються лише на основі персональних знань виконавців процесу. Зазначені особливості процесного опису є характерними для найнижчого, першого рівня процесної зрілості. Для даного рівня зрілості, упорядкованість подій в рамках одного екземпляру бізнес-процесу зазвичай не задається і журнал подій містить події з різних екземплярів різних бізнес-процесів:

$$E = \{e_{i,j} : (\forall i \forall j)\exists t_{i,j} \in e_{i,j}\}. \quad (2)$$

Вирішення задача вилучення знань для удосконалення моделей бізнес-процесів автоматизованим способом полягає у встановленні залежності в часі між довільними подіями $e_{i,j}$ та $e_{i,m}$ для кожного екземпляру бізнес-процесу. Причини такого підходу полягають в тому, що темпоральна інформація є обов'язковою для опису подій в журналі, а інформація про причини виконання записаної послідовності подій зазвичай не може бути вилучена із журналу. Неможливість встановлення причин виконання дій із записів журналу є наслідком того, що кожна подія містить лише запис про поточні значення властивостей. Властивості, які були причинами для виконання відповідної дії, окремо в журналі не виділяються. Проте слід зазначити, що темпоральні залежності між подіями журналу є відображенням причинно-наслідкових зв'язків між відповідними діями процесу, оскільки зафіксована в журналі послідовність робіт бізнес-процесу визначається знанням про каузальні залежності між діями з отримання фінального продукту або послуги.

Темпоральна упорядкованість подій в журналі відображає як явні причинно-наслідкові залежності, які були враховані в апріорно заданій моделі БП, так і неявні залежності, які відображають персональні знання виконавців. Виявлення цих темпоральних знань дає можливість знайти неточності та невідповідності в існуючій моделі бізнес-процесу.

Таким чином, порівняння явних та неявних знань на основі виділених темпоральних залежностей створює умови для побудови удосконаленої моделі бізнес-процесу з урахуванням розбіжностей між явними та неявними знаннями.

Тобто необхідно знайти набір темпоральних знань $\{f_m^j\}$, що зв'язують $j-$ті та $m-$події в єдину послідовність, від початку і до завершення виконання бізнес-процесу. Вилучення вказаних залежностей дає можливість перейти від неупорядкованої множини подій, яка містить події з декількох екземплярів бізнес-процесу, до множини упорядкованих послідовностей подій для кожного екземпляру бізнес-процесу.

У випадку, якщо журнал подій містить упорядковані послідовності для кожного екземпляру БП, то задача вилучення знань полягає у виявленні таких залежностей $f_{i,m}^{i,j}$ між подіями $e_{i,j}$ та $e_{i,m}$ що були відсутні в апріорно заданій процесній моделі.

Кожна подія $e_{i,j}$ задається множиною значень змінних $v_{i,j}^l$. Ця множина значень може бути представлена вектором $\bar{V}_{i,j}$, причому початковим елементом $v_{i,j}^{(0)}$ даного вектору є мітка часу, тобто:

$$e_{i,j} \equiv \overline{V_{i,j}} \Big| v_{i,j}^{(0)} = t_{i,j}. \quad (3)$$

Можливість наведеного векторного опису подій журналу обґрунтовується тим фактом, що кожен запис в журналі подій містить інформацію про значення одних і тих же змінних, що відображають властивості події. Ці властивості записані у одному і тому ж порядку для кожної події. Тому індекс елемента у векторі $\bar{V}_{i,j}$ однозначно визначає конкретну властивість події.

Відповідно до (1) – (3), темпоральні залежності $f_{i,m}^{i,j}$ між подіями $e_{i,j}$ та $e_{i,m}$ між визначають послідовність переходів між роботами бізнес-процесу з урахуванням як дій, що безпосередньо слідують одна за одною, так і дій, між якими виконуються інші дії.

Для кожного екземпляру процесу ці події записуються послідовно. Однак порівняння цих залежностей для декількох екземплярів бізнес-процесу дає можливість встановити послідовне, паралельне або альтернативне виконання дій БП. Вказані відмінності виконання визначаються темпоральними правилами

f_m^j, які поєднують залежності $f_{i,m}^{i,j}$ із записів про реалізацію декількох екземплярів бізнес-процесу.

Таким чином, задача вилучення знань щодо БП полягає у знаходженні темпоральних правил f_m^j, що відображають знання про можливі послідовності виконання робіт бізнес-процесу.

Задача побудови моделей бізнес-процесів передбачає виконання таких кроків:

– побудова узагальненого процесного опису діяльності підприємства;

– побудова моделей окремих бізнес-процесів;

– встановлення зв'язків між бізнес-процесами.

На першому кроці вирішення даної задачі експертами в предметні області виділяються основні та допоміжні процеси, а також бізнес-процеси розвитку та управління. Потім ці бізнес-процеси визначаються на різних рівнях деталізації, зазвичай з урахуванням поточної організаційної структури підприємства.

На другому кроці визначаються ключові елементи процесної моделі:

– роботи (дії) бізнес-процесу;

– потік робіт, який фактично представляє собою алгоритм виконання БП;

– об'єкти, з якими оперує бізнес-процес; до таких об'єктів зазвичай відносять комплектуючі, вузли, матеріали, обладнання, а також елементи організаційної структури та виконавці.

– власник процесу, який керує розподілом ресурсів для БП, а також повноваження власника;

– входи та виходи БП, а також відповідні постачальники та клієнти;

– знання щодо умов та обмежень на виконання бізнес-процесу, які зазвичай представляються у формі бізнес-правил.

На третьому кроці вирішення даної задачі зв'язки між процесами зазвичай встановлюються через спільних постачальників/клієнтів. На даному етапі враховується, що бізнес-процеси можуть бути зовнішніми та внутрішніми (в межах організації). В другому випадку зв'язки між процесами встановлюються з урахуванням обмежень, що накладаються організаційною структурою підприємства.

З урахуванням наведених структурних характеристик бізнес-процесу, при першій побудові моделі БП спочатку виділяються бізнес-правила, що обумовлюють причинно-наслідкові зв'язки між діями процесу, роботи (дії) процесу та об'єкти, з якими ці дії оперують. Також задається розподіл відповідальності для власника та виконавців процесу. В подальшому визначається потік робіт. При визначенні послідовності робіт враховуються умови та обмеження, які задаються бізнес-правилами. Оскільки бізнес-правила не завжди враховують неявні знання виконавців, то проводиться уточнення моделі з виявленням темпоральних знань на основі аналізу логів формалізованих і неформалізованих бізнес-процесів.

При уточнені моделі бізнес-процесу з використання отриманих в результаті виконання задачі вилучення знань темпоральних правил ці темпоральні правила порівнюються з бізнес-правилами та відомими обмеженнями, закладеними у потоці робіт. За результатами порівняння виконується уточнення моделі бізнес-процесу або ж дій виконавців. Дії виконавців регламентуються в рамках послідовності робіт або окремих робіт бізнес-процесу.

Розглянуті особливості задач впровадження процесного управління обумовлюють наступні етапи методу автоматизованої побудови прототипу бізнес-процесу з використанням темпоральних знань.

Фаза 1. Вилучення темпоральних знань.

Етап 1. Формування темпоральних залежностей $f_{i,m}^{i,j}$ щодо послідовностей виконання окремих екземплярів бізнес-процесу.

На даному етапі встановлюються темпоральні залежності між парами подій журналу. Для побудови залежностей використовуються мітки, що містять абсолютне значення часу. Залежності використовують відносне значення часу (тобто темпоральну упорядкованість).

Для журналу на першому рівні процесної зрілості належність подій до одного і того ж бізнес-процесу встановлюється за спільним переліком атрибутів подій.

Етап 2. Формування темпоральних правил f_m^j на основі об'єднання залежностей $f_{i,m}^{i,j}$.

На даному етапі однакові залежності з різних екземплярів бізнес-процесу об'єднуються у темпоральне правило. Залежності вважаються еквівалентними, якщо вони характеризуються однаковими векторами атрибутів $\bar{V}_{i,j}$.

Етап 3. Формування підмножин $F^{(\rightarrow)}$ $F^{(\|)}$ правил f_m^j, що обумовлюють відповідно послідовне та паралельне (або альтернативне) виконання робіт бізнес-процесу.

На даному етапі використовуються мітки абсолютного часу $t_{i,j}$, що дає можливість порівняти у часі реалізацію правил f_m^j у різних екземплярах бізнес-процесу. У випадку, якщо пара правил f_m^j та f_g^k виконувались у різному порядку в різних екземплярах БП, то вони відносяться до підмножини $F^{(\|)}$ паралельних правил. В іншому випадку, якщо для всіх відомих екземплярів БП правила виконувались в одній послідовності, то ці правила належать до $F^{(\rightarrow)}$.

Фаза 2. Побудова моделі бізнес-процесу.

Етап 4. Уточнення журналу подій бізнес-процесу з урахуванням темпоральних правил.

На даному етапі виконується формування або уточнення послідовності подій в журналі.

На першому рівні процесної зрілості виконується формування множини послідовностей подій для відомих екземплярів бізнес-процесу. В результаті

із журналу виду (2) формується журнал виду (1), що містить упорядковані послідовності подій E_i.

На вищих рівнях процесної зрілості в уточненому журналі може бути виділено підмножину послідовностей подій, яка містить нетипові (відсутні в цільовій моделі) темпоральні залежності.

Етап 5. Формування моделі бізнес-процесу, що виконується, методами process mining.

На даному етапі на основі аналізу журналу подій створюється граф послідовностей робіт, що містить імплементовані у виконаних екземплярах бізнес-процесу темпоральні правила.

Результатом методу є прототип бізнес-процесу, що складається із графу потоків робіт та темпоральних правил, які відображають умови та обмеження на виконання дій БП, що фактично були реалізовані при багаторазовому виконанні бізнес-процесу.

Отримана модель розглядається як прототип з тієї причини, що вона не містить формалізованих умов вибору тієї чи іншої послідовності робіт. Для визначення таких каузальних правил потрібно доповнити темпоральні правила інформацією про ті значення змінних $v_{i,j}^{(k)}$, які були умовами вибору відповідних дій.

Розглянемо приклад імплементації запропонованого методу.

Вхідні дані методу містять журнал, кожна подія якого визначається такими ключовими атрибутами: країна; організація; виконавець; роль виконавця; дія процесу; ступінь важливості дії для БП; продукт, що випускає бізнес-процес; поточний стан дії; мітка часу.

Журнал містить інформацію про виконання процесів сервісного обслуговування міжнародної компанії.

Приклад опису події в журналі наведено на рис. 2.

```
<event>
    <string key="org:group" value="V30"/>
    <string key="resource country" value="France"/>
    <string key="organization country" value="fr"/>
    <string key="org:resource" value="Anne Claire"/>
    <string key="organization involved" value="Org line A2"/>
    <string key="org:role" value="A2_4"/>
    <string key="concept:name" value="Queued"/>
    <string key="impact" value="Medium"/>
    <string key="product" value="PROD582"/>
    <string key="lifecycle:transition" value="Awaiting
                Assignment"/>
    <date key="time:timestamp" value="2010-04-
                06T16:45:07+02:00"/>
</event>
```

Рис. 2. Приклад опису події у журналі

Вхідні дані є процесно-орієнтованими і відповідають рівню 2 зрілості процесного управління.

На фазі 1, при виконані етапу 1 методу формуються темпоральні залежності виду (4) на одній послідовності подій з урахуванням абсолютного значення часу:

$$\langle event"2010-04-06T16:45:07+02:00"\rangle \rightarrow \quad (4)$$
$$\langle event"2010-04-08T12:52:23+02:00"\rangle$$

При реалізації етапу 2 формуються знання у формі темпоральних правил виду (5).

$$\langle event\,2741\rangle \rightarrow \langle event\,2749\rangle. \quad (5)$$

Ці правила задають порядок у часі для подій з унікальними $id = 2741$ та $id = 2749$. На відміну від (4), використовується відносний порядок подій. Вказані ідентифікатори визначають подію незалежно від конкретної послідовності E_i. Всі варіанти такої події на різних послідовностях E_i мають однаковий набір значень атрибутів (рис. 2).

Результатом етапу 3 є:

– множина з правил виду (5), які виконуються лише послідовно;

– множина з підмножин правил (5), які виконуються лише паралельно.

Приклад елементу множини паралельних (або альтернативних) правил має вигляд (6).

$$\left\{\begin{array}{l}\langle event\,2741\rangle \rightarrow \langle event\,2749\rangle,\\ \langle event\,2741\rangle \rightarrow \langle event\,2751\rangle\end{array}\right\}. \quad (6)$$

Згідно (6), після події $\langle event\,2741\rangle$ля частини екземплярів БП відбувалась подія $\langle event\,2749\rangle$, а для іншої частини – подія $\langle event\,2751\rangle$. Така пара правил свідчить про можливість розпаралелювання робіт, або ж про ситуацію вибору однієї із можливих подальших послідовностей робіт.

В подальшому правило (6) може бути використано аналітиком при порівнянні фактичної і апріорної моделі БП. Тобто якщо у апріорно заданій моделі бізнес-процесу одна із альтернатив відсутня, то це свідчить про застосування персональних знань виконавців при реалізації бізнес-процесу.

На фазі 2, у відповідності до задач удосконалення БП, в складі логу залишаються, наприклад, послідовності робіт, які містять набори правил (6).

На етапі 5 формується граф моделі процесу з використанням відомих методів process mining [12].

Порівняння ваг дуг графу із правилами виду (6) дає можливість встановити тип відношень для пари правил: альтернатива або розпаралелювання.

Таким чином, отримані в результаті використання даного методу темпоральні правила дають можливість виявити послідовності дій, які виникли внаслідок застосування персональних знань виконавців та в подальшому врахувати ці звання при удосконаленні процесної моделі.

Висновки

Виконано структуризацію задач процесного управління з урахуванням необхідності постійного удосконалення бізнес-процесів при підвищенні рівня процесної зрілості. Процесне управління передбачає вирішення задач управління бізнес-процесами та впровадження процесного підходу на підприємстві. Автоматизація задачі впровадження процесного підходу передбачає аналіз журналів подій бізнес-процесу та виявлення знань про темпоральну упорядкованість дій процесу у часі, що потребує формування темпоральних знань.

Запропоновано метод побудови моделі бізнес-процесу, що фактично виконується на підприємстві, з використанням темпоральних знань. Метод містить етапи формування темпоральних правил ,що визначають послідовність дій процесу, виділення підмножин правил послідовного та паралельного виконання дій, а також формування графу потоків робіт бізнес-процесу. В практичному плані метод створює умови для постійного удосконалення процесних моделей шляхом виявлення та використання неявних знань виконавців, які відображені у темпоральних правилах.

Список літератури:

[1] M. Dumas, M. La Rosa, J. Mendling, and H. A. Reijers, *Fundamentals of Business Process Management*, 1st ed. Berlin, Heidelberg: Springer-Verlag, 2018. doi: 10.1007/978-3-662-56509-4.

[2] A. Gadatsch, "Introduction to Business Process Management," in *Business Process Management*, Springer, Wiesbaden, 2023, ch.1. doi: 10.1007/978-3-658-41584-6_1.

[3] R. Dijkman, S.V. Lammers, and A. de Jong, "Properties that influence business process management maturity and its effect on organizational performance," Inf Syst Front, vol. 18, pp. 717-734, 2016. doi: 10.1007/s10796-015-9554-5.

[4] S. Chalyi and I. Bogatov, "Technology of automated construction of a business process prototype model based on pre-processing of event logs," Bulletin of National Technical University KhPI Series System Analysis Control and Information Technologies, pp. 57-63, 2020. doi: 10.20998/2079-0023.2020.02.10.

[5] Brocke J., Rosemann M. Handbook on business process management. Berlin, Springer-Verlag Publ., 2015. 709 p.

[6] O. Kalynychenko, S. Chalyi, Y. Bodyanskiy, V. Golian, and N. Golian, "Implementation of search mechanism for implicit dependences in process mining," in Proc. 2013 IEEE 7th Int. Conf. on Intelligent Data Acquisition and Advanced Computing Systems (IDAACS), 2013, doi: 10.1109/IDAACS.2013.6662657.

[7] Sergii, C., Ihor, L., Aleksandr, P., Ievgen, B. (2018). Causality-based model checking in business process management tasks. 2018 IEEE 9th *International Conference on Dependable Systems, Services and Technologies* (DESSERT). doi: https://doi.org/10.1109/dessert.2018.8409176

[8] Левикін В. М., Чала О.В. Метод підтримки управлінських рішень в умовах невизначеності на основі темпоральних знань. *Біоніка інтелекту*. 2018. № 2(91). С. 54-59.

[9] Чала О. В. Побудова темпоральних правил для представлення знань в інформаційно-управляючих системах. *Сучасні інформаційні системи*. 2018. Том 2, № 3. С. 54-59. DOI: 10.20998/2522-9052.2018.3.09.

[10] Chala O. Models of temporal dependencies for a probabilistic knowledge base. *Econtechmod. An International Quarterly Journal*. 2018. Vol. 7, No. 3. P. 53 - 58.

[11] W. M. P. van der Aalst, Process Mining: Discovery, Conformance and Enhancement of Business Processes, 1st ed. Berlin, Heidelberg: Springer-Verlag, 2011. doi: 10.1007/978-3-642-19345-3.

[12] C. dos Santos Garcia, A. Meincheim, E. R. Faria Junior, M. R. Dallagassa, D. M. V. Sato, D. R. Carvalho, E. A. P. Santos, and E. E. Scalabrin, "Process mining techniques and applications – A systematic mapping study," *Expert Systems with Applications*, vol. 133, pp. 260-295, 2019. doi: 10.1016/j.eswa.2019.05.003.

Надійшла до редколегії 12.09.2023

УДК 378.147:004 DOi 10.30837/ bi.2023.1(99).04

В.І. Жеребкін[1], С.Г. Удовенко[2], Л.Е. Чала [1], О.Є. Гриньова [1]
[1]ХНУРЕ, м. Харків, Україна, v.zherebkin@gmail.com, ORCID iD: 0000-0002-6616-293X
[2]ХНЕУ ім. С. Кузнеця, м. Харків, Україна, serhiy.udovenko@hneu.net,
ORCID iD: 0000-0001-5945-8647
[1]ХНУРЕ, м. Харків, Україна, larysa.chala@nure.ua, ORCID iD: 0000-0002-9890-4790
[1]ХНУРЕ, м. Харків, Україна, olena.hrynova@nure.ua, ORCID iD: 0000-0001-5945-8647

АНАЛІЗ СКЛАДНОСТІ ТА ПОБУДОВА КОНЦЕПТУАЛЬНИХ ГРАФІВ МАСОВИХ ВІДКРИТИХ ОНЛАЙН-КУРСІВ

Досліджено проблему персоналізованого навчання, що здійснюється в масових відкритих онлайн-курсах (MBOK). Запропоновано підхід до планування сценарію онлайн-навчання та побудови технології автоматичної генерації персоналізованих шляхів навчання. На відміну від стратегії ручного оцінювання складності навчання, цей підхід дозволяє автоматично обчислювати загальну суму балів знань на основі моделі складності та концептуального графу онлайн-навчання. Розглянуто особливості реалізації запропонованого підходу для MBOK з використанням навчальної платформи Open Edx за результатами оцінювання стану навчання студентів у режимі реального часу. Наведено приклади побудови моделі складності та концептуального графу онлайн-навчання для ресурсів курсу «Штучний інтелект». Отримані результати сприятимуть реалізації персоналізованого планування сценаріїв навчання студентів для MBOK на основі своєчасної діагностики стану успішності навчання.

МОДЕЛЬ СКЛАДНОСТІ НАВЧАННЯ, КОНЦЕПТУАЛЬНИЙ ГРАФ, НАВЧАЛЬНА ПЛАТФОРМА OPEN EDX, ПЛАНУВАННЯ СЦЕНАРІЇВ НАВЧАННЯ, ДІАГНОСТИКА СТАНУ УСПІШНОСТІ

V.I Zherebkin, S.G. Udovenko, L.E. Chala, O.E. Hrynova. Complexity analysis and construction of conceptual graphs of mass open online courses. The problem of personalized training carried out in mass open online courses (MOOC) was studied. An approach to online learning scenario planning and construction of technology for automatic generation of personalized learning paths is proposed. In contrast to the manual learning difficulty assessment strategy, this approach automatically calculates a total knowledge score based on the difficulty model and online learning conceptual graph. The specifics of the implementation of the proposed approach for vocational education and training using the Open Edx educational platform based on the results of real-time assessment of students' learning status were considered. Examples of building a complexity model and a conceptual graph of online learning for the resources of the course "Artificial Intelligence" are given. The obtained results will contribute to the implementation of personalized planning of students' learning scenarios for MVOK based on timely diagnosis of the state of educational success.

LEARNING COMPLEXITY MODEL, CONCEPTUAL GRAPH, OPEN EDX LEARNING PLATFORM, LEARNING SCENARIO PLANNING, SUCCESS STATUS DIAGNOSTICS

Вступ

В національних пріоритетах суспільного розвитку України відбуваються зміни, що обумовлені її переходом до європейського освітнього простору. Це призводить до зростання ускладнення умов функціонування навчальних закладів та розвитку нових форм навчання [1].

Персоналізоване навчання сприяє успішності оволодіння знаннями студентами внаслідок використання спеціально підібраних навчальних каркасів, що пропонуються в масових відкритих онлайн-курсах (MBOK). MBOK, які набувають популярність серед студентів через низький поріг реєстрації та вільний час навчання, певним чином інтегрують можливості соцмереж, колекцій відкритих освітніх ресурсів та експертної підтримки у відповідних галузях [2]. Студенти мають свободу та можливість вибирати актуальні та якісні курси. Однак деякі перешкоди, такі як низькі показники завершення курсу, низькі показники успішності та низька ефективність навчання, перешкоджають розвитку MBOK.

Сучасні платформи дистанційного навчання дають можливість відстежувати та записувати усі події онлайн-користувачів під час перегляду відеофрагментів, включаючи повторення, швидке перемотування вперед і пропуск. Після перегляду відео студенти мають виконати тестову вправу з розділу.

Більшість сучасних систем рекомендацій персоналізованих шляхів навчання застосовуються до конкретного курсу. Більше того, існуючі платформи MBOK не можуть забезпечити персоналізовані послідовності навчання студентів навіть на певному курсі [3]. Таким чином, доцільним є розробка підходу, який сприятиме реалізації персоналізованого планування навчального шляху на основі своєчасної діагностики стану навчання студентів у сценаріях навчання MBOK. З огляду на це тема роботи є актуальною і має практичну значущість.

Метою роботи є побудова моделі складності та персоналізації сценаріїв онлайн-навчання для MBOK на основі платформи Open Edx.

Відповідно до поставленої мети, необхідно вирішити наступні завдання:

– порівняльний аналіз існуючих MBOK та відповідних платформ навчання;

– дослідження особливостей планування персоналізованих сценаріїв навчання студентів в MBOK;

– побудова моделі складності та концептуального графу навчання для MBOK;

– побудова алгоритму персоналізації сценаріїв навчання студентів для MBOK.

1. Сучасні технології та платформи побудови MBOK

У 2001 році Масачусетський технологічний інститут (MIT) приступив до створення першого великого репозиторія відкритих освітніх ресурсів (ВОР) в рамках проекту OpenCourse Ware. ВОР це розміщені у відкритому доступі матеріали, призначені для безкоштовного використання в процесі навчання, автори яких дали згоду на їх вільне використання і модифікацію. Модель відкритої освіти полягає в тому, щоб відкрити перед студентами максимальні можливості набуття знань і навичок незалежно від географічних, соціально-економічних та інших факторів. За минулі роки у світі з'явилася велика кількість дистанційних курсів, заснованих на відкритому змісті, та навіть виникли університети електронного навчання. У 2005 році з'явилася ідея швидкого створення електронних навчальних курсів за рахунок використання слайдів, мультимедійних презентацій, флеш-технологій – сформувався термін «Швидкі технології дистанційного навчання».

У 2008 році Джордж Сіменс і Стівен Даунс провели відкритий дистанційний курс «Коннективізм та сполучні знання», присвячений проблемам нової теорії навчання – коннективізму. Для того, щоб охарактеризувати число тих, хто навчається на курсі (більше 2000 слухачів), Дейв Корм'ї і Брайан Олександр запропонували термін MBOK – «Масовий відкритий онлайн-курс» (або MOOC – від англ. Massive open online course). Концепція створення MBOK передбачає інтеграцію переваг соцмереж, колекцій відкритих освітніх ресурсів та підтримки експертів у відповідній галузі [4]. В абревіатурі MBOK «масовий» означає кількість учасників курсів, а також можливості курсу з точки зору дозволу доступу до великої кількості заходів. Джордж Сіменс визначив «масовий» як: «Все, що є достатньо велике, щоб можна було отримати субкластери самоорганізованих інтересів. Відповідним еталоном може бути, наприклад, число Робіна Данбара (150 осіб), тобто максимум, після якого група починає створювати менші дроби» [5]. «Відкритий» зазвичай означає безкоштовний доступ до окремих курсів, і іноді це також стосується відкритої платформи вмісту. «Онлайн» стосується можливості доступу до MBOK через Інтернет. «Курс» означає організацію викладення у заданий інтервал часу конкретної дисципліни з предметної області, яка містить набір ресурсів з чітко визначеними цілями та результатами. До кінця 2013 року більшість провідних університетів почали пропонувати різновиди MBOK на базі різних навчальних платформ.

Платформа edX. Останнім часом поширення набуває використання платформи edX для створення онлайн-курсів. Ця платформа вже охоплює чималу кількість курсів з різних галузей, що організовуються престижними університетами з можливістю отримання відповідного сертифіката. Вона передбачає необхідність підключення до учасників до Інтернету, отримання учасниками навчальних матеріалів, консультації з викладачами, та можливість оцінювання набутих знань. Існує чимало досліджень, що стосуються використання MBOK на основі платформи edX. Наприклад, в роботі [5] пропонується в рамках MBOK реалізувати творчі заходи з оцінкою індивідуальних особливостей курсів та комунікативною діяльністю учасників курсу, що передбачає спілкування між викладачами та студентами курсу, а також серед самих учасників. Основне ядро платформи edX містить такі функції:

– інтерактивні відеолекції з субтитрами та індексацією на субтитрах;

– онлайн-тести різних типів (вікторини з вбудованим відео, практичні сесії, проміжний та випускний іспити тощо);

– віртуальні лабораторії з інтерактивним інтерфейсом для перегляду користувачем результатів моделювання;

– створення та контроль календарного графіку курсів;

– багатомовна підтримка;

– дискусійні форуми;

– поточні звіти про хід роботи та інші види вбудованої аналітики;

– різні види систем оцінювання поданих завдань (задачі з відкритими відповідями; самооцінювання);

– електронні листи та засоби повідомлень для зареєстрованих студентів з наданням результатів атестації;

– реєстрація та зняття з курсу;

– підтримка трафіку спілкування з користувачами у певний час.

Слід зазначити, що існують також деякі інші платформи, що використовуються для створення MBOK (зокрема, Moodle, CourseSites by Blackboard, Udemy, Versal тощо).

Платформа Moodle. Moodle – це система управління навчанням з відкритим кодом (LMS – Learning Management System), яка дозволяє користувачам створювати та пропонувати онлайн-курси [6, 7]. Цю платформу, зазвичай, простіше встановити, ніж EdX. Moodle підходить для організацій, яким потрібна повнофункціональна настроювана LMS. Втім продуктивність системи Moodle погіршується при наявності великої кількості студентів. Moodle дозволяє: стимулювати, розширювати та записувати дискусії поза класом; тестувати та звітувати про навчання

студентів, заохочувати студентів до навчання і самоперевірки; здійснювати проведення іспитів; швидко збирати відгуки студентів; сприяти обміну, зберіганню та поширенню знань; організовувати дискусійні форуми; завантажувати необхідні файли; створювати миттєві повідомлення тощо.

В табл. 1 наведено характеристики деяких популярних безкоштовних платформ, що використовуються для створення MBOK.

Таблиця 1

Характеристики деяких популярних безкоштовних платформ створення MBOK

Платформи	Максимальний розмір групи	Аналітика та можливість бренду
Edx	300000	+
Moodle	10000	+
CourseSites by Blackboard	Безлімітний	-
Udemy	Безлімітний	-
Versal	Безлімітний	-

Таблиця свідчить про певні переваги платформ edX та Moodle.

Однак існують додаткові причини вибору платформи edX для створення онлайн-курсів.

Наприклад, edX дозволяє користуватися перевагами масштабу інфраструктури хмарних обчислень Google (на сайті http://mooc.org/ EdX позиціонується як безкоштовна хмарна пропозиція для будь-якої освітньої організації, яка хоче пропонувати відкриті курси).

Крім того, університет (або інший навчальний заклад, що пропонує курс) може вибрати варіант локально розміщення екземпляра Edx і не нести витрати на хостинг та забезпечення потрібної пропускної здатності. Це має знизити бар'єри для створення (та пошуку потенційними учасниками) відкритих курсів.

Найбільш перспективним є використання варіанта платформи edX з відкритим кодом, відомого під назвою Open edX. Платформа Open edX надає технологію програмного забезпечення для онлайн системи управління онлайн навчанням, що дозволяє викладачам створювати динамічні курси. Виділимо загальні принципи навчання за допомогою Open edX: вільний графік навчання, змішана модель навчання та нетворкінг. Ця навчальна платформа використовується також в деяких проектах глобальних організацій, таких як Microsoft і IBM.

Проектом Open edX керує некомерційна організація NP, що очолюється Гарвардським університетом та Масачусетським технологічним інститутом. зосереджуватиметься на інклюзивному навчанні та освіті. NP співпрацює з навчальними закладами, урядами та іншими організаціями для розробки та оцінки нових підходів до навчання та педагогіки; інвестує в нові моделі навчання, а також сприяє впровадженню передового досвіду в освітньому континуумі. NP планує підтримувати інновації в існуючих структурах навчанні та розвивати платформи для навчання нового покоління.

2. Структура платформи Open edX

Розглянемо докладніше архітектуру та компоненти платформи Open Edx, що передбачається використовувати для реалізації персоналізованого варіанту навчання студентів в рамках MBOK.

Платформа Open Edx складається з системи керування вмістом (CMS) для розробників курсів і системи керування навчанням (LMS) для студентів. Основні функції, доступні для викладачів і студентів, визначені в кодовій базі платформи Edx. Крім того, існують незалежні додатки для розгортання (IDA), які покращують функціональні можливості платформи.

Серверний код у проекті Open edX написаний на мові Python з фреймворком веб-додатків Django.

Система керування навчанням LMS є ключовою частиною проекту Open edX. LMS також надає інформаційну панель інструктора, до якої користувачі, які мають роль адміністратора або персоналу, можуть отримати доступ, вибравши Інструктор. LMS використовує ряд сховищ даних. Курси зберігаються в MongoDB, а відео подаються з YouTube або Amazon S3. Дані для кожного студента зберігаються в базі MySQL. Коли студенти проходять курси та взаємодіють з ними, події публікуються в конвеєрі аналітики для збору, аналізу та звітності.

Код Django на стороні сервера в LMS використовує програмний додаток Mako для генерації шаблонів інтерфейсу. Код на стороні браузера написаний переважно на JavaScript. Частини коду на стороні клієнта використовують фреймворк Backbone.js, що переміщує більшу частину бази коду для використання користувачем цієї платформи.

Проект Open edX має просту першу сторінку для перегляду курсів, а також окрему домашню сторінку та сайт пошуку курсів, який не є відкритим вихідним кодом.

Курси Open edX складаються з блоків XBlocks. Відзначимо, що існує можливість написання користувачами нових XBlocks, які дозволять викладачам розширити набір компонентів для своїх курсів. Платформа Open edX також містить кілька модулів XModules (попередників XBlocks), що генерують певні типи вмісту курсу для користувачів як для створення, так і для навчання. Наприклад, існують XModules для відео та для фрагментів HTML

XModules створюють складний набір проміжних сутностей поверх XBlocks (усі XModules зараз

успадковуються від XBlock). В перспекиві планується їх вилучення в окремі репозиторії або перетворення в XBlocks, що має спростити час виконання завдань Open edX.

На додаток до XBlocks є кілька способів розширити поведінку курсу, зокрема, розробники навчального курсу з використанням LMS можуть вбудовувати інструменти LTI, щоб інтегрувати різні засоби навчання в курс Open edX.

В цілому процес створення онлайн-курсу з використанням засобів Open edX можна поділити на такі етапи, як: розробка робочої програми (тематичний план), проекту сценарію онлайн-курсу (розширений план курсу), сценарію онлайн-курсу (наповнення матеріалом проекту сценарію) і технічний перенос сценарію в Open edX Studio.

Studio – це інструмент платформи Open edX, який використовуєтться для технічної реалізації запланованої структури курсу (з використанням інтерактиної взаємодії викладачів зі студентами, відеоресурсів ресурси, поточного аналізу складності завдань тощо). За його допомогою можна керувати розкладом курсів, визначати членів команди курсу, встановлювати політику оцінювання та оприлюднювати конкретний курс.

Кожен курс проект Open edX дозволяє представити у вигляді деревовидної структури блоків, яку можна змінювати під час створення або модифікації курсу. Мінімальний набір блоків складає: одну главу; одну секцію; один урок; одне завдання.

На рис. 1. представлений приклад типової структури курсу.

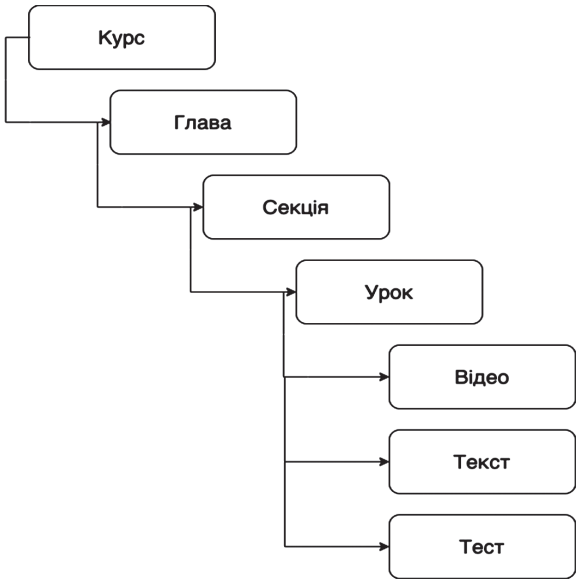

Рис. 1. Типова структура курсу на платформі Open edx

Для зберігання даних курсу в Open edX використовується нереляційна база даних MongoDB, а всі операції взаємодії виконуються через додатковий модуль Modulestore. Головні функції, які виконує цей модуль: зберігає пов'язані з курсом дані Xmodule і XBlock в різних серверних програмах; кешує дані; надає інтерфейс сховища ключів/значення для цих даних; впроваджує систему версій для курсів; забезпечує засоби для перенесення даних курсу з одного сервера modulestore до іншого.

Після запуску сервера LMS ModuleStore завантажує в пам'ять кожен блок кожного курсу. Він доступний тільки в режимі для перегляду і не дозволяє користувачам змінювати курс без перезапуску сервера.

БД Mongo поділяє курс на індекс курсу, структуру курсу та визначення Xblock або наповнення курсу. Індекс курсу – це словник, в якому зберігаються ідентифікатори курсу. Кожен ідентифікатор курсу свідчить про структуру курсу. Вказівник курсу підтримує декілька гілок курсу, причому для кожного курсу є гілка як для опублікованої, так і для запланованої версії курсу. Таким чином, курс має деревовидну структуру даних, яка складається з різних блоків.

Важливою частиною платформи Open edX є модуль аналізу навчання OXALIC, призначений для представлення огляду діяльності студента з використанням кількох різних методів обробки даних та візуалізації. Основна мета OXALIC – надати різним групам зацікавлених сторін (головним чином викладачам та дослідникам), корисні оцінки даних, які збираються під час онлайн навчання. Це дослідження зосереджується на одному з можливих застосувань цієї згенерованої інформації, а саме на групуванні студентів за їхньою активністю та залученням до MBOK.

Огляд архітектури інструмента представляє потік даних відстеження в системі. Спочатку дані відстеження збираються і передаються в систему. Потім ці дані зберігаються в базах даних, одна для генерування графіків, а інша для загальної обробки даних, що призводить до отримання кількох частин інформації про курс. Нарешті, ця інформація форматується і представляється користувачам у вигляді веб-сторінок з різною статистикою та графіками.

Щоб зрозуміти корисність існуючої системи, можна виділити кілька груп зацікавлених сторін:

– інструктори;
– студенти;
– конструктори курсів;
– власники платформи;
– дослідники.

Основна група, яка отримує найбільшу користь – це інструктори. Використовуючи цей інструмент, вони можуть спостерігати прогрес студентів та їхню взаємодію з різними частинами курсу. На основі цієї інформації інструктори можуть керувати прогресом студентів і вживати заходи, щоб допомогти студентам покращити їхні результати. Це досягається завдяки візуалізаціям, які генеруються шляхом обробки

даних, зібраних під час поточного та попереднього проходження курсу.

Друга група – це самі студенти. Для них формується інформаційна панель зі зведеною персоналізованою інформацією про поточні результати навчання на курсі. Також може бути представлена додаткова інформація, де надаються рекомендації, прогнози та різні види аналізу, які допоможуть студентам краще планувати хід навчання та виправляти потенційні проблеми.

Третя група – дизайнери курсів. Спостерігаючи за інформацією, яку генерує інструмент, розробники курсів можуть оцінити ефективність створеного ними курсу з метою його можливого покращення протягом наступного періоду навчання. Цю інформацію також можна використовувати для планування та створення нових курсів.

Власників платформ також можуть аналізувати дані, згенеровані модулем OXALIC для концептуального доналаштування структури всієї платформи, а також обсягу та типів даних, які генеруються та зберігаються. Таким чином можна потенційно підвищити як ефективність аналітики, так і інформативність потоків даних.

Нарешті, дослідники також можуть використовувати дані, згенеровані цим інструментом (зокрема, агреговані та відфільтровані дані). Це може заощадити час для дослідників на отримання додаткової інформації, яка вже була автоматично відфільтрована та уточнена на основі цілей дослідження

3. Попередня обробка даних та побудова моделі складності онлайн навчання

Дані курсу та навчання, що зберігаються на платформах MBOK, в основному складаються з двох частин: даних про ресурси курсу та даних про навчальну поведінку студентів. Кожен курс містить, зазвичай, чимало розділів з відео та вправами, що обумовлює наявність ієрархічної схеми ресурсів курсу.

Дані в кожному з розділів (глав) курсу можуть бути визначені таким чином: заголовки та субтитри відео з курсу; тестові завдання; обов'язкові зв'язки між точками знань.

Платформи MBOK відстежують та записують усі події онлайн-користувачів під час перегляду відео курсів, включаючи повторення, швидке перемотування вперед і пропуск. Після перегляду відео студенти мають виконати вправу з розділу. Навчальні дані кожного студента зображені таким чином: поведінка при перегляді відео; продуктивність виконання вправи; коментарі та відповіді в секції коментарів.

Схему ієрархії змісту ресурсів курсу MBOK, що, зазвичай використовується в процесі масового відкритого онлайн-навчання, наведено на рис. 2.

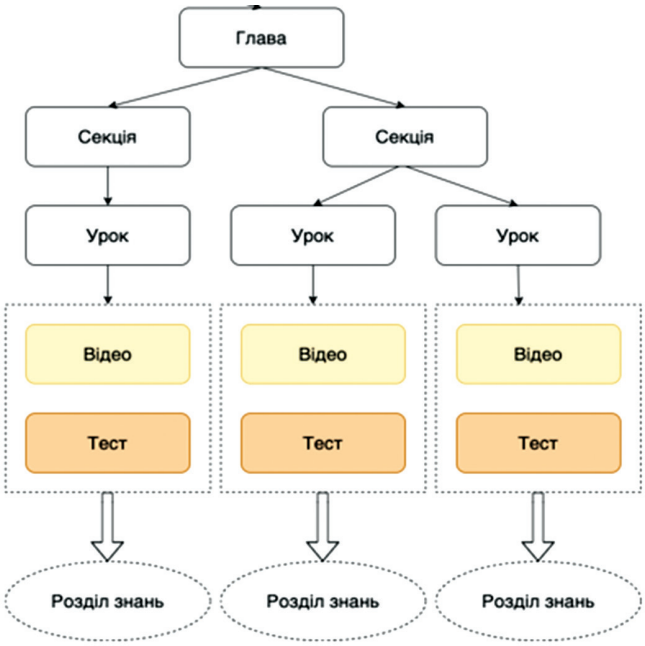

Рис. 2. Типова структура курсу на платформі Open edX

Попередня обробка даних. Для розробки алгоритму персоналізації курсів навчальні дані пропонується попередньо обробляти в наступній послідовності:

– виділення ключових слів з заголовків відео, субтитрів відео та вправ до розділів;

– класифікація вправ (порівняння ключових слів для вправ з розділу з ключовими словами заголовків і субтитрів відео). Кожна вправа класифікується на точки знань із найбільшою кількістю зустрічей ключових слів;

– нормалізація результатів тестових завдань.

Курси MBOK незалежні один від одного, але це не сприяє запиту та зберіганню даних і може значно збільшити кількість вимог до анотацій при подальшій обробці. Тому, задля збагачення асоціації курсів, їх групували в системи класифікації предметів. Відповідно до галузей класифікації та кодексу дисциплін три експерти позначають галузь, до якої належить курс. Щоб контролювати детальність класифікації, всі курси позначено до другого рівня, утворивши таким чином дерево дисциплін курсу.

Багато популярних курсів пропонуються неодноразово. Хоча вони можуть бути трохи змінені, велика кількість ресурсів курсу дублюється, що, очевидно, призводить до таких проблем, як розрідженість даних і зайве маркування. Тому було проведено початкову фільтрацію на основі такої інформації, як назва курсу, викладач і школа, щоб знайти повторні курси. Потім точно поєднано відео та вправи, використовуючи їх тексти змісту. Для кожного ресурсу (відео та вправи) його позначено ccid. Ресурси з тим самим ccid означають ті самі ресурси, які повторюються в повторних курсах. Обсяг дедуплікованих ресурсів уточнено до 10,3% від початкового розміру.

Інтеграція поведінки студентів. Платформа Open edX дає можливість відстежувати та записувати усі події онлайн-користувачів під час перегляду відеофрагментів, включаючи повторення, швидке перемотування вперед і пропуск. Після перегляду відео студенти мають виконати вправу з розділу. Навчальні дані щодо кожного студента структуруються таким чином:

– поведінка при перегляді відео;

– продуктивність виконання вправи;

– коментарі та відповіді в секції коментарів.

Дані про перегляд відео студентами – це логи «HeartBeat» з інтервалом у п'ять секунд, тобто кожні п'ять секунд система записує, яке відео цей студент зараз переглядає та яку позицію він/вона вивчає. Такі дані не зручні у використанні, тому далі проводиться обробка даних поведінки при навчанні в сегменти перегляду, тобто кожен сегмент вказує, який сегмент відео учень дивиться (час початку та закінчення) і швидкість відтворення в цей час. Якщо у відео є стрибок або пауза, воно вважається новим сегментом.

Оцінка складності онлайн навчання. На основі аналізу загальної ефективності студентів у сценаріях навчання MBOK цей розділ створює модель оцінки для вимірювання складності конкретних розділів знань. Вхідними параметрами моделі складності на основі розділів знань є середні бали тесту за вправи всіх учнів, які вивчали розділ знань. Результатом моделі складності знань є рівень складності знань, для оцінки якого пропонується використовувати такий параметр:

$$diff(j) = w_1\left[1-\bar{s}(j)\right] + w_2\bar{r}(j) + w_3\bar{c}(j), \qquad (1)$$

де $\bar{s}(j)$, $\bar{r}(j)$ і $\bar{c}(j)$ – середня оцінка тесту, середня кількість повторних переглядів відео та середня кількість коментарів j-го розділу знань відповідно; w_1, w_2 і w_3 – ваги вхідних параметрів $\bar{s}(j)$, $\bar{r}(j)$ і $\bar{c}(j)$ відповідно.

Очевидно, що значення $diff(j)$ прямо пропорційне складності засвоєння j-ї компоненти знання.

Значення параметрів $\bar{s}(j)$, $\bar{r}(j)$ і $\bar{c}(j)$ пропонується розраховувати таким чином (згідно з методом аналітичної ієрархії (AHP) визначення ваг в моделі оцінок [6]):

$$\bar{s}(j) = \frac{\sum_{i=1}^{N_j^{history}} s_{ij}}{N_j^{history}}; \qquad (2)$$

$$\bar{r}(j) = \frac{\sum_{i=1}^{N_j^{history}} r_{ij}}{N_j^{history}}; \qquad (3)$$

$$\bar{c}(j) = \frac{\sum_{i=1}^{N_j^{history}} c_{ij}}{N_j^{history}}, \qquad (4)$$

де $N_j^{history}$ – коефіцієнт, що відповідає: а) загальній кількості студентів, які засвоїли j-й розділ знань (для рівняння (2)); б) загальній кількості переглядів відео j-го розділу знань на всьому курсі (для рівняння (3)); в) загальній кількості коментарів щодо j-го розділу знань на всьому курсі (для рівняння (4)); s_{ij}, r_{ij} та c_{ij} – оцінка тесту, кількість повторних переглядів відео та кількість коментарів i-го учня для j-го розділу знань відповідно.

Як приклад, для вхідних параметрів $\bar{s}(j)$, $\bar{r}(j)$ і $\bar{c}(j)$ був використаний метод аналітичної ієрархії (MAI) з метою визначення ваги моделі оцінок професорами, які працюють в університетах у Китаю, для кількісної оцінки труднощів оволодіння знаннями.

В табл. 2 наведено ваги моделі складності знань, отримані з використанням MAI.

Таблиця 2

Вагові коефіцієнти моделі складності (1)

w_1	w_2	w_3
0.633	0.260	0.107

Для цієї моделі коефіцієнт консистенції (CR) дорівнює 0,033, що менше 0,1. Це означає, що результат пройшов тест на консистенцію [8].

Для підтвердження працездатності розглянутої моделі були використані дані, отримані з набору даних MOOCCubeX. З цього набору даних для подальшого тестування було застосовано результати тестів студентів з тестовими завданнями, поведінку студентів під час перегляду відео та дані про коментарі студентів.

Для розрахунку складності були використані власноруч згенеровані набори даних, які можна вважати нормально розподіленими. В табл. 3 наведено приклад набору даних з оцінками студентів в кожному окремому розділі знань та середнім балом за весь курс.

У даній вибірці окрім даних з оцінками також присутня інформація про характер навчання (тобто чи є курс платним або безкоштовним), можливість отримати сертифікат та його тип, а також актуальний статус студента (чи користувач записаний на курс).

В табл. 4 наведено приклад набору даних з додатковими параметрами, які необхідні для розрахування складності курсу, та окремого розділу знань (середня кількість переглядів відео у кожній секції та середня кількість коментарів у кожній секції).

Після отримання необхідних даних, можна розрахувати складність окремо для кожного розділу знань.

В табл. 5 наведено приклад обчислення складностікожного окремого розділу знань з початкового курсу.

Таким чином, в наведеному прикладі найлегшим за складністю є перший розділ знань з індексом складності, який дорівнює 0,04. Найскладнішим з усього курсу є передостанній, шостий розділ знань з індексом складності, який дорівнює 0,2.

Таблиця 3

Приклад набору даних з оцінками студентів по курсу

	Student ID	Grade	Assignment Checklist 1: Assignment Checklist for Module 1	Assignment Checklist 2: Assignment Checklist for Module 2	Assignment Checklist 3: Assignment Checklist for Module 3	Assignment Checklist 4: Assignment Checklist for Module 4	Assignment Checklist 5: Assignment Checklist for Module 5	Assignment Checklist 6: Assignment Checklist for Module 6	Assignment Checklist 7: Assignment Checklist for Module 7	Enrollment Track	Verification Status	Certificate Eligible	Certificate Delivered	C
0	3269	0,78	0,85	0,75	0,90	0,80	0,65	0,70	0,70	audit	NaN	N	N	
1	3262	0,87	1,00	0,80	0,90	0,70	1,00	0,80	0,75	audit	NaN	N	N	
2	3261	0,90	0,90	1,00	0,85	0,90	1,00	0,75	0,80	audit	NaN	N	N	
3	3244	0,69	0,80	0,65	0,60	0,60	0,75	0,75	0,70	audit	NaN	N	N	
4	3228	0,88	1,00	0,75	0,90	0,70	1	0,90	0,85	audit	NaN	N	N	
5	3217	0,96	1,00	1,00	1,00	1,00	0,90	0,85	0,90	audit	NaN	N	N	
6	3214	0,93	0,90	1,00	0,85	1,00	0,85	1,00	0,80	audit	NaN	N	N	
7	3213	0,73	0,90	0,85	0,75	0,60	0,65	0,60	0,75	audit	NaN	N	N	
8	3212	0,93	1,00	0,80	1,00	0,90	0,90	1	0,85	audit	NaN	N	N	
9	3209	0,95	1,00	1,00	1,00	0,75	1,00	0,95	1,00	audit	NaN	N	N	

Таблиця 4

Приклад набору даних з додатковими параметрами моделі

	Subsection	Comments amount	Video repeats	Video Repeats Rate	Comments Rate
0	Assignment Checklist 1: Assignment Checklist f...	25,00	50,00	0,05	0,12
1	Assignment Checklist 2: Assignment Checklist f...	32,00	80,00	0,08	0,15
2	Assignment Checklist 3: Assignment Checklist f...	38,00	194,00	0,20	0,18
3	Assignment Checklist 4: Assignment Checklist f...	21,00	78,00	0,08	0,10
4	Assignment Checklist 5: Assignment Checklist f...	43,00	223,00	0,23	0,20
5	Assignment Checklist 6: Assignment Checklist f...	29,00	188,00	0,20	0,13
6	Assignment Checklist 7: Assignment Checklist f...	27,00	146,00	0,15	0,13

Таблиця 5

Результати обчислення складності розділів курсу

	Subsection	Comments amount	Video repeats	Video Repeats Rate	Comments Rate	Avg Grade	Difficulty
0	Assignment Checklist 1: Assignment Checklist f...	25	50	0,05	0,12	0,97	0,04
1	Assignment Checklist 2: Assignment Checklist f...	32	80	0,08	0,15	0,88	0,11
2	Assignment Checklist 3: Assignment Checklist f...	38	194	0,20	0,18	0,88	0,15
3	Assignment Checklist 4: Assignment Checklist f...	21	78	0,08	0,10	0,84	0,13
4	Assignment Checklist 5: Assignment Checklist f...	43	223	0,23	0,20	0,87	0,16
5	Assignment Checklist 6: Assignment Checklist f...	29	188	0,20	0,13	0,79	0,20
6	Assignment Checklist 7: Assignment Checklist f...	27	146	0,15	0,13	0,8	0,18

Подібні обчислення виконуються для кожного розділу знань у кожному курсі, що дає можливість, після побудови концептуального графу спиратися на індекс складності з метою подальшої модифікації контенту курсу для кожного студента

4. Побудова концептуального графу онлайн навчання

Структура концептуального графу. Концепти відносяться до галузей знань, які викладаються в окремих курсах (наприклад, «Дерево двійкового пошуку» курсу «Структура даних»). Ці поняття є підсумком навчальних ресурсів з точки зору знань. Останнім часом використання концептуальних графів набуло поширення в багатьох темах, пов'язаних з адаптивним навчанням (зокрема, це навчання у MBOK, організація даних на основі концептів знаннями та когнітивне моделювання тощо) [9]. Окремою проблемою засвоєння концептів є виділення понять та їх зв'язків із навчальних текстів [10]. Попередній етап збору інформації про курс надає необхідні текстові ресурси, включаючи субтитри відео, опис курсів, викладачів, навчальних програм, а також зміст вправ. Однак високоякісне одержання концепту з цих

текстів є складним завданням, адже працезатратні ручні анотації є недостатні для формування масштабних ресурсів MBOK. Щоб вирішити цю проблему, можна використовувати, наприклад, контрольований метод дрібного виділення концептів, а потім застосовувати інтерактивний механізм спільного навчання для виявлення необхідних зв'язків між поняттями з мінімальною анотацією. Крім того, можна використовувати лише тексти відеосубтитрів як джерело, оскільки тексти інших ресурсів занадто короткі для сприйняття концепту.

Розглянемо деякі аспекти формування концептів для подальшої побудови концептуального графу:

– обрання кандидата (кандидати на концепт обираються як об'єднання найбільш релевантних варіантів);

– формування ключових фраз (усі отримані фрази є витягом з анотованих якісних текстів. Вони зберігаються в таблиці речень для кожного відео, що відображається в його субтитрах як кандидат на концепт);

– зв'язування об'єктів (зв'язування об'єктів має на меті виявити посилання на них з зовнішньої бази знань. Для кожного відео виконується зв'язування сутностей за допомогою XLink і обираються пов'язані сутності як кандидати на концепт, що може допомогти обрати концепти з масштабної бази знань Xlore [11]);

– розпізнавання іменованих об'єктів (використовуються попередньо навчені мовні моделі для вилучення нерелевантних концептів. Схему навчання можна розглядати як розпізнавання названої сутності з однією категорією. Для кожної дисципліни анотуються концепти із субтитрів (120–150 відео 20–30 випадкових курсів). Потім налаштовується класифікаційна мережа RoBERTa, після чого об'єднуються заголовки відео перед субтитрами як підказки, розділені спеціальним маркером [SEP].

Приклад структури концептуального графу навчання з використанням сукупності виділених концептів наведена на рис. 3.

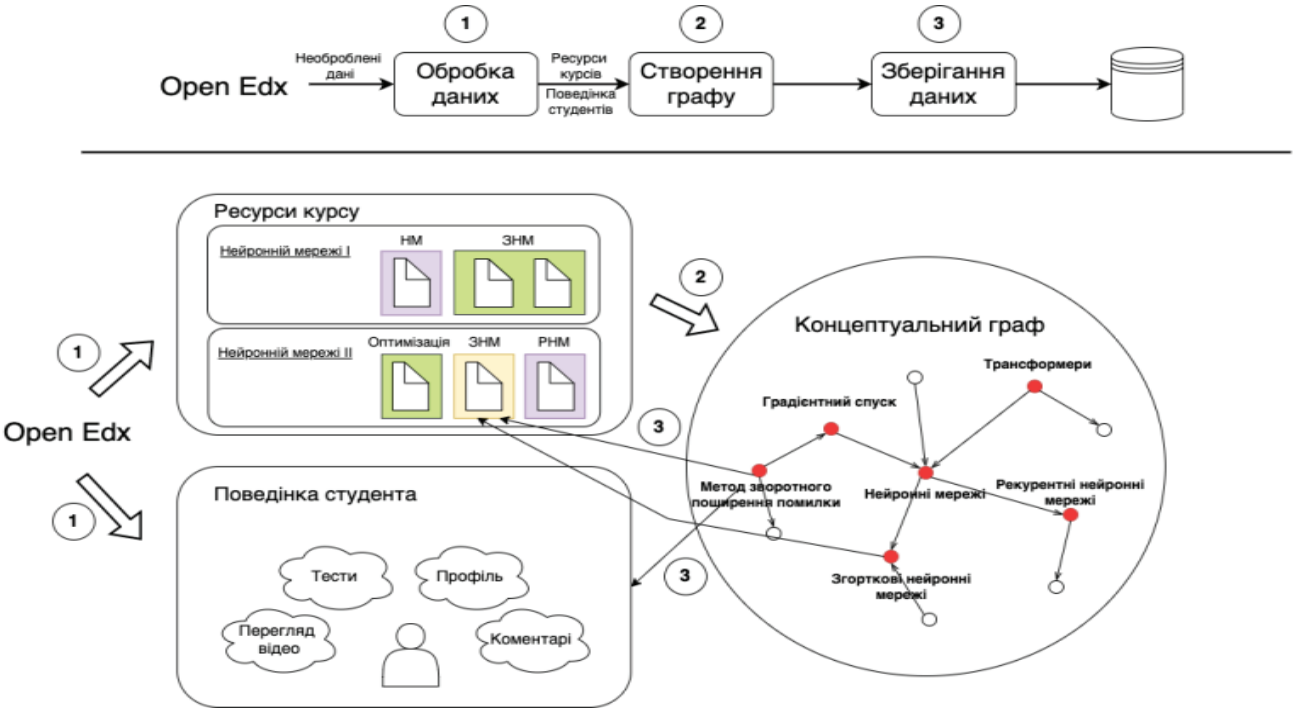

Рис. 3. Структура концептуального графу

Оцінювання стану навчання студентів згідно з концептуальним графом. Розглянемо процес динамічного оцінювання стану навчання студентів згідно з концептуальним графом курсу. Аналіз поведінки студентів MOOK під час перегляду відео фрагментів та проходження відповідних тестів дозволяє оцінювати поточний рівень засвоєння ними знань з відповідних розділів. Опис рекомендованих параметрів моделі оцінювання рівня засвоєння знань студентами наведено в табл. 6.

Таблиця 6

Параметри моделі оцінювання рівня засвоєння розділів курсу

Змінна	Тип даних	Опис
s_{ij}	Float	Нормований бал i-го студента за тестове завдання j-го розділа
r_{ij}	Bool	Кількість перемотувань вперед i-м студентом (або пропусків під час перегляду) відео j-го розділу
c_{ij}	Int	Оцінка опанування i-м студентом j-го розділу курсу

На рис. 4 наведено схему алгоритму оцінки статусу (поточного рівня засвоєння знань) студента. Детальний процес виглядає наступним чином:

– студенти вивчають початкові розділи;

– студенти дивляться відео, а потім виконують контрольні тестові завдання з розділу;

– статуси студентів оцінюються на основі алгоритму оволодіння знаннями;

– відповідні розділи знань рекомендуються студентам на основі оцінки поточного статусу їхнього навчання.

Розглянемо докладніше наведену схему оцінювання поточного статусу (поточного стану засвоєння знань) студента (далі – просто стану).

Ненормальна поведінка включає швидке перемотування вперед або пропуск під час перегляду відео, а нормальна поведінка не передбачає таких операцій.

Згідно зі схемою, наведеною на рис.4, стан засвоєння студентом окремих розділів курсу поділяється на чотири рівні (відповідно: невивчений, незасвоєний, недостатньо засвоєний та засвоєний). «Невивчений» (стан 1) означає, що нормований бал оцінки результатів проходження тесту студентом нижчий за 0.6 (наприклад, він перемотував вперед або пропускав відео під час перегляду, що не є нормальною поведінкою). Студенту мають бути присвоєні відповідні бали знань для перевірки. «Незасвоєний» (стан 2) означає, що нормалізований результат проходження тесту нижчий за 0.6, але все відео переглядалось без перемотування вперед або пропуску. Студенту мають бути присвоєні відповідні бали знань для перевірки. «Недостатньо засвоєний» (стан 3) означає, що нормалізований результат проходження тесту вище за 0.6, але менше ніж 0.8. Студенту також слід присвоїти відповідні бали знань для перевірки. «Засвоєний» (стан 4) означає, що нормалізований результат проходження тесту перевищує 0,8 і студент отримав більшу частину балів знань. Студенту при цьому слід призначити новий розділ для вивчення.

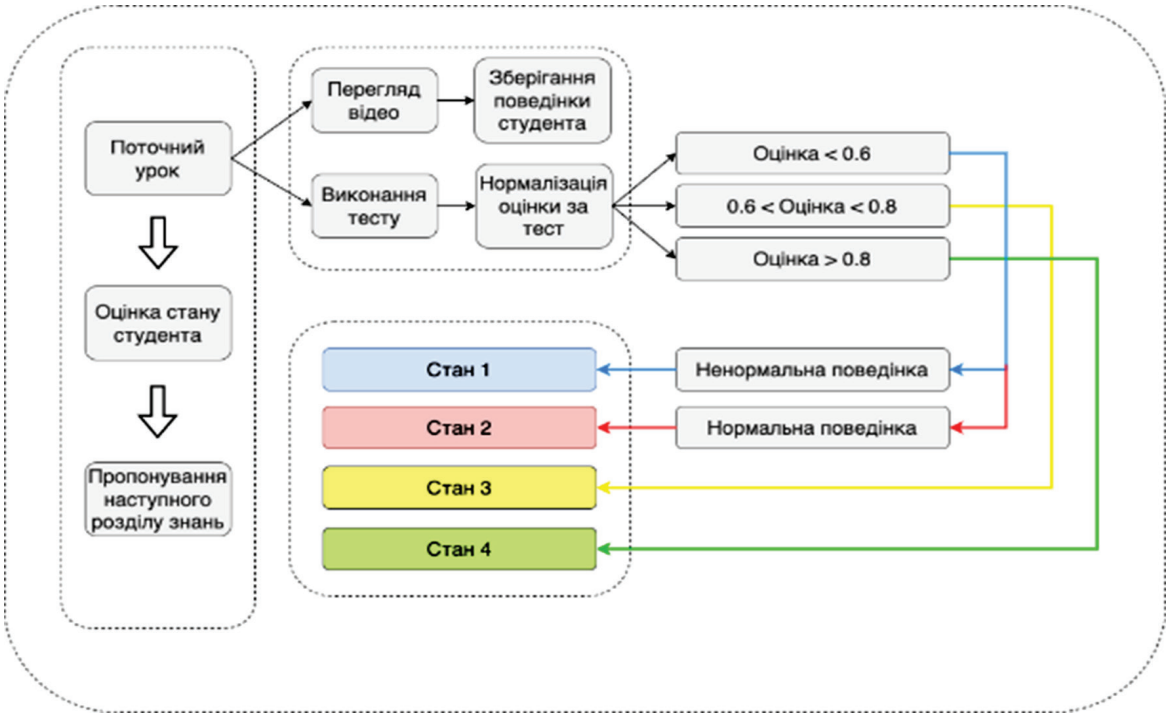

Рис. 4. Схема оцінювання поточного статусу студента

Фрагмент сценарію (шляху) можливого варіанту проходження студентом розділів курсу, передбачений рекомендаціями використання платформи Open edX, наведено на рис. 5.

Рис. 5. Приклад стандартного сценарію проходження розділів курсу

На рис. 5 застосовано такі позначення поточного статусу студента після чергового тестування за розділами: «невивчений» – I; «незасвоєний» – II; «недостатньо засвоєний» – III; «засвоєний» – IV.

В роботі [12] запропоновано модифікувати розглянутий стандартний шлях у межах певного курсу. Експерименти з алгоритмом динамічного планування шляху навчання були засновані на інтерфейсах прикладного програмування (API), наданих Університетом Цінхуа, платформі XuetangX MOOC і наборі даних MOOCCubeX. Студенти дотримувалися

фундаментальної логічної послідовності знань (тобто початкової послідовності шляху) на початку експериментів, оскільки оволодіння розширеними знаннями залежить від оволодіння необхідними знаннями. Отже, підхід до персоналізованого планування шляху навчання, запропонований у цьому дослідженні, використовував стратегію зворотного зв'язку для формування персоналізованого шляху навчання на основі станів навчання.

Схему модифікованої послідовності навчання зі зворотним зв'язком наведено на рис. 6 (позначення поточного статусу на цій схемі відповідають позначенням на рис. 5).

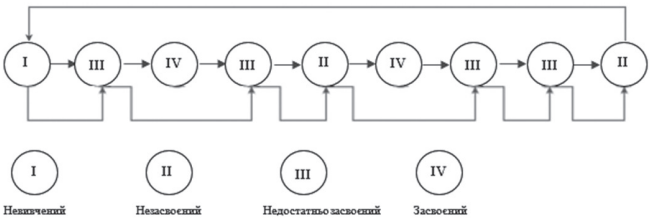

Рис. 6. Приклад модифікованого сценарію проходження розділів курсу

Коли студенти вважають, що закінчили вивчення розділу, їхній статус оволодіння знаннями автоматично оновлюється за результатами поточного тестування. Статуси присвоюються студенту згідно з отриманими балами та фіксуються у журналі навчання.Бали знань на різних рівнях навчання розташовуються на основі передумовного відношення, а бали знань на одному рівні розташовуються за зростанням. Після ознайомлення зі списком балів знань у журналі студент може знову виконати тест (з можливістю оновлення стану). Якщо в поточному розділі статус студента ще відповідає стану I (невивчений), стану II (незасвоєний) або стану III (недостатньо засвоєний), то процес повертається до відповідного кроку (кола на схемі). Якщо ж у поточному розділі статус студента відповідає стану IV (засвоєний), то процес повертається до першого кроку і продовжується до наступного запропонованого розділу курсу.

5. Персоналізація сценаріїв навчання студентів для MBOK

Необроблені дані, отримані від Open edX, мають чимало недоліків, зокрема, низьку релевантність аналізованих курсів, дублювання деяких ресурсів та надмірно детальну схему контроля процесу засвоєння матеріалу студентами. Для подолання цих проблем доцільно розробити окремі рішення для покращення якості даних. Спираючись на попередньо побудовані моделі оцінювання складності навчання та поточного статусу студента, а також на розробку концептуального графу для кожного розділу навчального курсу, було створено алгоритм персоналізації курсів, що модифікує процес навчання студента. Окремі кроки реалізації цього алгоритму можна описати таким чином:

– студент проходить онлайн курс за визначеною послідовністю;

– коли студент успішно проходить тест за поточною темою, то його стан автоматично оновлюється на основі схеми оволодіння знаннями;

– якщо за результатом тесту стан засвоєння студентом поточної теми (або низки тем) – «засвоєний», то він має продовжити навчальну послідовність, в іншому випадку студенту буде запропоновано пройти теми з іншого курсу, який має меншу складність, попередньо прораховану за допомогою моделі складності;

– якщо студент досягне необхідного рівня після тесту з тем з меншим рівнем складності, то він повернеться до базового варіанту, щоб пройти попередній тест. Коли студент не досягає потрібного стану оволодіння знаннями, то наступний тест буде з курсу з більш низьким рівнем складності.

Персоналізоване навчання сприяє успішності оволодіння знаннями студентами внаслідок використання спеціально підібраних навчальних каркасів. На рис.7 наведено приклад сценарію онлайн навчання за пропонованим алгоритмом, який може автоматично генерувати персоналізовані шляхи навчання з використанням платформи Open edX за результатами оцінювання стану навчання студентів у режимі реального часу (позначення поточного статусу на рис 7 відповідають позначенням на рис.5 та рис.6). Алгоритм передбачає можливість адаптації сценарію навчання з переходами на різні рівні складності курсу (курс1, курс2, курс 3 – на рис.7).

Варто зазначити, що даний алгоритм має враховувати не лише складність певного розділу знань та стани засвоєння матеріалу, а також і економічні аспекти (коли під час проходження першого курсу в алгоритм пропонує пройти додатковий розділ знань з іншого курсу, є вірогідність того, що студент вирішить придбати і цей курс, саме тому цей алгоритм має бути більш привабливим для інтеграції на такі платформи як Open Edx).

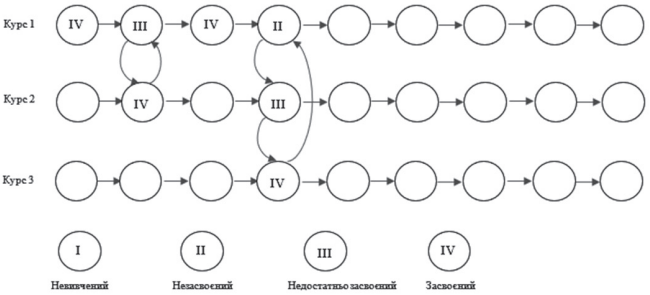

Рис. 7. Приклад визначення персоналізованого сценарію проходження розділів курсу

Також в алгоритмі передбачений ліміт на глибину рекомендацій щодо сценарію проходження курсу, щоб уникнути ситуації, коли деякі студенти мають проходити курс занадто довго.

6. Результати експериментального дослідження

В експериментах з динамічним плануванням контенту курсу за допомогою запропонованого алгоритму (на прикладі курсі «Штучний інтелект») були використані інтерфейси прикладного програмування (API). Студенти, що приймали участь в експериментах, дотримувалися початкової послідовності засвоєння тем з подальшим використанням стратегії зворотного зв'язку для формування персоналізованого шляху навчання на основі автоматизованого оцінювання станів навчання.

Для побудови концептуального графа дисципліни «Штучний інтелект» анотувалися концепти із субтитрів курсів цього напрямку. Потім налаштовувалася класифікаційна мережа RoBERTa для всіх відеосубтитрів і вибиралися фрагменти, впевненість яких перевищувала 0,85, як кандидати на концепти. Експериментальні результати показують, що метод RoBERTa-NER витягує більш дрібні концепти, ніж стандартний пошук і зв'язування сутностей.

Завдяки побудові концептуального графа ми можемо збагатити асоціацію гетерогенних ресурсів MBOK та інтегрувати більше типів зовнішніх ресурсів. Для неоднорідних MOOK-ресурсів проводиться анотація концепту, щоб зв'язати їх у концептуальному графі.

На рис. 8 наведено приклад опису ресурсу курсу «Штучний інтелект», що пропонується для онлайн навчання з використанням платформи Open edX.

CourseId	CName	Field	ChapterId	VideoId	VideoText	ExerciseId	ProblemId	ProblemText	PType
C_1729	Artificial Intelligence	CS	L_4522	V_59697	AI is the intelli -gence exhibited by machines or software.	Ex_7552	Pm_14512	Which of the follow -ing are the research fields of AI? **A**:...	0
						Ex_7554	Pm_14520	What is the AI method to represent information by symbols and their relationship?	2

Рис. 8. Приклад опису ресурсу курсу «Штучний інтелект»

У прикладі показано метаінформацію, назву курсу, номер завдання, відео, виділені концепти та відповідні проблеми в курсі «Штучний інтелект».

На рис.9 наведено приклад відображення поведінки студента в процесі онлайн навчання.

UserId	CourseId	CName	VideoId	ExerciseId	ProblemId	Answer	Comment	Reply	Time
U_112	C_1729	Artificial Intelligence	V_59697	\	\	\	Is the Turing test complete?	\	2020-04-20 16:57:50
			\	Ex_7552	Pm_14512	A	\	\	2020-04-21 10:14:13
			\	Ex_7554	Pm_14520	Symbolicism	\	\	2020-04-23 14:29:06
			V_59703	\	\	\	\	I think under -standing and intelligence are two things...	2020-04-26 21:35:45

Рис. 9. Приклад відображення поведінки студента в процесі онлайн навчання (курс «Штучний інтелект»)

У прикладі показано поведінку під час перегляду відео, поведінку вправ та поведінку коментування та відповіді студента U_112 з курсу «Штучний інтелект».

Існуючі ресурси MOOK зберігають лише дуже слабкі зв'язки через структуру курсу. Пов'язуючи різні ресурси з детальним концепт-графом, є можливість збагатити їхні зв'язки на рівні знань. Оскільки концепти витягуються з субтитрів відео, відео природно анотуються тонкими поняттями. Проте інші типи ресурсів досі не мають концептуальних анотацій. Для кожного курсу його концепти є об'єднанням його відео концептів.

Для оцінки ефективності запропонованого підходу було класифіковано поведінку групи студентів у онлайн-навчанні з курсу «Штучний інтелект» на дві категорії: ефективна поведінка та неефективна поведінка. Поведінка, що сприяла отриманню незасвоєних і недостатньо засвоєних знань (згідно з персоналізованим сценарієм навчання), була визначена як ефективна поведінка. В іншому разі поведінка студента визначалася як неефективна. Ефективність навчального шляху була визначена наступним чином

$$EBR = \frac{1}{N}\sum_{i=1}^{N}\left[1-\frac{U_i^{rep}}{U_i}\right], \quad (5)$$

де N – кількість студентів; U_i – довжина навчального шляху i-го студента; U_i^{rep} – кількість повторних засвоєнь знань навчального шляху i-го студента.

Після розрахунку середній показник ефективної поведінки у групі з 10 студентів склав близько 90%.

Це вказує на те, що запропонований алгоритм динамічного планування шляху навчання може точно виявити ефективні сценарії та підвищити ефективність навчання в МООК.

Висновки

Існуючі платформи MBOK не завжди можуть забезпечити персоналізовані послідовності навчання студентів. В даній роботі запропоновано підхід, який сприятиме реалізації персоналізованого планування навчального шляху на основі своєчасної діагностики стану навчання студентів.

Перевагою розглянутого алгоритму, окрім врахування складності кожного модуля та використання інформації про поведінку студента для аналізу засвоєння знань, є також те, що персоналізований підхід до планування схеми засвоєння навчального матеріалу може сприяти використанню різних варіантів курсів у побудові сценаріїв онлайн-навчання.

Перспективним напрямком розвитку запропонованого підходу є використання останніх досягнень в дизайні онлайн навчання та методів розширення платформи Open edX на основі генеративних штучних нейронних мереж.

Список літератури

[1] Арешонков В.Ю.(2020) Цифровізація вищої освіти: виклики та відповіді. Вісник НАПН України, № 2(2). С. 1-6.

[2] Курінний А., Вольвач В., Дарій В. (2017). Створення та розробка онлайн курсу на платформі open edX. Медична освіта. №2. 37-40 Режим доступу:https://doi.org/10.11603/me.2414-5998.2017.2.7608

[3] Мала І. (2022). Дистанційне навчання як дієвий інструмент управлінської освіти. Вчені записки Університету «КРОК», 2(66), 132–151. https://doi.org/10.31732/2663-2209-2022-66-132-151

[4] Massive Open Online Course (MOOC) (2020). http://www.prnewswire.com/newsreleases/massive-open-online-course-mooc-market-size-to-grow-from-usd-183billion-in-2015-to-usd-850-billion-by-2020–

[5] Belinskiy.A. (2021) Exploring engagement profiling in MOOCs through Learning Analytics: The Open edX Case, 24-27 p.

[6] Niknam M., Thulasiraman P. (2020). LPR: A bio-inspired intelligent learning path recommendation system based on meaningful learning theory. Educ. Inf. Technol. 2020, 25, 3797–3819.

[7] Hailong Jin, Chengjiang Li, Jing Zhang, Lei Hou, Juanzi Li, and Peng Zhang (2018). XLORE2: Large-scale cross-lingual knowledge graph construction and application. Data Intelligence 1, 77–98.

[8] Jing Zhang, Yixin Cao, Lei Hou, Juanzi Li, and Hai-Tao Zheng (2017). XLink: An Unsupervised Bilingual Entity Linking System. In Chinese Computational Linguistics and Natural Language Processing Based on Naturally Annotated Big Data, Maosong Sun, Xiaojie Wang, Baobao Chang, and Deyi Xiong (Eds.). Springer International Publishing, Cham, 172–183.

[9] Удовенко С., Чала Л. (2022) Алгоритм персоналізації контенту навчальних курсів на основі платформи Open Edx. Збірник наукових праць за матеріалами Всеукраїнської науково-методичної інтернет конференції «Актуальні проблеми освітньо-виховного процесу та шляхи їх вирішення в умовах сучасних викликів». Харків: ХНАДУ. .398-401. URL: https://fmab.khadi.kharkov.ua/index.php?id=1281&no_cache=1

[10] Удовенко С., Чала Л., Гриньова О. (2019) Метод аналізу звязків між концептами предметних онтологій. Матеріали XVII Міжнародної науково-практичної конференції «Математичне та програмне забезпечення інтелектуальних систем». Дніпро. 264-265 Режим доступу: http://mpzis.dnu.dp.ua/wp-content/uploads/2019/12/MPZIS_2019.pdf

[11] Jing Zhang, Yixin Cao, Lei Hou, Juanzi Li, and Hai-Tao Zheng (2017). XLink: An Unsupervised Bilingual Entity Linking System. In Chinese Computational Linguistics and Natural Language Processing Based on Naturally Annotated Big Data, Maosong Sun, Xiaojie Wang, Baobao Chang, and Deyi Xiong (Eds.). Springer International Publishing, Cham, 172–183.

[12] Rohloff T.; Sauer D.; Meinel C. (2019) On the acceptance and usefulness of personalized learning objectives in MOOCs. In Proceedings of the the Sixth ACM Conference on Learning@Scale, Chicago, IL, USA, 24–25 June 2019; pp. 1–10.

Надійшла до редколегії 18.10.2023

УДК 656+004.9 DOi 10.30837/ bi.2023.1(99).05

С. М. Неронов[1], Г. А. Плєхова[2], М. В. Костікова[3]
[1]ХНАДУ, м. Харків, Україна, sernikner@gmail.com, ORCID iD: 0000-0003-2381-1271
[2]ХНАДУ, м. Харків, Україна, plehovaanna11@gmail.com, ORCID iD: 0000-0002-6912-6520
[3]ХНАДУ, м. Харків, Україна, kmv_topaz@ukr.net, ORCID iD: 0000-0001-5197-7389

НЕЙРОМЕРЕЖЕВА СИНЕРГЕТИКА ТА NEURONET АВТОМОБІЛЬНОГО ТРАНСФЕРУ

Проведено дослідження з доведення доцільності перенесення віртуального управління автотранспортом в хмарну середу як з технічної так і з економічної точки зору клієнтури транспортних підприємств. З метою розгортання клієнт сервісної технології рухомого складу автомобільного трансферу були використані інструментальні засоби – Data center ХНАДУ, Internet Google, Information Services. Науковий результат полягає у доведенні доцільності розробки програмної платформи програмного забезпечення автомобільних комп'ютерних систем за технологією Web та переходу до IT Industry 4.0.

ТРАНСФЕР, IT-ІНДУСТРІЯ, WEB, CLOUD COMPUTING, КОНКУРЕНТНА СПРОМОЖНІСТЬ, ДЕЦЕНТРАЛІЗОВАНІ СХОВИЩА ДАННИХ.

S. M. Neronov, G. A. Pliekhova, M. V. Kostikova. Neural network synergy and Neuronet of car transfer. A study was conducted to prove the feasibility of transferring the virtual management of motor vehicles to the cloud environment both from the technical and economic point of view of the clientele of transport enterprises. In order to deploy the customer service technology of the rolling stock of the car transfer, tools were used - Data center of the Khnadu, Internet Google, Information Services. The scientific result consists in proving the expediency of developing a software platform for automotive computer systems based on Web technology and the transition to IT Industry 4.0.

TRANSFER, IT-INDUSTRY, WEB, CLOUD COMPUTING, COMPETITIVE CAPACITY, DECENTRALIZED DATA STORAGE.

Вступ

У зв'язку з постійним інформаційним розвитком суспільства та його промислової складової нові транспортні системи і машини досягли високого інформаційного рівня досконалості. Відповідно з'явилося нове протиріччя між стрімким розвитком засобів та методів інформатизації складних об'єктів і систем та гетерогенним характером існуючих підсистем та ланок транспортного комплексу України. Розв'язання цього протиріччя дозволить на всіх рівнях транспортної інфраструктури поліпшити обслуговування мешканців міст і регіонів, удосконалити перевізні процеси, уникнути існуючих негативних впливів. Як наслідок, усуваються: збої в організації руху, незадовільний стан шляхів сполучень, нераціональне використання коштів, що виділяються на ремонт, експлуатацію та облаштованість транспортних магістралей. Це буде сприяти підвищенню безпеки руху, покращенню якості транспортних послуг, забезпеченню комфорту пересування людей та збереженню вантажу.

Контент відповідних сервісів повинен базуватися на просторово-часовій орієнтації, алгоритмізації та маршрутизації рухомого складу підприємств та організацій, що забезпечують перевізні процеси. Однак сьогодні необхідність такого контенту віртуальна логістика лише декларує, тому що не може усунути відповідні вартісні обмеження. Тому необхідно зосередити увагу саме на впровадженні та імплементації основних положень віртуального управління перевізними процесами.

Метою виконання цієї роботи є доведення доцільності перенесення віртуального управління автотранспортом в хмарну середу як з технічної так і з економічної точки зору клієнтури транспортних підприємств. Об'єкт дослідження – просторово-часова орієнтація учасників дорожнього руху. Предмет дослідження – Cloud Computing автомобільного трансферу. Завдання – розгортання клієнт сервісної технології рухомого складу автомобільного трансферу. Інструментальні засоби – Data center ХНАДУ (файловий архів ХНАДУ www.files.khadi.kharkov.ua); Internet Google, Information Services.

В основі дослідження вибір хмари, імплементація гіпотези про зниження втрат, що пов'язані з розвитком комп'ютерного ресурсу автотрансферу, з Industry 4.0 в задачах підвищення конкурентної спроможності дорожніх транспортних підприємств України.

Головне в удосконаленні перевізного процесу для ланцюга виробника, промисловості, перевізника, отримувача є задача надання учасникам перевізного процесу, особам, що приймають рішення з віртуального управління транспортними та дорожніми організаціями, інформацію про дорожні ситуації. Рішення має інструментальний засіб – Internet сайт, когнітивній комп'ютерної технології Web прийняття рішень щодо раціональної організації автомобільного трансферу (будь-якого пересування пасажирів або вантажу у просторово-часовому просторі перевізних процесів) з урахуванням стану дорожнього середовищу. На відміну від існуючого стану логістики,

основних законів, правил та принципів розвитку ІТ-індустрії передбачається інтерактивний моніторинг як автомобілю, так і учасників перевізного процесу, саме дороги.

Науковий результат полягає у науковому обґрунтуванні механізму самоорганізації синергетичного об'єднання комп'ютерних ресурсів усіх учасників дорожнього руху в єдиному інформаційному просторі глобальної мережі Internet – від окремої транспортної машини до корпоративного рівня віртуальній логістики.

Ця робота є дослідженням основних законів, тверджень та правил створення автомобільних комп'ютерних систем (АКС). Вона є основою застосування на транспорті віртуальній логістики та розвиток ІТ-індустрії. Це не тільки організація надпотужних розподілених обчислень, але й така спільна робота користувачів, що надає можливості достатньо повного використання Cloud Computing. Для цього ідеально підходить сучасна Internet-технологія типу Web 2.0 (Web 1-4).

В основі полягає модель оптимізації інфраструктури (ІО) Microsoft з використанням досвіду, накопиченого як ІТ-індустрією, так і самої Microsoft. Модель ІО є послідовністю чотирьох рівнів (або фаз) поступово зростаючої технологічної зрілості: «Базовий», «Стандартизований», «Раціоналізований», «Динамічний». Для будь-яких підприємств та організацій, фірм (далі просто компаній) з інфраструктурою рівня «Базовий» характерні ручні локалізовані процеси, мінімальне централізоване керівництво. Компанію, інфраструктура якої знаходиться на рівні «Стандартизований», можна охарактеризувати як таку, що має керовану інфраструктуру. Організації з інфраструктурою рівня «Раціоналізований», як правило, вже відіграють значну роль у підтримці й розширенні бізнесу. Організації з інфраструктурою рівня «Динамічний» мають чітку уяву про стратегічне значення інфраструктур для ефективності бізнесу й конкурентоспроможності.

ІТ-відділи орієнтуються на потреби бізнесу та керуються ними. Досвід та наукові досягнення першої в Україні наукової школи з синергетики, мехатроніки та телематики ХНАДУ дозволяють визначити рівень ІТ використанням Cloud Computing та Web 1-4 як «Когнітивний» бізнесовий рівень синергетичній комп'ютерної зрілості будь-яких ІТ-компаній.

Основним є базування на принципах правильного просторово-часового співвідношення спеціальних та універсальних рішень Макімото з урахуванням закону Амдала та відомого твердження Мура. Розробники проекту виконали складну імплементацію існуючого транспортного порталу ХНАДУ у новий логістичний портал інформаційний сайт агрегатор можливих маршрутів згідно особливостям перевізних процесів в умовах стохастичного попиту клієнтури транспортних та дорожніх підприємств.

Основним джерелом цієї розробки є постановка задачі на підвищення конкурентної спроможності транспортних підприємств в умовах розвитку ІТ-індустрії віртуального управління перевізними процесами. У дослідженнях [1, 2] визначені проблеми інтеграції транспортних застосувань з створенням інтелектуальних транспортних систем. Результати експериментів на платформі транспортного засобу [1] доводять як надійно виявляти транспортні засоби в реальних транспортних середовищах. У статті [2] визначається стратегія роботи програмного забезпечення, оцінюється відповідний контент. Віртуальне управління розглянуто як основа для використання основних положень розробки інформаційних комунікаційних технологій (ІКТ) саме у транспортних застосуваннях [3, 4]. У дослідженні [3] наведено використання для цього нечіткої логіки. У статті [4] також висвітлені рішення проблеми застосування нейроматематики у транспортних додатках. Загально теоретичні та прикладні задачі інформаційної підтримки цієї розробки висвітлені результатами [5, 6]. У статті [5] наведено дані про зв'язки трафіку.

Доведення твердження про необхідність удосконалення існуючих систем інформаційної системи підтримки прийняття рішень є результатом досліджень [7, 8]. На відміну від цього пропонується не простий модельний підхід [7], удосконалення маркетингових взаємовідношень [8], а синергетичне об'єднання внутрішньої та зовнішньої телематики рухомого складу перевізника.

Цільова настанова передбачає підвищення ефективності віртуального управління перевізними процесами за рахунок синергетичного об'єднання внутрішньої та зовнішньої телематики рухомого складу усіх учасників перевізного процесу та ринку транспортних послуг. Завдання спрямовано на своєчасне прийняття рішень та інтерактивний моніторинг умов руху, раціональне розподілення комп'ютерних ресурсів учасників транспортного процесу для зниження витрат. Також досягається раціональне сполучення єдиного інформаційного простору та використання Cloud Computing у віртуальному управлінні перевізними процесами. Результат повинен полягати в усуненні протиріч наявності загальних вартісних обмежень та потрібних комп'ютерних ресурсів з раціональної організації клієнт-серверної технології перевізного процесу. Своєрідною вільною нішею відповідних розробок є синергетичний, інформаційний розвиток ринку транспортних послуг.

Спроба визначити таке об'єднання запропонована у прикладних дослідженнях [9] та статті, що присвячена оптимізації трафіку рухомого складу [10].

Вона полягає у доведенні доцільності розробки програмної платформи програмного забезпечення автомобільних комп'ютерних систем (АКС) за технологією Web та переходу до IT Industry 4.0. Основним питанням є системне адміністрування автомобільної телематики в задачах розподілу комп'ютерного ресурсу, саме з професійного становлення аналітики вимог, або бізнесової аналітики, системного адміністратора корпоративної телематики транспортних або дорожніх підприємств [11 – 17].

1-й етап – постановка задачі на створення теоретичних основ розподілення комп'ютерних ресурсів між учасниками дорожнього руху, користувачами автомобільних доріг. Декларування аксіоматики, дослідження закономірностей розвитку телематики на автомобільному транспорті та визначення основних принципів як використати новітню мережеву технологію Cloud Computing у транспортних та дорожніх організаціях. 2-й етап – пропозиції з доведення досяжності, спостерігаємості і створення клієнтської частини телематики транспортної (дорожньої) організації, автомобілю – засобів інформаційної взаємодії учасників дорожнього руху. Фізичне, імітаційне моделювання, тестування, верифікація комп'ютеризації інформаційних процесів оцінки дорожніх ситуацій. Інтерактивності транспортних процесів стали експериментальним підтвердженням шляхів втілення у транспортних та дорожніх організаціях інформаційно-комунікаційної технології управління наземним транспортом. Цьому передувало розробка та створення внутрішньої автомобільної телематики, інтерактивна система реєстрації, оцінки та накопичення, узагальнення даних про оперативну ситуацію і середовище дорожнього руху. 3-й етап – доведення достовірності висловлених принципів, закономірностей втілення у транспортних системах інформаційно-комунікаційних технологій спостереження та моніторингу транспортних ситуацій.

Це доведення базувалося на прикладі оцінки інвестиційної привабливості та впровадження в державних підприємствах, задіяних в утриманні автомобільних доріг, інформаційно-комунікаційної технології огляду автомобільних доріг.

Ці данні також корисні для розгляду синергії інформаційної діяльності IT-фахівців, соціалізації учасників дорожнього руху.

1. Транспортна інфраструктура

Сучасна транспортна інфраструктура міст та регіонів є сукупністю інтелектуальних систем планування та моделювання транспортних мереж, керування дорожнім рухом та телематичними комплексами, які надають оперативну інформацію про стан дорожнього середовища та дозволяють синергетичне взаємодіяти із всіма учасниками дорожнього руху. Для розвитку та експлуатацію транспортної інфраструктури потрібні потужні комп'ютерні ресурси. Можливості їх удосконалювання гальмує брак коштів, що властиво практично усім місцевим органам самоврядування.

Розподілені комп'ютерні системи дають можливість вже сьогодні отримати такі ресурси за рахунок використання принципу розподілу апаратних, програмних складових мереж, паралельної роботи декількох користувачів. Кластерні рішення, віртуалізація програмно-апаратних засобів LAN, оптимізація навантаження вузлів мереж – напрямки отримання «додаткових» комп'ютерних ресурсів. Розглянемо формальний опис архітектури такої системи. Вона складається з узагальненого інформаційного простору G, який є аналогом WAN. До цього простору входять N_i – локальні мережі LAN. Усі мережі є сукупністю логічних (логічно неподільних L_{ij}) та фізичних (фізично неподільних a_{ijl}) вузлів.

Історично склалося, що у великих містах обчислювальні мережі будь-яких стабільно існуючих підприємств, організацій різних профілів будувалися по мірі фінансування та удосконалення можливостей придбання комп'ютерного обладнання. Практика впровадження нових технологій випереджувала науково-технічне обґрунтування, оцінку ефективності проектних рішень та узагальнення результатів, яких було досягнуто.

Поступово такі мережі перетворювалися із простих обчислювальних комплексів до взаємно пов'язаних систем корпоративного рівня, які мають такі комп'ютерні резерви, що забезпечують рішення поточних завдань перетворення та подання користувачам необхідної інформації. Розглянемо концептуальне обґрунтування отримання додаткових комп'ютерних ресурсів для розвитку транспортної інфраструктури великого міста або регіону за рахунок доступу до таких комп'ютерних систем.

Будь-які комп'ютерні ресурси організацій та підприємств, які стабільно розвиваються, мають і тенденції розвитку комп'ютерних ресурсів. У їх обчислювальному середовищі можливості координації використання гетерогенних розподілених ресурсів покладають на GRID-технології, як найбільш прості реалізації Cloud Computing. Вони надають можливості застосування різноманітних ресурсів: обчислювальних, накопичення даних та комунікаційних. Слід зазначити, що надійність та продуктивність окремих систем може бути порівняно невеликою, але користувач такої розподіленої системи отримує єдину надійну та продуктивну платформу для обчислень, отримання доступу до баз даних та знань, а також може зберігати свої дані та користуватися різноманітними комунікаційними технологіями. Розвиток топології обчислювальних мереж практично проходить три

рівня: Intragrid (внутрішні GRID) → Extragrid (зовнішній GRID, що об'єднують вже декілька організацій) → Intergrid (глобальні системи, які об'єднують вже багато організацій, партнерів, кластерних рішень). Саме цей рівень GRID повинен відповідати за розвиток транспортної інфраструктури великого міста або регіону. Звичайно таке об'єднання координується GRID-системою, а відповідна віртуальна мережа дозволяє технічно об'єднати розрізнені внутрішні мережі та кластери у єдиний інформаційний простір, що координується вже єдиною GRID-технологією, яка надається користувачу у вигляді єдиної віртуальної платформи. Відповідна топологія на нижньому рівні – це окрема обчислювальна лабораторія, у якій користувач застосовує механізм віртуальної ЕОМ і має доступ до Intergrid-ресурсів.

Своєрідне занурення автоматизованих робочих місць окремої транспортної WEB-лабораторії, інформаційного відділу підприємства до комп'ютеризованого простору існуючих у великому місті окремих систем, що мають Intergrid-ресурси дозволяє отримати значний зиск від використання залучених таким чином додаткових комп'ютерних ресурсів.

У цьому випадку немає необхідності створення великої корпоративної мережі для підтримки транспортної інфраструктури. Однак, для визначення можливості отримання такого зиску потрібно прогнозувати розвиток відповідної окремої комп'ютерної системи, визначити її програмну платформу, операційну систему. Якщо проаналізувати таку тенденцію використання в існуючих комп'ютерних мережах різних операційних систем, то можна стверджувати про перевагу Microsoft Windows (Windows XP). Інші версії цієї операційної системи використовуються менш частіше. Це обумовлено тим, що старі версії працюють на комп'ютерах, які мають невелику продуктивність обчислень. Новітні версії операційних систем впроваджуються порівняно повільно, завдяки необхідності підвищення кваліфікації користувачів. Також можна затверджувати, що новітні операційні системи, як правило, встановлюються на нове обладнання, що теж опосередковано впливає на популярність їх застосування.

Однак, більшість реалізацій розподілених додатків та GRID-систем, або Cloud Computing систем зараз реалізовано для платформи Linux. Тому доцільним є розглянути ці системи у якості альтернативи чи додаткової операційної системи. Згідно досвіду попередніх етапів цього дослідження висловимо наступне.

У середовищі GRID-технологій з практичної точці зору найбільш зручним та багатофункціональним буде дистрибутив Instant-GRID (http://www.Instant-grid.org). Він базується на системі Knoppix (Live-CD) та системи Globus Toolkit (програмне забезпечення проміжного рівня – middleware, що забезпечує можливість застосування GRID-технологій). Система Instant-grid має зручний інтерфейс користувача, який оснований на використанні технологій подання даних web-браузером. Ця система поєднує як програми, що виконуються у консольному режимі, так і графічні додатки. У дистрибутиві реалізовано багато рішень Globus Toolkit, наприклад: компонент WS-GRAM для управління задачами, що вирішуються у системі, система GRIDFTP для обміну файлами та система віддаленого входу до машин-клієнтів по захищеному каналу.

Таким чином, можна визначити концепцію застосування програмного забезпечення для рівня LAN, що входить до складу розвиненої WAN мережі універсального призначення. Взагалі впровадження GRID-технологій потребує використання мережевої операційної системи. На робочих станціях можна виконати подвійну установку операційних систем. Одна – буде із родини Windows, а інша – на базі вільної операційної системи типу Linux.

Розвиток гетерогенних комп'ютерних ресурсів передбачає принцип конкуренції, що забезпечує «виживання» найбільш ефективних зв'язків згідно властивості самоорганізації. У такому випадку можливо, що в комп'ютерних мережах – подібно тому, як це відбувається в інших синергетичних системах, – виникнуть нові більш ефективні об'єднання комп'ютерів, та розподіл завдань між окремими комп'ютерами.

Навіть візуальний аналіз результатів моніторингу окремих вузлів будь-якої мережі доведе, що їх навантаження досить нерівномірне у розрізі доби, тижня, місяцю. Твердження про своєрідну рекуперацію резервів комп'ютерних потужностей, що полягає у одночасному використанні комп'ютерів для вирішення завдань різних користувачів, потребує паралельної обробки інформації, синергетики та самоорганізації вузла LAN. Такий вузол, включно з програмною платформою, є синергетичним комп'ютером логічного рівня мережевої системи, принцип роботи якого засновано як на розподілених обчисленнях, так і визначенні віртуального комп'ютерного середовища користувачів. У цілому це організаційно-технічна система, яка повинна базуватися на принципах структурної стійкості, структурного збігу, автономності та ефективної реалізації у сенсі, який було визначено раніше у дослідженнях [18, 19]. Із впровадженням GRID-технології ми одержуємо не просто інформаційну систему, а своєрідний інтелектуальний регулятор, що сполучує переваги систем програмного керування із адаптивними системами, що працюють на основі синтезу керуючого впливу. Такі властивості регулятора обумовлюють надання властивості аналога інтелекту програмно-апаратної системи забезпечення функціонування складових GRID-технології. Вона буде складовою частиною інтелектуальної

технології управління рухом наземного транспорту великих міст та регіонів. З одного боку така система буде входити до простору WAN, LAN, а з іншого – в неї занурено LAN мережу транспортної організації.

Основною вимогою впровадження новітніх технологій є забезпечення цілісності програмного комплексу, що забезпечує виконання наукових розрахунків, моделювання та обробки експериментальних даних. У цьому комплексі обчислювальна мережа транспортної організації являє собою уніфіковане програмно-апаратне середовище, у якому паралельно виконуються програмні модулі: офісні додатки, програмні засоби Internet, системи автоматизації проектування, моделювання та спеціальні програмні комплекси. Однак, після більш ретельного аналізу багатьох задач, що вирішуються як у транспортній організації або будь-якої промисловій слід висловити твердження про великі обсяги комп'ютерного навантаження каналів, перш за все сполучень з Internet.

2. Математичний опис та формалізація

Інформація у транспортних системах існує у різному цифровому уявленні та графічному й мультимедійному вигляді. Цьому уявленню відповідає просторово-часове існування цифрового контенту. Інформаційна взаємодія користувачів відповідної IT-інфраструктури транспортних організацій основана на електронному документообігу на рівні локальної (LAN) або глобальної мережі (WAN) із виходом до Internet. Розгляд властивостей цієї взаємодії потребує врахування як сукупної продуктивності вузлів відповідної мережі, так і фінансових витрат на забезпечення відповідних енергоресурсів та сервісного обслуговування. Тому застосування та використання єдиного інформаційного простору транспортних організацій потребує оптимізації багатокритеріальної мережевої системи. Найбільш привабливою для такої оптимізації є концепція web 2.0, що є логічним продовженням розвитку колективного користування ресурсами Internet. Застосування технологій web 2.0 знімає потреби у програмуванні як для виконання задач обробки інформації відповідного електронного документообігу, так і при створенні учасниками руху своїх особистих web-ресурсів. Однак обчислення та рішення задач для транспортних додатків потребують наявності потужних комп'ютерних систем.

Вирішення проблеми апаратного забезпечення продуктивних обчислень можливо за рахунок отримання «додаткових» ресурсів на базі існуючих великих комп'ютерних систем та корпоративних мереж за рахунок застосування новітніх Cloud-технологій. Ці технології надають засоби для організації єдиного обчислювального середовища у гетерогенних розподілених системах. Особливість формування єдиного інформаційного простору з використанням Cloud-технологій полягає не тільки у технічній організації розподілених обчислень гетерогенного середовища, але й у створенні певної соціальної структури, до якої залучаються користувачі цієї системи – учасники руху. Тому GRID або Cloud є організаційно-технічною системою, що має високий рівень автоматизації. Операторне уявлення вузла такої системи можна представити, як

$$h_i(\tau) = H[h_x(t), h_y(t), h_z(t), \tau], \qquad (1)$$

де $h_i(\tau)$ — динамічна функція, що відповідає процесу функціонування i-го об'єкта в досліджуваній системі на часовому інтервалі τ.

Під об'єктом у (1) слід розуміти логічний i-вузол LAN. Реальна інтерпретація $h_i(t)$ – продуктивність (навантаження) або пропускна здатність вузла GRID, якій відповідає система змінних: $x(t)$ – навантаження логічних (фізичних) вузлів LAN; $y(t)$ – ємність пам'яті вузлів LAN; $z(t)$ – пропускна здатність вузлів LAN.

Впровадження GRID-технологій (далі будемо визначати такі технології просто як WEB) надає не просто інформаційну систему, а своєрідний інтелектуальний WEB-регулятор, що поєднує переваги систем програмного керування із адаптивними системами, які працюють на основі синтезу керуючого впливу. Така інтелектуальна технологія управління рухом наземного транспорту великих міст та регіонів буде входити до простору WAN транспортної корпорації. Поряд з цим у неї будуть «занурені» локальні мережі різних транспортних організацій, що входять до відповідної корпорації (об'єднання). У силу того, що LAN та WAN, як правило, є гетерогенними системами, мають різнорідну фізичну структуру, їхній опис виконується за допомогою узагальнених методів.

Математично це вимагає застосування замість звичайних арифметико-логічних співвідношень, засобів функціонального аналізу: теорії операторів і операторні співвідношення.

Висновки

Для математичного опису процесів, динаміки зміну стану вузлів LAN та WAN можна застосувати загальні операторні залежності, поняття метричного простору та визначення існування мережі у метричному просторі часових перетворень. Ймовірності зміни станів фізичних вузлів LAN, узагальнення поняття стану складної системи на вузли WAN та застосування основних положень теорії обслуговування за аналогією «обслуговування верстатів», індикаторні функції стану окремого вузла дозволяють оцінити основні характеристики GRID-системи: пропускну здатність – $z(t)$, навантаження – $x(t)$, ємність пам'яті – $y(t)$, як динамічні функції.

Застосування для формального аналізу GRID-технологій основних положень теорії операторів є справедливою.

Практичний результат: рекомендації з використання Cloud Computing (хмарних обчислень) для створення єдиного інформаційного простору транспортних послуг без зайвих капітальних витрат на створення спеціальної IT-інфраструктури, що є основою підвищення конкурентній спроможності транспортних та дорожніх організації.

Список літератури:

[1] Guo J. Preceding Vehicle Detection and Tracking Adaptive to Illumination Variation in Night Traffic Scenes Based on Relevance Analysis / Guo J., Wang J., Guo X., Yu C., & Sun X. // Sensors. – 2014. – 14 (8). – P. 15325–15347.

[2] Aksjonov A. Design and Simulation of the Robust ABS and ESP Fuzzy Logic Controller on the Complex Braking Maneuvers / Aksjonov A, Augsburg K, Vodovozov V. // Applied Sciences. 2016. – 6 (12):382. – 18 p.

[3] Клец Д. М. Научные основы системного обеспечения маневренности автомобиля с применением новых принципов действия и элементов искусственного интеллекта / Д. М. Клец // Зб. наук. пр. ПолтНТУ. – 2013. – № 1 (36). – С. 113–123.

[4] Bodyanskiy Y. V. Adaptive learning of an evolving cascade neo-fuzzy system in data stream mining tasks / Bodyanskiy Y. V., Tyshchenko O. K., Kopaliani D. S. // Evolving Systems. – 2016. – 7 (2). – P. 107–116.

[5] Karel Z. Assistance System for Traffic Signs Inventory / Karel Z, Tomáš K, David P, Vytečka Marcel // Acta Universitatis Agriculturae et Silviculturae Mendelianae Brunensis. – 2015 – 63 (6). – P. 2197–2204.

[6] Thüm T. Secure and Customizable Data Management for Automotive Systems: A Feasibility Study / Thüm T, Schulze S, Pukall M, Saake G, Günther S. // ISRN Software Engineering. – 2012. Vol. 2012. – 8 p.

[7] Naumov V. Evaluation of freight forwarder risk to transportation market entry / V. Naumov // Eastern-European Journal of Enterprise Technologies. – 2015. – Vol. 4/3 (76). – P. 28–31.

[8] Naumov V. Definition of the optimal strategies of transportation market participators / V. Naumov // Transport Problems: an International Scientific Journal. – 2012. – Vol. 7. – Is. 1. – P. 43–52.

[9] Kovac M. Innovative applications of vehicle / Kovac M., Leskova A. // Journal of Systems Integration. – 2012. – 3 (4), P. 51–60.

[10] Li S. Research on the Method of Traffic Organization and Optimization Based on Dynamic Traffic Flow Model / Li S, Wang G, Wang T, Ren H // Discrete Dynamics in Nature and Society. – 2017. – 10 p.

[11] Свідоцтво авторського права на твір № 63190 «Транспортна телематика (презентація професійної діяльності зі створення автомобільних комп'ютерних систем – АКС)» / О. П. Алексієв, В. О. Алексієв. – Дата реєстрації 24.12. 2015, Бюл. № 15.

[12] Свідоцтво авторського права на твір науково-практичного характеру № 63148. «SYNERGETICS» (презентація синергетичного підходу до створення автомобільних комп'ютерних систем) / О. П. Алексієв, В. О. Алексієв. – Дата реєстрації 24.12. 2015, Бюл. № 15.

[13] Свідоцтво авторського права на твір науково-практичного характеру № 63149. «Автомобільна мехатроніка» (Термінологічний довідник-електронний ресурс з автомобільної мехатроніки) / О. П. Алексієв, В. О. Алексієв. – Дата реєстрації 24.12. 2015, Бюл. № 15.

[14] Свідоцтво авторського права на твір науково-практичного характеру № 63189 «Інформаційна технологія створення автомобільних комп'ютерних систем (конспект лекцій з створення автомобільних комп'ютерних систем – АКС ІТ АКС)» / О. П. Алексієв, В. О. Алексієв. – Дата реєстрації 24.12. 2015, Бюл. № 15.

[15] Свідоцтво авторського права на твір науково-практичного характеру № 63188 «Автоніка» (презентація результатів дослідження зі створення автомобільних комп'ютерних систем) / О. П. Алексієв, В. О. Алексієв. – Дата реєстрації 24.12. 2015, Бюл. № 15.

[16] Свідоцтво авторського права на твір науково-практичного характеру № 63192 «Вступ до системної інженерії (навчально-методичний посібник для підготовки системних інженерів з автомобільних комп'ютерних систем)» / О. П. Алексієв, В. О. Алексієв. – Дата реєстрації 24.12. 2015, Бюл. № 15.

[17] Свідоцтво авторського права на твір науково-практичного характеру № 63193 «Розподілені телематичні транспортні системи (презентація постановки задачи розподілених телематичних транспортних систем) «DISTRYBYTED TELEMATICS» / О. П. Алексієв, В. О. Алексієв. – Дата реєстрації 24.12. 2015, Бюл. № 15.

[18] Алексієв В. О. Мехатроніка, телематика, синергетика у транспортних додатках: навчально-методичний посібник / В. О. Алексієв, О. П. Алексієв, О. Я. Ніконов. – Харків: ХНАДУ, 2011. – 212 с.

[19] Алексієв В. О. Інтерактивний моніторинг автомобільних доріг / В. О. Алексієв, О. П. Алексієв, А. А. Видмиш, В. О. Хабаров – Харків: ХНАДУ, 2012. – 160 с.

Надійшла до редколегії 22.06.2023

УДК 004.8 DOi 10.30837/ bi.2023.1(99).06

Д.С. Суворов[1], І.В. Афанасьєва[2], К.Г. Онищенко[3], О.В. Калиниченко[4]

[1]ХНУРЕ, м. Харків, Україна, daniil.suvorov@nure.ua, ORCID iD: 0009-0008-0083-1978

[2] ХНУРЕ м. Харків, Україна, iryna.afanasieva@nure.ua, ORCID iD: 0000-0003-4061-0332

[3]ХНУРЕ, м. Харків, Україна, kostiantyn.onyshchenko@nure.ua, ORCID iD: 0000-0002-7746-4570

[4]ХНУРЕ, м. Харків, Україна, olga.kalynychenko@nure.ua, ORCID iD: 0000-0003-1466-3967

ВПЛИВ РОЗМІРУ КАДРУ НА РОЗПІЗНАВАННЯ ЕМОЦІЇ ЗА МОВЛЕННЯМ

У задачі розпізнавання емоції за мовленням, як і у більшості задач машинного навчання розпізнавання за звуком, використовується так званий фреймінг. Це процес поділу вихідного аудіосигналу на кадри певного розміру, кожен з яких оброблюється окремо. У цій статті представлено порівняння впливу розміру кадрів на результат розпізнавання емоції на прикладі CNN мережі. Для експериментів використовувався набір CREMA-D із аугментаціями, використовуючи додавання шуму, розтягування у часі та зміну висоти тону. В ході досліджень вдалося досягти точності розпізнавання в 98,8% із використанням динамічного розміру кадру.

АУДІО, ЕМОЦІЇ, КАДР, МАШИННЕ НАВЧАННЯ, НЕЙРОННІ МЕРЕЖІ, РОЗПІЗНАВАННЯ, PYTHON, TENSORFLOW

D.S. Suvorov, I.V. Afanasieva, K.G. Onyshchenko, O.V. Kalynychenko. The effect of frame size on speech emotion recognition. Speech emotion recognition task, as well as most audio recognition machine learning tasks, uses the so-called framing. This is the process of dividing the original audio signal into frames of a certain size, each of which is processed separately. This article presents a comparison of the effect of frame size on the emotion recognition result using a CNN network as an example. For the experiments, the CREMA-D dataset was used with the augmentations using noise adding, time stretching, and pitch shifting. We managed to achieve a recognition accuracy of 98.8% using dynamic frame size.

AUDIO, EMOTIONS, FRAME, MACHINE LEARNING, NEURAL NETWORKS, RECOGNITION, PYTHON, TENSORFLOW

Вступ

З активним зростанням технологій штучного інтелекту, ці ж технології набувають широкого поширення на різні сфери життя людини. Однією з таких галузей є психологічний аналіз стану людини. Існують різні підходи до такого аналізу, проте найбільшого поширення наразі набули методи розпізнавання емоції за текстом, з використанням міміки та пози людини, а також за мовленням [1]. Саме розпізнавання за мовленням є темою поточного дослідження.

Подібний аналіз дає змогу за невеликим уривком запису мовлення людини визначити емоцію, з якою людина говорила. Подібний підхід може мати деякі переваги, пов'язані з мовним розмаїттям і віковою варіацією. Розроблена модель на основі однієї мови (наприклад, англійської) може бути всього лише донавчена з використанням додаткового набору даних іншої мови (наприклад, української). Однак, навіть без додаткового розширення вибірки, створена модель уже може працювати з різноманітними мовами (хоча й можуть бути певні винятки, пов'язані з культурними особливостями, ідеологією і просто специфічною говіркою).

Проте, подібного роду аналіз представляє досить перспективний підхід для різних систем, таких як розумні будинки або системи екстреного реагування. Але не варто забувати, що для повноцінного емоційного аналізу необхідно використовувати багатофакторний аналіз, який би містив кілька джерел, що давало б більш об'єктивну оцінку (візуальна інформація, інформація ЧСС і так далі).

У задачі розпізнавання емоції за мовленням важливим моментом у вилученні параметрів з аудіосигналу є поділ цього сигналу на фрагменти [2], кожен з яких окремо обробляється. Саме цей аспект попередньої обробки аудіо і буде детально розглянуто в статті, щоб отримати повне уявлення впливу розміру таких фрагментів на точність моделі.

1. Опис предметної галузі

Перш ніж перейти до опису вилучення параметрів з аудіо, для початку розглянемо звук загалом [3]. У нашому звичайному (аналоговому) світі, звук є безперервною хвилею (див. рис. 1). Однак, для обробки за допомогою ЕОМ необхідно аналоговий звук оцифрувати.

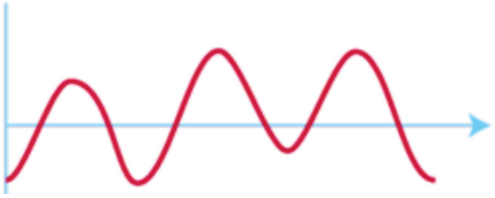

Рис. 1. Аналогова звукова хвиля

Це відбувається за допомогою АЦП — аналогово-цифрового перетворення. Тоді сигнал починає виглядати трохи інакше (див. рис. 2).

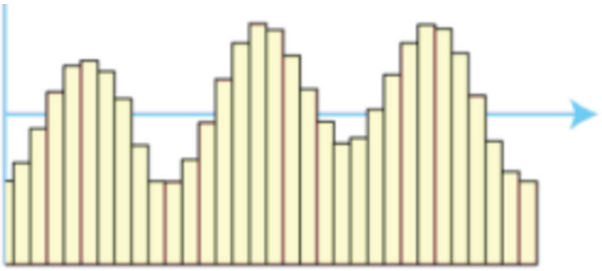

Рис. 2. Цифрова звукова хвиля

У чому ж полягає процес АЦП і чому сигнал стає ступінчастим (дискретним)? Завдяки записуючим пристроям та аналогово-цифровим перетворювачам звукова хвиля зчитується з певною частотою, вимірюваною в Герцах (Гц). Наприклад, 10 Гц означає, що перетворювач 10 разів на секунду зчитує звукову хвилю. Але для оцифрування звуку 10 Гц — це дуже мало. Мало тому, що дуже велика частина інформації буде загублена. Тому поширеними частотами є 22,050 Гц або 44,100 Гц. Така частота дає змогу перетворювати аналоговий сигнал у досить якісну цифрову версію. Частота 22,050 Гц часто використовується в машинному навчанні, оскільки дає змогу захопити достатньо деталей у звуці, водночас отримуючи файли відносно невеликого розміру.

Отже, уявімо, що ми зчитали сигнал із частотою 22,050 Гц. Тепер файл із такою частотою (її називають sample rate) лежить у сховищі комп'ютера. Тепер із ним можна працювати. По ходу опису методу обробки, розглянемо специфічну для звуку термінологію.

Ми вже говорили про таке поняття, як фреймінг — розбиття аудіосигналу на фрагменти. Ці фрагменти називаються кадрами. Або, по-іншому, оскільки в цифровому вигляді сигнал — це послідовність семплів, то кадр — це підпослідовність цих семплів. Насамперед необхідно зрозуміти, для чого застосовується цей фреймінг.

Звук — це непостійний сигнал. Однак багато методів аналізу сигналу (зокрема, за допомогою перетворення Фур'є) призначені для інтерпретації тільки постійних сигналів. Тому, щоб застосувати методи до звуку, ми працюємо з кадрами. Тривалість кадру вибирається залежно від уявлення про те, як швидко змінюється зміст сигналу. Передбачається, що в кадрі сигнал постійний, і тому ми можемо застосувати до нього подібного роду аналіз. Таким чином, фреймінг є обов'язковим інструментом при аналізі природних звукових сигналів.

Отже, перший крок в обробці сигналу ми розібрали — розбиття на кадри. Але й у цього процесу є особливість. А саме перекриття (див. рис. 3).

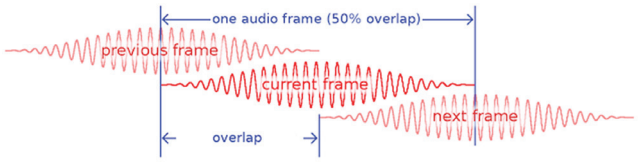

Рис. 3. Перекриття кадрів

Вигода перекриття проявляється на наступних етапах обробки. Але в загальних рисах перекриття дає змогу:

– зберігати залежність сигналу при переході від кадру до кадру

– зберігати дані сигналу після застосування windowing function

Якщо узагальнювати, то перекриття завжди використовується під час фреймінгу. Для розміру кадру і перекриття використовуються такі терміни як frame size і hop size (або frame length та hop length) (див. рис. 4).

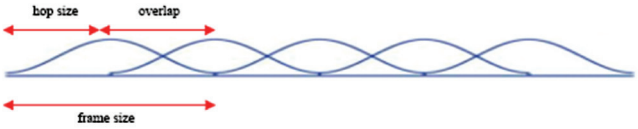

Рис. 4. Frame size, hop size та overlap

Якщо уявити, що ми здійснюємо фреймінг за допомогою ковзного вікна, то hop size — це крок зсуву ковзного вікна. Досить часто hop size роблять удвічі меншим за розмір кадру, що дає змогу зберегти достатньо інформації в обох кадрах.

Розглянемо згадану вище windowing function [4]. Зараз мимохідь згадаємо, що в цій роботі будуть використані виключно Мел-частотні кепстральні коефіцієнти. Це означає, що необхідно перевести сигнал у часово-частотну область, для чого застосовується перетворення Фур'є (а саме STFT). У цьому процесі може відбуватися так звана spectral leakage, для усунення якої і використовується windowing function.

Найпоширенішими такими функціями є Hamming і Hann (див. рис. 5 і 6).

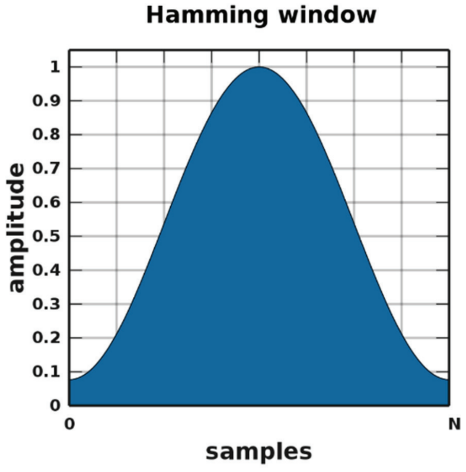

Рис. 5. Hamming windowing function

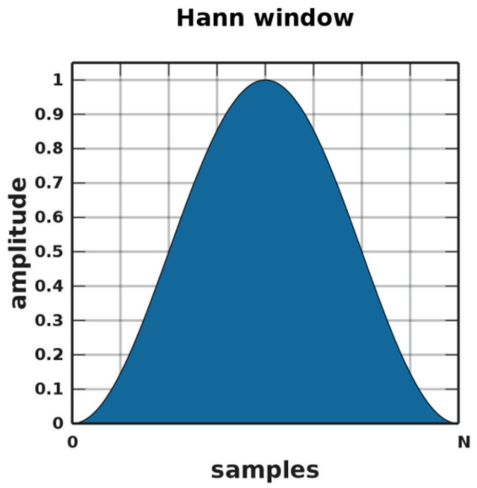

Рис. 6. Hann windowing function

Отже, чому ж без windowing function не вийде адекватного перетворення?

Більшість реальних аудіосигналів неперіодичні, тобто реальні аудіосигнали, як правило, не повторюються в точності протягом будь-якого заданого проміжку часу. Однак математика перетворення Фур'є припускає, що перетворюваний сигнал є періодичним.

Ця невідповідність між припущенням Фур'є про періодичність і реальним фактом, що аудіосигнали, як правило, неперіодичні, призводить до помилок у перетворенні, які й називаються spectral leakage та проявляються у вигляді неправильного розподілу енергії за спектром потужності сигналу.

Щоб дещо пом'якшити такі помилки в перетворенні можна попередньо помножити сигнал на windowing function, розроблену спеціально для цієї мети.

Після застосування, сигнал загасає на краях (див. рис. 7). Тут проявляється друга перевага фреймінгу. Через те, що після множення на windowing function дані сигналу на краях значно загублені, за рахунок перекриття під час аналізу всі дані вихідного сигналу будуть враховані.

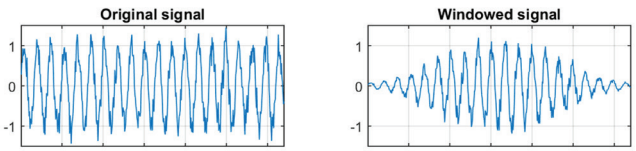

Рис. 7. Сигнал після застосування windowing function

На цьому етапі ми маємо для кожного аудіосигналу набір кадрів із перекриттям і застосованою windowing function. Саме час розглянути параметри аудіосигналу, оскільки наступним кроком буде їх вилучення.

Під час обробки аудіо виокремлюють кілька типів параметрів [5]. Кожен із типів отримують зі свого подання сигналу (часового, частотного або часово частотного).

Базовим поданням (тим, в якому спочатку представлений звуковий сигнал) є часове подання. Це те, як ми звикли бачити сигнал (див. рис. 8). Вісь x відповідає за час, а вісь y — за амплітуду сигналу. З такого подання можна отримати часові параметри, до яких належать amplitude envelope (максимальна амплітуда кадру), root mean square (середньоквадратичне значення амплітуди) тощо. І хоча подібні параметри дають деяке уявлення про характер сигналу, їх абсолютно недостатньо для повного аналізу аудіо. У всякому разі без додаткових параметрів.

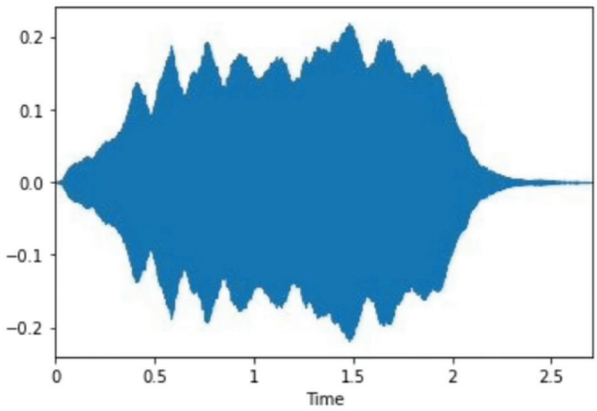

Рис. 8. Сигнал у часовому поданні

Наступне подання називається частотним. Аналогічно до часового подання, проте вісь x відповідає за частоту, а вісь y — за енергію частоти («кількість» частоти в сигналі) (див. рис. 9).

Тобто завдяки такому поданню можна проаналізувати частотний склад усього сигналу та отримати різні спектральні параметри: amplitude spectrum, spectral centroid, spectral bandwidth тощо. Ці параметри вже більш інформативні та можуть показувати набагато більше даних про сигнал, але є ще більш описове подання.

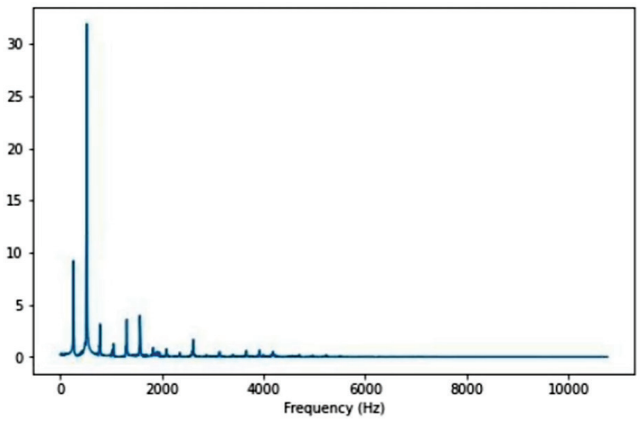

Рис. 9. Сигнал у частотному поданні

Часово-частотне подання має вигляд спектрограми (див. рис. 10). Це складніша структура, яка, однак, дає змогу отримати інформацію про частоту в конкретний проміжок часу. Іншими словами, це те саме частотне представлення, але для дуже маленьких

фрагментів аудіосигналу, а не всього сигналу загалом.

Саме таке подання є найбільш описовим і дає змогу отримувати найрепрезентативніші параметри аудіо, такі як спектрограма, Мел-спектрограми, MFCCs (Мел-частотні кепстральні коефіцієнти). Саме останній параметр найчастіше використовується в подібного роду обробці аудіо, зокрема в завданні розпізнавання емоції за мовленням. Результати з використанням цих коефіцієнтів є найвищими серед подібних досліджень.

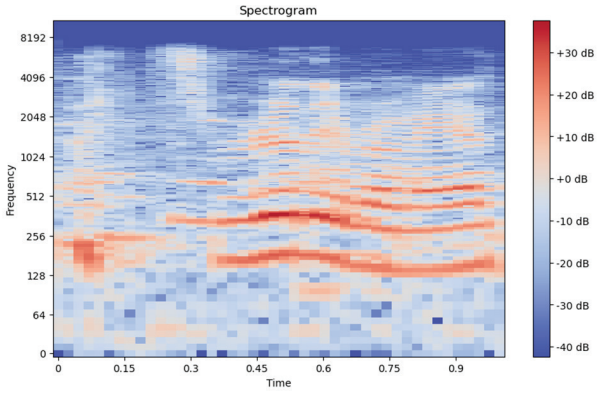

Рис. 10. Сигнал у часово-частотному поданні

Повернемося до того, що ми на даний момент зробили з сигналом. У нас є часове подання сигналу, розділене на кадри з перекриттям і застосованою windowing function. Щоб отримати MFCCs, необхідно кожен кадр перевести в часово-частотне подання. Для цього використовується вже згаданий раніше механізм під назвою перетворення Фур'є.

Перетворення Фур'є (FT) — це метод, який дає змогу розкласти складний сигнал на його базові частоти. В основі цього методу полягає ідея, що будь-який складний сигнал можна уявити як суму кількох простіших сигналів із різними частотами.

Простіше кажучи, можна уявити, що у нас є складний звук, наприклад, музична мелодія або голос. Перетворення Фур'є розбиває цей звук на його складові частини — від найнижчих до найвищих звукових частот, які присутні в цьому сигналі.

Є кілька варіацій цього перетворення: FT, FFT, DFT, STFT. Але всі вони необхідні для однієї цілі — розбити сигнал на частотні складові.

У нашому випадку, необхідний саме STFT (Short-Time Fourier Transform). Цей різновид перетворення дає змогу розбивати сигнал на частоти саме для невеликих фрагментів — кадрів.

Більш детально розглянемо перетворення Фур'є [6]. Можна виділити 3 етапи:

1. Для кожної частоти перетворення Фур'є обчислює комплексні експоненти, які є основними функціями синуса і косинуса

2. Комплексні експоненти множаться на значення вихідного сигналу. Це відбувається для кожної з розглянутих частот

3. Результати множення комплексних експонент на вихідний сигнал підсумовуються для кожної частоти. Це створює спектральні компоненти, що представляють амплітуди і фази кожної частоти у вихідному сигналі.

Таким чином на виході ми отримуємо в графічному поданні спектрограму (у разі застосування STFT) (див. рис. 10).

Перед тим, як отримати MFCCs є ще кілька кроків. Але перед цим розглянемо, що таке Мел і для чого ця шкала використовується в аудіообробці.

Шкала Мел використовується в аудіообробці для представлення частот в більш інтуїтивно зрозумілому вигляді. Вона ґрунтується на сприйнятті звукових частот людиною, відображаючи нелінійний спосіб сприйняття звуку. Тобто не всі частоти рівномірно розподілені. Людський слух більш чутливий до частот нижче 1000 Гц, ніж до частот вище. У шкалі Мел частотні діапазони, що відповідають низьким частотам, розтягнуті, а ті, що відповідають високим частотам, стиснуті. Це дає змогу краще враховувати особливості сприйняття звуку людиною.

Шкала Мел [7] була розроблена на основі досліджень психоакустики [8], яка вивчає сприйняття звуку людиною. Таким чином, вона краще відповідає реальному сприйняттю частоти, ніж лінійні шкали. І одними з найпопулярніших галузей, де використовується ця шкала, є аналіз аудіоданих.

Отже, щоб отримати необхідні MFCCs, потрібно, для початку, перетворити отриману після STFT спектрограму в Мел-спектрограму.

Перетворення частоти f у герцах на частоту m у Мелах представлено у формулі 1.

$$m = 2595 log_{10}\left(1 + \frac{f}{700}\right), \quad (1)$$

де f – частота в Герцах; m – частота в Мелах.

Частотний діапазон від мінімальної частоти $f_{\min}$ до максимальної частоти $f_{\max}$ ділиться на n Мел фільтрів. Ці фільтри розташовуються на шкалі Мел рівномірно.

Кожен Мел-фільтр являє собою трикутний фільтр, який охоплює певний діапазон частот. Центр фільтра відповідає центральній частоті на шкалі Мел, а його межі сходяться до нуля на сусідніх центральних частотах.

Припустимо, у нас є спектрограма $S(f, t)$, де f — частота, а t — час. Для кожного фільтра $H_i(f)$ і кожної часової точки t розраховується енергія у фільтрі шляхом підсумовування значень спектрограми, помножених на значення фільтра. Тоді за формулою 2 можна отримати матрицю i-го фільтру у певний час.

$$M(i,t) = \sum_{f} S(f,t) * H_i(f), \quad (2)$$

де M – матриця; i – номер фільтру; t – час; $S(f, t)$ – спектрограма; $H_i(f)$ – Мел-фільтр.

Для поліпшення сприйняття й аналізу до отриманих значень енергії застосовується логарифмічна шкала.

Таким чином, ми отримуємо Мел-спектрограму (див. рис. 11).

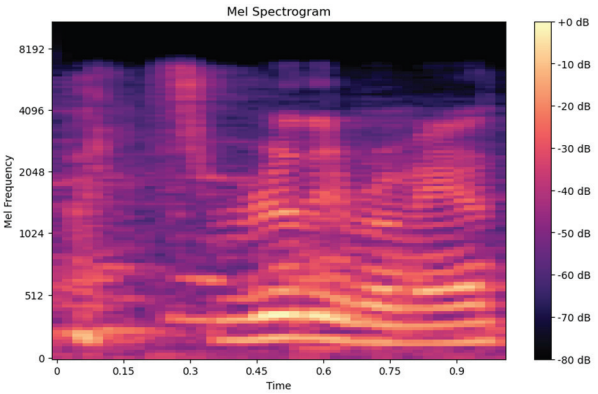

Рис. 11. Мел-спектрограма

Подальші кроки є фінальними для отримання Мел-частотних кепстральних коефіцієнтів. Ці коефіцієнти — це набір параметрів, які являють собою компактний та інформативний опис аудіосигналу, що відображає особливості людського сприйняття звуку. У графічній репрезентації MFCCs можна зобразити у вигляді спектрограми (див. рис. 12).

Отже, для отримання MFCCs необхідно до Мел-спектрограми застосувати дискретне косинусне перетворення (DCT). Це математичне перетворення, яке використовується для перетворення послідовності чисел у набір коефіцієнтів. Воно схоже на перетворення Фур'є, але замість використання комплексних експонентів використовує косинуси. У чому ж перевага такого підходу?

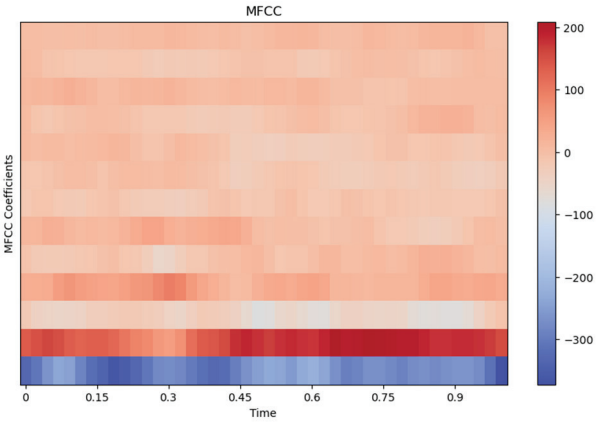

Рис. 12. Представлення MFCCs у вигляді спектрограми

DCT зменшує кореляцію між ознаками, що допомагає поліпшити продуктивність алгоритмів машинного навчання, які використовують ці ознаки.

Також, оскільки DCT концентрує інформацію в декількох коефіцієнтах, можна використовувати тільки перші кілька коефіцієнтів, відкидаючи інші без значної втрати інформації.

MFCCs, отримані з використанням DCT, менш схильні до шуму та викривлень, що робить їх надійнішими для завдань розпізнавання мови та звуків.

Щодо кількості коефіцієнтів MFCC, можна сказати таке. Невелика кількість коефіцієнтів дає змогу зменшити розмір даних і підвищити швидкість обробки таких даних. Однак, значна частина даних за невеликої кількості коефіцієнтів може бути загублена.

Велика ж кількість збільшує розмір даних, зменшує швидкість опрацювання, проте може містити набагато більше потрібної інформації. Таким чином варто досліджувати оптимальну кількість коефіцієнтів для конкретного завдання для отримання найкращих результатів.

Підбиваючи підсумки, ми описали всі основні етапи роботи з аудіо. Визначили необхідну термінологію, розглянули кроки отримання параметрів з аудіосигналу для подальшого їхнього опрацювання в нейронній мережі. Ми визначили роль фреймінгу та кадрів у процесі обробки і подальші експерименти спрямовані на вивчення впливу розміру кадрів, а також розміру зсуву/перекриття на якість моделей, які можна отримати.

2. Інструменти розробки

Для проведення експериментів з метою аналізу впливу розміру кадру на точність моделі нейронної мережі насамперед використовували бібліотеку librosa, що має великий функціонал для роботи з аудіо, зокрема:

– зчитування аудіо
– вилучення параметрів
– аугментація

Використовувалася мова програмування Python і середовище розробки JupyterLab. А конфігурація системи, на якій проводилося навчання моделей, така:

– NVIDIA RTX 3060 12GB
– Ryzen 5 3600X
– RAM 32GB 3200MHz

3. Експериментальні дослідження

Проведення експериментів почалося з вилучення параметрів аудіо, про які йшлося раніше, а саме MFCCs.

Як набір даних для задачі було обрано CREMA-D — великий набір аудіо- та візуальних даних для задачі розпізнавання емоції за аудіо [9]. Трохи деталей про цей набір:

– містить аудіозаписи понад 90 професійних акторів, кожен з яких зачитує 12 фраз на 4 рівнях емоційності

– містить записи 6 емоцій (злість, щастя, відраза, страх, смуток і нейтральний стан)

– має рівномірний розподіл емоцій у наборі та містить загалом 7,442 аудіозаписи

Цей набір є одним із найбільш репрезентативних серед усіх проаналізованих наборів знайдених у відкритому доступі для задачі розпізнавання емоції за мовленням. Цей набір балансує між достатньою кількістю даних і якістю цих даних. Для дослідження було взято 4 емоції (щастя, злість, смуток і нейтральний стан), чого буде достатньо для поставленої задачі.

Отже, перед безпосередньо проведенням експериментів було проведено підготовку, до якої входить кілька кроків:

1) вилучення MFCCs із записів
2) аугментація даних
3) створення архітектури моделі згорткової нейронної мережі

Вилучення звукових параметрів було проведено за схемою, описаною раніше. Однак, уже тут для кожного експерименту була своя особливість. Оскільки метою дослідження є вивчення впливу розміру кадру на точність моделі, а розмір кадру вказується вже на цьому етапі, було підготовлено такі набори параметрів:

– кадр 2048, зсув 1024 (перекриття 50%)
– кадр 2048, зсув 512 (перекриття 75%)
– кадр 1024, зсув 512 (перекриття 50%)
– кадр 1024, зсув 256 (перекриття 75%)
– кадр 512, зсув 256 (перекриття 50%)
– кадр 512, зсув 128 (перекриття 75%)
– кадр динамічний, кадр в 2 рази менший за розмір кадру (перекриття 50%)

Таким чином, було отримано 7 наборів параметрів (MFCCs). Далі будемо позначати кожен із цих наборів як набір із розміром кадру X/динамічним і перекриттям N%.

Одночасно з вилученням параметрів було проведено аугментацію для кожного набору параметрів.

Аугментація — це особливий спосіб розширення набору даних, який використовує вже наявні дані для створення нових шляхом застосування до цих даних спеціальних операцій.

Одними з найпоширеніших таких операцій (методів аугментації) для аудіоданих є:

– додавання шуму
– розтягнення і стиснення в часі
– зміна висоти тону

Метою даних експериментів не є вивчення впливу різних методів аугментації на якість одержуваних моделей, тому в експериментах цієї роботи використовувалися всі три методи. Тобто для кожного набору параметрів вибірку CREMA-D було розширено шляхом додавання шуму до кожного запису, прискорення та сповільнення запису, підвищення та зниження висоти тону.

Важливо зазначити, що через фіксований розмір кадру не кожен запис може цілком потрапити у фінальний набір. Наприклад, якщо наш аудіозапис складається з 1300 семплів, a frame size і hop size дорівнюють 256 і 128 відповідно, то тільки 10 кадрів можна вилучити з такого запису (128 * 10 = 1280 семплів). 20 семплів запису просто не потраплять у навчання. Подібне можна вирішити, «дорощуючи» необхідну кількість семплів для цілого кадру, наприклад, нулями. Тоді весь запис буде враховано в навчанні, однак також буде враховано і додані нулі, які не є частиною запису, що, теоретично, може негативно впливати на якість моделі.

Проте в цій роботі не було використано додавання нулів. Однак, був придуманий альтернативний спосіб врахування повного аудіозапису для навчання.

Оскільки моделі, які будуть використані, вимагають фіксований розмір даних на вхід, необхідно задовольнити цю умову. Якщо в першому підході обмеження виходять від розміру кадру і фіксованої тривалості аудіозапису, який можна отримати, то наступний підхід не вимагає встановлювати обмеження на тривалість запису, хоча і передбачає приблизно однакову тривалість. Таким чином було придумано підхід динамічного фреймінгу. У такому разі фіксованою стає кількість кадрів, яку необхідно вилучити з аудіозапису. У цьому дослідженні кількістю кадрів було встановлено 128. Це дало змогу отримувати кадри розміром від 500 семплів до 2000-3000 семплів, що є доволі адекватним розміром, співмірним із тим, що було використано у наборах з фіксованим розміром кадру.

Таким чином, набори параметрів із фіксованим розміром кадрів враховують записи не повністю, тоді як набори з динамічним розміром кадрів — повністю. Варто нагадати, що розмір кадру виходить з тієї думки, що впродовж усього кадру сигнал статичний і не змінюється. Це і буде перевірено.

Отже, після отримання наборів параметрів з урахуванням аугментації, було розроблено архітектуру моделі нейронної мережі на основі згорткових шарів [10] (див. рис. 13).

Модель містить усі основні компоненти базової згорткової мережі. Згорткові блоки (повторюваний набір шарів, де відбувається операція згортання) складаються з:

– згортковий шар (розмір фільтрів 5 на 5)
– шар нормалізації (дає змогу прискорити навчання)
– шар об'єднання (який виділяє найбільш значущі патерни характеристик у даних)
– шар відсіву (відключає деякі нейрони під час навчання, що перешкоджає перенавчанню нейронної мережі та сприяє підвищенню якості моделі)

Модель містить 6 таких згорткових блоків, після яких йде шар Flatten, для перетворення параметрів після згорткових шарів у вектор. Після якого

йде повнозв'язний шар мережі з активацією ReLU. Останнім шаром мережі є шар із 4 нейронів, кожен з яких відповідає одній з емоцій. Цей останній шар із функцією активації Softmax відповідає за безпосередньо визначення ймовірності класу аудіозапису [11].

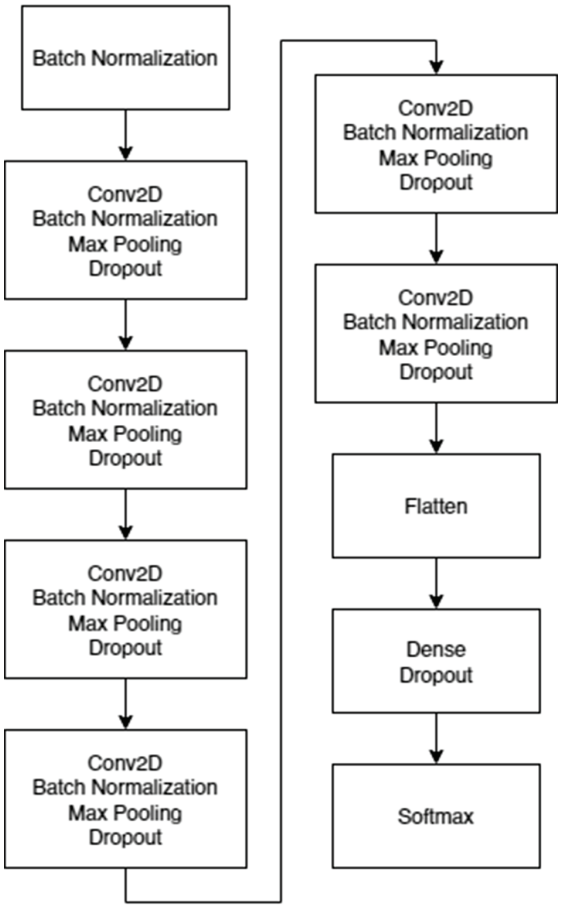

Рис. 13. Графічне зображення архітектури моделі нейронної мережі

Варто описати методику оцінювання моделей.

По-перше, для зняття метрик, які будуть описані нижче, використовувався підхід K-Fold Cross Validation. Цей метод оцінювання моделі користується широкою популярністю в машинному навчанні й дає змогу найбільш об'єктивно оцінити якість математичної моделі. Цей метод також менш чутливий до нерівномірного розподілу класів у наборі, якщо такий є.

Основна суть цього підходу оцінки полягає в тому, що набір даних розбивається на K частин. Далі протягом K ітерацій 1 з K частин виступає як тестова вибірка для моделі, а на решті K-1 частин виконується навчання. Таким чином усі дані використовуються для тестування, при цьому зберігається їхня незалежність.

На кожній із цих K ітерацій знімають метрики, які потім усереднюють і описують оцінку якості моделі на наборі даних.

У цьому дослідженні використовували популярні стандартні метрики під час оцінювання моделей нейронних мереж:

– accuracy
– precision
– recall
– f1-score

F1-score (або F-міра) є однією з найефективніших метрик, оскільки враховує і precision, і recall.

Отже, усі підготовчі етапи описано. Далі було проведено безпосередньо експерименти, які складаються з:

– навчання нейронної мережі з використанням методу K-Fold Cross Validation
– зняття та усереднення метрик

Таким чином, було проведено 7 експериментів для кожного набору параметрів. Для більш компактного відображення результатів зобразимо в таблиці тільки f1-score для всіх 7 наборів.

Таблиця 1 містить значення F-міри для чотирьох емоцій на яких проводилися експерименти.

Також до комірок таблиці застосоване умовне форматування, щоб наочно було видно кольором, як саме змінюється якість моделі в залежності від розміру кадру.

Варто також сказати, що, не дивлячись на рівномірність даних у наборі, після вилучення параметрів для тих утворених наборів параметрів із фіксованим розміром кадру розподіл екземплярів класів (емоцій) вже не такий рівномірний, оскільки не усі аудіозаписи змогли задовольнити умови розміру кадру та тривалості запису (аудіозапис повинен складати не менше 2 секунд).

Таблиця 1

Порівняння якості моделей розроблених при різному розмірі кадру

	frame 2048		**frame 1024**		**frame 512**		**dynamic frame**
overlap	**50%**	**75%**	**50%**	**75%**	**50%**	**75%**	**50%**
anger	0,958	0,956	0,954	0,962	0,953	0,960	0,995
happiness	0,922	0,918	0,913	0,930	0,908	0,922	0,989
sadness	0,934	0,943	0,933	0,945	0,925	0,940	0,986
neutral	0,911	0,921	0,906	0,924	0,899	0,916	0,981
avg	0,931	0,935	0,927	0,940	0,921	0,934	0,988

Тож, для фіксованих кадрів найбільша кількість екземплярів для тестування була для емоцій злості та суму, нейтральний стан йшов на останньому місці.

Отже, які висновки можна зробити із наданої таблиці?

По-перше, видно, що значення f-міри для фіксованого розміру кадру майже не відрізняються для усіх варіацій кадрів. Тобто для усіх них точність моделей тримається на рівні 92-94%.

По-друге, можна побачити, що збільшення перекриття (або зменшення hop size) дійсно позитивно впливає на якість моделей. Для усіх проведених експериментів якість моделей при перекритті на 75% вище за моделі із перекриттям 50%.

По-третє, якщо розглядати лише експерименти із фіксованим розміром кадру, то розмір кадру в 1024 семпли з перекриттям 75% демонструє найвищу якість моделі.

Розглянемо порівняльний графік отриманих результатів (див. рис. 14).

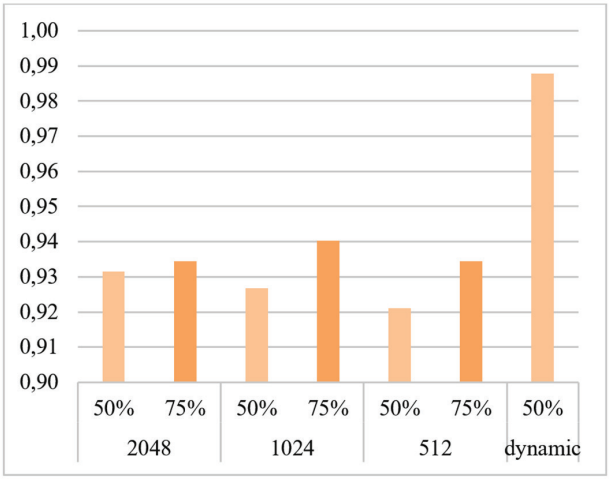

Рис. 14. Значення f1-score при різному розмірі кадру та різному перекритті кадрів

З діаграми чітко видна перевага більшого перекриття. Також видно, що динамічний розмір кадру відіграє значну роль і дозволяє підвищити якість отриманої моделі на 5%. Хоча це значення може здатися невеликим, із динамічним розміром ми отримали модель, яка пропонує точність розпізнавання майже 99%. Звісно, на чотирьох емоціях.

Висновки

У статті було проаналізовано різні механізми, які використовуються під час обробки звуку. Були розглянуті особливості та підходи попередньої обробки аудіосигналу для подальшого їхнього аналізу з використанням нейронної мережі. Як нейронну мережу використовували згорткову мережу, яку самостійно було спроектовано. Результати показали, що фіксований розмір кадру демонструє показники точності значно нижчі, ніж динамічний. При цьому різниця в точності між усіма протестованими (фіксованими) розмірами кадрів дуже незначна. Це може свідчити про те, що для завдання розпізнавання емоції за мовленням розмір кадру відіграє меншу роль для точності, на відміну від повного охоплення аудіосигналу для навчання.

Як подальші дослідження можуть бути проаналізовані інші типи звукових характеристик та їхній вплив на точність, а також вплив типів аугментації (їхній внесок) на загальну точність навченої моделі.

Список літератури

[1] What is speech emotion recognition? – klu. Design, Deploy, and Optimize LLM Apps with Klu – Klu.ai. URL: https://klu.ai/glossary/speech-emotion-recognition (дата звернення: 13.04.2024).

[2] Bevor Sie zu YouTube weitergehen. URL: https://www.youtube.com/@ValerioVelardoTheSoundofAI (дата звернення: 06.03.2024).

[3] Valerio Velardo — The Sound of AI. Understanding audio signals for machine learning, 2020. YouTube. URL: https://www.youtube.com/watch?v=daB9naGBVv4 (дата звернення: 21.03.2024).

[4] Windowing signals – telecommunication engineering. Telecommunication Engineering – My WordPress Blog. URL: https://telecommunicationengineering.softecks.in/535/ (дата звернення: 20.05.2024).

[5] Valerio Velardo — The Sound of AI. Types of audio features for machine learning, 2020. YouTube. URL: https://www.youtube.com/watch?v=ZZ9u1vUtcIA (дата звернення: 03.04.2024).

[6] Valerio Velardo — The Sound of AI. Short-Time fourier transform explained easily, 2020. YouTube. URL: https://www.youtube.com/watch?v=-Yxj3yfvY-4 (дата звернення: 20.05.2024).

[7] Mel. Simon Fraser University. URL: https://www.sfu.ca/sonic-studio-webdav/handbook/Mel.html (дата звернення: 27.04.2024).

[8] Minard A. Psychoacoustics: understanding the listening experience. Ansys Blog. URL: https://www.ansys.com/blog/understanding-psychoacoustics/ (дата звернення: 11.03.2024).

[9] GitHub — cheyneycomputerscience/crema-d: crowd sourced emotional multimodal actors dataset (CREMA-D). GitHub. URL: https://github.com/CheyneyComputerScience/CREMA-D (дата звернення: 17.05.2024).

[10] Basic CNN architecture: explaining 5 layers of convolutional neural network | upgrad blog. upGrad blog. URL: https://www.upgrad.com/blog/basic-cnn-architecture/ (дата звернення: 09.02.2024).

[11] Emotional speech recognition using deep neural networks. PubMed Central (PMC). URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8877219/ (дата звернення: 26.05.2024).

Надійшла до редколегії 28.08.2023

УДК 656::004.942::355.41 DOi 10.30837/ bi.2023.1(99).07

С. М. Неронов

ХНАДУ, м. Харків, Україна, sernikner@gmail.com, ORCID iD: 0000-0003-2381-1271

МОДЕЛІ ДОСЛІДЖЕННЯ ЛОГІСТИКИ ПЕРЕВЕЗЕНЬ У ПЕРІОД ВОЄННОГО СТАНУ

Ставиться та вирішується задача моделювання та планування логістичних процесів перевезень у період воєнного стану країни. Відокремлюються особливості транспортування вантажів та людей, які по'вязані з наявністю військових загроз (обстріли, прильоти ракет, атаки дронів, тощо). Актуальність теми дослідження обумовлена створенням оригінальних оптимізаційних моделей для планування перевезень в умовах особливого стану. Велику увагу приділено часу та ризикам перевезень в умовах військових загроз. Наукова новизна дослідження пов'язана зі створенням оригінальних оптимізаційних моделей, які дозволяють аналізувати та планувати логістику перевезень військових вантажів на передову та евакуацію населення до тилу.

ЛОГІСТИКА, ПЕРЕВЕЗЕННЯ, ВІЙСЬКОВА ЗАГРОЗА, ОПТИМІЗАЦІЙНА МОДЕЛЬ, ЛОГІСТИЧНІ ПОКАЗНИКИ

S. M. Neronov. Research models of transportation logistics in the period of martial law. The task of modeling and planning logistics processes of transportation during the period of the country's martial law is set and solved. The features of the transportation of goods and people are separated, which are connected with the presence of military threats (shelling, missile attacks, drone attacks, etc.). The relevance of the research topic is due to the creation of original optimization models for planning transportation in special conditions. Much attention is paid to the time and risks of transportation in conditions of military threats. The scientific novelty of the study is related to the creation of original optimization models that allow for the analysis and planning of the logistics of transporting military cargo to the front and evacuating the population to the rear.

LOGISTICS, TRANSPORTATION, MILITARY THREAT, OPTIMIZATION MODEL, LOGISTICS INDICATORS

Вступ

Воєнний стан країни змусив переглянути логістичні процеси перевезень [1 – 9]. З'явились нові напрямки в логістиці, які необхідно дослідити для ефективного планування перевезень в умовах військових загроз. Особливо важливими є напрямки логістики, які пов'язані з транспортуванням озброєння та військової техніки (ОВТ) на передову та перевезення (евакуація) населення до тилу з прифронтових районів. Тому, актуальна тема запропонованої публікації, в якій створюються оптимізаційні моделі для раціонального вибору шляхів перевезень в умовах воєнного стану. Метою дослідження є створення моделей для прикладної інформаційної технології дослідження логістичних процесів транспортування вантажів та людей у період воєнного стану країни. Завдання, які вирішуються у роботі:

– створення оптимізаційних моделей для планування перевезень ОВТ на передову;

– створення оптимізаційних моделей для планування евакуації населення до тилу.

1. Оптимізаційна модель планування перевезень озброєння та військової техніки на передову

Одним з актуальних завдань, яке пов'язане з проведенням ефективних оперативно-тактичних дій на полі бою, є формування необхідних запасів озброєння та військової техніки (ОВТ) на передову. Лінія фронту включає актуальні військові локальні зони (ВЛЗ), в яких проводяться активні бойові дії. Необхідно сформувати у ВЛЗ потрібні запаси ОВТ для проведення успішних оперативно-тактичних дій. Тому, актуально завдання пошуку відносно безпечних шляхів постачання ОВТ на передову в умовах військових загроз. Для вирішення поставленої задачі будемо використовувати метод цілочисельного (булевого) програмування. Введемо змінну x_{ijk}:

$$x_{ijk} = \begin{cases} 1, & \text{якщо обрано } j\text{-й шлях постачання ОВТ до } i\text{-ї ВЛЗ з } k\text{-м складом логістичних компонент (перевалка, складування, розподіл, тимчасова зупинка тощо);} \\ 0, & \text{в іншому випадку.} \end{cases} \quad (1)$$

При цьому необхідно щоб: $\sum_{j=1}^{n_i}\sum_{k=1}^{m_j} x_{ijk} = 1$, що означає обов'язковий вибір конкретного шляху постачання ОВТ до i-ї ВЛЗ з k-м складом логістичних компонент, де N – кількість ВЛЗ на лінії фронту; m_j – кількість можливих складів логістичних компонент на j-у шляху постачання; n_i – кількість можливих шляхів постачання військових вантажів до i-ї ВЛЗ.

Введемо основні логістичні показники для оцінки та вибору можливого варіанту транспортування ОВТ на передову:

1. R – ризики постачання ОВТ на передову, в умовах військових загроз.

2. T – час потрібний для постачання ОВТ на передову.

3. W – запаси озброєння, які формуються на передовій для виконання актуальних оперативно-тактичних завдань військового керівництва.

Представимо показники R, T, W з урахуванням змінних x_{ijk}:

$$R=\sum_{i=1}^{N}\sum_{j=1}^{n_i}\sum_{k=1}^{m_j} r_{ijk}x_{ijk}\,, \tag{2}$$

де r_{ijk} – ризик доставки військових вантажів до i-ї ВЛЗ по j-у шляху постачання з k-м складом логістичних компонент.

$$T=\sum_{i=1}^{N}\sum_{j=1}^{n_i}\sum_{k=1}^{m_j} t_{ijk}x_{ijk}\,, \tag{3}$$

де t_{ijk} – час, потрібний на транспортування військових вантажів в i-у ВЛЗ за j-м шляхом з k-м можливим складом логістичних компонент.

$$W=\sum_{i=1}^{N}\sum_{j=1}^{n_i}\sum_{k=1}^{m_j} w_{ijk}x_{ijk}\,, \tag{4}$$

де w_{ijk} – кількість партій ОВТ, які можна пересовувати за j-м можливим шляхом постачання з k-м складом логістичних компонент до i-ї ВЛЗ.

Сформуємо оптимізаційні моделі для рішення завдання формування запасів ОВТ на передовій для проведення ефективних бойових дій на полі бою.

1. Мінімізація ризиків формування запасів ОВТ в умовах дій військових загроз:

$$\min R,\ R=\sum_{i=1}^{N}\sum_{j=1}^{n_i}\sum_{k=1}^{m_j} r_{ijk}x_{ijk}\,, \tag{5}$$

при виконанні обмежень:

$$T\le T^*,\ T=\sum_{i=1}^{N}\sum_{j=1}^{n_i}\sum_{k=1}^{m_j} t_{ijk}x_{ijk}\,, \tag{6}$$

де T^* – допустимий (запланований) час постачання ОВТ на передову.

$$W\ge W^*,\ W=\sum_{i=1}^{N}\sum_{j=1}^{n_i}\sum_{k=1}^{m_j} w_{ijk}x_{ijk}\,, \tag{7}$$

де W^* – запас ОВТ, який необхідно сформувати для виконання актуальних оперативно-тактичних завдань військового керівництва.

2. Максимізація запасів ОВТ на передовій для проведення успішних бойових дій:

$$\max W,\ W=\sum_{i=1}^{N}\sum_{j=1}^{n_i}\sum_{k=1}^{m_j} w_{ijk}x_{ijk}\,, \tag{8}$$

при виконанні обмежень:

$$R\le R^*,\ R=\sum_{i=1}^{N}\sum_{j=1}^{n_i}\sum_{k=1}^{m_j} r_{ijk}x_{ijk}\,, \tag{9}$$

де R^* – допустимий ризик постачання ОВТ в умовах дій військових загроз.

$$T\le T^*,\ T=\sum_{i=1}^{N}\sum_{j=1}^{n_i}\sum_{k=1}^{m_j} t_{ijk}x_{ijk}\,. \tag{10}$$

Можлива багатокритеріальна постановка оптимізаційної задачі з використанням показників R, T, W.

У цьому випадку необхідно сформувати комплексний показник:

$$K=\alpha_R\overset{\vee}{R}+\alpha_T\overset{\vee}{T}+\alpha_W\overset{\vee}{W}\,, \tag{11}$$

де α_R, α_T, α_W – «ваги» показників R, T, W, $\alpha_R+\alpha_T+\alpha_W=1$.

$\overset{\vee}{R}$ – пронормований показник R:

$$\overset{\vee}{R}=\alpha_R\frac{R-R_{min}}{R^*-R_{min}}\,, \tag{12}$$

де $R_{\min}$ – мінімальне значення показника R після його оптимізації.

$\overset{\vee}{T}$ – пронормований показник часу постачання:

$$\overset{\vee}{T}=\alpha_T\frac{T-T_{min}}{T^*-T_{min}}\,, \tag{13}$$

де $T_{\min}$ – мінімальне значення часу T після його оптимізації.

$\overset{\vee}{W}$ – пронормований показник W:

$$\overset{\vee}{W}=\alpha_W\frac{W_{max}}{W^*_{max}}\,, \tag{14}$$

де $W_{\max}$ – максимальне значення запасу ОВТ після його оптимізації.

Необхідно знайти:

$$\begin{aligned}
&\min K=\alpha_R\overset{\vee}{R}+\alpha_T\overset{\vee}{T}+\alpha_W\overset{\vee}{W}=\\
&=\alpha_R\frac{R-R_{\min}}{R^*-R_{\min}}+\alpha_T\frac{T-T_{\min}}{T^*-T_{\min}}+\alpha_W\frac{W_{\max}-W}{W_{\max}-W^*}=\\
&=\frac{\alpha_R}{R^*-R_{\min}}\sum_{i=1}^{N}\sum_{j=1}^{n_i}\sum_{k=1}^{m_j} r_{ijk}x_{ijk}+\\
&+\frac{\alpha_T}{T^*-T_{\min}}\sum_{i=1}^{N}\sum_{j=1}^{n_i}\sum_{k=1}^{m_j} t_{ijk}x_{ijk}-\\
&-\frac{\alpha_W}{W_{\max}-W^*}\sum_{i=1}^{N}\sum_{j=1}^{n_i}\sum_{k=1}^{m_j} w_{ijk}x_{ijk}-\\
&-\frac{\alpha_R R_{\min}}{R^*-R_{\min}}-\frac{\alpha_T T_{\min}}{T^*-T_{\min}}+\frac{\alpha_W W_{\max}}{W_{\max}-W^*}.
\end{aligned} \tag{15}$$

2. Оптимізаційна модель планування перевезень озброєння та військової техніки на передову

Сучасна війна призвела до евакуації населення з прифронтової зони до тилу. Виникли міграційні процеси, для яких необхідно створювати логістичні ланцюги евакуації. Тому, актуальне дослідження евакуаційних потоків, для оцінки здатності транспортної мережі, та виконувати заплановані перевезення людей у тимчасові місця проживання (ТП). При плануванні процесів евакуації, необхідно сформувати множину місць (M), які здатні приймати населення, з їх можливостями щодо забезпечення соціальних потреб. Далі, необхідно сформувати шляхи перевезення людей, в умовах ризиків (R) військових загроз, оцінити вартість (W) та спланувати час (T) евакуації. Сформуємо

оптимізаційну модель, за допомогою якої можна визначити раціональні шляхи евакуації населення (F) з прифронтової зони до можливих місць тимчасового проживання в умовах воєнного стану країни. Введемо цілочисельну (булеву) змінну x_{plk}:

$$x_{plk} = \begin{cases} 1, \text{ якщо буде проведено перевезення людей до } p\text{-го місця проживання за допомогою } l\text{-го шляху транспортування з } k\text{-м складом логістичних компонент (тимчасова зупинка, перехід з одного шляху на інший, розподіл потоків евакуації, тощо);} \\ 0, \text{ в іншому випадку} \end{cases} \quad (16)$$

В якості основних логістичних показників евакуаційного процесу, будемо розглядати:

1. Час, потрібний на евакуацію людей (T).
2. Вартість процесу евакуації населення (W).
3. Ризики військових загроз (R).
4. Кількість населення, яке буде евакуйовано (F).

З урахуванням змінних x_{plk}, логістичні показники евакуації населення мають вигляд:

$$T = \sum_{p=1}^{M}\sum_{l=1}^{m_p}\sum_{k=1}^{n_l} t_{plk} x_{plk}\ , \quad (17)$$

де m_p – кількість можливих шляхів евакуації населення до p-го місця ТП; n_l – кількість можливих складів логістичних компонент для їх використання на l-у шляху перевезень; t_{plk} – час, потрібний для переміщення людей до p-го місця ТП з урахуванням l-го обраного шляху евакуації та k-го складу логістичних компонент.

$$W = \sum_{p=1}^{M}\sum_{l=1}^{m_p}\sum_{k=1}^{n_l} w_{plk} x_{plk}\ , \quad (18)$$

де w_{plk} – вартість перевезення людей до p-го можливого місця ТП з урахуванням обраного l-го шляху перевезення та k-го складу логістичних компонент.

$$R = \sum_{p=1}^{M}\sum_{l=1}^{m_p}\sum_{k=1}^{n_l} r_{plk} x_{plk}\ , \quad (19)$$

де r_{plk} – ризик перевезення людей, в умовах військових загроз, в p-е можливе місце ТП з урахуванням обраного l-го шляху транспортування та k-го складу логістичних компонент.

$$F = \sum_{p=1}^{M}\sum_{l=1}^{m_p}\sum_{k=1}^{n_l} f_{plk} x_{plk}\ , \quad (20)$$

де f_{plk} – кількість населення, яке буде спрямоване в p-е місце ТП з урахуванням обраного l-го шляху транспортування та k-го складу логістичних компонент.

Можливі такі постановки оптимізаційної задачі, які пов'язані з евакуацією населення до тилу:

1. Мінімізувати час, потрібний на евакуацію населення:

$$\min\ T,\ T = \sum_{p=1}^{M}\sum_{l=1}^{m_p}\sum_{k=1}^{n_l} t_{plk} x_{plk}\ , \quad (21)$$

з урахуванням обмежень:

$$W \le W^*,\ W = \sum_{p=1}^{M}\sum_{l=1}^{m_p}\sum_{k=1}^{n_l} w_{plk} x_{plk}\ , \quad (22)$$

$$R \le R^*,\ R = \sum_{p=1}^{M}\sum_{l=1}^{m_p}\sum_{k=1}^{n_l} r_{plk} x_{plk}\ , \quad (23)$$

$$F \ge F^*,\ F = \sum_{p=1}^{M}\sum_{l=1}^{m_p}\sum_{k=1}^{n_l} f_{plk} x_{plk}\ , \quad (24)$$

де W^* – запланована вартість процесу евакуації населення; R^* – допустимий ризик процесу евакуації, який пов'язаний з можливими діями військових загрозо; F^* – запланована кількість населення, яка буде евакуйована з прифронтової зони до тилу.

2. Максимізувати кількість населення, яке буде евакуйоване з прифронтової зони до тилу:

$$\max\ F,\ F = \sum_{p=1}^{M}\sum_{l=1}^{m_p}\sum_{k=1}^{n_l} f_{plk} x_{plk}\ , \quad (25)$$

з урахуванням обмежень:

$$T \le T^*,\ T = \sum_{p=1}^{M}\sum_{l=1}^{m_p}\sum_{k=1}^{n_l} t_{plk} x_{plk}\ , \quad (26)$$

$$W \le W^*,\ W = \sum_{p=1}^{M}\sum_{l=1}^{m_p}\sum_{k=1}^{n_l} w_{plk} x_{plk}\ , \quad (27)$$

$$R \le R^*,\ R = \sum_{p=1}^{M}\sum_{l=1}^{m_p}\sum_{k=1}^{n_l} r_{plk} x_{plk}\ , \quad (28)$$

де T^* – запланований час на евакуацію населення.

Мінімізувати ризики евакуації населення:

$$\min\ R,\ R = \sum_{p=1}^{M}\sum_{l=1}^{m_p}\sum_{k=1}^{n_l} r_{plk} x_{plk}\ , \quad (29)$$

з урахуванням обмежень:

$$T \le T^*,\ T = \sum_{p=1}^{M}\sum_{l=1}^{m_p}\sum_{k=1}^{n_l} t_{plk} x_{plk}\ , \quad (30)$$

$$W \le W^*,\ W = \sum_{p=1}^{M}\sum_{l=1}^{m_p}\sum_{k=1}^{n_l} w_{plk} x_{plk}\ , \quad (31)$$

$$F \ge F^*,\ F = \sum_{p=1}^{M}\sum_{l=1}^{m_p}\sum_{k=1}^{n_l} f_{plk} x_{plk}\ . \quad (32)$$

Можлива багатокритеріальна постановка оптимізаційної задачі евакуації населення. Для цього введемо комплексний критерій у вигляді адитивного складу логістичних показників T, W, R, F:

$$Q = \alpha_T \overset{\vee}{T} + \alpha_W \overset{\vee}{W} + \alpha_R \overset{\vee}{R} + \alpha_F \overset{\vee}{F}\ , \quad (33)$$

де α_T, α_W, α_R, α_F – «ваги» показників T, W, R, F, $\alpha_T+\alpha_W+\alpha_R+\alpha_F=1$; $\overset{\vee}{T}, \overset{\vee}{W}, \overset{\vee}{R}, \overset{\vee}{F}$ – пронормовані значення показників T, W, R, F:

$$\overset{\vee}{T} = \frac{T - T_{min}}{T^* - T_{min}}\ , \quad (34)$$

$$\overset{\vee}{W} = \frac{W - W_{min}}{W^* - W_{min}}\ , \quad (35)$$

$$\overset{\vee}{R} = \frac{R - R_{min}}{R^* - R_{min}}, \quad (36)$$

$$\overset{\vee}{F} = \frac{F_{max}}{F^*_{max}}. \quad (37)$$

Необхідно мінімізувати комплексний критерій Q:

$$\min Q, Q = \alpha_T \overset{\vee}{T} + \alpha_W \overset{\vee}{W} + \alpha_R \overset{\vee}{R} + \alpha_F \overset{\vee}{F} = \frac{\alpha_T}{T^* - T_{min} \sum_{p=1}^{M} \sum_{l=1}^{m_p} \sum_{k=1}^{n_l} t_{plk} x_{plk}} + \frac{\alpha_W}{W^* - W_{min} \sum_{p=1}^{M} \sum_{l=1}^{m_p} \sum_{k=1}^{n_l} w_{plk} x_{plk}} + \frac{\alpha_R}{R^* - R_{min} \sum_{p=1}^{M} \sum_{l=1}^{m_p} \sum_{k=1}^{n_l} r_{plk} x_{plk}} - \frac{\alpha_F}{F^*_{max} \sum_{p=1}^{M} \sum_{l=1}^{m_p} \sum_{k=1}^{n_l} f_{plk} x_{plk}} - \frac{\alpha_T T_{min}}{T^* - T_{min} - \frac{\alpha_W W_{min}}{W^* - W_{min} - \frac{\alpha_R R_{min}}{R^* - R_{min} + \frac{\alpha_F F_{max}}{F^*_{max}}}}}, \quad (38)$$

де $T_{\min}$, $W_{\min}$, $R_{\min}$, $F_{\max}$ – екстремальні значення показників після їх оптимізації.

Висновки

У роботі проведено дослідження логістичних процесів перевезень у період воєнного стану країни. Відокремлені актуальні напрямки дослідження, які пов'язані з транспортуванням військових вантажів на передову, а також евакуації населення з прифронтової зони до тилу. Сформовані основні логістичні показники, які необхідно використовувати для оцінки процесів перевезень в умовах дій військових загроз (час перевезень, ризики перевезень, вартість перевезень, кількість населення, яке евакуюється). Створені оптимізаційні моделі для вибору раціональних шляхів перевезень на передову та до тилу. Проведена локальна оптимізація логістичних показників, з урахуванням обмежень за допустимим часом та ризиком перевезень. Створені багатокритеріальні моделі для пошуку компромісних рішень логістики перевезень.

Використані математичні методи та моделі: системний аналіз, методи транспортної логістики, цілочисельна (булева) оптимізація, багатокритеріальна оптимізація, методи експертного оцінювання.

Наукова новизна дослідження пов'язана зі створенням комплексу оригінальних оптимізаційних моделей, за допомогою яких можна аналізувати та планувати логістику перевезень озброєння та військової техніки на передову та евакуацію населення до тилу.

Запропонований підхід є основою для створення прикладної інформаційної технології планування логістики перевезень як на передову, так і до тилу, з урахуванням можливих військових загроз, у період воєнного стану країни.

Список літератури:

[1] Федорович О. Є., Западня К. О., Іванов М. В. Використання прецедентного підходу для формування плану заходів щодо підвищення конкурентоспроможності підприємства, що розвивається / О. Є. Федорович, К. О. Западня, М. В. Іванов // Радіоелектронні і комп'ютерні системи. – 2016. – № 1 (75). – С. 114–118.

[2] Федорович О. Є., Прончаков Ю. Л. Метод формування логістичних транспортних взаємодій для нового портфелю замовлень розподіленого віртуального виробництва / О. Є. Федорович, Ю. Л. Прончаков // Радіоелектронні і комп'ютерні системи. – 2020. – № 2 (94). – С. 102–108

[3] Федорович О. Є., Сломчинський О. В.,. Пуйденко В. А. Дослідження логістики управління виробництвом високотехнологічної продукції віртуального підприємства / О. Є. Федорович, О. В. Сломчинський, В. А. Пуйденко // Авіаційно-космічна техніка і технологія. – 2018. – № 4 (148). – С. 107–115.

[4] Федорович О. Є., Гайденко О. А., Пуйденко В. А. Планування вантажних перевезень в уммовах підвищених ризиків / О. Є. Федорович, О. А. Гайденко, В. А. Пуйденко // Авіаційно-космічна техніка і технологія. – 2017. – № 6 (141). – С. 98–102.

[5] Федорович О. Є., Уруський О. С., Лутай Л. М., Западня К. О. Оптимізація життєвого циклу створення нової техніки в умовах конкуренції та стохастичної поведінки ринку збуту високотехнологічної продукції / О. Є. Федорович, О. С. Уруський, Л. М. Лутай, К. О. Западня // Авіаційно-космічна техніка і технологія. – 2020. – № 6 (166). – С. 80–85.

[6] Алексієв О. П., Алексієв В. О.,. Неронов С. М Мультиагенти у віртуальному управлінні транспортним процесом / О. П. Алексієв, В. О. Алексієв, С. М. Неронов // Вісник Харківського національного автомобільно-дорожнього університету – 2023. – Вип. 100. – С. 15–18.

Надійшла до редколегії 14.09.2023

УДК 656::004.942 DOi 10.30837/ bi.2023.1(99).08

Л. М. Козачок[1], С. М. Неронов[2], Г. А. Плєхова[3], М. В. Костікова[4], К. В. Плеша[5]

[1]ХНАДУ, м. Харків, Україна, LarisaK2010@ukr.net, ORCID iD: 0000-0002-5246-4240

[2]ХНАДУ, м. Харків, Україна, sernikner@gmail.com, ORCID iD: 0000-0003-2381-1271

[3]ХНАДУ, м. Харків, Україна, plehovaanna11@gmail.com, ORCID iD: 0000-0002-6912-6520

[4]ХНАДУ, м. Харків, Україна, kmv_topaz@ukr.net, ORCID iD: 0000-0001-5197-7389

[5]kirplesha@gmail.com, ORCID iD: 0000-0002-9908-558X

МАТЕМАТИЧНЕ МОДЕЛЮВАННЯ ТА ДОСЛІДЖЕННЯ ТРАНСПОРТНИХ ПОТОКІВ ТА ПРОЦЕСІВ ТРАНСПОРТНИХ СИСТЕМ МІСТ

Ставиться та вирішується задача математичного моделювання та дослідження транспортних потоків та процесів транспортних систем міст. Розглядаються зв'язані роботи, методи, моделювання інтенсивності транспортного потоку як процесу авторегресії першого порядку, знаходження оцінок параметрів моделі процесу зі стандартною нормальною випадковою складовою.

МОДЕЛЮВАННЯ, ТРАНСПОРТНА МЕРЕЖА, ТРАНСПОРТНИЙ ПОТІК, АВТОРЕГРЕСІЯ, ПАРАМЕТР, ДЕЦЕНТРАЛІЗОВАНІ СХОВИЩА ДАННИХ,СТИСНЕННЯ ЗОБРАЖЕНЬ

The problem of mathematical modeling and research of transport flows and processes of transport systems of cities is set and solved. Related works, methods, modeling of traffic flow intensity as a first-order autoregression process, finding estimates of process model parameters with a standard normal random component are considered.

MODELING, TRANSPORT NETWORK, TRANSPORT FLOW, AUTOREGRESSION, PARAMETER, DECENTRALIZED DATA STORAGE, IMAGE COMPRESSION

Вступ

Транспорт – це переміщення людей та товарів з одного місця в інше з використанням різних транспортних засобів у різних інфраструктурних системах.

Якщо сільське господарство та промисловість вважаються тілом країни, можна сказати, що транспорт може бути нервами та венами економіки.

Збільшення кількості транспортних засобів за рахунок збільшення доступності більшої кількості пунктів призначення дозволяє краще досягати виконання економічних цілей та інтересів людей, але тягне за собою нові витрати як на індивідуальному, так і на соціальному рівні. Необхідно керувати розвитком та роботою транспорту, створювати транспортні системи, планувати їх функціонування. Таким чином, змінюючи доступність, транспорт надає форму розвитку територій.

Вхідні, вихідні дані та кінцеві результати транспортування для моделей дослідження транспортних перевезень представляють традиційні ресурси (інфраструктура), робоча сила, необхідна для виробництва та обслуговування транспортних засобів, земля, що споживається інфраструктурою, витрати енергії та інше. Проектування систем, що дозволяють організувати транспортні перевезення, як пасажирські, так і вантажні, знизити чи мінімізувати споживання енергії, вимагає мислення, що виходить поза традиційних дисциплінарних кордонів. Необхідний аналіз процесів, що відбуваються в транспортних системах, моделювання окремих частин функціонування транспорту щодо аналізу різних величин, що характеризують транспортні перевезення та транспортні мережі, з використанням яких транспортні перевезення здійснюються. А також безліч завдань вимагають вирішення з використанням побудованих моделей та написанням алгоритмів здійснення процесів з транспортних перевезень, планування та управління роботою транспорту. Зважаючи на використання великих обсягів даних для вирішення поставлених завдань застосовується комп'ютерна техніка, а також розроблені комп'ютерні системи та технології.

Транспортна мережа – мережа магістральних вулиць та доріг, оснащених лініями громадського транспорту.

1. Зв'язані роботи

Якість функціонування будь-якої системи визначається кількісними показниками ефективності її роботи. Розроблені відповідні методи визначення різних показників, які характеризують процеси, що відбуваються у системах, це дає змогу знаходити найефективнішу організацію цих процесів.

При вирішенні завдань дослідження потоку автотранспортних засобів враховується два випадкові фактори – час надходження автомобіля до транспортної мережі, як до системи та час обслуговування – перебування у мережі. Важливою характеристикою роботи системи, що моделюється для дослідження, є інтенсивність потоку подій, тобто інтенсивність надходження подій до системи, інтенсивність вхідного потоку, яка лише в окремих випадках розглядається

як постійна величина, а в основному змінюється з часом. Саме ця характеристика і досліджується у цій роботі.

Велика кількість процесів, що спостерігаються або досліджуються, добре описуються моделлю авторегресії, тобто моделлю, у якій стан динамічної системи у даний момент часу лінійно залежить від попередніх станів цієї ж динамічної системи. Таким чином, авторегресійна модель – це модель часового ряду, рівнями якого є значення деякої досліджуваної величини або характеристики, яка використовується для аналізу – виявлення тенденції, періодичності, сезонності та інших особливостей значень та для прогнозування змін величини чи характеристики. Часто часові ряди, які погано описуються авторегресійною моделлю, при певній модифікації значень часового ряду (наприклад, логарифмування, потенціювання або розгляд приростів часового ряду деякого порядку) значно краще накриваються цією моделлю [1, 2].

Однак класична авторегресійна модель не враховує періодичність або сезонність повною мірою, тому що її параметри постійні і не змінюються з часом, від цього її застосування обмежене. Цього недоліку позбавлена загальніша модель авторегресійного типу [3]. Йдеться про моделі періодичної авторегресії, параметри якої змінюються у часі на кожному кроці, але повторюються через певний період. Подібні моделі можуть застосовуватись в аналізі різних часових рядів, у яких спостерігається виражена сезонність.

У моделі стан процесу лінійно залежить від свого попереднього стану. При цьому значення параметрів з наступним кроком змінюються, але повторюються з певною періодичністю. Тобто модель є лінійною, і вона може враховувати періодичність часового ряду. Також у модель додана випадкова складова, що є випадковими величинами, розподіленими за певним законом.

2. Методи

Дорожній трафік можна розглядати як потік автомашин, автотранспортних засобів, що приймають участь у дорожньому русі та заїжджають у транспортну мережу, що розглядається, і рухаються ділянками транспортної мережі. Потоком подій називається послідовність подій, які відбуваються одна за одною у випадкові моменти часу. Потік подій можна зобразити у вигляді послідовності випадкових точок t_1, t_2, ..., t_n, ... на вісі *Ot*, де n – кількість подій у потоці, а точки відображають моменти часу настання подій, в наших дослідженнях це моменти часу заїзду автомобілів на ділянки доріг, що входять у транспортну мережу, що розглядається. Випадкові інтервали часу між цими точками визначаються за формулами: $T_1 = t_2 - t_1$, $T_2 = t_3 - t_2$, ..., $T_{n-1} = t_n - t_{n-1}$,... .

Нехай випадкова величина $X(t,\Delta t)$ – число подій, що з'явилися на проміжку $(t, t+\Delta t)$, у якості подій розглядається заїзд автомобіля у транспортну мережу, а для чисельного відображення фіксується момент часу цього заїзду. Якщо обчислювати границю відношення математичного очікування цієї випадкової величини до довжини інтервалу часу, що розглядається, при прямуванні цієї довжини до нуля і якщо ця границя існує, то вона називається інтенсивністю ординарного потоку подій у момент t, в наших дослідженнях це інтенсивність потоку автомобілей у транспортній мережі:

$$\lambda(t) = \lim_{\Delta t \to 0} \frac{M\left(X(t,\Delta t)\right)}{\Delta t}. \qquad (1)$$

Таким чином, інтенсивністю потоку називається середнє число подій, які відбуваються за одиницю часу.

3. Моделювання інтенсивності транспортного потоку як процесу авторегресії першого порядку

Розглядається модель знаходження значень рівнів часового ряду інтенсивності транспортного потоку на певному маршруті та на певній трасі вуличної дорожньої мережі міста.

Якщо стан моделі, а тобто і значення часового ряду інтенсивності транспортного потоку можна записати у вигляді лінійної функціональної залежності наступного значення від попереднього, параметри якої теж залежать від часу і при цьому змінюються з певною періодичністю, то ми матимемо авторегресію першого порядку.

Розглянемо один із самих зручних у застосуванні варіантів моделі: порядок моделі дорівнює одному, період моделі дорівнює двом.

Тоді модель має вигляд:

$$\lambda_t = a(t)\lambda_{t-1} + \varepsilon_t^{\infty}, \qquad (2)$$

де ε_t^{∞} – послідовність незалежних випадкових величин таких, що математичне очікування дорівнює 0, дисперсія позитивна.

Параметр $a(t)$ періодичний із періодом $T = 2$. Тобто, $\forall k > 0 : a(t) = a(t + kT)$.

$\sigma^2(t)$ так само періодичний із періодом T. Види розподілів ε_t та ε_{t+kT} однакові.

Запишемо значення рівнів інтенсивності у вигляді більш придатного для знаходження значень

$$\lambda_t = a(t)\lambda_{t-1} + \varepsilon_t. \qquad (3)$$

Введемо вектори:

$$\Lambda_t = \begin{pmatrix} \lambda_{t-1} \\ \lambda_t \end{pmatrix}, \; E_t = \begin{pmatrix} 0 \\ \varepsilon_t \end{pmatrix}. \qquad (4)$$

Побудуємо матрицю $A(t)$:

$$A(t) = \begin{pmatrix} 0 & 1 \\ 0 & a(t) \end{pmatrix}. \qquad (5)$$

Отже,

$$\Lambda_t = A(t)\Lambda_{t-1} + E_t,\ \Lambda_t = A(t)A(t-1)\Lambda_{t-2} + A(t)E_{t-1} + E_t,$$

$$\Lambda_t = \begin{pmatrix} \lambda_{t-1} \\ \lambda_t \end{pmatrix}, E_t = \begin{pmatrix} 0 \\ \varepsilon_t \end{pmatrix}. \quad (6)$$

Розглянемо метод послідовного статистичного аналізу при оцінці параметрів моделі авторегресії: спочатку для випадку одиничної дисперсії, потім для випадку зі змінною дисперсією.

Значення параметрів моделі змінюються з кожним кроком процесу, що розглядається, при цьому в якості наступних кроків беруться наступні моменти, значення параметрів потім повторюються з певною періодичністю.

Особливістю таких оцінок є мінімізація обсягу вибірки за гарантованої якості оцінювання. Але на початку будуються непослідовні оцінки шляхом найменших квадратів. Особлива увага приділяється розподілу послідовних оцінок параметрів моделі.

4. Знаходження оцінок параметрів моделі процесу зі стандартною нормальною випадковою складовою

Розглянемо модель періодичної авторегресії першого порядку з періодом два:

$$\lambda_t = a(t)\lambda_{t-1} + \varepsilon_t\ ,t \geq 0\,, \quad (7)$$

де ε_t – послідовність незалежних стандартних нормальних величин.

Вважатимемо, що параметр $a(t) = a(1)$, коли t непарно і $a(t) = a(2)$, коли парно.

Тепер маємо:

$$\begin{cases} \lambda_{2t+1} = a^{(1)}\lambda_{2t} + \varepsilon_{2t+1}, \\ \lambda_{2t} = a^{(2)}\lambda_{2t-1} + \varepsilon_{2t}. \end{cases} \quad (8)$$

$$\lambda_{2t} = a^{(1)}a^{(2)}\lambda_{2t-1} + a^{(1)}\varepsilon_{2t} + \varepsilon_{2t+1}. \quad (9)$$

Тепер представимо λ_t наступним чином:

$$\lambda_t = \left(s_1(t)a^{(1)} + s_2(t)a^{(2)}\right)\lambda_{t-1} + \varepsilon_t\ , \quad (10)$$

де $s_1(t) = \dfrac{1-(-1)^t}{2}$, $s_2(t) = \dfrac{1+(-1)^t}{2}$.

При цьому виконуються наступні рівності:

$$s_1(t)s_2(t) = 0\,,\ \ s_1^2(t) = s_1(t)\,,\ s_2^2(t) = s_2(t)\,. \quad (11)$$

У результаті маємо таку систему різницевих рівнянь:

$$\lambda_t = s_1(t)a^{(1)}\lambda_{t-1} + s_2(t)a^{(2)}\lambda_{t-1} + \varepsilon_t\,. \quad (12)$$

Розглянемо певну кількість моментів часу від початку моменту розгляду інтенсивності та отримаємо оцінки для параметрів моделі за допомогою методу найменших квадратів. Шукаємо значення параметрів, при яких сума квадратів змін інтенсивності мінімальна, а випадкові відхилення моделі теж прямують до мінімуму: $t = \overline{1,N}$, $a^{(1)}$, $a^{(2)}$

$$\sum_{t=1}^{N}\left(\lambda_t - \left(s_1(t)a^{(1)}\lambda_{t-1} + s_2(t)a^{(2)}\lambda_{t-1}\right)\right)^2 \to \min\,. \quad (13)$$

Знайдемо частинні похідні функції, що оптимізується, та прирівняємо їх до нуля, тому що потрібна умова існування екстремуму функції двох змінних – це рівність нулю частинних похідних функції по змінним $a^{(1)}$, $a^{(2)}$ а:

$$-2\sum_{t=1}^{N}\left(\lambda_t - \left(s_1(t)a^{(1)}\lambda_{t-1} + s_2(t)a^{(2)}\lambda_{t-1}\right)\right)s_1(t)\lambda_{t-1} = 0\,, \quad (14)$$

$$-2\sum_{t=1}^{N}\left(\lambda_t - \left(s_1(t)a^{(1)}\lambda_{t-1} + s_2(t)a^{(2)}l_{t-1}\right)\right)s_2(t)\lambda_{t-1} = 0\,. \quad (15)$$

Складемо і розв'яжемо систему рівнянь:

$$\begin{cases} \sum_{t=1}^{N}\left(\lambda_t - \left(s_1(t)a^{(1)}\lambda_{t-1} + s_2(t)a^{(2)}\lambda_{t-1}\right)\right)s_1(t)\lambda_{t-1} = 0\,, \\ \sum_{t=1}^{N}\left(\lambda_t - \left(s_1(t)a^{(1)}\lambda_{t-1} + s_2(t)a^{(2)}\lambda_{t-1}\right)\right)s_2(t)\lambda_{t-1} = 0\,. \end{cases} \quad (16)$$

$$\begin{cases} \sum_{t=1}^{N}s_1(t)\lambda_t\lambda_{t-1} = \sum_{t=1}^{N}\left(a^{(1)}s_1^2(t) + a^{(2)}s_1(t)s_2(t)\right)\lambda_{t-1}^2, \\ \sum_{t=1}^{N}s_2(t)\lambda_t\lambda_{t-1} = \sum_{t=1}^{N}\left(a^{(1)}s_1(t)s_2(t) + a^{(2)}s_2^2(t)\right)\lambda_{t-1}^2. \end{cases} \quad (17)$$

Використовуючи властивості, можемо записати $s_1(t)$, $s_2(t)$:

$$\begin{cases} \sum_{t=1}^{N}s_1(t)\lambda_t\lambda_{t-1} = \sum_{t=1}^{N}a^{(1)}s_1(t)\lambda_{t-1}^2, \\ \sum_{t=1}^{N}s_2(t)\lambda_t\lambda_{t-1} = \sum_{t=1}^{N}a^{(2)}s_2(t)\lambda_{t-1}^2. \end{cases} \quad (18)$$

Звідси отримуємо вирази для оцінок:

$$a_N^{(1)} = \frac{\sum_{t=1}^{N}s_1(t)\lambda_t\lambda_{t-1}}{\sum_{t=1}^{N}s_1(t)\lambda_{t-1}^2}\,,$$

$$a_N^{(2)} = \frac{\sum_{t=1}^{N}s_2(t)\lambda_t\lambda_{t-1}}{\sum_{t=1}^{N}s_2(t)\lambda_{t-1}^2}\,. \quad (19)$$

Таким чином, ми отримали оцінки параметрів моделі інтенсивності руху на певній ділянці траси за N моментами часу розгляду інтенсивності, але при цьому самі параметри залежать від часу та змінюються, що дає пристосування моделі до реальних умов, перерахунок параметрів при отриманні нових даних моделі.

Розглянемо тепер послідовні оцінки параметрів процесу, що виходять за допомогою деякої модифікації оцінок, отриманих методом найменших квадратів.

Вивчення інтенсивності руху проводяться до певного моменту зупинки, таким чином обсяг вибірки може бути випадковою величиною, і в момент зупинки досліджень можна визначити оцінку параметрів моделі. Позначимо $t_1(n)$ – момент зупинки розгляду процесу, оскільки параметр моделі в загальному випадку змінюється відповідно часу, то відповідно

до моменту часу зупинки ми отримаємо точкову оцінку параметра моделі авторегресії $a^{(1)}_{t_1(n)}$.

Ми розглядаємо дві складові процесу та побудови його моделі, для іншої частини побудови моделі процесу момент зупинки та відповідна точкова оцінка будуть $t_2(n)$, $a^{(2)}_{t_2(n)}$, таким чином знімаємо залежність параметрів від часу та отримаємо точкові оцінки параметрів.

Особливістю таких оцінок є мінімізація обсягу вибірки за гарантованої якості оцінювання.

$$t_1(n)=\inf\left\{N:\sum_{t=1}^{N}s_1(t)\lambda^2_{t-1}\ge n\right\},$$
$$t_2(n)=\inf\left\{N:\sum_{t=1}^{N}s_2(t)\lambda^2_{t-1}\ge n\right\}. \tag{20}$$

$$a^{(1)}_{t_1(n)}=\frac{1}{n_1}\sum_{t=1}^{t_1(n)}s_1(t)\sqrt{\gamma^{(1)}_t}\lambda_{t-1}\lambda_t,$$
$$a^{(2)}_{t_2(n)}=\frac{1}{n_2}\sum_{t=1}^{t_2(n)}s_2(t)\sqrt{\gamma^{(2)}_t}\lambda_{t-1}\lambda_t. \tag{21}$$

Таким чином, кожному моменту часу відповідає певна оцінка параметрів моделі, тому обчислюючи послідовно оцінки, ми приходимо до значення оцінки, що нас задовольняє, а йому відповідає певний момент зупинки, тобто ми можемо далі не спостерігати за процесом, а на основі вибірки даних, що відповідає моменту зупинки, будувати адекватну модель з певною точністю.

Визначимо величини, яких не вистачає:

$$n_1=\sum_{t=1}^{t_1(n)}s_1(t)\sqrt{\gamma^{(1)}_t}\lambda^2_{t-1},\ n_2=\sum_{t=1}^{t_2(n)}s_2(t)\sqrt{\gamma^{(2)}_t}\lambda^2_{t-1}. \tag{22}$$

При цьому виконуються наступні рівності, які визначають значення величин

$$\gamma^{(1)}_t=\begin{cases}1,\ \text{якщо}\ 1\le t<t_1(n),\\ \eta^{(1)}(n),\ \text{якщо}\ t=t_1(B).\end{cases}$$
$$\gamma^{(2)}_t=\begin{cases}1,\ \text{якщо}\ 1\le t<t_2(n),\\ \eta^{(2)}(n),\ \text{якщо}\ t=t_2(n).\end{cases} \tag{23}$$

$\eta^{(1)}(n)$ та $\eta^{(2)}(n)$ можна визначити, вирішивши наступні рівняння:

$$\sum_{t=1}^{t_1(n)}s_1(t)\frac{\lambda^2_{t-1}}{\sigma^2_t}+\eta^{(1)}(n)\frac{\lambda^2_{t_1(n)-1}}{\sigma^2_t}=n,$$
$$\sum_{t=1}^{t_2(n)}s_2(t)\frac{\lambda^2_{t-1}}{\sigma^2_t}+\eta^{(2)}(n)\frac{\lambda^2_{t_2(n)-1}}{\sigma^2_t}=n. \tag{24}$$

Тоді визначаємо величини наступним чином:

$$\eta^{(1)}(n)=\frac{1}{\lambda^2_{t_1(n)-1}}\left(n-\sum_{t=1}^{t_1(n)-1}s_1(t)\lambda^2_{t-1}\right),$$
$$\eta^{(2)}(n)=\frac{1}{\lambda^2_{t_2(n)-1}}\left(n-\sum_{t=1}^{t_2(n)-1}s_2(t)\lambda^2_{t-1}\right). \tag{25}$$

Враховуючи те, що ε_k – незалежні стандартні нормальні випадкові величини, які відображають випадкову складову побудованої моделі, ми можемо записати ймовірність відхилення точкових оцінок параметрів моделі:

$$P_{a^{(1)}}\left\{\frac{n_1}{\sqrt{n}}\left(a^{(1)}_{t_1(n)}-a^{(1)}\right)\le x\right\}=\Phi(x),\ x\in(-\infty;+\infty),$$
$$P_{a^{(2)}}\left\{\frac{n_2}{\sqrt{n}}\left(a^{(2)}_{t_2(n)}-a^{(2)}\right)\le x\right\}=\Phi(x),\ x\in(-\infty;+\infty). \tag{26}$$

При цьому, $\Phi(x)=\frac{1}{2\pi}\int_{-\infty}^{x}e^{-\frac{t^2}{2}}dt$, а величини мають нормальний закон розподілу з наступними параметрами $a^{(1)}_{t_1(n)}$, $a^{(2)}_{t_2(n)}N\left(a^{(1)},\frac{n_1^2}{n}\right)$, $N\left(a^{(2)},\frac{n_2^2}{n}\right)$.

Тобто послідовні оцінки параметрів періодичної авторегресії мають нормальний розподіл із середнім рівним істинному значенню параметрів та дисперсією, яка залежить від параметру n. Саме ця властивість і перевірятиметься у моделюванні.

Висновки

Таким чином, робимо висновок, що немає підстав відхиляти гіпотезу про нормальний закон розподілу величин $y^{(1)}_t$, $y^{(2)}_t$ при заданому рівні значущості.

Серед усіх значень моментів зупинки при певному моделюванні визначимо середнє значення моменту зупинки, що буде наближено визначати оптимальний обсяг множини значень інтенсивності руху для досліджень: $\overline{t_1}(n)=615$, $\overline{t_2}(n)=688$.

Список літератури:

[1] Николайчук Я. М., Возна Н. Я., Пітух І. Р. Проектування спеціалізованих комп'ютерних систем. – Тернопіль: ТНЕУ, 2010.– 392 с.

[2] Бізнес-моделювання й управління потоками робіт і документообігом в економічних системах: Монографія / В. С. Пономаренко, І. О. Золотарьова, С. В. Мінухін та ін. – Харків : ХНЕУ ім. С. Кузнеця, 2010. – 270 с.

[3] Кундрат А. М., Кундрат М. М. Науково-технічні обчислення засобами MathCAD. Навч. посібник. Рівне: НУВГП, 2014. – 252 с.

Надійшла до редколегії 10.10.2023

УДК 004.42 DOi 10.30837/ bi.2023.1(99).09

Vladyslava Korovaina[1], Dmytro Kolesnykov[2], Oleksii Nazarov[3], Nataliia Nazarova[4]

[1] ХНУРЕ, м. Харків, Україна, vladyslava.korovaina.cpe@nure.ua

[2] ХНУРЕ, м. Харків, Україна, dmytro.kolesnykov@nure.ua,
ORCID iD: 0000-0002-4901-6869

[3] ХНУРЕ, м. Харків, Україна, oleksii.nazarov1@nure.ua,
ORCID iD: 0000-0001-8682-5000

[4] ХНУРЕ, м. Харків, Україна, Ukraine, nataliia.nazarova@nure.ua,
ORCID iD: 0009-0007-7816-7088

PREDICTIVE MODEL FOR ASSESSING PERFORMANCE OF STUDENTS

The object of research is the process of developing a predictive model for assessing the performance of university students based on the results of current studies. The purpose of the study is to build a predictive model of students' session results depending on the estimated parameters of current performance. The main problems in this area were analyzed, and goals were set for their direct implementation, fragmented preliminary data processing to build a machine learning model. Various machine learning models were built and the qualitative indicators of each model were evaluated. After selecting the optimal model, a graphical user interface for the predictive model was created. A predictive model of university students academic performance was created, as well as a graphical interface for its use. The significant factors in predicting student performance have been identified.

PREDICTING STUDENT PERFORMANCE, MACHINE LEARNING MODEL, DATA SET, CLASSIFICATION

Коровайна В.С., Колесников Д.О., Назаров О.С., Назарова Н.В. Прогностична модель оцінювання успішності студентів. Об'єктом дослідження є процес розробки прогнозної моделі для оцінки успішності студентів університету за підсумками поточного навчання. Мета роботи – побудова прогнозної моделі підсумків сесії студентів залежно від оціночних параметрів поточної успішності. Було проведено аналіз основних проблем у цій галузі, поставлено цілі для їх безпосереднього виконання, зроблено попередню обробку даних для побудови моделі машинного навчання. Було побудовано різні моделі машинного навчання та оцінено якісні показники кожної моделі. Після вибору оптимальної моделі було створено графічний інтерфейс користувача прогнозної моделі. Було створено прогнозну модель успішності студентів вишу, а також графічний інтерфейс для її використання. Визначено значущі фактори під час прогнозування успішності у студентів.

ПРОГНОЗУВАННЯ УСПІШНОСТІ СТУДЕНТІВ, МОДЕЛЬ МАШИННОГО НАВЧАННЯ, НАБІР ДАНИХ, КЛАСИФІКАЦІЯ

Introduction

The level of student success in higher education is a form of diagnosis and prediction of the level of commitment of a future specialist. In turn, student success is an indicator of the performance of the higher education institution in solving educational tasks. In order to solve these tasks as efficiently as possible, constant objective evaluation, adjustment and management are required. However, management is impossible without forecasting. Therefore, it is necessary to predict the performance of students at all stages of education.

Having information about those students who are most likely to have academic debt by the end of the semester if they do not change the current trend, we can influence students and thereby improve their academic performance.

The purpose of this thesis is to create a predictive model of the academic performance of NURE students. The availability of such a model will allow us to pay more attention to students who are at risk of having a large number of debts in academic disciplines and, as a result, will be candidates for expulsion. Early identification of such students will allow for more detailed and personalized work with them to help them manage their academic workload.

Based on the objective, this work includes the following tasks: review the literature on the subject, study the methods and algorithms used, clean and prepare the initial data, develop a predictive model, test the results, create a graphical user interface for the module for predicting student performance.

1. Subject area analysis

Education plays one of the most important roles in any country. The quality of education in a given society largely determines the pace of its economic and political development and its moral state.

The rapid development of information technologies makes it possible to automate many areas of human activity and increase their efficiency, and education is no exception. This paper will focus on creating a predictive model of student performance based on current grades using data mining technologies.

The measure of the quality of education received by a particular student is his or her grades in the subjects passed. When we talk about an educational institution,

one of the measures of the quality of education it provides is the aggregate of its students' grades. Timely measures to help students who cannot cope with the academic load are one of the main parts of educational work in higher education institutions that affect the quality of education in the institution.

Recently, many different changes have been made to improve the quality of education in higher education institutions. For example, the transition from a traditional grading system to a point system eliminates the possibility that a student who has not attended classes throughout the semester will simply come and take an exam. To be allowed to take the exam, they must earn a certain number of points, which in turn requires them to attend classes and complete current assignments.

This approach has an effective impact on the understanding of the study material. According to many studies, information that has been studied over a long period of time is retained for a long time, while "cramming" the night before an exam can only lead to a good result on the exam, which will eventually be passed, but the student will have no residual knowledge of the subject. In addition, there is a milestone control, a date set in the middle of the semester when a certain part of the material must be passed.

However, knowing the peculiarities of student life, many students still leave everything until the last minute. Some students manage to pass the course, while others are left with a debt for another semester. The problem is that the single dean's office of Tomsk Polytechnic University starts its control activities only after the student has already incurred debts.

Therefore, these measures cannot be called preventive. The main task of the predictive model is to identify such students and conduct certain talks with them before the problem of academic debt arises. At present, this measure is partly implemented by the fact that each group of freshmen has a tutor, and this tutor accompanies each group until the second year, and the schedule includes such an event as the "tutor's hour", where he analyzes the students' progress.

This is an excellent practice and should not be abandoned, but the human factor comes into play. It is not always possible for a tutor to convey to students the importance of attending classes. The introduction of a new system for predicting end-of-semester debt based on current academic performance is an important step in automating the educational process.

In this way, a single dean's office will be able to take control measures against at-risk students much earlier than a real problem arises. In this way, it will be possible to help students successfully complete the semester and develop professional skills, and to expel those who are not interested in studying earlier, freeing up places for those who really want and are ready to receive knowledge.

The main focus is on the forecasting system, because it is not only about monitoring the attendance of students. The data analysis shows that the final result of the semester is influenced not only by one factor of attendance, but by a whole range of different data. For example, there is a whole category of students who are already working in their field, mostly masters students and final year students. These students do not always have the opportunity to attend classes, and yet they show excellent results at the end of the session because they understand many aspects of the profession at a higher level than their classmates.

In fact, a huge number of factors affect a student's performance, and some of the most important are motivation to study, morale, relationships with classmates, and so on, but thanks to the fact that the system will identify problem students and will be able to work with each of them in detail, it will be possible to identify what problems exist with this particular student who risks ending the semester with a large amount of debt. It is possible to identify such parameters as motivation, determination, and psychological data, but this requires a large number of tests to be conducted on a regular basis. These methods are not very effective due to the complexity of their implementation, as well as the verification and interpretation of the results.

Once it is clear that this problem is indeed relevant, it is worth considering the existing methods for solving it.

At present, there are no publicly available materials devoted to the implementation and use of a system for predicting student performance in higher education institutions. Although this idea is not innovative, since about 2010 there have been articles about predicting the performance of applicants, predicting the performance of students in a particular course, where the following parameters are usually taken as initial data: the level of current knowledge of the subject, the number of absences, intermediate control, grades in the courses that provide the discipline.

Next, we will consider machine learning algorithms used to solve similar problems in the area of student performance prediction. The most commonly used methods and algorithms will be presented, as well as a description of existing works with specific tasks for which these methods are used.

Some sources use the clustering method, in our opinion this is not quite correct, since we already have the right classes, namely the number of debts, so in this case we should use classification methods. Clustering refers to unsupervised machine learning methods, and classification and regression refer to supervised learning, since they take markers for each class as input. Cluster analysis methods for estimating the final grade in a subject have been applied in [1].

2. K-Nearest Neighbors Algorithm

The K-nearest neighbor algorithm (KNN) is a type of supervised machine learning algorithm that can be used for both classification and regression prediction tasks. It is one of the easiest algorithms to understand, but it has been proven effective in a variety of tasks and is not only used for educational purposes. The Nearest Neighbors algorithm is also called a "lazy" classifier because it does not build a model during training, but simply stores data. All computations are started only when it is necessary to classify new data.

The essence of this algorithm is that the prediction of new data values is based on their proximity to already labeled data in the training set. In other words, if a new data point has 4 points of class A and 1 point of class B among its nearest neighbors, then this new point will be defined as class A. Thus, the k-nearest neighbors algorithm has 2 most important parameters, namely the distance metric (Euclidean, Manhattan, or Hamming) and the number of neighbors we will consider. A visual representation of the algorithm is shown in Fig. 1.

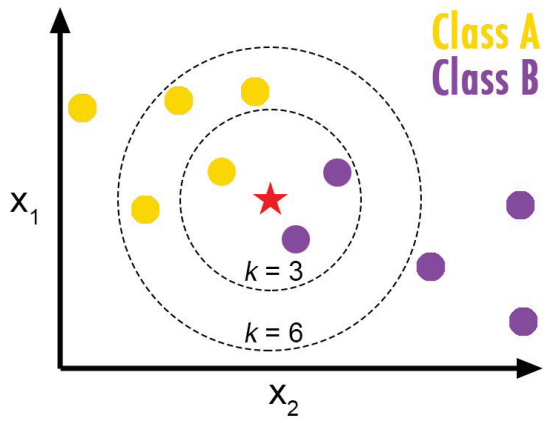

Fig. 1. KNN Algorithm Operation

This figure shows an array of source points color-coded according to their class membership. The asterisk marks the point to be classified. All points are plotted in two dimensions, along the X1 and X2 axes. If the set number of nearest neighbors is 3, then the unlabeled object belongs to class B, if it is 6, then it belongs to class A.

This algorithm has many nuances, for example, we can add weights to the voting data depending on the proximity to our unlabeled object. It is these nuances that make the KNN algorithm relevant for solving a wide range of tasks.

The benefits of this algorithm are:

– Easy to understand and interpret.
– Works well with non-linear data.
– It is a universal algorithm, suitable for both classification and regression tasks.
– It has relatively high accuracy.

Disadvantages of this algorithm:

– Large memory consumption to store all the data, unlike algorithms that use model building.
– Sensitivity to data size.
– Slow prediction for large amounts of data.
– High sensitivity to data noise.

In [2], this algorithm is used to classify the grade of an individual student in each subject based on his or her previous grades and the grades of students of previous courses with the most similar parameters in these subjects. This article presents a fairly high accuracy of the algorithm for this task, with a maximum grade prediction error of 0.55 points. However, only 307 students of one subject were considered, and there was no question of implementing this methodology in the university system.

This algorithm is also used in [3], where a student's grade for an exam in a given subject is predicted based on his or her grades in previous subjects, attendance, and midterm grades.

3. Support Vector Method

Support Vector Machines (SVM) are a family of similar learning algorithms for solving classification and regression problems. It is one of the most common learning methods belonging to the family of linear classifiers. One of the characteristics of the support vector method is that it consistently reduces the empirical classification error and increases the variance. Therefore, this method is also called the maximum distance classification method.

The basic idea of the support vector method can be illustrated by an example: there are points on a plane that are labeled into 2 classes that are linearly separated. In this case, the resulting function will be the plane that separates these classes. However, it is possible to draw many hyperplanes that separate these classes. To find the optimal hyperplane, you need to find the maximum sum of the normal vectors from class A and class B. A visual representation of this method can be seen in Fig. 2. In this figure, the reference vectors are perpendicular to the normals.

The formal description of this method is as follows - suppose we have an instructive example:

$$\{(X_1, C_1), (X_2, C_2), ..., (X_i, C_i)\},$$

where X_i – is a p-dimensional real vector; C_i – the value 1 or -1 that the class takes.

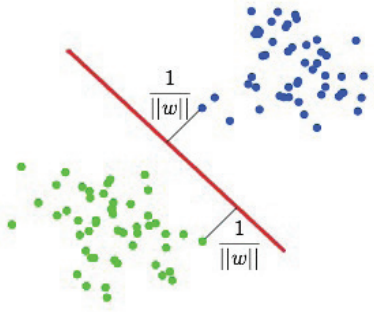

Fig. 2. Support Vector Method

The support vector method builds a classification function in the form of:

$$F(x) = sign([w, x] + b), \quad (1)$$

where [,] – scalar multiplication; w – is a normal vector to the separating hyperplane; b – auxiliary parameter.

So we can write it all down in the form of an optimization problem that has one solution, and only one:

$$\left\{\|w\|^2 \to \min; c_i\left(w \cdot x_i - b\right) \geq 1, 1 \leq i \leq n\right\}. \quad (2)$$

This problem is solved by quadratic programming and using Lagrange multipliers.

The case when there are 2 separate classes has been considered. In practice, the classes are almost always not linearly separable, and the task is to classify more than 2 classes. To solve the problem with linearly inseparable classes, we allow the classifier to make an error on the training set. Let's write the equation for this assumption.

$$\begin{gathered}\left\{\frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}\xi_i \to \min_{w,b,\xi_i};\right. \\ \left. c_i\left(w \cdot x_i - b\right) \geq 1;\ \xi_i \geq 0,\ 1 \leq i \leq n\right\}.\end{gathered} \quad (3)$$

where C – method setting parameter, ξ_i – is the value of the allowable error.

To solve multi-class problems, the generalized support vector method is used, since the transition to classification into many classes is made by splitting into 2 classes, such as the corresponding class and the non-corresponding class. This strategy is also called "One vs. All" and is used to apply binary classifiers to multi-class tasks.

Advantages of the algorithm:

– The problem is well studied and has a single solution.

– The principle of the optimal separating hyperplane leads to a reliable classification.

– Equivalent to a two-layer neural network, where the number of neurons in the hidden layer is automatically determined as the number of support vectors.

Disadvantages:

– Instability to noise, outliers in the initial data directly affect the construction of the separating hyperplane.

– No feature selection.

– It is necessary to select methods for constructing kernels and rectifying spaces separately for each task.

This algorithm is used in [3], where a student's grade for an exam in a given subject is predicted based on his grades in previous subjects, attendance, and midterm grades.

4. Neural Networks

In addition to the above methods, neural networks are also used to predict student performance. There are a number of references to the possibility of using neural networks to solve this problem, but there is no information about the actual implementation of such predictive models.

Neural networks are mathematical models based on the organizational and functional principles of biological neural networks. Neural networks are trained rather than programmed in the conventional sense. During training, neural networks are able to recognize and generalize complex relationships between input and output data. Once trained, the network is able to predict future values for a given sequence based on a series of past values.

An illustration of how a neural network works is shown in Figure 3. A neural network consists of neurons, layers, and synapses. Neurons are shown as nodes of different colors. All nodes of the same color belong to the same layer of the neural network. Synapses are lines that connect neurons in one layer to neurons in another layer. Synapses have only one parameter, the weight.

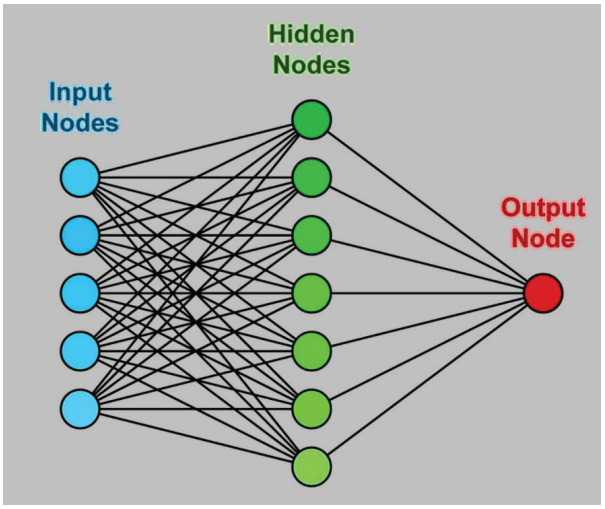

Fig. 3. Neural Network

Each neuron performs a specific mathematical function, so it receives a set of values as input and a single value as output. Thus, the output is a specific value produced by a previously trained neural network.

Advantages:

– Resistance to input noise.

– Self-learning and creativity. Ability to solve tasks that cannot be solved by other algorithms.

– Adaptation to changes, retraining.

Disadvantages:

– For large networks, it is impossible to estimate the network training time even approximately in advance.

– Difficulty in interpreting the result.

– Approximation of the obtained answer.

Let's take a closer look at the use of a neural network to solve the problem of predicting student performance. In [4], a neural network model is trained to predict whether a student will be promising. The task of binary classification of applicants is solved using input data such as school number, grades in physics and math, and parents' occupations. In [5], a neural network is used to predict grades in a computer science course. Article [6] discusses the task of classifying students based on NMT results.

Thus, we have reviewed the algorithms most commonly used to solve the problem of predicting student performance based on different initial data and solving different problems, whether it is performance in a particular subject or a general picture of performance in all disciplines. The considered examples of prediction are not systematic, but are only attempts to come closer to solving this problem. Obviously, to solve this problem successfully, it is necessary to apply several methods and compare their results.

5. Data preprocessing

The quality of machine learning models is highly dependent on the quality of the underlying data. However, real-world data is often poorly structured, with missing values, noise, and incorrect values. If the data is not well prepared, no amount of tuning of machine learning algorithms can ensure high predictive accuracy of these models. Data preparation before analysis takes up about 80 percent of a machine learning specialist's time, but this work is necessary.

In this paper, the methodology described in [7] is used to prepare the data.

To familiarize ourselves with the data and prepare it for further analysis, we will use the Python programming language and its libraries. The choice of the Python programming language is due to its high performance in data processing, simplicity, and a large number of libraries for machine learning. Python is one of the best languages for working with data. The following Python libraries are used in this paper.

– NumPy. This library adds support for large multidimensional arrays and matrices, as well as high-level commands for mathematical functions with very high performance on these arrays [7].

– Pandas. A data processing and analysis software library.

– Seaborn. A data visualization library based on another Python library, Matplotlib.

– Scikit - learn. An open source library for machine learning.

– PyQt5 is a set of extensions to the Qt graphical framework for the Python programming language, made in the form of a Python extension. This library implements almost all the capabilities of Qt and allows you to create a graphical interface for programs written in Python.

First of all, in order to work with data using Python, you need to convert it into a format that this language and its libraries can work with. In this case, this format is DataFrame from the Pandas library.

After the conversion, you can familiarize yourself with the main characteristics of the original dataset.

You can see that some attributes in the tuple have null values, for greater clarity you should see which attributes they are in.

You can see that the attribute "Disciplines in which unsatisfactory grades were received" itself suggests the presence of "missings", while "Profile" and "Country" are most likely missing values as a result of filling the database.

Next, we will analyze the data in the simplest terms in order to understand the initial vector of data preparation.

First of all, let's look at the number of debtors (and therefore the quality, which we will evaluate by the number of debts).

In order to take into account this attribute (the number of unsatisfactory grades), it is necessary to further divide this range of attribute values into groups, since the difference between a student with no debts and one debt is greater than between students with 18 and 17 debts, where it is not essential.

It was decided to divide students into 3 groups according to the number of debts:

– Group 1 "successful" - 0 debts

– Group 2 "debts" - from 1 to 6 debts

– Group 3 "many debts" - from 7 and more (maximum 18 debts)

In order to facilitate the analysis of this criterion, an additional column was created in the dataset indicating the group to which the student belongs.

6. Machine learning model

Thus, after analyzing all the parameters, the following were identified as having a greater impact on student performance, namely

1. Type of study
2. Qualification
3. Course of studies
4. Specialty
5. Academic leave (valid) - and no / /.
6. Total hours of absence in the semester
7. Total hours of classes in the semester

These parameters will be used in machine learning with the teacher to classify the student according to the number of debts.

Let's look at the correlations between the selected parameters.

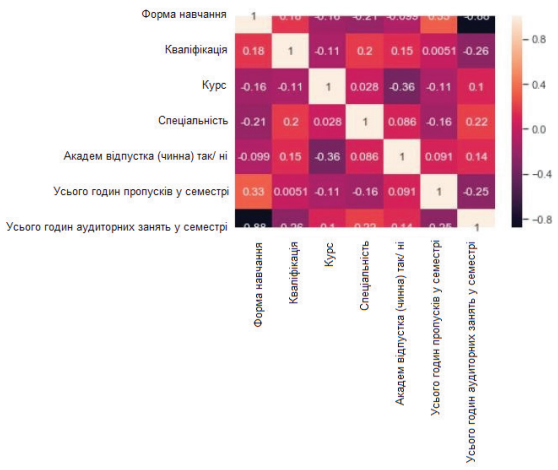

Fig. 4. Feature Correlation

To start training the model, we need to convert all the parameters into numerical representations. To do this, we will use the LabelEncoder function, which is already available in Python from the sklearn.preprocessing library.

After the conversion, we get the following: X is the data frame of all parameters, already converted to a numeric value, Y is the labels.

To start building the model, we divide our data into training and test samples.

```
from sklearn.model_selection import train_test_split
x_train, x_test, y_train, y_test = train_test_split(X, Y, test_size=0.3, random_state=0)
```

The first classifier used is logistic regression.

```
from sklearn.linear_model import LogisticRegression
lr = LogisticRegression(C=1000.0, random_state=0 )
lr.fit(x_train,y_train)
```

```
LogisticRegression(C=1000.0, class_weight=None, dual=False, fit_intercept=True,
                   intercept_scaling=1, l1_ratio=None, max_iter=100,
                   multi_class='warn', n_jobs=None, penalty='l2',
                   random_state=0, solver='warn', tol=0.0001, verbose=0,
                   warm_start=False)
```

```
print(lr.score(x_train, y_train))
print(lr.score(x_test, y_test))
```

```
0.6668915500084331
0.6336088154269972
```

```
pred_y = lr.predict(x_test)
from sklearn import metrics
# Model Accuracy, how often is the classifier correct?
print(metrics.classification_report(pred_y, y_test))
```

	precision	recall	f1-score	support
борги	0.84	0.60	0.70	1732
багато боргів	0.48	0.74	0.58	433
успішно	0.40	0.69	0.51	376
accuracy			0.63	2541
macro avg	0.57	0.67	0.59	2541
weighted avg	0.71	0.63	0.65	2541

This method showed that there is some dependency and that the parameters were chosen correctly. Next, we will try other types of classification and compare the results.

Let's use the support vector method.

```
from sklearn import svm
clf = svm.SVC()
clf.fit(x_train,y_train)
```

```
SVC(C=1.0, cache_size=200, class_weight=None, coef0=0.0,
    decision_function_shape='ovr', degree=3, gamma='auto_deprecated',
    kernel='rbf', max_iter=-1, probability=False, random_state=None,
    shrinking=True, tol=0.001, verbose=False)
```

```
print(clf.score(x_train, y_train))
print(clf.score(x_test, y_test))
```

```
0.8794063079777366
0.7197953561589925
```

```
pred_y = clf.predict(x_test)
from sklearn import metrics
# Model Accuracy, how often is the classifier correct?
print(metrics.classification_report(pred_y, y_test))
```

	precision	recall	f1-score	support
борги	0.86	0.68	0.76	1564
багато боргів	0.54	0.85	0.66	421
успішно	0.64	0.74	0.69	556
accuracy			0.72	2541
macro avg	0.68	0.76	0.70	2541
weighted avg	0.76	0.72	0.73	2541

Here we can see that the support vector method is prone to overfitting, as there is a large 10 percent difference between the training and test sets in label detection. However, the result on the test set is almost 10 percent higher than the classifier based on logistic regression.

Using the 3rd Classifier "Random Forest"

```
from sklearn.ensemble import RandomForestClassifier
clas = RandomForestClassifier(max_depth=15, random_state=0)
clas.fit(x_train,y_train)
```

```
RandomForestClassifier(bootstrap=True, class_weight=None, criterion='gini',
                       max_depth=15, max_features='auto', max_leaf_nodes=None,
                       min_impurity_decrease=0.0, min_impurity_split=None,
                       min_samples_leaf=1, min_samples_split=2,
                       min_weight_fraction_leaf=0.0, n_estimators=10,
                       n_jobs=None, oob_score=False, random_state=0, verbose=0,
                       warm_start=False)
```

```
print(clas.feature_importances_)
```

```
[0.01858244 0.03035815 0.12210839 0.2179611  0.07945399 0.31382953
 0.21770641]
```

```
print(clas.score(x_train, y_train))
print(clas.score(x_test, y_test))
```

```
0.8775510204081632
0.7898465171192444
```

```
pred_y = clas.predict(x_test)
from sklearn import metrics
# Model Accuracy, how often is the classifier correct?
print(metrics.classification_report(pred_y, y_test))
```

	precision	recall	f1-score	support
борги	0.82	0.78	0.80	1311
багато боргів	0.74	0.85	0.79	579
успішно	0.77	0.76	0.77	651
accuracy			0.79	2541
macro avg	0.78	0.80	0.79	2541
weighted avg	0.79	0.79	0.79	2541

Random Forest showed the best result, although the spread between the training and test samples is quite large. The accuracy of the test sample prediction is 79 percent. This is quite a high accuracy rate. Also, random forests with different maximum depths were tested, and experience showed that a depth greater than 15 does not significantly affect the test set prediction, which affects the accuracy.

Thus, we can conclude that the task of determining student performance based on such parameters as

- Type of study
- Qualification
- Course of study
- specialization
- Academic leave (valid) - and not /
- Total hours of absences in the semester
- Total hours of classes in the semester

The analysis also showed that the most significant contribution to the model is made by 3 parameters, namely

- Total absenteeism hours in the semester
- Total classroom hours in the semester
- Course.

In the future, additional parameters, such as the student's hobbies, participation in the social life of the university, can also improve the accuracy of the model.

Conclusions

As a result of the research practice, the following tasks were accomplished:

1. Analysis of methods used to solve similar problems in the field of education.

2. Preliminary data processing was carried out.

3. Built a machine learning model capable of predicting student performance at the end of the semester.

4. Identified the most significant features in determining student performance.

5. Developed a graphical user interface for the student performance prediction model.

Since the preprocessing of the data used labels to indicate "student performance," the type of machine learning was supervised. Three supervised machine learning models were tested: logistic regression, support vector machine, and random forest. The random forest classifier performed best on the test sample. This algorithm is optimal because it has the following advantages.

In addition, a graphical interface was created for the module to predict student performance at the end of the semester based on current grades.

Conflict of Interest

The authors declare no conflict of interest.

References

[1] В.О. Шевченко. Прогнозування успішності студентів на основі методів кластерного аналізу // Вісник ХНАДУ. 2015. №68. – С. 15-18.

[2] В.О. Шевченко. Інформаційна технологія формування індивідуальних траєкторій самостійної роботи студентів // Вісник НТУ "ХПІ» 2015. №21(1130). – С. 76-83.

[3] В.О. Шевченко, А.І. Кудін. Алгоритмізація процедури кластерного аналізу для прогнозування успішності студентів // Автомобіль і електроніка. Сучасні технології, 11/2017. – С. 64-67.

[4] O. Haitan, O. Nazarov. Hybrid approach to solving of the automated timetabling problem in higher educational institution // Системи управління, навігації та зв'язку, 2020, випуск 2(60). – С. 60-69.

[5] Al-Shehri H. et al. Student performance prediction using support vector machine and k-nearest neighbor //2017 IEEE 30th Canadian Conference on Electrical and Computer Engineering (CCECE). – IEEE, 2017. – С. 1-4.

[6] Breazley, D. Python Cookbook, Third Edition / D. Breasley, B. K. Jones. – USA: O'Reilly Media, 2013. – 688 p.

[7] McKinney, W. Python for Data Analysis. – USA: O'Reilly Media, 2013. – 453 p.

The article was delivered to editorial stuff on the 22.03.2023

УДК 004.42 DOi 10.30837/ bi.2023.1(99).10

Anhelina Shemrikovych[1], Oleksandr Samantsov[2], Oleksii Nazarov[3], Nataliia Nazarova[4]

[1] ХНУРЕ, м. Харків, Україна, anhelina.shemrikovych@nure.ua

[2] ХНУРЕ, м. Харків, Україна, oleksandr.samantsov@nure.ua,ORCID iD: 0000-0002-4788-4144

[3] ХНУРЕ, м. Харків, Україна, oleksii.nazarov1@nure.ua, ORCID iD: 0000-0001-8682-5000

[4] ХНУРЕ, м. Харків, Україна, nataliia.nazarova@nure.ua, ORCID iD: 0009-0007-7816-7088

STUDY OF ALGORITHMS FOR OPTIMIZATION OF ENERGY MANAGEMENT IN TRANSPORTATION SYSTEMS FOR REDUCTION OF ENVIRONMENTAL IMPACT

The object of the research is the technology of optimization algorithms for energy consumption management in transport systems. The purpose of the work is research and analysis of the effectiveness of optimization algorithms for reducing environmental impact, selection of criteria and methods for comparative analysis. The existing algorithms for optimizing energy consumption management in transport systems were considered, their advantages and disadvantages and principles of operation were investigated, methods of comparison were described and demonstrated, and formulas for calculating numerical indicators were proposed.

TRANSPORT SYSTEMS, OPTIMIZATION ALGORITHMS, ENERGY MANAGEMENT, ECO-FRIENDLY ALGORITHMS

Шемрикович А.Д., Саманцов О.О., Назаров О.С., Назарова Н.В. Дослідження алгоритмів оптимізації енергоменеджменту в транспортних системах для зменшення впливу на навколишнє середовище. Об'єктом дослідження є технології алгоритмів оптимізації керування енергоспоживанням у транспортних системах. Метою роботи є дослідження та аналіз ефективності алгоритмів оптимізації для зменшення екологічного впливу, виділення критеріїв та методів для проведення порівняльного аналізу. Розглянуто існуючі алгоритми оптимізації керування енергоспоживанням у транспортних системах, досліджено їх переваги та недоліки та принципи роботи, описано та продемонстровано методи порівняння та запропоновано формули для обчислення числових показників.

ТРАНСПОРТНІ СИСТЕМИ, АЛГОРИТМИ ОПТИМІЗАЦІЇ, ЕНЕРГОМЕНЕДЖМЕНТ, ЕКОЛОГІЧНО ЧИСТІ АЛГОРИТМИ

Introduction

In the intricate tapestry of our modern world, the veins of trade and communication are intricately woven through maritime and road transportation systems. While these arteries of trade and mobility are essential to global prosperity, they also carry a burden of environmental impact.

The constant demand for movement, whether of goods or people, has cast a deep shadow on our planet, manifesting itself in the rapid growth of carbon emissions and a significant environmental burden.

The specter of climate change looms, demanding a reassessment of how we move around our planet. The oceans teem with ships carrying goods across continents, and the roads pulse with vehicles carrying people and goods. But the fuel that powers these journeys is often too costly for our environment.

This study examines algorithms for optimizing energy management in transportation systems from an environmental impact perspective.

The analysis derives the main evaluation criteria and forms a comparison model.

The goal of this work is to identify methodologies that minimize energy consumption without compromising the integrity of transportation systems.

In order to get a result that would satisfy the goal, it is necessary to solve a series of the following problems:

– analyze existing energy management optimization algorithms;

– determine the methods and criteria for comparing algorithms;

– model the conditions for conducting experiments to evaluate the algorithms;

– use the obtained data to measure the selected metrics for comparing algorithms;

– analyze the results obtained;

– formulate recommendations for the use of the algorithms;

– propose a possible extension of the study and characterize the relevance of the work in the future.

The subject of the study is the effectiveness and feasibility of using algorithms to optimize energy management in transportation systems in terms of environmental impact. The subject of the study are algorithms for optimization of energy management.

The research methods are measurement of performance indicators by criteria and their calculation using the proposed mathematical formulas for calculating each of the indicators. The results of the study can be successfully used in the creation or further analysis of new algorithms for optimization of energy management, or in the selection of the most optimal algorithm under existing conditions.

1. Analysis of the subject area

Gasoline, the quintessential energy source for marine and road transportation, is the cornerstone of global mobility. But behind its ubiquitous role lies a shadow — a history of environmental impacts that reverberate across oceans and urban landscapes.

In maritime trade and transportation, gasoline takes its place among a variety of fuels. While it contributes to the energy needs of ships, its environmental impact requires careful consideration. While gasoline-powered ships emit relatively low levels of sulfur oxides (SOx) and particulate matter, they contribute to the global greenhouse gas burden. Burning gasoline releases carbon dioxide (CO2), adding to the complex matrix of marine emissions that affect the climate and the delicate balance of marine ecosystems.

Next, let's talk about gasoline in automobiles. The history of gasoline's impact on road transportation is similar to its role at sea. As the primary fuel for internal combustion engines in cars, trucks and buses, gasoline plays a key role in ensuring mobility. But that confidence comes at a price.

Gasoline-powered vehicles contribute significantly to urban air pollution by emitting nitrogen oxides (NOx) and volatile organic compounds (VOCs). These emissions not only degrade air quality, but also contribute to respiratory health problems, especially in densely populated urban areas.

Next, the cumulative impact. The cumulative environmental impact of both marine and road transport fuels transcends geographic boundaries. While road transport typically affects local air quality, maritime transport extends its impact over large expanses of water, affecting coastal regions and the high seas. The cumulative emissions of CO2, methane, NOx and other pollutants paint a grim picture — a story of environmental impact that goes beyond the convenience and necessity of transportation.

On to mitigation and solutions. Tackling the pollution caused using gasoline in transportation requires a multipronged approach. Tighter regulations enforce emissions standards and encourage the development of cleaner engine technologies. The transition to electric vehicles, hybrid systems, and research into sustainable alternative fuels offer a glimmer of hope for reducing the environmental impact of gasoline. Innovations in engine efficiency and emissions control offer promising ways to reduce pollution while maintaining mobility.

Finally, the balance between mobility and responsibility. Gasoline, an indispensable energy source, requires a delicate balance between progress and environmental stewardship. As we move toward a transportation-dependent future, the imperative is not to deny mobility, but to innovate greener solutions. Using cleaner fuels, improving engine technology, and fostering a collective commitment to reducing our dependence on gasoline are important steps toward a harmonious coexistence of mobility and environmental responsibility. This balance holds the promise of a cleaner and healthier planet for future generations and is shown in Fig. 1.

Environmental Impact Metrics	Marine Transport (per year)	Auto Transport (per year)
Carbon Dioxide (CO2) Emissions	120 million tons	420 million tons
Sulfur Oxides (SOx) Emissions	5,000 tons	2,000 tons
Particulate Matter (PM) Emissions	300 tons	1,500 tons
Nitrogen Oxides (NOx) Emissions	2,000 tons	6,000 tons
Volatile Organic Compounds (VOCs)	150 tons	300 tons

Fig. 1. Balance between mobility and responsibility

2. Problem statement

After analyzing the subject industry, its main needs and existing problems, it is necessary to analyze what algorithms exist for optimizing energy management and formulate criteria for evaluating these algorithms. Possible criteria for evaluating the performance of each algorithm may include:

– Energy savings. The effectiveness of the algorithm in conserving energy while maintaining or improving performance, and the ability of the algorithm to minimize fuel consumption during transport operations;

– Environmental Impact Reduction. The ability of the algorithm to reduce greenhouse gas emissions (CO2, NOx, SOx, VOCs) associated with transportation activities, and the impact of the algorithm on reducing pollutants that contribute to air and water pollution; — Operational Performance. How efficiently the algorithm uses resources, improving vehicle/ship performance while reducing energy consumption;

– Scalability and adaptability. How well the algorithm performs when applied to different scales of transportation systems, from individual vehicles/ships to entire fleets, the performance of the algorithm under different conditions, including weather, traffic, and work shifts;

– Real-time implementation. The speed of the algorithm to provide optimized solutions for dynamic changes in the environment and operations;

– Computational requirements of the algorithm for real-time implementation in transportation systems.

– Economic Efficiency. Costs associated with implementing and supporting the algorithm in transportation systems, ability of the algorithm to provide significant environmental benefits compared to the cost of implementation;

– Durability and reliability. Resilience of the algorithm to uncertainties and unexpected scenarios in

transportation operations, consistency and accuracy of the algorithm to provide optimized solutions over time;

– Compliance with regulatory requirements. The extent to which the algorithm helps transportation systems meet environmental and emissions standards;

– Ease of use and integration. Ease of integration of the algorithm into existing transportation systems;

- User Adaptability. Convenience of the algorithm for transport operators and decision makers.

Taking into account all of the above criteria and analyzing the subject area, the following tasks need to be solved as part of the study of algorithms for optimizing energy management in transportation systems to reduce environmental impact:

– Review existing algorithms for optimizing energy management in transportation systems,

– Select those that can be used to reduce environmental impact;

– Prioritize the above evaluation criteria;

– Analyze and organize the algorithms according to the above evaluation criteria;

– Formulate an experimental plan to obtain experimental data, create software test environments for measurements for each of the criteria and algorithms;

– Conduct the experiment, analyze the results, and document the results;

– Provide recommendations and analysis results for the use of specific energy management optimization algorithms.

3. Overview of the main algorithms used for optimization

These optimization algorithms offer a variety of approaches to managing energy consumption in transportation systems, providing solutions for route optimization, resource allocation, vehicle scheduling, and energy efficient operation. Each algorithm has its own strengths and applications, and their choice often depends on the specific needs and constraints of the transportation context.

The most commonly used algorithms are:

Linear Programming. Used in route optimization, resource allocation, and planning in transportation systems. It aims to maximize or minimize a linear objective function subject to linear constraints. It is used to optimize transportation logistics and distribution.

Genetic Algorithms (GA). Used in vehicle routing, fleet optimization, and energy-efficient vehicle design. GA mimics the processes of natural selection to constantly evolve solutions. This is important for finding optimal solutions to complex transportation and logistics problems.

Ant Colony Optimization (ACO). Used to find the shortest routes in transportation networks and optimize traffic flow. ACO simulates the behavior of ants foraging for food and guides algorithms to find optimal routes and paths. It is effective in solving routing and resource allocation problems.

Particle Swarm Optimization (PSO). Used in route optimization search, vehicle scheduling, and energy-efficient vehicle routing. PSO models the social behavior of organisms by iteratively optimizing possible solutions. It is useful for solving complex optimization problems in transportation systems.

Heating Simulation. Used in vehicle routing, energy efficient routing, and scheduling. Simulated annealing mimics the annealing process in metallurgy to find optimal solutions by taking worse solutions first before approaching the optimal one.

Dynamic Programming. Used for optimal control of vehicle operation and energy-efficient routing. Dynamic programming breaks down complex problems into simpler subproblems that are suitable for finding optimal solutions over time, such as in route planning and energy management.

Heuristic Algorithms. Used in vehicle routing, traffic flow optimization, and fleet management. Heuristic algorithms, including methods such as nearest neighbor, insertion, and expansion, provide approximate solutions to transportation optimization problems.

Metaheuristic algorithms. Used in vehicle scheduling, energy-efficient routing, and fleet optimization. Metaheuristic algorithms include a variety of methods such as tabu search, genetic algorithms, and simulated annealing. These methods provide high-level strategies for finding solutions efficiently.

4. Linear Programming

Linear programming is a cornerstone in the field of energy management, providing a structured mathematical approach to optimizing resource use, streamlining operations, and reducing environmental impact. In the dynamic landscape of transportation, where efficiency is key, and sustainability is imperative, linear programming is becoming a key tool for navigating the complex interplay between energy consumption, operational efficiency, and environmental protection.

At its core, linear programming is a mathematical technique that seeks to optimize an objective function subject to a set of linear constraints. In the context of energy management, this technique is becoming a catalyst for optimizing fuel consumption, minimizing emissions, and improving energy efficiency in transportation systems. Linear programming models provide a systematic framework for decision making, helping to allocate resources while meeting operational constraints.

Linear programming plays a key role in determining the most efficient routes for vehicles, ships, or transportation networks. By taking into account variables such as distance, fuel consumption, and time constraints, it helps determine the optimal paths that minimize energy consumption and meet operational requirements.

The efficient use of resources such as fuel, time, and vehicle capacity is critical to transportation systems. Linear programming models help to optimally allocate these resources across fleets or routes, ensuring that energy consumption is minimized without compromising performance.

By optimizing schedules, minimizing downtime, and balancing load factors, linear programming helps improve vehicle efficiency. This makes it easier to make strategic decisions to reduce energy waste and increase overall productivity.

Linear programming algorithms provide near-optimal solutions to energy management problems, enabling accurate resource allocation and operational planning.

These models enable real-time decision making by providing information for route planning, resource allocation, and operational planning. By optimizing fuel consumption and reducing emissions, linear programming makes a significant contribution to reducing the environmental impact of transportation systems.

While linear programming provides powerful tools for managing energy consumption, it is not without limitations. It assumes linear relationships between variables and constraints, which can oversimplify the complexity of real-world transportation systems. Future developments aim to address these limitations by integrating nonlinear models and advanced optimization techniques to accurately model more complex transportation dynamics.

Linear programming is a fundamental pillar in the quest for efficient and sustainable energy management in transportation systems. Its application to route optimization, resource allocation, and operations planning paves the way for reducing energy consumption, minimizing environmental impact, and increasing efficiency, ultimately leading transportation systems to a future where progress is seamlessly integrated with environmental responsibility.

5. Genetic algorithms

Genetic algorithms (GAs) represent an advanced approach to solving complex optimization problems, and their application to energy management in transportation systems heralds a transformative paradigm. Based on the principles of natural selection and evolutionary processes, GAs offer innovative solutions that optimize resource utilization, reduce energy consumption, and mitigate the environmental impact of various transportation modes.

GAs mimics the process of natural selection, using the principles of selection, reproduction, and mutation to iteratively evolve solutions to complex problems. In the area of energy management:

GAs excels at finding optimal routes for vehicles, taking into account factors such as fuel economy, traffic conditions, and time constraints. By developing and refining potential solutions, they identify routes that minimize energy consumption while meeting operational requirements.

These algorithms help optimize fleet performance by determining the best vehicle configuration, scheduling, and resource allocation to minimize energy loss across the fleet.

GAs play an important role in the development of energy-efficient vehicles, optimizing engine performance, aerodynamics, and vehicle weight to improve fuel efficiency and reduce overall energy consumption.

GAs explores large solution spaces and provide near-optimal solutions to complex problems of optimizing many variables in transportation systems. They adapt to changing environments and dynamic conditions, making them suitable for real-time decision making in transportation operations.

GAs contributes to innovative solutions by exploring unconventional paths and configurations that may be overlooked by human-designed algorithms. The computational requirements of GAs can be intensive, requiring significant computing resources. Tuning the parameters for optimal performance and convergence is challenging but offers opportunities for improvement. Combining the algorithms with other optimization methods, such as neural networks or metaheuristic algorithms, increases their efficiency and effectiveness.

The evolution of GA continues, with promising advances aimed at addressing current limitations and further optimizing energy management in transportation systems. Future developments will focus on improving scalability, increasing convergence performance, and integrating GA with new technologies to achieve even greater efficiency and sustainability.

Genetic algorithms are emerging as a pioneering force in the revolution of energy management in transportation systems. Their application in route optimization, fleet management, and vehicle design heralds a future where transportation is not only efficient, but also environmentally sustainable. By mimicking the evolutionary processes of nature, GAs are leading us to a greener, more energy-efficient future of transportation, where innovation and nature-inspired algorithms work together to reduce environmental impact and increase operational efficiency.

6. Ant colony optimization

Ant Colony Optimization (ACO) is a powerful biological algorithm that mirrors the behavior of ants foraging for food. In the field of energy management in transportation systems, ACO is emerging as a transformative force, offering innovative solutions that optimize routes, reduce fuel consumption, and minimize environmental impact.

ACO algorithms model the behavior of ants as they communicate and navigate to find the shortest path to food sources. This approach uses pheromone trails and heuristic information to iteratively converge on optimal solutions.

ACO algorithms are ideal for finding the most energy-efficient routes for vehicles or ships. By mimicking the communication between ants using pheromones, these algorithms determine the paths that minimize fuel consumption, taking into account factors such as distance, traffic, and energy efficiency.

In urban transportation systems, ACO helps optimize traffic flow by identifying routes that minimize congestion, reduce idle time, and optimize traffic signals to improve fuel economy. ACO helps optimize the allocation of resources across transportation networks by scheduling deliveries and vehicle routes to minimize energy consumption and optimize resource utilization.

ACO explores multiple paths and configurations, providing near-optimal solutions to complex transportation optimization problems. ACO adapts to dynamic and changing conditions, making it suitable for real-time decision making in transportation operations. By mimicking the self-organization and decentralized decision-making of ants, ACO promotes innovative solutions in energy management. ACO can be a computationally intensive process that requires parameter optimization for efficiency. Fine-tuning of ACO parameters is critical for optimal convergence, opening opportunities for further research and improvement.

As technology advances, ACO algorithms are expected to continue to evolve. Future developments aim to reduce computational complexity, increase scalability, and integrate ACO with new technologies to improve transportation efficiency and sustainability.

Ant Colony Optimization is an innovative way to rethink energy management in transportation systems. Its application to route optimization, traffic flow management, and resource allocation heralds a future where transportation operations are not only efficient, but also environmentally conscious. Inspired by the principles of organization in nature, ACO algorithms pave the way for a sustainable, energy-efficient transportation ecosystem where biological algorithms guide us to reduce our environmental impact and optimize our energy consumption.

7. Particle Swarm Optimization

Particle Swarm Optimization (PSO) is a biological algorithm that models the social behavior of organisms, including the flocking and swarming patterns observed in birds and fish. In the field of energy management in transportation systems, PSO is emerging as a dynamic and effective tool that provides innovative solutions to optimize routes, increase fuel efficiency, and reduce environmental impact.

PSO algorithms are based on the collective behavior of organisms in a swarm. Individuals (particles) within the swarm cooperate and communicate by exchanging information and seeking optimal solutions through iterative movement in the solution space.

PSO is well suited for determining energy-efficient routes for vehicles, ships, or transportation networks. By simulating the movement of particles, these algorithms identify paths that minimize fuel consumption, taking into account factors such as distance, traffic, and energy efficiency.

By optimizing operations, PSO helps determine the best vehicle configuration, scheduling, and resource allocation to minimize fleet energy consumption.

PSO contributes to the development of energy-efficient vehicles by optimizing engine performance, aerodynamics, and vehicle weight, resulting in improved fuel efficiency and reduced energy consumption.

Benefits and Impacts.

PSO explores a wide variety of paths and configurations, providing near-optimal solutions to complex transportation optimization problems.

PSO adapts to changing environments and evolving conditions, making it suitable for real-time decision making in transportation operations.

By mimicking swarm behavior, PSO uses collective intelligence to find innovative solutions to manage energy consumption.

Optimization of PSO parameters is critical for convergence and efficiency. PSO can be a computationally intensive process that requires optimization for scalability and performance.

Integrating PSO with complementary optimization techniques can increase its effectiveness in complex transportation systems.

As technology advances, PSO algorithms are expected to continue to evolve. Future developments will address computational complexity, increase convergence, and integrate PSO with new technologies to improve transportation efficiency and sustainability.

Particle Swarm Optimization represents an advanced approach to revolutionize the management of energy consumption in transportation systems. Its application in route optimization, fleet management, and vehicle design provides a glimpse into a future where transportation is not only efficient, but also environmentally sustainable. By following the natural principles of cooperation, PSO algorithms pave the way for a more sustainable and energy efficient transportation ecosystem, where innovative algorithms guide us to reduce our environmental impact and optimize energy consumption.

8. Annealing simulation

Simulated Annealing (SA) is a powerful optimization method inspired by the physical process of annealing in metallurgy. It is a versatile algorithm used in a wide variety of fields, including energy management in transportation systems. SA provides a unique approach to solving complex optimization problems to minimize energy consumption, optimize routes, and reduce environmental impact.

The SA algorithm mimics the annealing process in metallurgy, where metals are heated and gradually cooled to reduce defects and produce a more stable structure. Similarly, SA gradually approaches optimal solutions, allowing for inferior decisions from time to time to avoid local optima.

SA can find energy-efficient routes for vehicles or ships. By exploring and gradually cooling the system, the algorithm identifies paths that minimize fuel consumption, taking into account factors such as distance, traffic conditions, and energy efficiency.

By optimizing vehicle scheduling, SA helps minimize downtime, improve utilization, and reduce energy costs during transportation operations.

SA helps optimize resource allocation across transportation networks, planning deliveries and vehicle routes to minimize energy consumption and increase overall efficiency.

Benefits and Impact

SA explores multiple solutions, allowing you to identify near-optimal paths and configurations in complex transportation optimization problems.

SA adapts to changing conditions, enabling real-time decision making in transportation operations.

SA's ability to make worse decisions from time to time helps avoid getting stuck on local optimal solutions, leading to better overall results.

Parameter Tuning: Optimizing SA parameters is essential for achieving optimal convergence performance and solution quality.

SA can be computationally intensive, requiring optimization for scalability and efficiency.

Integrating SA with complementary optimization techniques can increase its effectiveness in solving complex transportation optimization problems.

As technology advances, SA algorithms continue to evolve. Future advances will address computational complexity, increase convergence, and integrate SA with new technologies to improve transportation efficiency and sustainability.

Simulated Annealing is a sophisticated tool for rethinking energy management in transportation systems. Its application to route optimization, vehicle scheduling, and resource allocation promises a future where transportation operations are not only efficient, but also environmentally conscious. By mimicking the annealing process, SA algorithms lead us to reduce environmental impact and optimize energy consumption, paving the way for a more sustainable and efficient transportation ecosystem.

9. Dynamic Programming

Dynamic Programming (DP) is a powerful mathematical optimization technique used in a variety of fields, including energy management in transportation systems. Known for its ability to solve complex problems by breaking them down into simpler subproblems, DP offers innovative solutions to optimize routes, reduce fuel consumption, and improve overall transportation efficiency.

At its core, Dynamic Programming solves a complex problem by breaking it down into smaller subproblems, solving each subproblem only once, and storing the solution. This bottom-up approach allows you to obtain optimal solutions from the optimal solutions of its subproblems.

DP is ideal for finding energy-efficient routes for vehicles, ships, or transportation networks. By considering converging subproblems, it identifies paths that minimize fuel consumption, taking into account variables such as distance, traffic conditions, and energy efficiency.

By optimizing vehicle performance, DP helps reduce idle time, optimize utilization, and streamline work schedules to minimize energy consumption during transportation operations.

DP helps to efficiently allocate resources such as fuel and time across transportation networks, plan deliveries and vehicle routes to minimize energy consumption, and increase overall efficiency. Benefits and Impact

DP ensures that optimal subproblem solutions contribute to overall optimal solutions by providing efficient solutions to complex optimization problems.

DP preserves subproblem solutions, reducing redundant computations and increasing computational efficiency.

DP adapts to changing conditions, making it suitable for real-time decision-making during transportation operations.

DP may face scalability and computational complexity issues for larger problems.

Balancing optimal solutions and computational efficiency requires careful consideration of tradeoffs.

The applicability of DP may be limited by computational resources and real-time constraints in dynamic transportation systems.

As technology advances, DP algorithms continue to evolve. Future developments are aimed at solving scalability problems, increasing computational efficiency, and integrating DP with new technologies to improve transportation efficiency and sustainability.

Dynamic programming is becoming the main tool for optimizing energy management in transportation systems. Its use in route optimization, resource allocation, and operations planning allows us to look to a future where transportation is not only efficient, but also environmentally conscious. By breaking down complex problems into manageable subproblems, DP algorithms guide us to reduce environmental impact and optimize energy consumption, creating a more sustainable and efficient transportation ecosystem.

10. Metaheuristic Algorithms

Metaheuristic algorithms represent a class of innovative and versatile optimization methods that go beyond traditional problem-solving techniques. Designed to solve complex optimization problems, including energy management in transportation systems, these algorithms provide dynamic and adaptive solutions to minimize fuel consumption, optimize routes, and reduce environmental impact.

Metaheuristics are high-level strategies that guide the exploration of solution spaces to find near-optimal solutions without guaranteeing the absolute optimum. These algorithms are characterized by their flexibility, adaptability, and ability to efficiently traverse large solution spaces.

Metaheuristic algorithms are ideal for finding energy-efficient routes for vehicles, ships, or transportation networks. Using strategies such as exploration and exploitation, these algorithms determine paths that minimize fuel consumption by taking into account various factors such as distance, traffic conditions, and energy efficiency.

When optimizing fleet operations, metaheuristics help determine the optimal vehicle configuration, scheduling, and resource allocation to minimize fleet energy consumption.

Metaheuristic algorithms help design energy-efficient vehicles by optimizing engine performance, aerodynamics, and vehicle weight, resulting in improved fuel efficiency and reduced energy consumption.

Benefits and Impact.

Metaheuristics can solve a wide range of optimization problems, providing tailored solutions in dynamic transportation systems.

These algorithms efficiently explore large solution spaces, providing near-optimal solutions to complex optimization problems.

Metaheuristics facilitate real-time decision making, enabling rapid response to changing conditions in transportation operations.

Optimization of metaheuristic parameters is critical for achieving optimal convergence rates and solution quality.

Metaheuristics can be computationally intensive, requiring optimization for scalability and efficiency.

Combining multiple metaheuristics or integrating them with additional optimization techniques can increase their effectiveness.

As technology advances, metaheuristic algorithms continue to evolve. Future advances will address computational complexity, increase convergence, and integrate these algorithms with new technologies to improve transportation efficiency and sustainability.

Metaheuristic algorithms are innovative tools for transforming the management of energy consumption in transportation systems. Their application in route optimization, fleet management, and vehicle design allows us to look into a future where transportation becomes not only efficient, but also environmentally conscious. By using high-level strategies to explore decision spaces, metaheuristics guide us to reduce environmental impact and optimize energy consumption, contributing to a more sustainable and efficient transportation ecosystem.

11. Heuristic Algorithms

Known for their simplicity and efficiency, heuristic algorithms serve as indispensable tools for solving optimization problems, including energy management in transportation systems. These algorithms provide practical and intuitive solutions to minimize fuel consumption, optimize routes, and reduce environmental impact, making them a valuable asset in the quest for efficient and environmentally friendly transportation.

Heuristics are problem-solving approaches that aim to find near-optimal solutions in a reasonable amount of time. They emphasize speed and practicality over guarantees of finding the absolute best solution, making them well suited to complex and dynamic systems such as transportation.

Heuristic algorithms excel at finding good enough routes for vehicles, ships, or transportation networks. Using intuitive rules and strategies, these algorithms determine the paths that minimize fuel consumption, taking into account factors such as distance, traffic conditions, and energy efficiency.

When optimizing fleet operations, heuristics help determine efficient vehicle configurations and plan and allocate resources to minimize fleet energy consumption.

The heuristic facilitates the efficient allocation of resources such as fuel and time among transportation networks, delivery schedules, and vehicle routes to minimize energy consumption and increase overall efficiency.

Benefits and Impact.

Heuristics provide simple solutions that are easy to implement and interpret, making them valuable for real-world applications.

These algorithms are fast, providing practical solutions in a reasonable time frame for dynamic transportation systems.

Heuristics adapt to changing conditions and uncertainties, making them suitable for rapid decision making in transportation operations.

Heuristics cannot always guarantee the best solution, but focus on acceptable, near-optimal solutions.

Trade-offs: The balance between solution quality and computational efficiency requires careful consideration.

Combining different heuristic approaches or integrating them with other optimization methods can increase their effectiveness.

As technology advances, heuristic algorithms continue to evolve and find new applications. Future developments aim to eliminate limitations, improve the quality of solutions, and integrate heuristics with new technologies to improve transportation efficiency and sustainability.

Heuristic algorithms serve as pragmatic tools to revolutionize energy management in transportation systems. Using intuitive rules and practical strategies, heuristics guide us to reduce environmental impact and optimize energy consumption, laying the foundation for a more sustainable and efficient transportation environment.

12. Rationale for Research Methods

Scientific research is the systematic analysis of phenomena and processes, studying their influence of various factors and interactions in order to arrive at convincing and useful solutions for science and practice. Research methods include the use of induction and deduction, analysis, synthesis, and comparison of both theoretical and practical aspects.

In this case, the theory explores algorithms for optimizing energy consumption in transportation systems, including their characteristics, principles of operation, possible implementations, advantages and disadvantages to improve system efficiency.

There are several research methods, but in this case an empirical approach was chosen to compare different algorithms for optimizing energy consumption in transportation systems. This method is the most appropriate because it requires real measurements. It allows us to determine which algorithms work better in practice and to determine their relative effectiveness in research.

The methodology of this study is a combination of methods used to describe the research. The main method chosen was the logical method of cognition, which is used to solve problems analytically, explain events and phenomena, describe problems and identify ways to solve them in empirical and theoretical tasks.

13. Comparison methods for energy saving criterion

Linear programming:

The energy efficiency formula for linear programming can focus on reducing energy consumption relative to the baseline or initial energy consumption.

$$Energy\_Eff = \\ = \frac{Init\_Energy\_Cons - Final\_Energy\_Cons}{Init\_Energy\_Cons} \times 100\% .$$

Reduce environmental impact:

$$Envir\_Imp\_Reduct(LP) = \\ = Init\_Imp(LP) - Final\_Imp(LP) .$$

Criteria: Operational performance

Execution Time: Evaluate the time it takes each algorithm to solve a given problem.

Record the time in milliseconds or seconds that the algorithms take to complete their tasks.

Solution Quality: Evaluate the quality or optimality of the solutions generated by each algorithm.

Define a quantitative quality metric specific to the problem domain (e.g., distance traveled, fuel consumption, etc.) or use objective metrics (minimization/maximization).

Composite metric: Create a composite metric that includes both lead time and solution quality.

Weight the metrics according to their relative importance.

$$Operat\_Perform = \frac{w1 \times Exec\_Time + w2 \times Solut\_Qual}{w1 + w2},$$

where $w1$, $w2$ represent the weights assigned to execution time and solution quality, respectively.

Measure and record the execution time of each algorithm for different problem sizes or scenarios.

Evaluate the quality of the solutions produced by each algorithm based on predefined metrics.

Combine execution time and solution quality using a composite metric formula to obtain an overall operational performance score for each algorithm.

Compare the aggregate scores of all algorithms to determine which algorithms perform better in terms of operational performance. This comprehensive evaluation helps you select the most effective algorithm(s) based on time efficiency and solution quality.

Scalability and Adaptability.

Scalability: Measure how algorithm performance changes with increasing problem size.

Evaluate runtime or memory consumption as problem size or complexity increases.

Adaptability: Evaluate how well the algorithm handles changes or variations in the problem without significantly degrading performance.

Test the performance of the algorithm in different scenarios or problem variations.

Applications: Linear programming, genetic algorithms, ant colony optimization, particle swarm optimization, annealing simulation, dynamic programming, heuristic algorithms, metaheuristic algorithms:

Evaluate algorithm performance metrics (execution time, memory usage) for problems of varying size or complexity.

Track how these metrics change as the problem scales, indicating the scalability of each algorithm.

Test the adaptability of the algorithms by making variations or changes to the problem parameters and observing how well they handle these changes without significantly degrading performance.

Analyze and compare the scalability and adaptability of each algorithm based on observed changes in performance as the size or complexity of the problem increases or under different variations of the problem scenarios. This evaluation will help you determine which algorithms scale well and adapt effectively to different situations.

Real-time implementation

Binary evaluation: Determine if the algorithm can satisfy real-time constraints.

Assign a binary value: 1 if the algorithm can be implemented in real time, 0 if not.

Applications: Linear programming, genetic algorithms, ant colony optimization, particle swarm optimization, annealing simulation, dynamic programming, heuristic algorithms, metaheuristic algorithms:

Estimate the execution time of each algorithm under different scenarios or problem sizes.

Set a threshold or benchmark for real-time implementation (e.g., execution time less than a certain limit is considered real-time).

If an algorithm's execution time consistently meets the defined threshold across all scenarios, mark it as real-time (1) or not (0).

Compare the binary score of each algorithm to determine its suitability for real-time implementation. Algorithms with a "1" are suitable for real-time execution, while algorithms with a "0" may not meet real-time constraints. This comparison will help you identify algorithms that are suitable for real-time applications.

Computing requirements for real-time implementation.

Time complexity: Measure the time complexity of an algorithm, typically expressed in Big O notation, to understand how its execution time grows with the size of the input data.

Space complexity: Estimate the space requirements of an algorithm by specifying the memory or storage it consumes as the problem size increases.

Applications: Linear programming, genetic algorithms, ant colony optimization, particle swarm optimization, annealing simulation, dynamic programming, heuristic algorithms, metaheuristic algorithms:

Analyze the time complexity of each algorithm, determining its efficiency relative to the size of the input.

Estimate the space complexity by understanding the memory or storage requirements as the problem grows.

Express the time and space complexity for each algorithm using Big O notation or appropriate mathematical expressions.

Compare the time and space complexity of the algorithms to determine their computational requirements for real-time implementation. Algorithms with lower time and space complexity (e.g., lower Big O values) are generally more suitable for real-time implementation in transportation systems due to their efficient use of resources. This comparison will help to identify algorithms suitable for real-time implementation.

The cost-effectiveness evaluation of algorithms involves evaluating their cost-effectiveness in achieving the desired improvements. Here is an approach to comparing algorithms based on the cost-effectiveness criterion:

Criteria: Cost Effectiveness

Cost: Estimate the costs associated with implementing and running each algorithm. This may include initial setup costs, computing resources, and maintenance costs.

Improvement Achieved: Measure the improvements or benefits achieved by applying the algorithm, such as reduced energy consumption, optimized decisions, or minimized operational costs.

Cost Effectiveness: Calculate the cost-effectiveness ratio, which indicates the cost-effectiveness of the algorithm in achieving the improvements:

$$Econom\_Eff = \frac{Cost}{Improv\_achiev}.$$

Applications: Linear programming, genetic algorithms, ant colony optimization, particle swarm optimization, annealing simulation, dynamic programming, heuristic algorithms, metaheuristic algorithms:

Estimate the cost of implementing and maintaining each algorithm in a given scenario or problem domain.

Measure the improvements achieved by applying each algorithm by quantifying the benefits or optimizations gained.

Calculate the cost-effectiveness ratio using a formula for each algorithm, taking into account the ratio of costs incurred to improvements achieved.

Compare the cost-effectiveness ratios of the algorithms to determine which algorithms provide the best cost-effectiveness in terms of improvements. Algorithms with lower cost-effectiveness ratios, indicating greater improvements at lower cost, are considered more cost-effective. This comparison will help you select the algorithms that provide the best balance of cost and benefit.

Applications: Linear programming, genetic algorithms, ant colony optimization, particle swarm optimization, annealing simulation, dynamic programming, heuristic algorithms, metaheuristic algorithms:

Estimate the cost of implementing and maintaining each algorithm in a given scenario or problem domain.

Measure the improvements achieved by applying each algorithm by quantifying the benefits or optimizations gained.

Calculate the cost-effectiveness ratio for each algorithm using a formula that takes into account the ratio of costs incurred to improvements achieved.

Compare the cost-effectiveness ratios of the algorithms to determine which algorithms provide the best cost-effectiveness in terms of improvements. Algorithms with lower cost-effectiveness ratios, indicating greater improvements at lower cost, are considered more cost-effective. This comparison will help you select the algorithms that provide the best balance of cost and benefit.

Durability and reliability.

Robustness: Measures the ability of an algorithm to consistently produce correct and reliable results across different scenarios or data sets.

Stability: Evaluate the stability of an algorithm by checking how sensitive it is to changes in input data or parameters. A stable algorithm provides consistent performance despite variation.

Applications: Linear programming, genetic algorithms, ant colony optimization, particle swarm optimization, annealing simulation, dynamic programming, heuristic algorithms, metaheuristic algorithms:

Run multiple trials with different data sets or scenarios to evaluate the consistency of results produced by each algorithm.

Introduce variations or perturbations in the input parameters to evaluate the stability of the algorithms.

Quantify reliability and stability metrics for each algorithm based on observed behavior, error rates, or deviations from expected results.

Comparative analysis: Compare the reliability and stability scores of algorithms to determine which ones consistently produce reliable results across different scenarios or data sets and exhibit stable behavior in response to changes.

Algorithms with higher consistency, lower error rates, and less sensitivity to input variations are considered stronger and more reliable.

This comparison helps identify algorithms that consistently produce reliable results and are less prone to bias or error, highlighting their strength and reliability.

Regulatory compliance.

Evaluate the result: Evaluate the results or solutions generated by algorithms against regulatory standards or constraints. This may include ensuring that solutions meet certain legal or security requirements.

Industry Standards: Analyze the extent to which the algorithm's results are consistent with industry guidelines, regulations, or best practices. For example, in transportation systems, algorithms must comply with safety protocols or environmental regulations.

Applications: Linear programming, genetic algorithms, ant colony optimization, particle swarm optimization, annealing simulation, dynamic programming, heuristic algorithms, metaheuristic algorithms:

Examine the results or solutions produced by each algorithm in the context of the regulatory requirements applicable to the transportation system or related domain.

Verify that the solutions provided by the algorithms comply with established regulations, safety standards, or industry norms.

Quantify the level of compliance achieved by each algorithm based on the alignment of its results with regulatory requirements.

Benchmark: Compare the level of compliance demonstrated by each algorithm to determine which produce results that better meet regulatory requirements or industry standards.

Algorithms that produce solutions that are closer to the required regulations or standards are considered more compliant.

This comparison helps to assess the degree to which the results of each algorithm meet the required regulatory requirements or industry standards in the area of transportation systems.

Convenience and Integration.

Ease of implementation: Evaluate the ease and simplicity of implementing each algorithm into existing systems or frameworks.

Compatibility: Evaluate how well the algorithm integrates with different platforms, technologies, or software architectures without requiring significant modifications.

Applications: Linear programming, genetic algorithms, ant colony optimization, particle swarm optimization, annealing simulation, dynamic programming, heuristic algorithms, metaheuristic algorithms:

Analyze the implementation process for each algorithm, taking into account the ease of adaptation into existing systems. This may include assessing the complexity of code integration or program dependencies.

Evaluate the compatibility of the algorithms with different software architectures or platforms. Algorithms that can be easily integrated with minimal customization are more convenient.

Quantify the level of usability and integration for each algorithm based on implementation complexity and compatibility metrics.

Benchmark: Compare the usability and integration scores of algorithms to determine which offer smoother integration processes and better compatibility with existing systems.

Algorithms with higher usability and integration scores, indicating easier implementation and seamless integration, are considered more usable and compatible.

This comparison will help select algorithms that are easier to implement and integrate into transportation systems or related structures, reducing the complexity of adoption and ensuring smooth integration.

User Adaptability.

User interface and interaction:

Evaluate the accessibility and usability of interfaces or tools associated with the implementation of these algorithms.

User training and support: Evaluate the ease of learning and using the algorithms, including the availability of documentation, tutorials, or support materials.

Applications: Linear programming, genetic algorithms, ant colony optimization, particle swarm optimization, annealing simulation, dynamic programming, heuristic algorithms, metaheuristic algorithms:

Evaluate the interfaces or tools provided with each algorithm, considering their intuitiveness, simplicity, and ease of use.

Analyze the availability and quality of support materials (documentation, tutorials, etc.) to facilitate user understanding and implementation.

Quantify user adoption for each algorithm based on UI usability metrics and the availability of comprehensive support materials.

Comparative analysis: Compare the usability scores of algorithms to determine which algorithms offer more user-friendly interfaces and better support resources.

Algorithms with higher usability scores, indicating easier-to-use interfaces and comprehensive support materials, are considered more user-friendly.

This comparison will help select algorithms that provide users with interfaces and resources that are easy to understand, learn, and implement, thereby increasing overall user adaptability.

Conclusions

In the course of this task, the subject area was analyzed and algorithms for optimizing energy consumption in transportation systems were considered. The advantages and disadvantages of the algorithms were described and evaluation criteria were proposed.

As a result, a research report was prepared, which included the formulated measurable criteria for comparing the algorithms, described in detail and argued the use of each of them and when they could be neglected. The key metrics chosen to compare the technologies were

– Energy savings;
– Reduction of environmental impact;
– Operational efficiency;
– Scalability and adaptability;
– Real-time implementation;
– Computational requirements of the algorithm for real-time implementation in transportation systems;
– Economic efficiency;
– Durability and reliability;
– Regulatory compliance;
– Ease of use and integration;
– User adaptability.

In the current study, we proposed relevant ways to measure each of the presented metrics, formulated criteria, and mathematical formulas used to calculate the numerical values of these metrics.

Conflict of Interest

The authors declare no conflict of interest.

References

[1] U.S. Bureau of Transportation Statistics. 2018. URL: https://www.bts.gov/transportation-economic-trends/tet-2018-chapter-2-contributioneconomy

[2] European Union. Statistical Pocketbook 2017: EU Transport in Figures; Publications Office of the European Union: Brussels, Belgium, 2017; ISBN 978-992-79-62311-0.

[3] Chen, G.; Wu, X.; Guo, J.; Meng, J.; Li, C. Global overview for energy use of the world economy: Household-consumption-based accounting based on the world input-output database (WIOD). Energy Econ. 2019, 81, 835–847.

[4] United Nations Department of Economics and Social Affairs. 2019. URL: https://population.un.org/wpp/ Download/ Standard/Population/

[5] Bektaş, T.; Ehmke, J.F.; Psaraftis, H.N.; Puchinger, J. The role of operational research in green freight transportation. Eur. J. Oper. Res. 2019, 274, 807–823.

[6] Juan, A.; Mendez, C.; Faulin, J.; De Armas, J.; Grasman, S. Electric vehicles in logistics and transportation: A survey on emerging environmental, strategic, and operational challenges. Energies 2016, 9, 86.

[7] Fan, Y.V.; Klemeš, J.J.; Walmsley, T.G.; Perry, S. Minimising energy consumption and environmental burden of freight transport using a novel graphical decision-making tool. Renew. Sustain. Energy Rev. 2019, 114, 109335.

[8] Dekker, R.; Bloemhof, J.; Mallidis, I. Operations Research for green logistics–An overview of aspects, issues, contributions and challenges. Eur. J. Oper. Res. 2012, 219, 671–679.

[9] Faulin, J.; Grasman, S.E.; Juan, A.A.; Hirsch, P. Sustainable Transportation: Concepts and Current Practices. In Sustainable Transportation and Smart Logistics; Elsevier: Amsterdam, The Netherlands, 2019; pp. 3–23.

[10] Neumann, F.; Witt, C. Combinatorial optimization and computational complexity. In Bioinspired Computation in Combinatorial Optimization; Springer: Berlin, Germany, 2010; pp. 9–19.

[11] Glover, F.W.; Kochenberger, G.A. Handbook of Metaheuristics; Springer Science & Business Media: Berlin, Germany, 2006; Volume 57.

[12] Psaraftis, H.N.; Kontovas, C.A. Speed models for energy-efficient maritime transportation: A taxonomy and survey. Transp. Res. Part C Emerg. Technol. 2013, 26, 331–351.

[13] Ríos-Mercado, R.Z.; Borraz-Sánchez, C. Optimization problems in natural gas transportation systems: A state-of-the-art review. Appl. Energy 2015, 147, 536–555.

[14] Yang, X.; Li, X.; Ning, B.; Tang, T. A survey on energy-efficient train operation for urban rail transit. IEEE Trans. Intell. Transp. Syst. 2015, 17, 2–13.

[15] Tranfield, D.; Denyer, D.; Smart, P. Towards a methodology for developing evidence-informed management knowledge by means of systematic review. Br. J. Manag. 2003, 14, 207–222.

The article was delivered to editorial stuff on the 4.05.2023

УДК 004:629:656:658 DOi 10.30837/ bi.2023.1(99).11

Б. С. Карпішен[1], С. М. Неронов[2], Г. А. Плєхова[3], М. В. Костікова[4], С. О. Петренко[5], О. О. Яценко[6]

[1]ХНАДУ, м. Харків, Україна, karpishen.bogdan@gmail.com, ORCID iD: 0009-0001-1790-9048

[2]ХНАДУ, м. Харків, Україна, sernikner@gmail.com, ORCID iD: 0000-0003-2381-1271

[3]ХНАДУ, м. Харків, Україна, plehovaanna11@gmail.com, ORCID iD: 0000-0002-6912-6520

[4]ХНАДУ, м. Харків, Україна, kmv_topaz@ukr.net, ORCID iD: 0000-0001-5197-7389

[5]rock11002@gmail.com

[6]iatsenkooleksii@icloud.com

МОДЕЛЬ ІНФОРМАЦІЙНО-КОМУНІКАЦІЙНОЇ СИСТЕМИ

У роботі запропоновано систему зв'язку в задачі попередження про можливе зіткнення. Для реалізації бездротової передачі та прийому інформації між автомобілями була використана технологія V2V. Результатом роботи є отримання інформації від підключених транспортних засобів в зоні дії системи та реакція інформаційної системи на оброблені дані у вигляді зовнішніх сигналів. Модель демонструє роботу системи за принципом «приймач-передавач» з використанням DSRC-зв'язку за допомогою програмного забезпечення MATLAB/ Simulink. Отримана модель дозволяє проводити різноманітні системні аналізи для подальшого вдосконалення інформаційно-комунікаційних систем в автомобілях.

V2V, ADAS, МОДЕЛЬ, СИСТЕМА, DSRC, ДЕЦЕНТРАЛІЗОВАНІ СХОВИЩА ДАННИХ,СТИСНЕННЯ ЗОБРАЖЕНЬ

B. S. Karpishen, S. M. Neronov, G. A. Pliekhova, M. V. Kostikova, S. O. Petrenko, O. O. Iatsenko. Model of the information and communication system. The paper proposes a communication system in the task of warning about a possible collision. V2V technology was used to implement wireless transmission and reception of information between cars. The result of the work is receiving information from connected vehicles in the area of the system and the response of the information system to the processed data in the form of external signals. The model demonstrates the operation of the transceiver system using DSRC communication using MATLAB/Simulink software. The resulting model allows for various system analyzes for further improvement of information and communication systems in cars.

V2V, ADAS, MODEL, SYSTEM, DSRC DECENTRALIZED DATA STORAGE, IMAGE COMPRESSION

Вступ

Стрімкий розвиток транспортних систем приніс багато зручностей у наше повсякденне життя, дозволяючи безпечно і надійно перевозити людей і вантажі всередині країни та за її межами. За оцінками, у світі налічується понад мільярд автомобілів, якими володіють люди. Передбачається, що ця кількість подвоїться протягом одного-двох десятиліть. Однак ряд питань, пов'язаних з цим зростанням, викликають занепокоєння. З точки зору безпеки, у 2021 році в аваріях на дорогах США загинуло понад 42 915 осіб (USDOT, 2023).

Зв'язок між декількома підключеними транспортними засобами (V2V) підвищує безпеку та ефективність наших транспортних систем.

Це досягається завдяки використанню систем управління дорожнім рухом, які покладаються на бортові датчики та зв'язок між транспортними засобами (V2V). Зв'язок в основному надає інформацію про стан (наприклад, прискорення, швидкість, місцезнаходження) переднього транспортного засобу або транспортних засобів у реальному часі (Z. Wang та ін., 2020).

Оцінки щодо важливості цієї технології:

Підвищення безпеки на дорогах: V2V-зв'язок дозволяє автомобілям обмінюватися інформацією про своє місцезнаходження, швидкість та інші параметри в режимі реального часу. Це допомагає уникнути аварій, зменшити кількість зіткнень та покращити реакцію на небезпеку. Наприклад, система може попереджати водіїв про можливі ризики, такі як аварійна ситуація на дорозі, перешкоди або небезпечний обгін.

Зменшення заторів та покращення транспортного потоку: Технологія V2V допомагає водіям обирати оптимальний маршрут і швидкість, щоб уникнути заторів. Це може покращити транспортний потік і скоротити час у дорозі.

Зменшення аварійності: Завдяки V2V-комунікації автомобілі можуть обмінюватися інформацією про небезпечні дорожні умови, такі як слизький асфальт, обмежена видимість або погана погода. Це допомагає водіям адаптувати свій стиль водіння до конкретних умов і знижує ризик аварій (H. Xie et al., 2022).

Попит на вдосконалені системи допомоги водієві (ADAS) (SAE, 2021) – ті, що допомагають виконувати завдання моніторингу, попередження, гальмування та керування – зростатиме протягом наступного десятиліття, що значною мірою зумовлено інтересом регуляторних органів та споживачів до програм безпеки, які захищають водіїв та зменшують кількість аварій.

В даний час системи ADAS розглядаються як галузь, що постійно розвивається, яка продовжує вдосконалюватися і розвиватися. Хоча на ринку вже існує багато різних систем ADAS, вони постійно вдосконалюються і доповнюються новими функціями. На етапі розробки системи виникають різні виклики і проблеми, такі як взаємодія між системами, безпека даних, відповідність нормативним вимогам, інтеграція з іншими системами і компонентами автомобіля, інтелектуальне управління даними та інші.

Розробка систем ADAS є складним і багатогранним процесом, який вимагає детального вивчення різних аспектів, від технічних можливостей до соціально-економічних і правових аспектів.

У цьому дослідженні ми зосереджуємося на послугах, які створюють або пов'язані з одноразовими повідомленнями, особливо з використанням платформи V2V-комунікацій.

1. Огляд літератури

На основі літературних джерел ми представляємо різні застосування V2V-зв'язку, класифіковані за двома широкими цілями: цілі безпеки та цілі, що не пов'язані з безпекою. Основними цілями в цій категорії є мінімізація проблем безпеки шляхом надання водієві вказівок або іншої інформації для запобігання або прогнозування дорожньо-транспортних пригод, таких як перед- або післяаварійні ситуації, прогнозування сліпих зон, допомога при перетині перехресть і т. д.

Обмін повідомленнями між транспортними засобами має на меті мінімізувати потенційні аварії та підвищити безпеку водіння за допомогою функцій допомоги водієві як для автономних, так і для неавтономних транспортних засобів. Крім того, зв'язок V2V посилить підтримку безпеки в п'ятирівневій автономії в автономних транспортних засобах, де поєднання штучного інтелекту, технологій транспортних засобів, Інтернету речей та комунікаційних можливостей прискорить масове впровадження автономних транспортних засобів у майбутньому (Ketut, 2021) .

Сьогодні завдяки зв'язку V2V за допомогою таких технологій, як DSRC на базі IEEE 802.11p і нових рішень 5G, стають можливими різні IFT, оскільки транспортний засіб може спілкуватися з транспортними засобами за межами свого безпосереднього оточення (SE Li et al., 2017).

Якщо брати характеристики самого каналу передачі, то існує можливість покращення зв'язку шляхом втручання в структуру протоколу 802.11p, яка охоплює поле, що має значення розміру інформаційного пакету. Робота (Yasser et al., 2021) зосереджена на розробці потужно адаптивної структури розміру пакетів, яка залежить від значення відношення сигнал/шум. Існують нейромережеві контролери, які навчаються за потенційними значеннями розміру пакета, що виводяться з рівняння, отриманого шляхом практичного тестування множинної залежності між частотою помилок і розміром пакета (J. Ploeg et al., 2015). Регулювання розміру інформаційного пакета призводить до зменшення частоти помилок при передачі пакетів.

У статті (Li, Keqiang & Bian та ін., 2020) запропоновано метод предиктивного управління на основі розподіленої моделі (DMPC) для управління системами з декількома транспортними засобами в комутованих топологіях зв'язку. Сформульовано оптимізаційну задачу з розімкненим циклом, до якої включено штрафи та обмеження на відхилення сусіда та самовідхилення для забезпечення стабільності. Алгоритм DMPC розроблено для систем з декількома транспортними засобами, що мають топологію зв'язку, яка комутується. В результаті створено контролер управління системою з декількома транспортними засобами в комутаційних топологіях зв'язку.

Наприклад, метою розробки адаптивного управління сигналами на основі людини (PB-ACA) було дослідження оптимальних планів сигналів на ізольованому з'єднанні. У цьому дослідженні, як децентралізоване координоване управління, алгоритм координованого управління сигналами на основі людини (C-PBC) дозволяє локальному контролеру на кожному перехресті в регіоні дорожньої мережі самостійно керувати PB-ACA на основі даних підключеного транспортного засобу в межах бездротового діапазону. для оптимізації планів сигналів на основі персоналу. Для кожного перехрестя діапазон зв''язку визначається як коло радіусом 250 м від центру перехрестя, і запланованe перехрестя може отримувати дані тільки в межах цього діапазону зв'язку (Zongyuan Wu, Ben Waterson & Bani Anvari, 2022) .

Зі стрімким розвитком технологій бездротового зв'язку поведінка водія перед автомобілем може передаватися на наступний транспортний засіб для покращення роботи системи.

У цій статті пропонується система попередження зіткнення попереду (FCW), яка виявляє наміри водія попереднього транспортного засобу і передає інформацію наступному транспортному засобу за допомогою технологій зв'язку V2V. Запропонований метод розпізнавання намірів водія забезпечує кращу продуктивність системи FCW і дає наступному транспортному засобу додатковий час для плавного гальмування (W. Yang, B. Wan and X. Qu, 2020).

У всіх цих випадках моделювання є важливим етапом розробки та тестування систем. У статті (Mo, Chunmei & Li, Yinong & Ling, Zheng, 2018) описано математичну модель аналізу V2V, обгону та мінімальної дистанції обгону з використанням нечіткої логіки. Експерименти були розроблені за допомогою PreScan / MATLAB. Модель показує ефективність створеного алгоритму.

У цій роботі система попередження зіткнень була створена з використанням інформаційно-комунікаційної моделі передачі даних на основі технології бездротового зв'язку за допомогою MATLAB / Simulink (The MathWorks, Inc., 2023) [1].

2. Матеріали та методи

Технологія Wi-Fi, відома як Dedicated Short Range Communication (DSRC) між кожним транспортним засобом, і технологія GPS, яка забезпечує детальне позиціонування шляхом обміну даними з аналогічно обладнаними транспортними засобами. DSRC – це спеціальний засіб зв'язку, призначений для транспортного засобу, який забезпечує зв'язок на невеликій відстані з сусіднім транспортним засобом або з навколишнім середовищем для досягнення спільної ситуації під час руху. DSRC використовує спектр 75 МГц для автомобільного зв'язку і використовує радіотехнологію на основі IEEE 802.11p з пропускною здатністю від 3 до 27 Мбіт/с (J. B. Kenney, 2011). Для забезпечення зв'язку V2V потрібно кілька компонентів:

1. DSRC – це спеціальна радіостанція, яка працює як приймач і передавач даних.

2. GPS-приймач (Cheng et al., 2007), що відповідає за визначення положення автомобіля у просторі та часі; ці дані будуть вхідними даними для DSRC.

3. OBU (On-Board Unit), який збирає дані про стан автомобіля, такі як швидкість, кут повороту керма, прискорення, стан гальм тощо. Він також встановлює додаток і екран для відображення інформації інтерфейсу.

Ми моделюємо комунікацію (V2V) і будуємо модель за принципом «приймач-передавач». Модель є невід'ємною частиною проекту і має заздалегідь підготовлену «сцену» (модель дорожнього покриття з перешкодами для передавача), «сценарій» (маршрут руху) та «акторів» (транспорт з визначеними характеристиками). У моделі використовується базове повідомлення DSRC про безпеку радіопередачі (BSM) (SAE J2735, 2022).

Комунікація покладається на характеристики каналу для визначення ймовірності успішного отримання повідомлення.

Підсистема Transmitter V2V генерує базове повідомлення безпеки (BSM) для кожного цільового транспортного засобу, використовуючи отриману інформацію про цього «актора». Передавач зчитує інформацію про «актора» і передає її через інерціальну навігаційну систему (INS) і глобальну навігаційну супутникову систему (GNSS) для накладення шуму на інформацію про «актора». Підсистема також трансформує інформацію про просторове розташування транспортних засобів-мішеней з декартових координат в географічні, використовуючи інформацію про «сцену». Потім підсистема генерує BSM для всіх транспортних засобів-мішеней. Блок SendMessage всередині передавача перетворює сигнал у Simulink-повідомлення і доставляє його до черги об'єктів. Черги організовані за принципом «перший прийшов – перший вийшов» (FIFO).

Підсистема приймача V2V реалізує поведінку приймача об'єкта транспортного засобу. Приймач отримує попередньо розраховані характеристики каналу як параметр маски та передану інформацію про BSM, сцену та автомобіль як вхідні дані. Коли передавач доставляє повідомлення до черги об'єктів, він запускає підсистему приймача V2V. Для кожного автомобіля-мішені приймач обчислює відстань від нього до свого автомобіля, а потім знаходить відповідну смугу пропускання, використовуючи попередньо розраховані характеристики каналу.

Коли пропускна здатність перевищує згенероване випадкове число, приймач отримує BSM і зберігає його на вихідній шині BSMOut. Черга FIFO моделює інтерфейс прийому повідомлень, який працює на основі доступності повідомлень.

Отримане повідомлення далі передається на блок обробки повідомлень, який перетворює вхідні дані BSM у фізичні величини і формує звіти про виявлення об'ктів для вхідних даних для відстеження декількох об'єктів (Карпишен Б. С., 2023).

Модель комунікації між транспортними засобами надана на рис. 1.

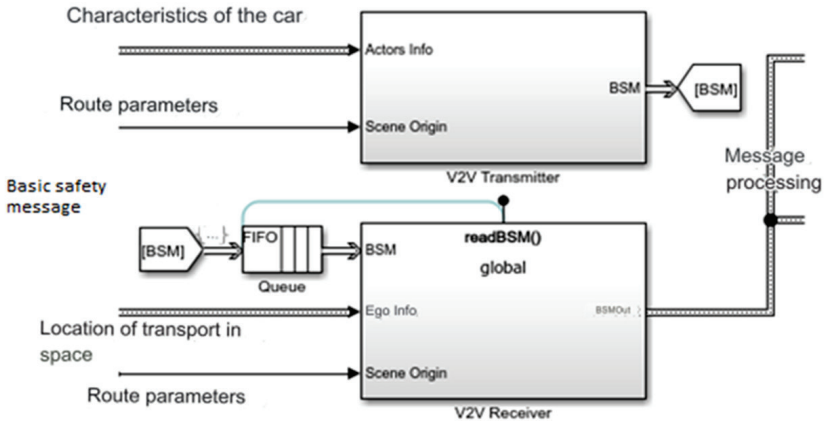

Рис. 1. Модель комунікації між транспортними засобами

3. Результати

Результати моделі показують залежність між відстанню від передавача до приймача та співвідношенням сигнал/шум (SNR) (Hasan Farahneh et al., 2020) для різних діапазонів передачі. У цьому прикладі порівнюється різниця в 50 і 150 м.

Залежність відстані та відношення сигнал/шум представлена на рис. 2.

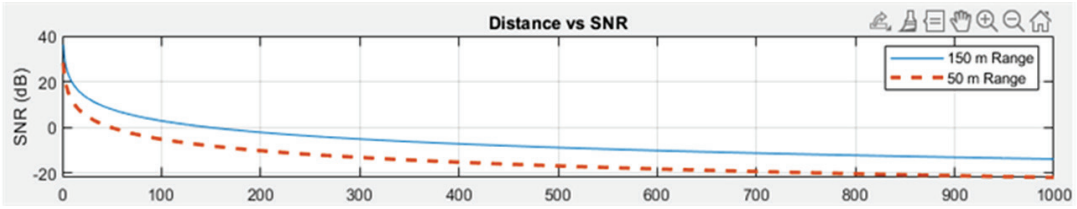

Рис. 2. Залежність відстані та відношення сигнал/шум

Ми також можемо побачити зв'язок між відстанню і пропускною здатністю для вказаного діапазону. Пропускна здатність означає очікувану ймовірність виявлення пакету. Коли дальність становить 150 м, графік показує, що ймовірність виявлення пакету майже 100 % до 150 м, а потім вона поступово зменшується, поки не досягне 0 % приблизно на 400 м, ймовірність виявлення пакету зменшується швидше і на відстані 150 м наближається до 0 %.

Залежність від пропускної здатності представлена на рис. 3.

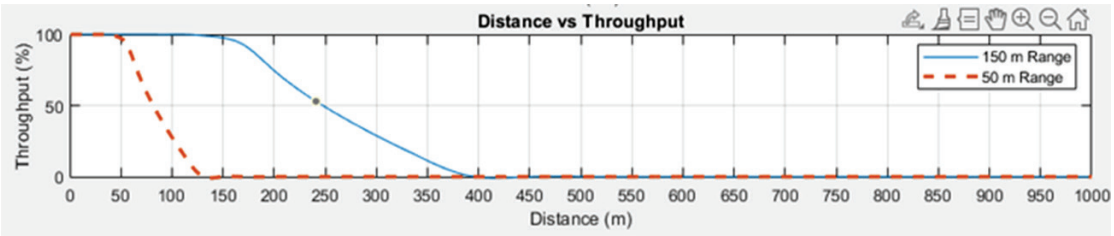

Рис. 3. Залежність від пропускної здатності

Під час роботи модель візуалізує важливі дані та видає наступну інформацію:

Співвідношення переданих та отриманих повідомлень, яке відображає кількість переданих та отриманих повідомлень на кожному часовому кроці.

Дані зв'язку V2V – відображає інформацію про передачу та прийом даних BSM та співвідношення сигнал/шум для кожного отриманого повідомлення.

Отримане BSM-повідомлення - показує широту, довготу, швидкість, курс, довготу і широту для кожної цілі, від якої отримано BSM-повідомлення.

Візуальне відображення отриманих та переданих повідомлень від «акторів» представлено на рис. 4.

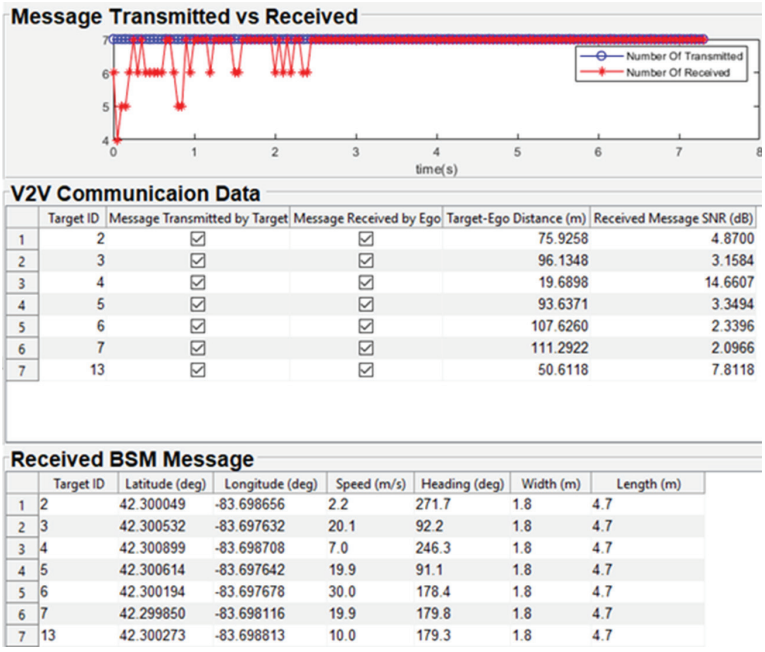

Рис. 4. Візуальне відображення отриманих та переданих повідомлень від «акторів»

Після отримання BSM для генерування попередження про можливе зіткнення розраховуються траєкторії руху цільового та сусідніх транспортних засобів з використанням їх поточного положення, швидкості та курсових кутів. Розрахункова траєкторія кожного транспортного засобу – це пряма лінія, що з'єднує початкове положення транспортного засобу та його прогнозоване положення через 20 секунд. Щоб оцінити ризик зіткнення, аналізатор перевіряє, чи перетинається передбачувана траєкторія сусіднього транспортного засобу з передбачуваною траєкторією цілі. У разі перетину аналізатор попередження зіткнення обчислює час прибуття «актора» в точку перетину і визначає абсолютну різницю між часом прибуття обох транспортних засобів в точку перетину.

Далі час прибуття «актора» і значення часового інтервалу порівнюються з відповідними заздалегідь визначеними пороговими значеннями. За результатами порівняння встановлюється відповідний рівень попередження.

Рівень попередження відповідно до часових умов наведений у табл. 1.

Таблиця 1

Рівень попередження відповідно до часових умов

Час прибуття «актора». Умова	Умова розриву в часі	Рівень попередження
ActArrrivalTime < minArrivalTime	timeGap < minTimeGap	Високий
ActArrrivalTime < minArrivalTime	timeGap >= minTimeGap	Помірний
ActArrrivalTime >= minArrivalTime	timeGap < minTimeGap	Низький
ActArrrivalTime >= minArrivalTime	timeGap >= minTimeGap	Низький

Середній та високий рівні попередження вимагають від водія обізнаності та реагування.

Підсумки та висновки

У цій роботі запропоновано систему зв'язку в задачі попередження про можливе зіткнення. Параметри зв'язку залежать від характеристик каналу для визначення ймовірності успішного отримання повідомлення. Для реалізації бездротової передачі та прийому інформації між автомобілями була використана технологія V2V. Результатом роботи є отримання інформації від підключених транспортних засобів в зоні дії системи та реакція інформаційної системи на оброблені дані у вигляді зовнішніх сигналів. Разом з тим, дослідження в даній роботі також прискорює використання технології V2V в області інтелектуальних транспортних засобів та покращує здатність інтелектуального транспортного засобу сприймати навколишнє середовище.

Що стосується підвищення ефективності систем допомоги водієві, то деякі рішення вже існують. У випадку з вирішенням задачі забезпечення передачі повідомлень для системи попередження фронтальних зіткнень були використані технології DSRC. Результати експериментів FCW показали, що система забезпечила більш раннє попередження, ніж попередній результат. Запропонована система не тільки надавала ранні попередження для запобігання зіткненням ззаду, але й сприяла більш ефективному гальмуванню (W. Yang, B. Wan and X. Qu, 2020).

Завдяки нещодавнім проривам у бездротових мережах управління, якість бездротового зв'язку можна контролювати передбачуваним чином (Zhang et al., 2014), що відкриває двері для спільного проектування бездротової автомобільної мережі (SE Li et al., 2017).

Завдяки спільному управлінню рухом декількох транспортних засобів, з'єднаних бездротовим зв'язком, можливі деякі або всі наступні переваги транспортної системи:

– Пропускна здатність дороги може бути збільшена за рахунок зменшення проміжків між транспортними засобами.

– Споживання енергії та викиди забруднюючих речовин можна зменшити за рахунок зменшення непотрібних змін швидкості та аеродинамічного опору транспортних засобів, що рухаються слідом.

– Потенційно підвищується безпека водіння, оскільки час виявлення та реагування скорочується порівняно з автомобілями з ручним керуванням.

– Комфорт споживачів можна покращити, оскільки поведінка системи краще реагує на зміни в дорожньому русі, а коротші інтервали руху можуть стримувати включення інших транспортних засобів (Z. Wang et al., 2020).

Модель демонструє роботу системи за принципом «приймач-передавач» з використанням DSRC-зв'язку за допомогою програмного забезпечення MATLAB/Simulink. Отримана модель дозволяє проводити різноманітні системні аналізи для подальшого вдосконалення інформаційно-комунікаційних систем в автомобілях.

Список літератури:

[1] Гоблик Н. М., Гоблик В. В. MATLAB в інженерних розрахунках. Комп'ютерний практикум. – Львів: Львівська політехніка, 2020.– 192 с.

Надійшла до редколегії 18.09.2023

УДК 004.912:81'374(043.3) DOi 10.30837/ bi.2023.1(99).12

Yevhen Kupriianov[1]

[1]NTU "KhPI", Kharkiv, Ukraine, Ukraine, eugeniokuprianov@gmail.com,
ORCID iD: 0000-0002-0801-1789

DEVELOPING SOFTWARE FOR COMPILING ELECTRONIC INFLECTIONAL DICTIONARY OF THE SPANISH LANGUAGE

The paper focuses on the technology of creating software for compiling an electronic inflectional Spanish dictionary using the framework of the L-systems theory by V.A. Shirokov. On its basis all Spanish words were classified into respective paradigmatic types, groups and classes is made, a formal model of the dictionary is built, and the structure of the database and the interface of the virtual lexicographic laboratory to work with the dictionary database are determined. The interface offers a number of functions, including adding, editing, and deleting words and paradigmatic classes. The developed database structure and data editing software tools contribute to the efficient organization of the process of creating a word-based Spanish dictionary. The created database can be successfully used in the study of inflection processes and phenomena.

L-SYSTEM, LEXICOGRAPHIC DATABASE, VIRTUAL LEXICOGRAPHIC LABORATORY, FORMAL MODEL

Купріянов Є. Розробка програмного забезпечення для укладання електронного словозмінного словника іспанської мови. Стаття присвячена технології створення програмного забезпечення для укладання електронного словозмінного словника іспанської мови з використанням апарату теорії Л-систем В.А. Широкова. На її основі вироблено словозмінну класифікацію іспанських мовних одиниць за парадигматичними типами, групами і класами, побудовано формальну модель словника, а також визначено структуру бази даних та інтерфейс віртуальної лексикографічної лабораторії для роботи з базою даних словника. Інтерфейс пропонує низку функцій, зокрема додавання, редагування та вилучення слів та парадигматичних класів. Розроблена структура бази даних та програмні засоби редагування даних сприяють ефективній організації процес створення словозмінного словника іспанської мови. Створена база даних може успішно використовуватись при дослідженні словозмінних процесів і явищ.

Л-СИСТЕМА, ЛЕКСИКОГРАФІЧНА БАЗА ДАНИХ, ВІРТУАЛЬНА ЛЕКСИКОГРАФІЧНА ЛАБОРАТОРІЯ, ФОРМАЛЬНА МОДЕЛЬ

Introduction

The modern period of computer linguistics development is marked by a shift in the research paradigm caused by factors unrelated to linguistics tasks, namely the rapid development of intellectual information and communication technologies, as well as natural language's rapid acquisition of technological status. These factors necessitate the development of appropriate linguistic resources that cover the widest possible range of language material and linguistic phenomena, which, in turn, requires the development of theoretical and linguistic basics for an integral description of the language system, oriented to use in computer linguistics and lexicography, as well as digital text information processing systems (machine translation, data and knowledge mining, conceptual and ontology-based).

These challenges present extremely important tasks for computer linguistics, particularly the creation of a universal system of digital lexicographic resources. The proceedings of the Ukrainian Language and Information Foundation of the National Academy of Sciences of Ukraine have made a substantial contribution to the problem of building integrated lexicographic objects based on formal models for Ukrainian and some other languages – the monographs "Information Theory of Lexicographic Systems", "Phenomenology of L-Systems", "Elements of Lexicography", "Computational Lexicography" (V. Shyrokov), "Linguistic and Technological Bases of Explanatory Lexicography" (V. Shyrokov, N. Zaika, et al.), "Grammatical Systems" (V. Shyrokov, I. Shevchenko, T. Liubchenko, K. Shyrokov), 5-volume set "Linguistic and Information Studies" (V. Shyrokov, et al.). These publications served as the scientific foundation for the creation of Ukraine's National Dictionary Base, the country's only linguistic property designated as a national treasure.

1. Related Works

As it is stated in [1], for information systems to function properly, the language is to be represented as a formal model. Prof. V. Shyrokov his team mates [2, 3] discuss in detail the problems of formal modeling of the inflectional system of a language and its representation in software tools, such as virtual lexicographic laboratories and electronic grammar dictionaries. The language system theory proposed by the researcher has been effectively applied to modeling the word change systems of Ukrainian, German, and other languages.

Many works [4-9] are focused on developing a database for different inflectional dictionaries. Among them are:

– **Database of Old Icelandic Inflections** (DOII) is a project that aims to describe the patterns of inflection in Old Icelandic through computer modeling, currently

underway at the Arni Magnusson Institute of Icelandic Studies (SÁM) at the University of Iceland. The inflection models form the basis of the database, and all headwords, regardless of which category they belong to, are assigned to one of them. Each noun inflection pattern consists of two main characteristics: stem type and case endings. The declined forms are entered manually on separate lines according to the inflectional structure.

– **Grammar Dictionary of the Polish Language** aims to provide the most complete description of the Polish language conjugation: a complete morphological characteristics and basic syntactic characteristics of Polish words; for each lexeme, all its declined forms are given with the meanings of all morphological categories (categories according to which the word is declined). In addition, the values of some syntactic features are given: gender for nouns, type for verbs, and required case for prepositions. Due to the large amount of data, the dictionary works using relational databases, which are a means of storing elements of the word-changing model. To build all the conjugated forms based on the dictionary data, other means, such as other tools, would be needed. The declined form in the developed model consists of a prefix, stem, ending, and suffix, which are controlled by several model objects. Each of these parts can be empty. However, since the mapping from endings to forms is universal, this complexity does not affect the process of adding new tokens or checking an existing description. These tasks can be successfully performed on a limited set of basic inflectional forms that are built from stem and ending only.

– **UniMorph** includes 23 meaning parameters and over 212 features. Meaning parameters are morphological categories such as person, number, tense, and mood. Each of them represents a coherent semantic space in inflectional morphology. They include: Participle, animation, aspect, case, comparison, definiteness, deixis, evidentiality, finiteness, gender, information structure, interrogative, mood, number, part of speech, person, polarity, politeness, switching, tense, valence, and voice. These dimensions contain a different number of features, from 2 for definiteness to 39 for case. Features represent the finest differences in meaning that are possible within a given dimension.

2. Method

All Spanish words are grouped into different paradigmatic types, i.e., sets of words with the same grammatical function and inflected according to the same sets of word-invariant parameters. Each paradigmatic type can cover several lexical and grammatical classes, i.e., sets of words united by common grammatical features. In turn, grammatical classes are divided into paradigmatic groups, i.e. groups of words representing a certain type of word-change paradigm (e.g. regular or irregular paradigm for verbs). Then, each paradigmatic group falls into paradigmatic classes, which are a set of words that use the same set of endings (quasi-flexions) during paradigm generation.

The inflection system of the Spanish language distinguishes the following paradigmatic types: nouns and adjectives T_1, verbs T_2, personal and reflexive pronouns T_3, articles T_4, and the irreducible or zero T_0, which includes the uninflected Spanish words, namely adverbs, prepositions, conjunctions, interjections. Each of these paradigmatic types is characterized by a certain set of inflection parameters.

The inflectional change of a certain word is traditionally called its inflectional paradigm, i.e., the set of word forms in all possible grammatical meanings. More formally, a word change is a correspondence (operator) P, according to which each word x corresponds to its word-changing paradigm $[x]$. Formally, this definition can be presented as follows:

$$x = [x];\ [x] = \{x^1, x^2, ..., x^N\},\ x^i = c(x) * f(x^i),$$

where x is a word, P is the conjugation operator, $[x] = \{x^1, x^2, ..., x^N\}$ is the complete inflectional paradigm of the word x; x^i, $i = 1, 2, ..., N$, are the components of the paradigm; $c(x)$ is the quasi-base (the unchangeable part of the lexeme x); $f(x^i)$ is the quasi-flexion of the component x^i. In this case, both the quasi-stem $c(x)$ and the quasi-flexion $f(x)$ can take on the following values:

– $c(x) = 0$ if the word has a suppletive form, such as the verb ser (to be) in the second and third person singular present tense of the active voice: eres, es; the verb caber in the first person singular present tense: quiero;

– $c(x) = x$ if x is an uninflected word, e.g., all pluralia tantum nouns (pantalones, gafas, cómicas, esposas, etc.), all prepositions (a, por, de, con, bajo), neuter pronouns (alguno, ello, lo), conjunctions (que, y, como); some nouns, e.g., those denoting days of the week (lunes, martes, miércoles, etc.), etc. etc.), words of Greek origin: análisis, artritis, crisis;

– $f(x) =$ x when all forms are complementary, as, for example, with the verb ir (to go) in all forms of the present tense: voy, vas, va, vamos, vais, van; the verb ser in the indefinite past tense: era, eras, era, éramos, erais, eran; the plural form of the masculine definite article: el – los.

Quasi-stem is a common term for an unchangeable component of a word that can include not just the derivational basis, but also a specific part of it (in some circumstances, only one letter), as well as the entire word. A quasi-flexion is a component of a word that varies during paradigm building. It can include:

– usual suffixes: *comer – com**o**, com**es**, com**e**; com**iera**s, com**iera**, com**iйramos***, etc.;

– a part of the stem: *plegar – pl**iego**, pl**iegas**, pl**iega**, pl**egamos**, pl**egбis**, pl**iegan***;

– the whole word form: *orden – **yrdenes**; ir – **voy, vas, va, vamos, vais, van***;

– ending: *gat**o** – gat**os**, mes**a** – mes**as***.

Here are examples of quasi-stem lengths in the table for the paradigm of the nouns *mano* and *orden*, as well as the verbs *bailar, tener*, and *ser*.

Table 1

Quasi-stems and their lengths

x	$c(x)$	$f(x)$	$x - f(x)$
mano			
mano (sing.)	*man-*	*-o*	3
manos (pl.)	*man-*	*-os*	3
orden			
orden (sing.)	0	*orden*	5
órdenes (pl.)	0	*órdenes*	5
bailar			
bailo (1 pers. sing.)	*bail-*	*-o*	4
bailas (2 pers. sing.)	*bail-*	*-as*	4
baila (3 pers. sing.)	*bail-*	*-a*	4
tener			
tengo (1 pers. sing.)	*t-*	*-engo*	4
tienes (2 pers. sing.)	*t-*	*-ienes*	4
tiene (3 pers. sing.)	*t-*	*-iene*	4
ser			
soy (1 pers. sing.)	0	*soy*	2
eres (2 pers. sing.)	0	*eres*	3
es (3 pers. sing.)	0	*es*	3

Moving on, let us now consider the issue of modeling the grammatical system of the Spanish language represented by word change. Formally, the inflection system can be interpreted as a L-system of a grammatical type (or grammatical system) where the lexicographic effect of inflection is induced:

$$LS^{GRAM} = \{I^{Q}(D), \Lambda^{\Lambda}(I^{Q}(D)), \mathrm{P}^{\Lambda}(I^{Q}(D)), F', C', H'\},$$

where LS^{GRAM} is L-system of grammatical type; $\Lambda^{\Lambda}(I^{Q}(D)) = \{x_0\}$ is a set of words in lemma form, $\mathrm{P}^{\Lambda}(I^{Q}(D)) = \{[x_0]\}$ is a set of paradigms, F' is an operator that establishes the relationship between a unit and its lemma form, C' is an operator that establishes the relation between a unit and its grammatical content (i.e., a paradigm), H' is an operator that correlates the lemma form of a word with a paradigm. In turn, the elements of the lexicographic system of the grammatical type $\Lambda^{\Lambda}(I^{Q}(D))$ and $\mathrm{P}^{\Lambda}(I^{Q}(D))$ can be decomposed using the recursive reduction mechanism $RR{\downarrow}[V(I^{Q}(D))]$, as shown in Fig. 1.

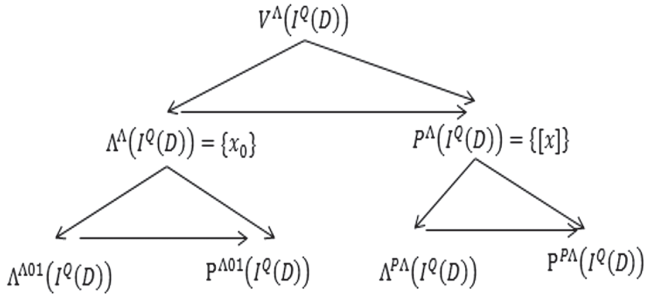

Fig. 1. L-system decomposition using recursive reduction

As a result of the recursive reduction, the set of descriptions $V^{\Lambda}(I^{Q}(D))$ is decomposed into smaller information elements containing relevant information about the described units of the Spanish language. The left part $\Lambda^{\Lambda}(I^{Q}(D))$ of the grammatical description includes a set of parameters $\Lambda^{\Lambda 01}(I^{Q}(D))$, that determine the place of the unit in the word-variable system of the Spanish language according to the classification described in [10], as well as $P^{\Lambda 01}(I^{Q}(D))$, i.e., the Spanish units in the canonical form that correspond to the parameters $\Lambda^{\Lambda 01}(I^{Q}(D))$. The right side of the grammatical system $P^{\mathrm{P}\Lambda}(I^{Q}(D))$, contains both all the word forms that make up the paradigm $P^{\mathrm{P}\Lambda}(I^{Q}(D))$ and the parameters representing the set of grammatical values $\Lambda^{\mathrm{P}\Lambda}(I^{Q}(D)) \equiv \Omega$. The components of the set of grammatical meanings of the Spanish language $\Omega = \{\Omega^1, \Omega^2, \Omega^3, \Omega^4, \Omega^5, \Omega^6, \Omega^7, \Omega^8, \Omega^9\}$ are:

$\Omega^1 = \{\omega_1{}^1, \omega_2{}^1, \omega_3{}^1, \omega_4{}^1\} \equiv \{m., f., m.\ y\ f., m.\ o\ f.\}$ is a set of grammatical meanings of the gender category, where *m* is masculine, *f* is feminine, *m y f* is common, and *m o f* is indefinite;

$\Omega^2 = \{\omega_1{}^2, \omega_2{}^2\} \equiv \{singular, plural\}$ denotes a set of grammatical meanings of the number category (*singular*, *plural*);

$\Omega^3 = \{\omega_0{}^3, \omega_1{}^3, \omega_2{}^3, \omega_3{}^3, \omega_4{}^3\} \equiv \{\varnothing, presente, pretérito\ perfecto\ simple, pretérito\ imperfecto, futuro\ simple\}$ is a set of grammatical meanings of the tense category (*present, imperfect past, simple perfect past, simple future*);

$\Omega^4 = \{\omega_1{}^4, \omega_2{}^4, \omega_3{}^4, \omega_4{}^4\} \equiv \{1a\ pers., 2a\ pers., 2a\ pers.\text{-}voseo, 3a\ pers.\}$ designates a set of grammatical meanings of the person category (*first, second, third*);

$\Omega^5 = \{\omega_1{}^5, \omega_2{}^5, \omega_3{}^5, \omega_4{}^5\} \equiv \{indicativo, condicional, subjuntivo, imperativo\}$ is a set of grammatical meanings of the mood category (*active, conditional, subjunctive, and imperative*);

$\Omega^6 = \{\omega_1{}^6, \omega_2{}^6, \omega_3{}^6, \omega_4{}^6\} \equiv \{nominativo, dativo, acusativo, preposicional\}$ means a set of grammatical values of the case category (*nominative, dative, accusative, local*).

$\Omega^7 = infinitivo$, $\Omega^8 = gerundio$, $\Omega^9 = participio$ are non-finite forms of the verb: infinitive, gerund, and participle.

Word forms are defined by complexes of grammatical meanings that together make up the grammatical state of a linguistic unit. These combinations of grammatical meanings, or members of sets, define a word form. An example of a fragment of the paradigm for the verb *hablar* (*to speak*) is shown in Table 2, and the parameters characterizing its grammatical state are indicated.

Table 2

Parameter values for the grammatical state of the verb *hablar*

Parameter chain	Parameter values in chain	Word form
$\{\langle\omega_1{}^4, \omega_1{}^2, \omega_1{}^3, \omega_1{}^5\rangle\}$	{⟨1ª pers., singular, presente, indicativo⟩}	*hablo*
$\{\langle\omega_2{}^4, \omega_1{}^2, \omega_1{}^3, \omega_1{}^5\rangle\}$	{⟨2ª pers., singular, presente, indicativo⟩}	*hablas / hablás*

Parameter chain	Parameter values in chain	Word form
$\{\langle\omega_3^4, \omega_1^2, \omega_1^3, \omega_1^5\rangle\}$	{⟨3[a] pers., singular, presente, indicativo⟩}	*habla*
$\{\langle\omega_1^4, \omega_2^2, \omega_1^3, \omega_1^5\rangle\}$	{⟨1[a] pers., plural, presente, indicativo⟩}	*hablamos*
$\{\langle\omega_2^4, \omega_2^2, \omega_1^3, \omega_1^5\rangle\}$	{⟨2[a] pers., plural, presente, indicativo⟩}	*habláis*
$\{\langle\omega_3^4, \omega_2^2, \omega_1^3, \omega_1^5\rangle\}$	{⟨3[a] pers., plural, presente, indicativo⟩}	*hablan*

3. Dictionary Database

The Spanish inflection dictionary database (as shown in Fig. 2) was developed using the conceptual model discussed above. Thus, information regarding paradigmatic kinds may be found in the "**par_types**" table:

– ID_partyp (paradigmatic type identifier),
– com (paradigmatic type, name),
– ac (comment, word change parameters).

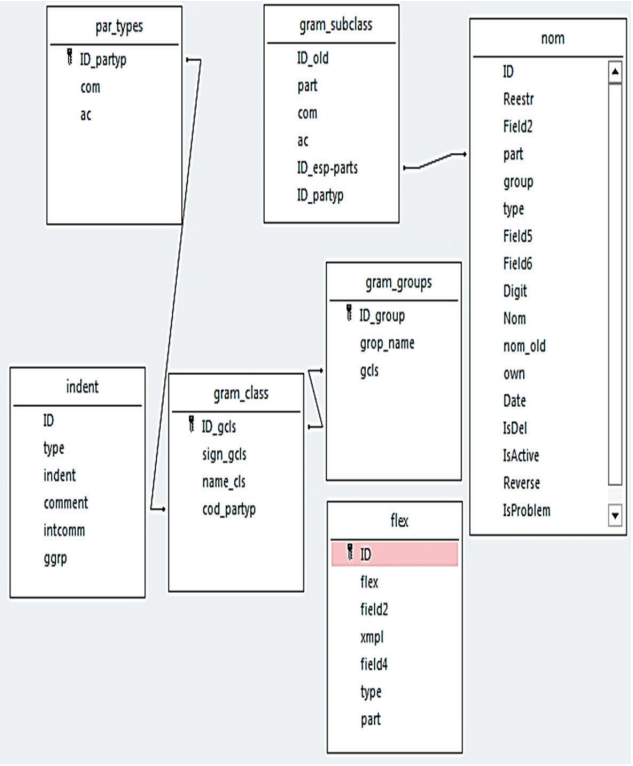

Fig. 2. Database scheme of the Spanish language inflectional dictionary

Parameters and identifiers of grammatical classes in the table **"gram_class"** include:

– ID_gcls (grammatical class identifier),
– sign_gcls (grammatical class signature according to the conceptual classification),
– name_cls (name of the grammatical class),
– cod_partyp (code of the paradigmatic type).

Table **"gram_groups"**, which contains information about grammatical groups and includes the following fields:

– ID_group (grammar group identifier),
– grop_name (grammar group name),
– gcls (the code of the grammar class to which the grammar group belongs).

Table **"indent"** includes the fields that the system uses to determine quasi-stems by paradigmatic classes:

– type (paradigmatic class number),
– indent (the number of letters to be cut off from the end of the word to obtain a quasi-stem),
– comment is the conventional name of the paradigmatic class in the conceptual model.

Table **"flex"** table contains sets of quasi-flexions for each case word, organized by the following parameters:

– Reestr (headword),
– Field2 (homonymy number),
– Part (grammatical class code),
– Group (grammatical group code),
– Type (paradigmatic class number).

4. Virtual Lexicographic Laboratory to handle Spanish Inflectional Dictionary

To work with the electronic dictionary database, a virtual lexicographic laboratory (VLL) was developed (Fig. 3), the functionality of which currently allows:

– viewing the grammar dictionary wordlist and the full paradigm of each word,
– search for words in the register, as well as display them in forward or reverse order,
– add, delete words from the word list,
– add, delete and edit paradigm classes,
– adding, deleting and editing quasi-flexions in paradigmatic classes,
– filtering the register by the following features (or a combination of them): part of speech, paradigmatic class, homonyms, etc.

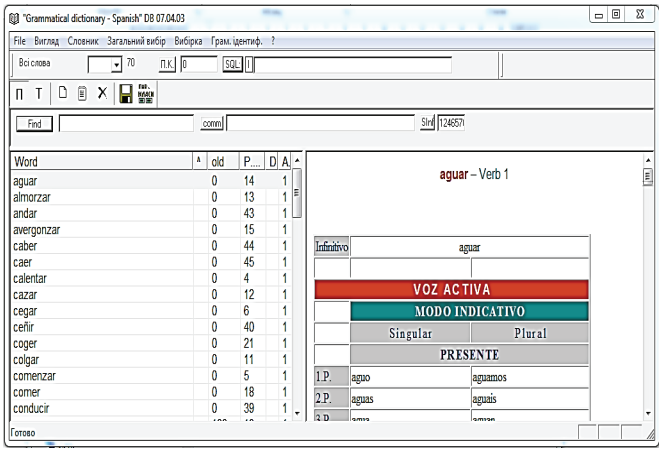

Fig. 3. VLL main window

The main window of the program is divided into three zones: the functional area, the register area, and the lexicographic information area. The functional area consists of the following sub-areas: the general menu, editing tools, tools for executing SQL queries, and the word search interface. The main menu is intended for selecting

the display modes of the dictionary, forming a selection according to certain parameters, and performing operations with the dictionary database file. The word list area represents the word list, indicating the number of the paradigmatic type, grammatical group and class, and paradigmatic class. In case of uninflected word, the number of paradigmatic class isn't given. Lexicographic information area is intended to display information on the word change of the word selected from the list (full inflection paradigm).

Paradigm classes, including inflections, are added using the dialog box (Fig. 4), which is opened by clicking the "Paradigmatic Classes" button on the editing panel. The left part of the window displays the numbers of the paradigmatic type, grammatical class, grammatical group, and paradigmatic class. In addition, technical parameters are displayed, such as the length of the quasi-stem of the word to which the quasi-reflection is to be attached (as shown in Table 2). The right part of the window contains a list of all quasi-flexions belonging to the paradigmatic class selected in the left part. The main functions are: searching for a particular paradigmatic class, creating a new paradigmatic class, and adding and removing flexions.

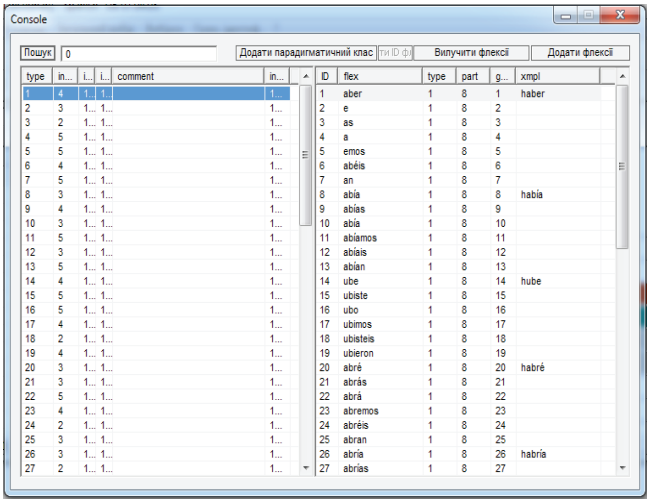

Fig. 4. The window for editing paradigmatic classes

To create a new paradigmatic class, click on the "Add paradigmatic class" button, where you need to enter the numbers of the corresponding paradigmatic type, grammatical class, grammatical group, and paradigmatic class.

Conclusions

The software (VLL) allows compilers to create, edit, and update electronic dictionaries for any language that has a developed word change system. The inflectional dictionaries created by the VLL tools provide the ability to automatically generate a paradigm for any word, display the whole inflection system together with all its paradigmatic types, groups and classes, and provide grammatical characteristics for any word form included in the paradigm. The developed database structure and software tools for data editing contribute to the efficient organization of the process of creating a word-by-word dictionary of the Spanish language. The created database can be successfully used in the study of word change processes and phenomena.

References

[1] Shyrokov V. Computer lexicography. – K: Naukova Dumka, 2011. – 351 p.

[2] Shyrokov et al. Computational linguistics studies. – V. 2.: Grammar systems. – K.: ULIF NASU, 2018. – 300 p.

[3] Shyrokov et al. Computational linguistics studies. – V. 3.: Explanatory lexicography. – B. 2: System semantics of explanatory dictionaries – K.: ULIF NASU, 2018. – 250 p.

[4] Johannsson E., Ingimundarson F. Describing inflectional patterns of nouns in Old Icelandic // CEUR. – 2022. – V. 3232. – P. 260-268.

[5] Chrzaszcz P. Enrichment of inflection dictionaries: automatic extraction of semantic labels from encyclopedic definitions // 9th International Workshop on Natural Language Processing and Cognitive Science. –. – 2012. – P. 106-119.

[6] Wolinski M. A Relational model of Polish inflection in grammatical dictionary of Polish // Human Language Technology. Challenges of the Information Society. – 2009. – №. 5603. – P. 1-11.

[7] Štěpankova B., Mikulova M., Hajič J. The MorfFlex Dictionary of Czech as a Source of Linguistic Data // Euralex XIX: Congress of European Association of Lexicography. – 2021. – P. 387-391.

[8] Arista J. M. Old English morphological inflection generation with UniMorph. Assessment with a relational database and training guidelines // Procesamiento del Lenguaje Natural. – 2022. – №. 68. – P. 61-70.

[9] Wolinski M., Kieras W. The online version of grammatical dictionary of Polish // Język Polski. – 2017. – 97(1). – P. 2589-2594.

[10] Kupriianov Ye. Lexicographic system of the Spanish Language: Phenomenology of Integral Description. K.: Naukova Dumka, 2018. – 254 p.

The article was delivered to editorial stuff on the 30.03.2023

УДК 004.42 DOi 10.30837/ bi.2023.1(99).13

Maksym Shulha[1], Dmytro Matvieiev[2], Oleksii Nazarov[3], Nataliia Nazarova[4]

[1] ХНУРЕ, м. Харків, Україна, maksym.shulha@nure.ua

[2] ХНУРЕ, м. Харків, Україна, dmytro.matvieiev@nure.ua, ORCID iD: 0000-0002-0622-8159

[3] ХНУРЕ, м. Харків, Україна, Ukraine, oleksii.nazarov1@nure.ua , ORCID iD: 0000-0001-8682-5000

[4] ХНУРЕ, м. Харків, Україна, nataliia.nazarova@nure.ua, ORCID iD: 0009-0007-7816-7088

RESEARCHED METHODS FOR SIMPLIFYING AND OPTIMIZING PARTICLES FOR PORTABLE GAMING DEVICES

The object of the research is particle simulation systems and their modules. The aim of the work is to conduct a study of the performance of particle simulation systems and the optimization process to improve it. The methods of development and design are the analysis of the problem area of research, the choice of a particle simulation system based on their comparison for further study and optimization. As a result of work, a game simulation project was developed to demonstrate the effectiveness of the proposed optimization methods.

VERTICES, OPTIMIZATION, SIMULATION, PARTICLES SYSTEMS, SHADERS, INSTRUCTIONS COUNT, UNREAL ENGINE

Шульга М.В., Матвєєв Д.І., Назаров О.С., Назарова Н.В. Дослідження методів спрощення та оптимізації частинок для портативних ігрових пристроїв. Об'єктом дослідження є системи симуляції часток та їх модулі. Метою роботи є проведення дослідження продуктивності систем симуляції часток та процес оптимізації для її покращення. Методами розробки та проектування є аналіз проблемної області дослідження, вибір системи симуляції часток на основі їх порівняння для подальшого вивчення та оптимізації. У результаті роботи було розроблено ігровий проект-симуляцію для демонстрації ефективності запропонованих методів оптимізації.

ВЕРТЕКСИ, ОПТИМІЗАЦІЯ, СИМУЛЯЦІЯ, СИСТЕМИ ЧАСТОК, ШЕЙДЕРИ, INSTRUCTIONS COUNT, UNREAL ENGINE

Introduction

Visual effects (VFX) are a technology that allows you to create realistic effects such as fire, smoke, water, sparks, and more using a large number of small objects called particles. These particles can have different properties, such as color, size, shape, speed, direction, etc., and interact with each other and the environment according to certain rules. Particle simulation systems are widely used in modern game engines to create immersive visual effects that enhance the immersion and atmosphere of the game.

The purpose of the study is to analyze and compare the methods of using particle simulation systems in modern game engines such as Unity, Unreal Engine, CryEngine and the possibility of using them in portable devices.

The study focuses on the processes of creating, optimizing and integrating VFX in game projects, as well as their impact on the graphical quality, performance and immersiveness of the game.

The study aims to identify the advantages and disadvantages of different approaches to VFX implementation and provide recommendations for their effective use.

In this article, we will review the basic principles and methods of particle simulation systems, as well as examples of their application in various game genres. We will also analyze the advantages and disadvantages of this technology, as well as the possibilities for its further development and improvement.

1. Description of subject area

The proliferation of tools and technologies for MPE development is reflected in the computer game industry [1], interactive cinema, and augmented and virtual reality applications. As development tools evolve, the need for dynamic model optimization comes to the fore. This is due to the fact that the simplification of the process of developing rather complex models has led to a decrease in developers' interest in multifactor optimization of existing models - it is easier to build a new model than to optimize an existing one.

The optimization process can be applied to almost all elements of modern engines. This is due to the fact that the engines provide as much functionality as possible to ensure compliance in the applications they are designed for. For applications such as the film industry or simulations, where models do not run in real time, optimization is not critical because it only affects the comfort of working with the project, not the final result, which can go through the rendering process at any time. The main area where optimization is critical is in real-time software applications. For this class of applications, the main requirement for the optimization process is to minimize optimization-related delays and maximize synchronization of all used engine elements, since they can run on different hardware and, in the future, be scaled to new portable devices. This fact attracts special attention because

mobility is developing quite actively, and therefore optimizing existing applications for new types of mobile devices becomes more important than ever.

The optimization process remains complex and situational in its parameters and includes many criteria that require an individual approach. Modern engines consist of modules that work together and can influence each other. Due to this fact, it is often impossible to disable them: for example, games adapted for virtual reality rarely use a standard interface module, which requires a lot of resources for initialization and computation. Also, the development team often does not include experienced specialists who can understand the structure of the engine code and disable a particular module in a way that does not affect other modules. In some engines, such as Unity, the engine code is not editable, requiring optimization at a higher level.

2. Statement of research problem

The article highlights the following tasks to be considered:

1) analysis and comparison of current tools for developing particle simulation systems for the game industry;

2) to compare the gaming industry with the film industry in terms of the creation and use of VFX;

3) review the theoretical foundations of particles and their classification according to various criteria;

4) analysis of existing methods for simplifying and optimizing particles for handheld gaming devices, such as reducing the number of particles, changing their shape and size, using textures or shaders, adapting to camera movement;

5) select software for modeling and visualizing particles using the selected simplification and optimization methods;

6) conclusions about the advantages and disadvantages of the considered methods of particle simplification and optimization for portable gaming devices.

This set of questions should be sufficient for a first overview of the industry, the VFX development tools [2] and the problems to be solved. The results will also help to adjust the process of analyzing the results and the methods that will be used to solve problems with the optimization of particle simulation systems.

3. Research Facilities

The game engines listed in the introduction will be used for the research. In the course of this research, we will analyze the existing modern engines, review their pros and cons, and then choose the one that suits us best.

Since we are focusing on portable devices, the most important parameter of these engines will be the ability to build projects for mobile devices using modern libraries.

Analyzing optimization for consoles is quite complicated, because it requires the presence of so-called "devkits" - consoles sent by their owners for development. This is impossible for the learning process. There are also personal computers, but the variability of their hardware is very large, which makes it impossible to adequately evaluate the results of optimization, as video cards of different generations have fundamental differences.

That's why a personal computer will be used to test intermediate results to speed up the analysis process to obtain relative results and formulate optimization methods. Further testing will be done on a modern smartphone, as this platform is available and meets the hardware limitations that can reflect the effectiveness of the optimization methods found.

4. Review and compare modern game engines

Choosing the right game engine is by far the most important decision a game developer has to make to get the best results from their product. The first puzzle a game developer has to solve when creating a game is which game engine to use to get the best user experience. A game engine helps to create not only classic arcade games [3] like Ping, Tetris, Snake, but also innovative and advanced leveling games like GTA and Assassin's Creed. Every game developer has heard of some incredible game engines like Unreal Engine, CryEngine and Unity3d. These are some of the most popular and leading game engines for creating advanced games today. Each of these 3 game engines has its own qualities and potential. We need to be clear about the nature of our project, for example, license budget, game platform, dimensions (2D or 3D), and so on. So, your goal should be clear before choosing the best game engine 2023 for our project.

Unity has many contributors in its community who can help us with the project right away. Another feature that makes it stand out as one of the best 3D game engines in 2023 is its ability to support a number of file formats used in leading 3D programs, including 3D Max, Blender, CINEMA, Maya, Softimage, and many others. In addition, game developers have access to more than 15,000 free and paid 3D models, audio, animations, editor extensions, materials, scripts, and shaders for use in game development. Unity 3D uses C# or JavaScript, which is more desirable than C++ because you don't have the hassle of switching from Java to C# compared to C++. Nevertheless, Unity 3D has a simple and fast interface and is light enough to run even on Windows XP with Service Pack 2 (SP2).

Second on the list is Unreal Engine 4, which is the latest engine released by one of the largest American video game and software development companies, Epic Games. Unreal Engine 4 is the successor of the Unreal Development Kit, commonly known as UDK in the gaming world. Unreal Engine 4 offers incredible graphics that add a realistic touch to the gaming experience with features such as advanced dynamic lighting. What makes

this game engine even more amazing is its new particle system, which has the ability to handle up to a million particles in a single scene.

In addition, UE4 is completely free to use, but you must pay a royalty of 5% of the money you make from your Unreal Engine 4 games. In short, Epic Games gets 5% of everything you make, whether it's in-app purchases, in-game advertising, or money you charge users to buy your game. However, the creators of Unreal Engine 4 allow developers to use the full version of Unreal Engine 4 for free if the revenue you make from your game is up to $3,000 per quarter.

In addition, Unreal Engine 4 uses Blueprint Visual Scripting technology, which allows you to create games using Blueprint. However, such games have some limitations. Among other things, the bad thing about this engine is that it is not capable of developing games for last generation consoles.

Finally on our list of the best game engines in 2023 is CryEngine. First introduced by major development company Crytek in the first Far Cry game, CryEngine is undoubtedly one of the most powerful and dominant game engines we have today. What makes CryEngine worthy of the list is its graphical capabilities, which eclipse those of Unity and are equivalent to what Unreal has. Although CryEngine is a heavy and powerful game engine, it takes a little time for the user to be able to use this platform effectively and is a little harder to understand for beginners who have not used other game engines before.

The CryEngine supports virtual reality and has amazing visual effects, including a three-dimensional fog and cloud visualization system that gives clouds a full 3D spatial visualization and realistic rendering for fog and weather effects. Furthermore, the best part of choosing the CryEngine platform for game development is that it does not require its users to pay royalties during game development. However, to get access to CryEngine, you need to pay a fixed amount, namely $9.90/- per month. In addition, CryEngine has a dedicated Q&A forum called CryEngine Answers that clears all your doubts and queries and helps you to have the best experience.

While comparing Unreal, Unity and CryEngine, we came across the best features that these game engines offer us. Comparing the performance of Unity and Unreal, we realized that Unity is a better platform for developing mobile and 2D/3D games, while Unreal is best suited for developing highly graphical and photorealistic games. This is the main difference between Unreal and Unity.

CryEngine, on the other hand, also has the ability to create games with high graphics. Furthermore, when comparing CryEngine to Unreal 2023 on the scale of providing next-generation platform features with a more attractive pricing model, CryEngine undoubtedly outperforms Unreal with its cost-effective structure. However, for a beginner, when it comes to CryEngine 5 vs. Unreal Engine 4 based on lightness and minimalism, despite the amazing features that CryEngine has, Unreal Engine 4 and Unity are definitely worth a try.

At the end of this analysis, we came to the conclusion that the most versatile and high-quality particle simulation system is implemented in Unreal Engine. Moreover, both systems of this engine have developed tools for optimization, not only for the creation of these particles. Therefore, further analysis should be done on this engine.

5. Industry analysis and Unreal Engine 4-5's place in it

The late 90's saw a monumental development in personal computers and the games that could be played on them. Graphics improved rapidly, and although ridiculous by today's standards, the development of 3D first-person graphics made video games much more immersive. The Internet was rapidly spreading to consumers, making it easier to participate in online games, join discussions on gaming forums, and even learn how to program games yourself. And the games released during this time are titles we still talk about with reverence: Doom, Sim City, Duke Nukem, Half-Life, StarCraft, Myst, and many others (see Fig. 1).

Fig. 1. Doom

But Unreal Tournament was in a league of its own: fast paced, large maps, and AI computer opponents that, for the first time, actually seemed intelligent. A community of players began to grow around the game, modifying it, making it their own, and sharing their creations with the community. All of this happened thanks to a game development company with an extremely bright future and the software they created called Unreal Engine.

Founded in 1991, Epic Games is one of the few software company success stories that has earned a rightful place in gaming history. Their first few games were commercially successful, but what really put them on the map came at the end of the millennium when they released Unreal [4] in 1998, followed by its sequel, Unreal Tournament, the following year.

The game's gameplay, graphics, and features set it apart from many other first-person shooters on the market. But

beyond the game itself, Epic Games made a decision that made the game legendary. They decided to ship the game with the same tool - the Unreal Editor - that is used to create levels and gameplay, allowing players to customize the game and make it their own. It's one of many game engines developed over the years, but it's far from the most successful or widely used.

"If you think of a technology that has stood the test of time, Unreal Engine is one of the biggest," says Jacob Feldman, a software and rendering solutions engineer at CoreWeave who moonlights as an authorized Unreal Engine instructor. "It has over 20 years of code, so it's a significant, complex product."

Realizing that they had a winner with the Unreal Engine, the Epic team began licensing the software so that other developers could build their own games on top of it. Fast forward to 2017, when the studio released Fortnite, a cultural sensation that would further cement the legacy of Epic and the Unreal Engine. More recently, the Unreal Engine has played a major role in another cultural moment that has propelled the tool into an entirely new industry: movie production [5] (see Fig. 2).

Fig. 2. Using UE4 in cinema

In 2019, Disney released the award-winning series The Mandalorian as part of the lineup for its new streaming service, Disney+. The show became an instant hit, continuing the long history of Star Wars productions that have pushed the technological boundaries of the film industry. "The Mandalorian" demonstrated a giant leap forward in set design by using real-time set rendering with the Unreal Engine, effectively making green screen obsolete.

"Unreal made a leap into broader industries like film after Fortnite came out," says Feldman. "It became clear that a game engine like Unreal, designed for flexibility, performance, and visual quality, had more applications than just gaming. This became especially clear as more powerful hardware began to enable things like real-time ray tracing and other visual effects that were simply not possible 5-10 years ago. The visual effects industry has some of the most stringent requirements of any visual medium, so Epic has made a serious commitment to supporting the kinds of tools that VFX studios need.

Filmmakers have always found ways to trick the audience's eyes into thinking they are seeing something larger than what is there - what the industry calls "set extension. In the early days of Hollywood, this involved literally drawing scenes that were then seamlessly integrated into the set to create the effect of infinite space.

Later, green-screen technology became the preferred method for creating large worlds. However, there are some significant problems with this method. It can be difficult for actors and crew on set to perform convincingly against an empty, bright green background. The actors cannot see the world around them (see Fig. 3) and therefore cannot react to it in a way that is convincing to the viewer.

But what if there was a way to create large worlds on command that actors could see right in front of them? It would be almost like throwing those actors into a video game and having them perform in that world, with all the views available to their own eyes.

Fig. 3. Shooting a movie

Mandalorian creator Jon Favreau approached the team at Epic Games to see if they could help him do this, essentially immersing his actors in a video game. Using massive high-definition LED screens surrounding their set, the Mandalorian team took the old Hollywood concept of painted sets and brought it into the twenty-first century. Now these painted sets were living, breathing environments with dynamic lighting and movement. The actors could see them and react to them. The physical elements in the foreground blend seamlessly into the virtual background.

But it gets even better-using game engine technology, you can tie camera movement to the background (see Fig. 4), allowing it to change with the camera to perfectly match the movements and provide smooth parallax movement that is indistinguishable from the real shot.

Fig. 4. The process of filming «Mandalorian»

In essence, they created the real-time graphics that underpin movie production. There is no need to clean up in post-production; these effects are done in the camera in real time.

"It's a huge help for the actors and the crew to be able to see their surroundings as they happen," says Feldman. "Today, an actor has to imagine what is happening around them on a green screen. This advancement allows them to see it."

In addition to making life easier for the actors, the real-time environment also makes the work of the post-production team infinitely less stressful. On the green screen, reflection is a constant battle. Any shiny material will reflect the green environment and ruin the VFX effect. These reflections must be removed, and then virtual light must be added to the tiny reflections in glasses, helmets, or pieces of metal in the environment. Real-time environments eliminate all of this because the reflections captured by the camera reflect the world as it should be in the story.

Epic Games took the wraps off its latest game engine, Unreal Engine 5, and pretty much broke the gaming-interested part of the Internet. Running on the PlayStation 5 developer hardware version, the results uploaded to Vimeo looked stunning at first, but when you realized they were real-time gameplay footage, they were truly impressive.

The visual quality is virtually unmatched by anything other than today's high-end pre-processed visual effects, with an incredible level of detail and truly photorealistic lighting [6]. What's even better, of course, is that the news about UE5 has lit up the broadcast graphics, movie effects, virtual studio, and pretty much anywhere else that high quality visual effects are used.

While other engines like Unity are available, Epic has done a nice job of combining a leading feature set with ease of use and a business model that encourages the use of UE in other software. For example, any game developer using Unreal commercially pays no licensing fees until gross software revenues reach $1 million (recently raised from $50,000), and this has helped to drive its widespread integration across a range of technologies. For non-game developers, it is 100% royalty free.

Its use in the graphics stack is similar to how connectivity to the IT industry as a whole has accelerated development in the broadcast sector; it allows a relatively small industry to leverage the development efforts of a much larger one, and the speed of change we're seeing as a result is impressive.

"When the new Unreal Engine 5 comes out, you won't be able to tell what's real and what's not," said Phil Ventre, vice president of sports and broadcast at Ncam Technologies. "The integration of game engines is democratizing the way companies use AR and VR, and it won't just be a tier-one technology for broadcasters in the future."

While the new demo was created in part to showcase some of the very smart new technologies in the upcoming PS5, such as the literally game-changing M.2 SSD, UE5 showcases some new technologies that are dramatically moving the goalposts for real-time CG work. Two in particular are worth mentioning: Nanite and Lumen.

Nanite is a new virtualized geometry [7] of micro polygons that essentially allows artists to create as much geometric detail as they want. It's translatable and scalable in real time, which is important to consider when you're planning an engine that will run on everything down to a smartphone (see Fig. 5).

Then there is Lumen. This is a fully dynamic global lighting system that responds instantly to changes in scene and lighting and is part of the secret to why the game's graphics look so good on a console. It's capable of reproducing diffuse bounce reflections with infinite bounce and indirect specular reflections in what Epic calls "vast, detailed environments on scales from kilometers to millimeters. It's also adaptable: Punch a hole in a wall and the scene will change to accommodate the light coming in through the hole.

Fig. 5. Demo scene on UE5

"The ability to light and render hundreds of millions of polygons in real time is a quantum shift that will change the way filmmakers interact with the images they create," said Miles Perkins, business development manager at Epic Games. "These new technologies will allow creatives to see the fullness of their vision without having to separate different parts of their shots, viewing animation separately from lighting, separately from environments and effects. Everything will be right in front of them, fully controllable. Filmmakers will be able to compose and light shots in real time, whether they are physical, virtual, or a combination of the two."

Perhaps one of the key points here is that the Unreal Engine, currently running at version 4.25, is already very good.

"We are currently using Unreal Engine 4 extensively in both pre-production and virtual production," said Hugh MacDonald, director of technology innovation at Nviz. "For pre-production, the real-time nature of UE4 means we can get incredibly high-quality images with minimal rendering time. We're also using it for virtual production,

which allows for better integration than we could have done in the past.

Nviz uses Unreal in two main tools that illustrate how it is being used for pre-production work in the industry and where it can go. The virtual camera system allows for virtual scouting of pre-production environments and allows directors and cinematographers to get hands-on experience with the camera, while the simulcast toolkit is tightly integrated with Unreal and allows the crew on set to preview what the shot will look like after visual effects are added in post.

"Unreal allows us to ensure that it's both flexible and high quality," says MacDonald. "From what we know of UE5 so far, the main leaps will be in geometry detail, as much higher resolution assets can be used and rendered. This is a much better fit for the movie VFX workflow, as it is hoped that assets will require less editing to make them engine ready. Fully dynamic lighting, which is lumen, will mean that there will be less need for lighting baking to get the same result. This will allow us to keep scenes fully dynamic, allowing us to adjust lighting in real time during production, if necessary, which is often requested on set as physical lighting changes from shot to shot."

New Creativity Along with the increase in quality, which Ventre of NCam compares to the leap from UE3 to UE4, UE5 opens up the tantalizing prospect of introducing both new ways of working and new ways of creatively exploring virtual spaces.

"The Unreal Engine is going to be a big part of the future of cinema," says CVP's Sam Mir. "It brings back the ability to get practical in-camera effects, whether it's interactive lighting on an actor or dynamic background changes in real time, even if they were created in a virtual space. The ability to get instant feedback on how something is going to look is invaluable."

This will find its way into many more living spaces than before. McDonald mentions theater and events, where video screens have become part of the interaction with lighting to create entirely new live spectacles, while virtual sets will take another leap in quality to become so indistinguishable from the real thing that only a live audience will influence the decision to use a physical set. Even these viewers will be able to see completely different shows, with mixed elements of augmented reality seamlessly integrated into the final TX. In the post-covid era, you can easily get three guests on the couch for a chat show from the comfort of your own home, and no one will be the wiser.

And, of course, there's the possibility that this could accelerate the development of live shots on LED screens, as pioneered by shows like The Mandalorian.

"While Unreal has been used a few times on LED screens during shoots, the new updates will hopefully take it much further and get a higher percentage of finished shots straight from the camera," MacDonald enthuses.

6. Particle simulation systems available in UE5

Cascade and Niagara [8] are two particle systems used in Unreal Engine to create effects such as fire, smoke, rain, snow, and sparks. Cascade is an older system introduced in Unreal Engine 3, while Niagara is a newer system added in Unreal Engine 4.25. Both systems have their advantages and disadvantages, which will be discussed below.

Cascade is based on the concept of emitters and modules. An emitter is an object that generates particles according to certain rules, and a module is a component that changes the properties of the particles, such as color, size, speed, rotation, etc. Cascade allows you to create complex effects from multiple emitters and modules and customize them using a graphical interface. Cascade also supports features such as GPU acceleration, particle collisions, lighting and shadow interaction, sprite animation, and more.

Niagara is based on the concept of systems and emitters. A system is an object that manages one or more emitters and their logic. An emitter is an object that generates particles according to certain rules, but in Niagara these rules can be specified using graph editors or scripting languages. Niagara allows you to create more flexible and dynamic effects with a variety of particle types, including mesh particles, line particles, ribbon particles, and more. Niagara also improves the performance and quality of effects through GPU optimization, fluid simulation, and interaction with engine functions.

A comparison of Cascade and Niagara can be seen as follows:

1) Cascade is easy to use and has many default settings for creating effects quickly.

2) Cascade has limitations on the number and type of particles.

3) Cascade does not allow you to change the logic of particle generation or create your own modules.

4) Niagara is difficult to use and requires more programming and math skills to create effects.

5) Niagara allows you to create many different types and shapes of particles.

6) Niagara allows you to modify the logic of particle generation or create your own functions and modules.

7) Niagara improves performance and effect quality through GPU optimization and fluid simulation.

7. Cascade

The basic and overarching concept of the cascade is that of modular particle systems. In some 3D effects packages, such as Maya, a particle system is created with most of the necessary properties for the behavior. The user then modifies these properties to achieve the desired result.

In Cascade, on the other hand, the particle system starts with just a few basic properties and a few behavior modules.

Each module represents a specific aspect of particle behavior and contains only the properties that control that behavior, such as color, birth position, motion, scaling, and many others. The user can then add or remove modules as needed to further define the behavior of the particles. Since only the modules for the desired behavior are added, there is no unnecessary computation of unnecessary properties.

Best of all, modules can be easily added, removed, copied, and even tried on emitters in the particle system, making complex setups very easy to achieve once the user is familiar with the available modules.

Some modules come standard with a particle emitter. When a new sprite emitter [9] - a key component of any particle system - is added to a particle system, it is always created with the following default modules:

1) Required - contains a variety of properties that are absolutely necessary for the particle system, such as the material applied to the particles, how long the emitter should emit particles, and many others.

2) Spawn - this module controls how fast the particles will appear from the emitter, whether they will appear in bursts, and any properties related to the particle birth time.

3) Lifetime - this parameter determines how long each particle will live after its birth. Without this module, particles will live indefinitely.

4) Initial Size - controls the size of the particle at the moment of its birth.

5) Initial Velocity - controls the movement of the particle at the moment of its birth.

6) Color Over Life - this module controls how the color of each particle will change during its life.

The Required and Spawn modules are permanent and cannot be removed from the emitter. All other modules can be removed at will.

Particle systems are also very closely related to the different materials and textures applied to each particle. The main task of the particle system itself is to control the behavior of the particles, while the specific look and feel of the particle system as a whole is often controlled by the materials.

There are many modules that can be added to particle emitters. To avoid confusion, these modules are divided into different categories. These categories include those listed in Table 1.

Two important concepts to keep in mind when working with particle modules are initial and lifetime. Initial modules are used to control some aspect of the particle at the moment of its birth. The Over Life or Per Life modules are used to allow some aspect of the particle to change during its lifetime.

Table 1

Categories of particle emission system modules

Category	Description
Acceleration	Modules that control how the acceleration of particles can be affected by, for example, drag forces.
Attraction	Modules that control particle movement by attracting particles to different points in space.
Camera	Modules that control the movement of particles in camera space, allowing the user to make them appear closer or farther away from the camera.
Collision	Modules that control how collisions between particles and geometry are handled.
Color	Modules that control the color of particles.
Event	Modules that control the activation of particle events, which in turn can trigger a variety of in-game reactions.
Kill	Modules that control the removal of particles.
Lifetime	Modules that control how long particles live.
Light	Modules that control the light emitted by particles.
Location	Modules that control the birthplace of particles relative to the location of the Actor emitter.
Material	Modules that control the material of the particles themselves.
Orbit	Modules that provide orbital behavior of the screen space to add movement to the effects.
Orientation	Modules that allow you to set the rotation axis of the particles.
Parameter	Modules that can be parameterized or controlled by external sources such as blueprints and wounds.
Rotation	Modules that control the rotation of particles.
RotationRate	Modules that control the change of rotation speed, such as spin.
Size	Modules that control the size of particles.
Spawn	Modules that add special particle appearance rates, such as spawning particles based on distance moved.
SubUV	Modules to display animated sprite sheets on a particle.
Velocity	Modules controlling each particle's speed.

For example, the Initial Color module allows you to set the color of the module at the time of birth, while the Over Life property allows the color of the particle to gradually change between the time of birth and the time of death.

If you change the property to a type of distribution that changes over time, some modules use "relative time" and some use "absolute time".

Absolute time [10] is essentially the time that the emitter contains. If you have an emitter set up for 3 cycles of

2 seconds, the absolute time for the modules in that emitter will go from 0 to 2 seconds three times.

The relative time is between 0 and 1 and indicates the lifetime of each particle.

As you work with Cascade to create your own particle effects, it's important to keep in mind how each object is related to the others. We've already discussed the concept of modules in this document, but they are only one component of a complete particle effect. In general, the components of a particle system are modules, emitters, particle systems, and emitter actors.

Just as there are many types of effects you'll want to create with your particles, there are also many types of emitters to help you create exactly what you need. Below is a list of the available emitter types:

Note that all emitters, regardless of type, are sprite emitters by default. Various Emitter Type Data modules are then added to the emitter to change its type to something else.

Not all aspects of a particle system can be defined in advance. Sometimes certain parts of the particle system's behavior need to be controlled or changed at runtime. For example, you may want to create a magic effect that changes color based on the amount of magic energy consumed during the spell. In such cases, you'll need to add parameters to the particle system.

A parameter is a type of property that can send or receive data to/from other systems, such as Blueprints, Matinee, Material, or many other sources. In Cascade, a parameter can be assigned almost any property, meaning that the property can be controlled from outside the particle system. For example, setting a parameter to control the rate of creation of a fire effect can allow the player to increase or decrease the amount of flame emitted during operation.

Conversely, there are parameter modules that can be added to a particle system that in turn can control other things in the level, such as the color used in a particular material.

Particle systems can easily become very expensive to compute. Even when using GPU particles, which offload much of the particle calculation to the GPU, it is important to consider the value of calculating particles that the player is too far away to see or properly evaluate.

For example, consider the fire particle system. If you look at it up closely, you can see the embers and sparks rising into smoke. But if you look at it from a distance of several hundred meters, those embers are so small that a monitor or screen cannot even reproduce them. So why calculate them?

This is where Levels of Detail (LOD) come into play. The LOD system [11] allows you to set your own distance ranges at which your particle system will automatically simplify. Each range represents a different LOD. Simplification comes in the form of reduced property values, disabled modules, or even disabled emitters. For example, in the bonfire example above, it would be ideal to completely disable the emitter that was adding sparks to the overall effect when the player was too far away to see them.

Your particle system can have as many LODs as you need, and you can manually control the ranges for each one.

Distributions are a set of data types for processing data in special ways, such as using a range for a value or interpolating a value along a curve. Whenever your particle system requires randomization or the ability to change some aspect of the particle over time, you will use a distribution to control that property.

Many properties found in Cascade modules can have different distributions applied to them. The actual value of the property is then set in the distribution.

8. Niagara

Why re-invent visual effects for Unreal Engine? Unreal Engine continues to expand its user base and is now used in many industries outside of game development.

Unreal Engine users are more diverse than ever, ranging from design students to small indie developers, to large professional studio teams, to individuals and companies outside the gaming industry. As they move forward, Epic's developers will not know everything about every industry that uses the Unreal Engine. They wanted to create a visual effects (VFX) system that would work for all users across all industries.

They wanted to create a new system that would give all users the flexibility to create the effects they wanted. Their goals for the new VFX system were:

1. Put complete control in the hands of the artists.
2. The ability to program and customize on every axis.
3. Better tools for debugging, visualization, and productivity.
4. Support for data from other parts of the Unreal Engine or external sources.

Total user control begins with access to data. Epic Games wants the user to be able to use any data from any part of the Unreal Engine, as well as data from other applications. So they decided to give the user everything.

In order to make all this data available to a user, you need to determine how someone can use the data. Namespaces provide containers for hierarchical data. For example, Emitter.Age [12] contains data about an emitter; Particle.Position contains data about a particle. A parameter map is a particle payload that contains all the attributes of a particle. This makes everything optional.

Any kind of data can be added as a particle parameter. We can add complex structures, transform matrices, or boolean flags. We can add these or any other data types and use them to simulate effects.

There are advantages to both the stack paradigm (as used in Cascade) and the graph paradigm (as used

in Blueprints). Stacks give users modular behavior and readability. Graphs give users more control over behavior. This new effect system combines the best of both paradigms.

Modules work in a graphical paradigm - we can create modules from HLSL in the script editor using a visual node graph. Modules interact with common data, encapsulate behavior, and are compiled together.

Emitters work in a stacked paradigm - they serve as containers for modules and can be stacked to create different effects. An emitter is single purpose, but it is also reusable. Parameters are passed from modules to the emitter level, but you can change modules and parameters in the emitter.

Like emitters, systems work in a stacked paradigm and also use a sequencer timeline to control the behavior of emitters in the system. A System is a container for Emitters. The system combines these emitters into a single effect. When editing a system in the Niagara editor, we can change and override any parameter, module or emitter in the system.

The particle simulation in Niagara works conceptually like a stack - the simulation flows from the top of the stack to the bottom, executing the modules in order. Importantly, each module is assigned to a group that describes when the module is executed. For example, modules that initialize particles or act when a particle appears belong to the Particle Spawn group.

Within each group, there may be several stages that are invoked at certain points in the system's life cycle. Emitters, systems and particles have Spawn and Update stages by default. Spawn stages are invoked in the first frame in which the group exists. For example, systems invoke their spawn stage the first time the system is created and activated on a level. Particles invoke their spawn stage whenever the emitter emits a particle, and spawn instructions are executed for each new particle created. Update stages are invoked in every frame where a system, emitter or particle is active.

There are also advanced steps such as events and simulation steps that can be added to the spawn and update flow. Events are called whenever a particle generates a new event and the emitter is configured to handle that event. Where possible, event handler steps occur in the same frame, but after the original event. Simulation steps are an advanced GPU feature. This feature allows you to run multiple sleep and update stages in sequence and is useful for building complex structures such as fluid simulations.

By adding each module to a stage (update, spawn, event, or simulation) in a group (system, emitter, or particle), we can control when the module runs and what data it processes. Stack groups are associated with namespaces that define what data the modules in that group can read or write.

For example, modules executing in the System group can read and write parameters in the System namespace but can only read from parameters that belong to the Engine or User namespaces. As execution moves down the stack from the System group to the Emitter group, modules executing in the Emitter group can read and write parameters in the Emitter namespace, but can only read from parameters in the System, Engine, and User namespaces. Modules in a particle group can only read from parameters in the System and Emitter namespaces.

Since modules in emitter groups can read parameters in the system namespace, a simulation relevant to all emitters can be performed once by modules in the system group, and the results of this simulation (stored in the system namespace) can be read by modules in the emitter group in each individual emitter. This continues with modules in the Particle group, which can read parameters in the System and Emitter namespaces.

In other respects, Niagara is very similar to Cascade. This system is more advanced, more flexible, and has endless possibilities for extending functionality, although this makes the process of creating particles more complicated. In this subsection, the main differences and architectural features between Cascade and Niagara have been presented, since the process of automated optimization in these systems is conceptually different.

9. Optimizing particle simulation systems

When creating a game, we can have a lot of variation in the FX workload depending on the composition of the scene. Sometimes it may be necessary to take steps to manage performance, such as dropping instances outside a certain range or instances that exceed a certain budget.

Effect Type resources allow you to configure a set of settings once and then apply them to a collection of Niagara systems.

The Effect Type object allows you to configure several different methods for selecting systems that exceed your budget. All of these options are available under the Budget Scaling heading.

Maximum global budget usage (see Fig. 6): This option allows you to set a budget above which any system will be discarded. Typically, this setting is set to a value between 0 and 1, corresponding to a percentage between 0 and 100%. You can set it to 1.5 if you want the system to be more permissive. This means that once a system reaches this percentage of your budget, it will be discarded. This is the best option if you value performance over appearance.

Maximum Distance Scale by Global Budget Utilization: This option allows you to customize the curve to determine how the distance at which you select systems decreases as your budget usage increases. For example, if your budget is very high, Niagara will only render those that are nearby, not those that are far away.

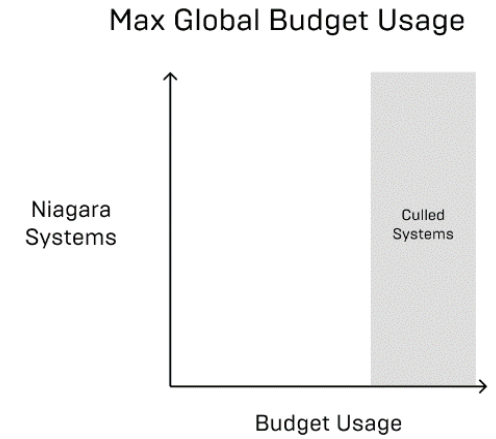

Fig. 6. Maximum global budget utilization

Maximum Instance Scale by Global Budget Usage: This option allows you to configure a curve that determines how the number of instances in your tier will decrease as budget usage increases. This will scale down all instances of all systems that are subject to this type of effect.

Maximum scaling of system instances by global budget usage: This option allows you to customize the curve (see Fig. 7) that determines how the number of instances at your tier decreases as budget usage increases. However, with this option, instead of dropping all instances on all systems, you drop a certain number of instances for each system.

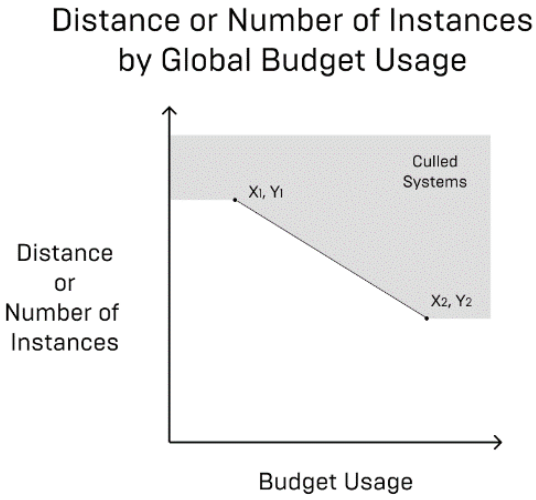

Fig. 7. Demonstration of the culling system operation

For each of these three parameters, which take the values Start X, Start Y, End X, and End Y, these values define a linearly interpolated curve. Anything above this curve is discarded. For an example of what the curve will look like, see the diagram below.

In the grand scheme of things, particle count plays a very small role in performance. Regardless of whether the screen is split or not, material complexity and screen coverage (overdraw) are your clear enemies when it comes to the overall cost of a given system. A simple emissive spark, with nothing more than a texture multiplied by vertex colors and connected to an emissive input in an unlit material, follows only a handful of instructions. You can create them in droves all day long, and the overall impact on your productivity is likely to be very small. Sprites are small, which means screen coverage is low, and the complexity of the material makes them cheap and fast to render. The number of vertices isn't really something you need to worry about in the long run unless you're reaching really extreme numbers (hundreds, thousands, or more).

A much bigger impact on overall performance is the number of instructions for your materials. For materials like fire and smoke, there are basically two ways to go. The first is to create a more complex material for your effect. In the fire example, you would create fewer sprites and let the randomness and complexity of the advanced material do the work of bringing the emitter to life. Another option is to use a cheaper material and spawn more sprites, keeping the overall cost the same but allowing more particles to do your job in achieving randomness, as opposed to a more complex material. Keep in mind that material costs decrease exponentially with distance (a quadrilateral drawn on the screen twice as far away from the camera costs 4 times less due to the fact that the total pixel area decreases exponentially with distance, reducing the number of pixels that are dragged).

In our case, we need to analyze how expensive our materials are, how many sprites we create, and how close to the screen we will approach these effects. These three properties are the main decision makers in terms of the complexity of the emitter, and they all need to be balanced.

In general, focus on reducing the complexity of your materials as a way to improve performance, and always be aware of potential drag when you are working with emitters as a whole. Don't get hung up on particle counts unless you are generating extreme numbers of particles, or you are generating meshes using mesh emitters that have extreme numbers of vertices.

Conclusions

After analyzing the industry, we came to the conclusion that Unreal Engine is the most suitable for our tasks. Unity is more popular among developers for portable devices, but at the same time it has a less developed VFX creation system. The system implemented in Unity has not been globally updated since the creation of the engine, and the capabilities implemented in it are not flexible enough to conduct a pure experiment and analyze possible optimization strategies. At the same time, Unreal Engine has the most advanced VFX creation system - Niagara, which provides the best opportunity to study and optimize particle simulation systems, as it allows not only to use ready-made solutions, but also to write code independently.

The comparison of Niagara and Cascade showed a fundamental difference in the principles of optimizing particle simulation systems. Only one principle is relevant now, as it was before - reducing the number of particles, but as practice shows, this is often not enough. The change in the principles of optimization of these systems shows that the search is still ongoing and the decisions made earlier by the engine developers are no longer relevant. This suggests that new solutions may also be imperfect, as they try to solve the optimization problem universally, rather than in the best way for mobile devices.

For this reason, further analysis of existing optimization methods and their improvement or creation of new ones is a relevant and promising task. To this end, we plan to study the principles of VFX, existing optimization methods, and conduct tests on portable devices.

REFERENCES

[1] *Матвєєв, Д.І., Лановий, О.Ф.* Методи спрощення опрацювання систем симуляції незалежних часток у середовищі Unreal Engine 4 // Èlektron. model. 2023, №45(2) с. 95-107. URL: https://doi.org/ 10.15407/emodel.45.02.095

[2] Cascade Particle Systems // Unreal Engine Documentation. URL: https://docs.unrealengine.com/ 4.26/en-US/RenderingAndGraphics/ ParticleSystems/

[3] Niagara Visual Effects // Unreal Engine Documentation. URL: https://docs.unrealengine.com/ 4.26/en-US/RenderingAndGraphics/Niagara/

[4] *Лановой А.Ф., Лановой А.А.* Моделирование поведения толпы на основе дискретно-событийного мультиагентного подхода // Східно-Європейський журнал передових технологій, 2014, №4(70), с. 52-57

[5] GPUSprites Type Data // Unreal Engine Documentation. URL: https://docs.unrealengine.com/ 4.26/en-US/RenderingAndGraphics/ParticleSystems/ Reference/TypeData/ GPUSprites/

[6] Collision Modules // Unreal Engine Documentation. URL: https://docs.unrealengine.com/4.26/en-US/ RenderingAndGraphics/ParticleSystems/Reference/ Modules/ Collision/

[7] Event Modules // Unreal Engine Documentation. URL: https://docs.unrealengine.com/4.26/en-US/Rendering AndGraphics/ParticleSystems/Reference/Modules/ Event/

[8] VFX Optimization Guide // Unreal Engine Documentation. URL: https://docs.unrealengine.com/ 4.26/en-US/RenderingAndGraphics/ParticleSystems/ Optimization/

[9] *Матвєєв, Д.І., Лановий, О.Ф.* Проблеми оптимізації графіки під пристрої віртуальної реальності // ΛΌГОΣ.ONLINE, 2020, №14. URL: http://eoi.citefactor.org/10.11232/2663-4139.14.04

[10] Особливості підготовки 3D моделей для використання у VR проектах // Science, Research, Development. URL: http://www.xneh.com.ua/ files/118_01_xi_2021.pdf#page=35

[11] Порівняння методів текстурування моделей для мобільних платформ // Science, Research, Development. URL: http://www.xneh.com.ua/ files/118_01_xi_2021.pdf#page=37

[12] Дослідження інструментів та засобів оптимізації 3D-графіки в комп'ютерних іграх та їх застосування до ігор у жанрі "First-person Shooter" // Електронний архів ХНУРЕ. URL: https:// openarchive.nure.ua/server/api/core/bitstreams/ e8582e45-10b9-44bf-aabc-cb0b120389ee/content

The article was delivered to editorial stuff on the 06.04.2023

УДК 004.42 DOi 10.30837/ bi.2023.1(99).14

Eduard Sheliemietiev[1], Yuriy Novikov[2], Oleksii Nazarov[3], Nataliia Nazarova[4]

[1] ХНУРЕ, м. Харків, Україна, eduard.sheliemietiev@nure.ua

[2] ХНУРЕ, м. Харків, Україна, yuriy.novikov@nure.ua, ORCID iD: 0000-0003-1910-3256

[3] ХНУРЕ, м. Харків, Україна, oleksii.nazarov1@nure.ua, ORCID iD: 0000-0001-8682-5000

[4] ХНУРЕ, м. Харків, Україна, nataliia.nazarova@nure.ua, ORCID iD: 0009-0007-7816-7088

INVESTIGATE METHODS FOR SMOOTH TRANSITIONS BETWEEN LEVELS OF DETAIL FOR EFFECTIVE VISUALIZATION OF 3D SPACES

The object of study is 3D-rendering. The subject of study is providing smooth transitions between LOD models. The purpose of the work is to increase rendering performance of large 3D-scenes considering smooth transitions between models with different levels of detail. The study examines existing algorithms for enabling smooth transitions between levels of detail considering computational complexity and visual appeal of said methods.

3D-RENDERING, ALPHA-BLENDING, COMPUTER GRAPHICS, GEOMORPHING, LOD, NOISE-BLENDING, POPPING EFFECT

Шелємєтьєв Е.О., Новіков Ю.С., Назаров О.С., Назарова Н.В. Дослідження методів плавних переходів між рівнями деталізації для ефективної візуалізації 3D-просторів. Об'єктом дослідження є 3D-рендеринг. Предметом дослідження є процес забезпечення плавного переходу між LOD моделями. Метою роботи є підвищення ефективності візуалізації великих 3D сцен із урахуванням плавності переходу між моделями з різним рівнем деталізації. В ході роботи виконується дослідження існуючих алгоритмів плавного переходу між рівнями деталізації з точки зору обчислювальної складності та візуального вигляду. Результати дослідження дозволяють прийняти рішення про застосування того чи іншого методу плавного переходу з урахуванням їх сильних та слабких сторін.

3D-РЕНДЕРИНГ, LOD, АЛЬФА-ЗМІШУВАННЯ, ГЕОМОРФІНГ, ЕФЕКТ ПОППІНГУ, КОМП'ЮТЕРНА ГРАФІКА, ШУМИ

Introduction

Levels of Detail (LOD) play an important role in improving the performance of 3D graphics applications by striking a balance between frame processing speed and user experience. The essence of this optimization method is that the further away from the virtual camera, the simpler the 3D models are drawn. This allows you to focus computing resources on objects that are more important to the user.

Since the camera position is often dynamic, some objects need to be detailed directly in the user's field of view. The problem arises of how to smoothly transition between different levels of detail — replacing simplified models with highly detailed ones and vice versa.

This paper explores methods for smoothing the transition between levels of detail to mitigate the notorious "popping" effect. This effect occurs when the transition between LODs results in abrupt and noticeable changes in the level of detail, disrupting the visual coherence of the scene.

By studying and comparing different approaches, this research aims to provide valuable insights into effective strategies for preventing or minimizing the popping effect, which will ultimately help to balance performance and optimal visual user experience.

Thus, the goal of the work is to improve the efficiency of the visualization of large 3D scenes, taking into account the smooth transition between models with different levels of detail.

The subject of the research is 3D visualization (rendering).

The subject of the study is the process of ensuring a smooth transition between LOD models.

To achieve this goal, it is necessary to solve the following tasks

– To analyze existing methods and algorithms for eliminating the popping effect when transitioning between levels of detail;

– Study the factors that affect the efficiency of drawing distant objects;

– Implement and compare methods for smooth transition between levels of detail;

– Draw conclusions about the appropriateness of using a particular approach.

Based on the results obtained, it will be possible to draw conclusions about

– Which algorithm should be used to ensure the highest performance;

– Which algorithm provides the smoothest transition;

– To what extent it is generally advisable to use a smooth transition between levels of detail compared to discrete LODs.

This work is relevant now and will be in the future as 3D visualization becomes more widespread and has many applications in the modern world.

1. Subject area analysis

Today, 3D visualization has many applications: video games, graphic design, visual effects, virtual reality, cinematography, visualization, virtual engineering, etc. These fields

require 3D graphics to look good (realistic or stylized) and to be efficient. Often these two requirements are in conflict, and application specialists must find a compromise between the quality of the image and the cost of generating it. Real-time 3D rendering is particularly challenging from this perspective when the cost of producing an image is expressed in terms of frames per second (FPS).

Often, a very effective way to increase the speed of visualization is to use 3D models with different levels of detail. The essence of this method is to draw simpler geometry and use faster shaders the further the object is from the camera.

For example, instead of rendering every detail of a very distant glass skyscraper, you can draw a stretched cube with a pre-created texture. With this approach, it's important to balance the amount of detail with the distance to the viewer, as you can either lose the effect of immersion in the 3D world or get a very long frame generation time.

When working with LOD, you should also pay attention to the smooth transition between levels of detail as the camera approaches the object. For example, if you decide to draw distant trees using a square with a texture stretched over it (the so-called billboard), when the virtual camera quickly approaches the forest, you can see how the 2D image of the tree suddenly turns into a 3D model (popping effect). This phenomenon has a negative impact on the user experience, as it reveals the falsity of the virtual world.

Several methods are used to eliminate the popping effect: LOD blending (smooth blending of two adjacent levels of detail using transparency or noise) and geomorphing (creating additional intermediate states of the model by bringing the geometry of one level of detail closer to another).

Although there are a number of methods available today, each approach has its advantages and disadvantages, and it is often necessary to find a balance that satisfies the requirements for performance and appearance.

2. Work Relevance

This work has significant relevance in the context of 3D graphics and virtual environments. Its results solve a critical problem that directly affects user experience and application performance.

Modern applications and 3D models strive for ever-increasing levels of detail. In the past, computers could draw a relatively small number of 3D polygons, but in recent years, computer-generated images (CGI) have become indistinguishable from real photographs. This is largely due to the use of highly detailed models.

And although modern hardware is capable of drawing large numbers of polygons, computing resources are not unlimited (especially in the context of real-time 3D rendering and on mobile platforms), so software developers still have to make compromises. LOD is still a must-have method for maintaining the illusion of continuity and optimal performance.

Users moving in virtual space often experience the "popping" effect, which negatively affects immersion. Choosing the wrong method to smoothly transition between levels of detail destroys the user experience, as the viewer's attention is directed not to what the designer or developer intended, but to the 3D object that is transitioning to a different level of quality. By exploring different methods to prevent this phenomenon, this paper aims to offer practical solutions that can be applied in a wide range of applications: from video games to architectural modeling and virtual reality.

In summary, the relevance of the research lies in its ability to improve the user experience in a variety of applications. Whether it is games, simulations, or resource-constrained platforms, the results of this research can positively impact image accuracy, performance, and user experience.

3. Problem statement

In accordance with the identified problems, we describe the task for research practice:

– Analyze existing methods and algorithms for eliminating the popping effect when switching between levels of detail;

– To study the factors influencing the efficiency of rendering distant objects;

– Perform the software implementation of the studied algorithms;

– Draw a conclusion about the appropriateness of applying a particular approach.

Based on the results obtained, it will be possible to draw conclusions about

– Which algorithm should be used to ensure the highest performance;

– Which algorithm provides the smoothest transition;

– To what extent it is generally advisable to use a smooth transition between levels of detail compared to discrete LODs.

Thus, the study will allow us to choose one or another algorithm to ensure a smooth transition between levels of detail, taking into account their strengths and weaknesses.

4. Justification of research methods and stages

The object of research is 3D visualization (rendering).

The subject of the research is the process of ensuring a smooth transition between LOD models.

The goal of the research is to improve the efficiency of the visualization of large 3D scenes, taking into account the smooth transition between models with different levels of detail.

The empirical scientific method "experiment" is used in this research: during a series of experiments the parameters of the selected algorithms for ensuring smooth transition between levels of detail (the value of the popping effect and performance) are determined.

Then the theoretical method of "analysis" will be used to decompose the obtained results separately from each other. As a result, it will be decided which algorithm is more suitable in a given situation and for what reasons.

This research consists of the following stages

1) Preparation of the study — definition of the purpose, specification of the subject and object of the study, formulation of the hypothesis and research methodology;

2) Experimental research and data processing — preparing and conducting experiments, processing the obtained data;

3) Analysis of research results — drawing conclusions about the feasibility of using certain algorithms for smooth transition between levels of detail;

4) Evaluation of the application of the research results.

5. Development of formal mathematical model of subject area

The study will compare the algorithms for ensuring a smooth transition between LOD models by two factors:

– the time of image generation;

– the importance of the popping effect.

Each algorithm studied is characterized by the following gray box mathematical model (see Figure 1): it is an expression of the complexity of two adjacent levels of detail between which a smooth transition is made (x1 and x2) and the distance of the object from the camera (d), as well as a random error ξ that we cannot control.

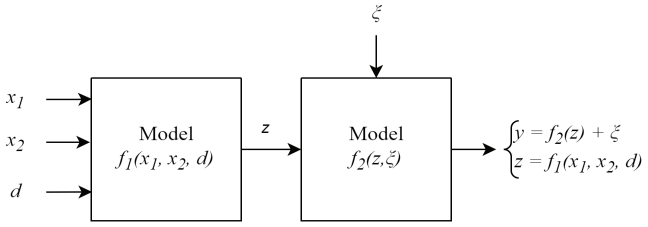

Fig. 1. Mathematical Model of the Gray Box for the Smooth Transition between Levels of Detail Algorithm

The function f1(x1,x2,d) is a controlled parameter that indicates the method of smooth transition between the levels of detail under study. The result of the function is the parameter z, which denotes a set of instructions to the GPU to draw 2 adjacent levels of detail at a given distance from the camera.

The function f2(z,ξ) denotes the environment that performs the drawing (the operating system and the GPU). This environment is uncontrolled and a black box. All we can do with it is pass input commands and data, and the output is the generated image.

Note that to simplify further calculations of the complexity of the x1 and x2 detail levels, the number of polygons in the model (triangles that form the object's edge) is taken without regard to the position of their normal relative to the camera (when using the back-face culling technique, polygons that are not visible from the camera, such as the back of a house, are cut off and do not participate in further steps of the graphical pipeline).

In the experiment comparing image generation times, y is the time (in milliseconds) to generate an image, and ξ is a random measurement error.

In the experiment to evaluate the significance of the popping effect, y is the difference between the reference image (the original high-poly model) and the image generated by the smooth transition algorithm between detail levels.

In this experiment, all random values (noise textures, etc.) are pre-generated, so the result of the experiment is completely deterministic (determined by x1, x2, d), and the error value ξ is zero.

After conducting the experiments, it will be necessary to conclude how well the algorithms meet certain criteria and to give examples of situations in which it is advisable to use a certain method of smooth transition between levels of detail.

6. Experimental methodology

To perform the experiments, we will create a software that consists of the following functional modules:

1) a module to measure the performance of the algorithms when displaying a large number of 3D models at different distances from the camera;

2) a module for measuring the popping effect when comparing the performance of the method under study with a reference (high-poly model).

Algorithms for smooth transition between the levels of detail to be studied:

– Alpha mixing;

– Mixing with noise;

– Geomorphing.

These methods are compared with rendering:

– The original high-poly model with no levels of detail;

– discrete levels of detail without smooth transition.

The experiment to measure the performance of the smooth transition algorithms will measure the time required to render a large number of models (50, 250, 500, 1,000, and 2,000 models) at different distances from the camera. The number of objects should be large enough because modern GPUs process large 3D scenes very quickly.

The objects are located along the x-axis in the range from the beginning of the camera's visible range to the distance required to activate the lowest level of detail at the same interval, which is equal to (1):

$$dx = (w - camera_x) / n, \quad (1)$$

where dx is the interval at which objects are placed in the scene, n is the number of objects, $camera_x$ is the camera coordinate, and w is the distance of the minimum level of detail.

Accordingly, the camera is oriented along the positive x-axis.

Also, occlusion culling should be turned off because the models overlap each other (if occlusion culling is turned on, the objects in front will overlap the models behind, and the latter will not be drawn).

As mentioned earlier, this experiment has an error ξ, so the time measurement should be performed several times (we assume that 3600 times is sufficient) and the average value should be taken.

Since the experiment is performed on a modern multiprocessing operating system, we assume that the time measurement errors are caused by other processes on the computer, so the average time is the time of the algorithm.

The magnitude of the popping effect is measured as follows:

1) The 3D model under study is placed on the scene at the minimum distance from the camera, so that the maximum level of detail is initially displayed;

2) the object is moved at a constant distance along the camera's line of sight, passing through all states of detail levels;

3) after each movement, the resulting image is stored;

4) after reaching the lowest level of detail, the object is gradually brought back in the same way;

5) Steps 1-4 are performed for the reference model and the specified algorithm, and the absolute difference between the pixel values of the images is found. The resulting image has only one channel, which is interpreted as black (value 0) and white (value 1). Since the input images are multichannel (RGB), the difference is the arithmetic mean of all channels;

6) for each difference image, the root mean square error (RMSE) is calculated [1].

As a result, we get a single number — how much the image generated by a particular method differs from the reference.

The image obtained in step 4 can be useful for visually assessing the difference between the algorithms.

As a result of the experiment you will find

– The fastest and slowest algorithms;

– which algorithm has the most pronounced popping effect.

Based on the results, it will be possible to draw conclusions about

– Which algorithm should be used to ensure the highest performance;

– Which algorithm provides the smoothest transition;

– the general advisability of using a smooth transition between levels of detail versus discrete LODs.

7. Nature of Experimental Errors and Uncertainty

The measurement error is present only when performing an experiment to evaluate the performance of algorithms for smooth transition between levels of detail. Despite the fact that the same set of data is passed to the program, the processing time is different each time.

This is due to the fact that the application is not running in a separate, isolated environment: many other processes and threads are running in parallel on the same device.

Hardware and software caching, code interpretation, and JIT compilation must also be taken into account. Because of this, the first experiment usually takes a little longer. Therefore, in practice, before measuring the execution time of a program, a so-called "warm-up" is performed (preferably with as many conditional transitions as possible). This increases the chance that subsequent experiments will have approximately the same runtime.

When calculating the size of the popping effect, operations are performed on floating point numbers (addition and division operations). It is known that this introduces an error in the resulting amount, but its relative value is very small and does not matter when comparing large numbers. Therefore, to simplify the experiment, its existence is allowed.

8. Alpha blending

One of the easiest methods to implement in software is the smooth transition method called alpha blending (LOD blending).

The essence of this method is to simultaneously display two levels of detail that are mixed using the alpha channel (transparency channel) according to the transition coefficient. This coefficient can take values from zero (the beginning of the transition) to one (the end). Conventionally, the method can be represented by the following formula (2):

$$\alpha_{result} = (1 - k) * \alpha_{prev} + k * \alpha_{next}, \quad (2)$$

where α_{result} — is the transparency of the resulting pixel on the screen, α_{prev} — is the transparency of the pixel of the previous detail level (where the transition started), α_{next} — is the transparency of the next (target) detail level, k is the transition coefficient.

An alpha channel value of one corresponds to a completely opaque pixel, and a value of zero corresponds to a completely transparent (hidden) pixel.

As the virtual camera moves and reaches the distance at which the transition should take place, the old model gradually fades out and a new model gradually appears in its place. When the transparency of the old model reaches zero, the old model is not drawn to save resources.

Visually, the principle of alpha blending is shown in Figure 2.

Fig. 2. Principle of alpha blending, transitioning from a low quality model to a high quality model

When implementing alpha blending, it is important to make the transition within a small distance range. For example, if you need to make a transition at a distance of 1m, the start and end of the transition should be 0.95m and 1.05m respectively.

Alpha blending can be combined with billboards because the virtual structure of the model is not important for this method. In the context of LOD optimization, billboards are usually used to improve the efficiency of visualizing the farthest objects (the lowest quality level).

Billboards are two-dimensional planes or sets of connected polygons that always face the camera, simulating a three-dimensional object. This technique is used to more efficiently represent distant or small objects in a scene, reduce computational load, and improve performance.

The advantage of billboards is that only two triangles are needed to represent a plane. Often, multiple planes are combined to create the illusion of a large object in space (tree leaves, bushes, clouds, etc.) [2].

This method of smooth transition has two major drawbacks.

First, visualizing two models at the same time is computationally intensive. The main reason for using detail levels is to reduce the number of polygons drawn simultaneously to speed up the creation of the image on the screen, but the transition in alpha blending requires drawing both models. As a result, this method can sometimes be detrimental to performance.

One way to deal with this drawback is to limit the number of objects that can transition between levels of detail at the same time. This helps to avoid jumps in the number of drawing function calls, which guarantees a more stable number of frames per second. Keep in mind that delaying the transition creates a delay that can be detrimental to the user experience.

Second, alpha blending is very noticeable to the viewer at close range: in certain situations, the model can look like a translucent ghost.

Alpha blending works very well at a distance from the camera, when adjacent layers of detail have a small number of polygons and the visual effect of blending is not noticeable enough.

9. Blending with noise

The smooth transition between levels of detail using noise is very similar to alpha blending.

However, in this method, the transparency of each pixel is a discrete value: either zero or one. In other words, this algorithm uses either fully transparent or fully opaque pixels to create a smooth transition.

Thus, the decision whether to display a particular pixel with x and y coordinates of each model is made based on the numerical value of a random variable (noise): if the value of the transition coefficient k exceeds the threshold set by the noise function, we use a pixel from the next level of detail, otherwise — the previous one (3):

$$\alpha_{result} = \alpha_{next}, \text{ if } k > f_{noise}(x, y); \quad \alpha_{result} = \alpha_{prev}, \textit{otherwise}, \tag{3}$$

where $f_{\text{noise}}(x, y)$ — is a noise function that returns a value between zero and one.

In computer graphics, noise is a pseudo-random value (one-dimensional or multidimensional) used to add detail to computer-generated images. Noise is very easy to compute, and its applications are almost limitless: from cloud and particulate visualization to ocean wave and tornado simulation [3].

For better performance, noise is pre-computed and stored as textures. This format is very convenient for GPUs as they are specialized in sampling data from textures.

So for a smooth transition, we need a two-dimensional noise value in the range of zero to one, stored in advance as a two-dimensional texture. Its size doesn't need to be large: the maximum size is the resolution of the generated image. Even if you compute a texture that is too small, it can be re-rendered, and this will minimize the user experience.

Noise texture sampling can be done at the processing stage instead of separately for each model to be drawn. This can speed up the algorithm and avoid noise repetition when objects are too close together.

The most popular noise functions include the following [4]:

a) white noise — returns pseudo-random values even for input values that are very close to each other;

b) gradient noise — interpolates white noise values and returns close noise values for close parameters;

c) pearlin noise — a subtype of gradient noise in which visual details have the same size; it is used to make computer graphics more realistic;

d) multilayer noise — uses a combination of gradient noise with different levels of scale and weight; allows you to get noise that has both high-frequency details and low-frequency details;

e) voronoi noise — returns a noise value that looks like a set of cells (or distances between cells).

A separate type of noise is dithering [5]. It is used to give a random variable the desired stylistic appearance. An example of a smooth transition using dithering is shown in Figure 3.

Fig. 3. Principle of noise blending, transition from low-quality model to high-quality model

The figure shows a useful feature of noise-based blending — at a certain value of the smooth transition threshold, only one model can be drawn at a time. This helps to reduce the load on the GPU. Of course, this trick is not suitable for all types of models and environments.

Another advantage of this method over alpha blending is that the designer has full control over the appearance of the transition: you only need to replace the noise texture, and you can make certain parts of the model transition faster than others; you can set the transition direction (from bottom to top, from edges to center, etc.).

Noise-based transitions between levels of detail have the same drawbacks as alpha blending: the need to draw two models at the same time and the obviousness of the transition when viewed up close.

In practice, however, the noise-based transition is more efficient in terms of performance [6].

The fact is that the most common way of working with translucent models is much slower than drawing opaque objects. Semi-transparent geometry often requires a separate graphics pipeline step from opaque objects, where the

depth test is disabled and transparent polygons are sorted from back to front [7].

The lack of a depth test causes every pixel to be redrawn (overdraw), which is very detrimental to performance as the GPU wastes time on a pixel that is not visible to the user anyway.

Sorting can be quite expensive when the number of model triangles is large, which is exactly the case when alpha blending is used (especially for models of levels of detail close to the virtual camera).

In practice, noise blending is often used for vegetation, which is essentially a reflection of random variables in nature [8].

10. Geomorphing

Geomorphing is another method of smooth transition between levels of detail. Its essence is to approximate a 3D model to create intermediate transition states.

The main operations of geomorphing are vertex splitting (new vertices are added to the model as the quality of the model increases) and edge collapse (some vertices are removed as the quality of the model decreases) [9].

During the transition, not only the number of vertices is changed, but also their position and other additional attributes: normal, color, texture coordinates, etc. (see Figure 4), which prevents popping.

Fig. 4. The principle of geomorphing, going from a low quality model to a high quality model

Most vertex attributes are linearly interpolated. Normals, which are 3-dimensional vectors with a length of one unit, are modified using directional interpolation to preserve their length.

When implementing geomorphing in software, it is important to consider the situation when vertex splitting or edge convolution is performed during an existing transition. In other words, vertices should be able to perform animations of transitions that overlap in time.

One way to optimize the geomorphing method is to perform a smooth transition only for the visible part of the model (view-dependent LOD control) [10]. This technique is useful for large models (both in size and number of vertices). The entire mesh is divided into parts (clusters) that have their own memory buffers. The decision of which cluster to draw on the screen is based on an optimization technique called frustum culling.

The term "frustum" refers to a pyramid-shaped viewport that covers the visible area of a 3D scene [11]. This truncated pyramid is obtained by projecting the camera perspective onto the near and far planes. Frustum culling selectively renders only those objects that fall within this truncated area, discarding those that are out of view.

By eliminating invisible objects early in the visualization pipeline, we significantly reduce the computational load and increase overall performance.

During the object removal process, each model in the scene is checked on a slice to determine if it intersects or lies outside the field of view. Several algorithms are used to quickly determine if an object is potentially visible, such as checking with spheres or Axis-Aligned Bounding Boxes (AABB). If an object is completely invisible, it is excluded from the rendering process, eliminating unnecessary geometry, lighting, and shadow calculations.

This optimization is especially useful in scenes with a large number of objects, allowing you to focus rendering resources on rendering only the models that are visible to the camera.

The main difficulty in dividing the model into clusters is ensuring a consistent and synchronized transition for vertices that are on the boundary of two clusters. If this implementation detail is overlooked, there may be breaks along the edges of the mesh parts, causing a "popping" effect in some cases.

The easiest way to implement this is to have the transition performed entirely on the CPU. A list of vertices that perform a smooth transition can be stored in RAM, and their attributes need to be updated every frame. However, it should be noted that this approach is not optimal for a large number of vertices, since after updating the data in CPU memory, it must be sent to the GPU, which can cause some delay [12].

Another approach is to store the high and low quality models in GPU memory. For a smooth transition, a weight parameter (from zero to one) is used to interpolate vertex parameters. Models of different quality can be preloaded with priority. Obviously, this method requires more video memory, but it significantly reduces the amount of data the CPU sends to the graphics card for each frame.

The advantage of the geomorphing method is that only one model is drawn at a time when switching between levels of detail. Also, a smooth change in the model is less noticeable to the viewer at any distance.

The disadvantages of this method are the complexity of the implementation and the need to pay special attention to the placement of the 3D model in memory. Thus, geomorphing becomes the key factor that determines how 3D rendering is performed.

11. Software implementation

The program for the study is developed using the C# programming language, the .NET 6 platform, and the MonoGame video game development framework (DirectX-based drawing).

Each of the studied smooth transition methods is represented by a separate class implementing the common ILodTransition interface. Each class receives models of detail levels, the degree of transition, the object transformation matrix, and a reference to the graphics pipeline for drawing.

Smooth transitions between detail levels are performed depending on the distance to the camera.

Alpha blending is implemented using a separate step for translucent objects and double-pass rendering. Because the last step of the Output-Merger (OM) graphics pipeline is rendering, all pixels with a depth value less than the depth of the z-buffer are not rendered at all. For this reason, semitransparent objects are usually drawn separately from opaque objects and then added to the final image with depth.

Double-pass rendering helps reduce the popping effect on certain models. The depth buffer does not work correctly when drawing translucent polygons of the same model, and sometimes transparent triangles can be seen through each other — resulting in regions of the model having different alpha values. When mixing levels of detail, it is important that all polygons of both models have the same transparency level, which is equal to one. If the transparency is less than one, the models are transparent. If the transparency is greater than one, the RGB color is usually different from the original model color, which is very noticeable to the user.

Therefore, each level of detail is drawn twice during the transition (a total of 4 draw calls). The first draw should only capture the depth value. Any changes in the RGB channels should be ignored. At this stage, a very simplified pixel shader is used that returns a simple color. This is done to avoid wasting time calculating the light and color of the pixels. In the second stage of drawing the model, the values of the RGBA channels are recorded based on the mixing factor (formula 4.1). Before that, you must disable writing to the z-buffer and leave it read-only. Before drawing the next model, the depth buffer must be cleared, otherwise the pixels of the second model will be discarded due to the depth remaining in memory:

```
private void DrawTransparentModelDoublePass(
 LodLevel lod, GraphicsDevice graphicsDevice, float alpha)
{
 graphicsDevice.Clear(
  ClearOptions.DepthBuffer, Color.Transparent, 1f, 0);
  graphicsDevice.DepthStencilState = DepthStencilState.
Default;
 graphicsDevice.BlendState = this.blendDepthOnly;
 this.mainMaterial.AlphaPass.Apply();
 foreach (var part in lod.Mesh.MeshParts) {
  graphicsDevice.SetVertexBuffer(part.VertexBuffer);
  graphicsDevice.Indices = part.IndexBuffer;
  graphicsDevice.DrawIndexedPrimitives(
   PrimitiveType.TriangleList,
   part.VertexOffset, part.StartIndex, part.PrimitiveCount);
 }
 using var bs = new BlendState {
  ColorSourceBlend = Blend.BlendFactor,
  ColorDestinationBlend = Blend.InverseBlendFactor,
  AlphaSourceBlend = Blend.BlendFactor,
  AlphaDestinationBlend = Blend.One,
  BlendFactor = new Color(alpha, alpha, alpha, alpha),
 };
 graphicsDevice.BlendState = bs;
graphicsDevice.DepthStencilState=DepthStencilState.Depth
Read;
 this.mainMaterial.MainPass.Apply();
 foreach (var part in lod.Mesh.MeshParts) {
  graphicsDevice.SetVertexBuffer(part.VertexBuffer);
  graphicsDevice.Indices = part.IndexBuffer;
  graphicsDevice.DrawIndexedPrimitives(
   PrimitiveType.TriangleList, part.VertexOffset,
   part.StartIndex, part.PrimitiveCount);
 }
}
```

The noise-based smooth transition implementation uses the built-in HLSL clip function to discard a given pixel from further processing. Such a pixel is not written to either the color texture or the depth buffer. This greatly simplifies image rendering, since you only need to draw models once, and you don't need a separate step for transparent objects. The rules by which each pixel is gradually cut off or appears are loaded as a 16x16 black and white texture.

Smooth transition based on geomorphing is implemented on the GPU. This helps to reduce the memory bus load, since the pre-prepared smooth transition mesh is loaded into video memory only once, and further interpolation between vertex attributes is performed in the vertex shader using the progress variable:

```
MainVertexShaderOutput GeomorphVS(
 in GeomorphVertexShaderInput input) {
    MainVertexShaderOutput output;
    float3 avgPos = lerp(
     input.StartPosition, input.EndPosition, Progress);
    float3 normal = lerp(
     input.StartNormal, input.EndNormal, Progress);
    output.Position = mul(float4(avgPos, 1.0),
                WorldViewProjection);
    output.Normal = normalize(normal);
    return output;
 }
```

The geomorphic feature set is built by finding the closest vertices between high- and low-quality models. Since 3D models often have vertices with the same positions but different additional attributes (normal, UV coordinates, etc.), a simple distance between vertex coordinates is not sufficient: the distance metric must also take into account normal vectors. If you ignore this property, triangles may have incorrect illumination and color during the transition, which negatively affects the smoothness of the transition.

The geomorphic mesh is stored in the cache using the Least Recently Used (LRU) strategy. When a transition is generated based on two models, it is added to the cache,

which allows it to be reused both during the next drawing call and during the same frame.

Creating a mesh is a slow operation for models with a large number of vertices. The slowest part is the calculation of the nearest vertices. Unfortunately, even with the use of parallelization, the delay in generation is quite noticeable. In practice, it is suggested to pre-calculate all the nearest vertices at the compilation stage of the program and just read them from the file. Another method is to create the model in the background thread and display it as soon as it is ready. In the software implementation under study, the delay is not critical, but it is taken into account when conducting experiments.

Analyzing smooth transition metrics

Let's measure the performance of discrete and smooth transition algorithms. The computer on which the measurements are performed has the following characteristics:

– Windows 10;
– Intel(R) Core(TM) i5-8250U 1.60GHz CPU;
– 8 GB OF RAM;
– NVIDIA GeForce MX250 GPU;
– 2 GB VRAM.

The resolution of the drawing window is 800x600 pixels.

The average time for drawing 3D space as a function of the number of models is shown in Figure 5.

It can be seen that the slowest algorithm for a relatively small number of objects (50-500 models) is alpha blending. However, the running time of this method is linear over the range studied, making it the fastest method for smooth transitions when visualizing 1000-2000 models.

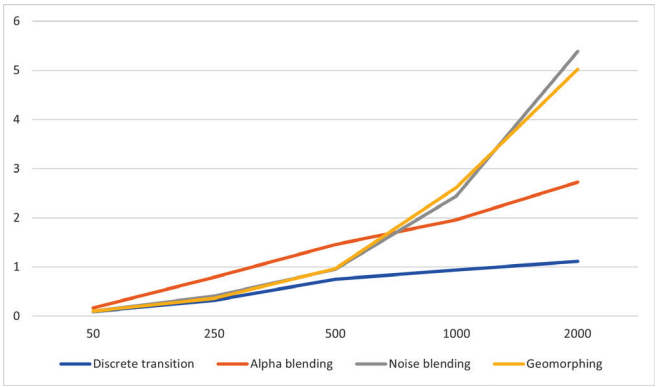

Fig. 5. Comparison of drawing times for smooth transition ethods

Geomorphing and noise blending have approximately the same runtime, which increases exponentially with the number of objects. Using the Windows Task Manager, running these methods on a large amount of data will load the GPU to 100% and the CPU to about 8%. It can be assumed that for a given hardware and software configuration, the GPU is the bottleneck for these methods.

The discrete transition between detail levels is the fastest. Let's analyze the magnitude of the popping effect for the studied algorithms for transitioning between levels of detail (see Figure 6).

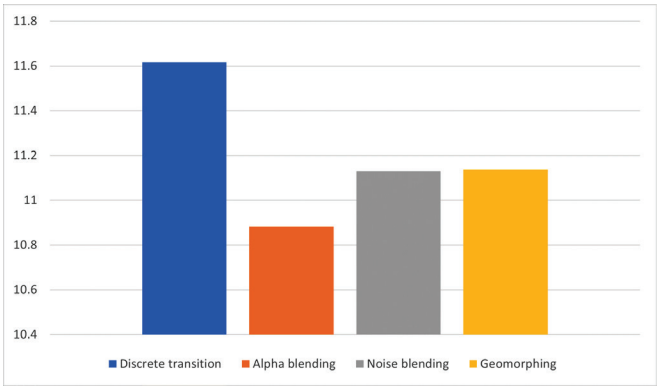

Fig. 6. Comparison of Popping Effect Levels

The figure shows that for the Stanford rabbit model under study, alpha blending is the best method for smooth transition. Noise-based transition and geomorphing have approximately the same amount of popping. It should be noted, however, that all of the algorithms studied are significantly smoother than discrete transitions.

Based on the results of the study, the following conclusions can be drawn about the feasibility of using smooth transition methods for the 3D model under study:

a) Alpha blending, although it gives the best visual appearance, is the slowest method in 3D spaces with small and medium number of objects. Alpha blending is the most efficient transition method when you need to display a transition for many objects at once;

b) Noise-based transition has an average level of performance, does not require much computing power for preparation, but is acceptable for drawing a limited number of models at the same time;

c) Geomorphing has about the same performance as noise-based blending, but it has one drawback: it requires the creation of a 3D mesh for a smooth transition.

Further research

The following areas of activity can be identified for further research on smooth transition methods:

– to study the work of the algorithms on a larger number of 3D models of different categories (architecture, landscape, vegetation, transportation, interiors, people, etc.);

– determine the effectiveness of each algorithm as a function of distance from the viewer;

– extend the metrics for evaluating the effect of popping (distances in CIELAB color space, graphical representation of popping, consideration of the statistical significance of the measurement);

– explore the possibility of automatically recommending transition methods depending on the type of model and the number of objects in 3D space to obtain the most effective values of the factors under study;

– to investigate the use of additional visual effects in combination with levels of detail (textures, normals, reflections, backlighting, ray tracing, etc.).

Conclusions

During the research internship, the students studied methods for smooth transitions between levels of detail (LOD) for three-dimensional graphics.

The focus was on finding a balance between performance and good-looking transitions. In particular, the research focused on eliminating the "popping" effect that disrupts visual consistency during transitions between LODs. A comprehensive analysis of existing methods and algorithms provided valuable information on strategies to prevent and minimize this effect.

It was found that alpha blending is the easiest method to implement for smooth transitions and reduces the popping effect the most, but is the slowest for a small number of models.

Noise blending has an average level of performance and a sufficient level of visual consistency in the image. This method also allows the most creative freedom in rendering objects and does not require any model preparation.

Geomorphing has the same performance and image consistency characteristics, but it is relatively difficult to implement and requires extensive calculations to prepare a 3D model for blending.

In practice, the results of this study are expected to help select algorithms that offer an optimal balance between computational efficiency and good looks, taking into account different circumstances. As 3D visualization continues to grow in popularity in a variety of applications, this work remains relevant and provides insights that can be used in future developments in the field.

References

[1] Jim Frost. Root Mean Square Error (RMSE). Statistics by Jim. URL: https://statisticsbyjim.com/regression/ root-mean-square-error-rmse/

[2] Anton L. Fuhrmann, Eike Umlauf, Stephan Mantler. Extreme Model Simplification for Forest Rendering. ResearchGate. URL: https://www.researchgate.net/ publication/221314842_ Extreme_Model_Simplification_ for_Forest_Rendering

[3] State of the Art in Procedural Noise Functions. [A. Lagae, S. Lefebvre, R. Cook, DeRose, G. Drettakis, D.S. Ebert, J.P. Lewis, K. Perlin, M. Zwicker]. URL: https://graphics.cs.kuleuven.be/publications/LLCDDELPZ10STARPNF/

[4] Noise Functions. Ronja's Shader Tutorials. URL: https://www.ronja-tutorials.com/noise.html

[5] The Importance of Dithering Technique Revisited with Biomedical Images – A Survey. [Liu Yue, P. Ganesan, B.S. Sathish, C. Manikandan, A. Niranjan, V. Elamaran, Ahmed Faeq Hussein]. ResearchGate. URL: https://www.researchgate.net/publication/329763540_The_Importance_of_Dithering_Technique_Revisited_With_Biomedical_Images-A_Survey

[6] Nithin Pranesh. Smoother LOD Transitions in Cesium for Unreal with Dithered Opacity Masking. 20.10.2022. Cesium. URL: https://cesium.com/blog/ 2022/10/20/smoother-lod-transitions-in-cesium-for-unreal/

[7] Transparency (or Translucency) Rendering. Nvidia Developer. URL: https://developer.nvidia.com/content/ transparency-or-translucency-rendering

[8] Benny Onrust, Rafael Bidarr, Robert Rooseboom, Johan van de Koppel. Procedural generation and interactive web visualization of natural environments. The 20th International Conference. ResearchGate. URL: https://www.researchgate.net/publication/ 300490331_Procedural_generation_and_interactive_ web_visualization_of_natural_environments

[9] Hugues Hoppe. Smooth View-Dependent Level-of-Detail Control and its Application to Terrain Rendering. Microsoft Research. URL: https://hhoppe.com/svdlod.pdf

[10] Pedro V. Sander, Jason L. Mitchell. Progressive Buffers: View-dependent Geometry and Texture LOD Rendering. Advanced Real-Time Rendering in 3D Graphics and Games. URL: https://advances. realtimerendering.com/s2006/ Chapter1-Out-of-Core_ Rendering_of_Large_Meshes_with_Progressive_Buffers.pdf

[11] Eun-Seok Lee, Byeong-Seok Shin. Vertex Chunk-Based Object Culling Method for Rreal-Time Rendering in Metaverse. 09.07.2023. MDPI. URL: https://www.mdpi.com/2079-9292/12/12/2601

[12] Data Transfer Matters for GPU Computing. [Yusuke Fujii, Takuya Azumi, Nobuhiko Nishio, Shinpei Kato, Masato Edahiro]. ResearchGate. URL: https://www.researchgate.net/publication/269197419_Data_Transfer_Matters_for_GPU_Computing

The article was delivered to editorial stuff on the 26.05.2023

УДК 004.42 DOi 10.30837/ bi.2023.1(99).15

Oleksii Kozel[1], Dmytro Kolesnykov[2], Oleksii Nazarov[3], Nataliia Nazarova[4]

[1] ХНУРЕ, м. Харків, Україна, oleksii.kozel.cpe@nure.ua

[2] ХНУРЕ, м. Харків, Україна, dmytro.kolesnykov@nure.ua,
ORCID iD: 0000-0002-4901-6869

[3] ХНУРЕ, м. Харків, Україна, oleksii.nazarov1@nure.ua,
ORCID iD: 0000-0001-8682-5000

[4] ХНУРЕ, м. Харків, Україна, nataliia.nazarova@nure.ua,
ORCID iD: 0009-0007-7816-7088

THEORETICAL FOUNDATIONS OF WEB SITE INTERFACE USABILITY ASSESSMENT

This paper presents how to redesign a website by applying a set of design principles to enhance the usability. The main objectives of the study are to find out the usability problems of the targeted interactive system in order to list out required suggestions to improve the website and to provide solutions by re-designing the existing interactive system. In order to implement the objectives of the project, we should initially evaluate the interactive system using usability evaluation. The outcome of the evaluation provides us information about the issues and requirements to design a new system. Based on the evaluation and its outcome, various methods will be used for resolving the problems while re-designing the website. This helps in identifying the problems which require usability improvements. The objective of this paper is to provide a conceptual framework and foundation for systematically investigating features in the Web environment that contribute to user satisfaction with a Web interface and uses motivation-hygiene theory to guide the identification of these features. Objects of research are generative systems with associative memory. Purpose is a development of a system for evaluating the hierarchy and heterogeneity of the interface of web pages using neural network technologies.

NEURAL NETWORK, MACHINE LEARNING, MEMORY, GENERATION, GENERATIVE MODELS, TEXT.

Козел О.Д., Колесников Д.О., Назаров О.С., Назарова Н.В. Теоретичні основи оцінки юзабіліті інтерфейсу веб-сайту. У цій статті представлено, як переробити веб-сайт, застосувавши набір принципів дизайну для покращення юзабіліті. Основними цілями дослідження є з'ясування проблем юзабіліті цільової інтерактивної системи для того, щоб сформулювати необхідні пропозиції щодо покращення веб-сайту та запропонувати рішення шляхом редизайну існуючої інтерактивної системи. Для того, щоб реалізувати цілі проекту, ми повинні спочатку оцінити інтерактивну систему за допомогою оцінки юзабіліті. Результати оцінки нададуть нам інформацію про проблеми та вимоги до проектування нової системи. На основі оцінки та її результатів будуть використані різні методи для вирішення проблем під час редизайну веб-сайту. Це допомагає виявити проблеми, які потребують покращення юзабіліті. Метою цієї статті є створення концептуальної основи для систематичного дослідження особливостей веб-середовища, які сприяють задоволеності користувачів веб-інтерфейсом, а також використання теорії мотивації та гігієни для визначення цих особливостей. Об'єктом дослідження є генеративні системи з асоціативною пам'яттю. Метою роботи є розробка системи оцінки ієрархічності та гетерогенності інтерфейсу веб-сторінок з використанням нейромережевих технологій.

НЕЙРОННА МЕРЕЖА, МАШИННЕ НАВЧАННЯ, ПАМ'ЯТЬ, ГЕНЕРАЦІЯ, ГЕНЕРАТИВНІ МОДЕЛІ, ТЕКСТ.

Introduction

The Internet has become a medium for a wide range of activities, including entertainment, communication, commerce, management, information sharing, and more. A website has become an integral part of any business, from retail to manufacturing. Social networks, personal business sites, web applications are prime examples that use web pages to display content. Over the past five years, the number of Internet users and the number of websites have increased significantly and are expected to continue to do so for a long time [1].

One of the most important criteria for successful business promotion on the Internet in terms of user experience has become customer acquisition and retention [2]. In the work related to the creation of an applied ontology for assessing the quality of user web interfaces, SEO was highlighted as one of the most important areas.

The author introduces the concept of element heterogeneity and describes its practical application [3].

From the business point of view, the user interface affects the quality of the provided services, creates a positive attitude towards the web service and leaves a desire to use it in the future. The quality of web interfaces is subject to increasing demands. User preference plays an important role [4]. Studies support this theory [5, 6, 7].

Technical aesthetics and ergonomics are applicable to the Web environment and are demanded by users. Research on the quality of user interface and its ergonomics is regularly conducted [6, 8-10], new theories are proposed, new tools are used to obtain reliable information.

The goal of the project is to develop a system for evaluating the hierarchicality and heterogeneity of web page interfaces using neural network technologies. To achieve this goal, it is necessary to solve the following tasks:

1) To study the subject area and conduct a comparative analysis of existing methods for evaluating hierarchicality and heterogeneity of the interface;

2) To develop a methodology for evaluating the hierarchicality and heterogeneity of the Web page interface;

3) design the architecture of a system for evaluating the hierarchy and heterogeneity of web pages using neural network technologies.

The project describes the theoretical justifications for creating a methodology for assessing quality based on the heterogeneity of elements and the creation of this methodology. It also describes the practical implementation of a quality assessment system based on current research in the field of UI/UX quality, using ISO standards, methods for assessing the heterogeneity of application components [8,11].

The system under development allows, based on the operation of a neural network, to determine the degree of compliance of the user interface with the established regulatory characteristics.

1. Subject area description

Today, evaluating the functional usability and visual appeal of a Web site is somewhat subjective and depends largely on human perception.

Due to differences in personal preferences and cultural backgrounds, different groups of website users can draw very different conclusions about the quality of the user interface. Therefore, it is difficult to perform an accurate and error-free usability evaluation using automated tools.

The implementation of the interface for working with an information system affects the success of that system: the user is interested in exploring the functionality, receives aesthetic pleasure, and feels comfortable if the implementation is based on general cultural principles and expectations. This affects both the duration of user interaction with the system and the level of user satisfaction after interaction with the system, and as a result, the desire to use the system in the future [6].

Since there are no other measures that provide a high level of reliability, user satisfaction is considered the most useful indicator of system success [10]. Satisfied users spend more time on a website and visit it more often. In general, user satisfaction can lead to audience retention and increased trust in the product. Therefore, it is important to improve the indicators that increase website satisfaction [3].

2. Modern methods for evaluating website usability

Modern usability assessment methods include a fairly wide range of methods and tools, ranging from user interviews and surveys to the use of sophisticated eye tracking devices and automated usability evaluation systems.

Modern usability evaluation methods can be divided into the following categories:

– Methods based on observation of user behavior,

– Methods based on self-evaluation of user behavior,

– Methods based on indirect user involvement [1,6].

In situations where users are directly involved, misinterpretation or incorrect answers to questions and low reliability can affect the reliability of the results. Expert methods, automated assessments, or process modeling not only take a long time to implement, but may also miss important issues and problems, reducing the reliability of the final assessment results.

Thus, based on the analysis of various usability assessment tools, it can be concluded that neither of the two existing assessment methods provides a complete, accurate and reliable usability assessment.

3. Analyzing Methods Used to Assess Interface Quality

Usability evaluation methods are divided into broad categories [4]:

1) Methods involving direct user participation:

– User observation — collecting information about the user's behavior and actions in the context of specific tasks while the user is working with the program.

– Critical event analysis — collecting data on specific events (positive or negative) that occurred during the user's work with the program.

– Performance measurements — collecting data on quantifiable performance characteristics to understand the impact of usability problems.

– Questionnaires — indirect evaluation methods that collect users' opinions about the user interface in specific questionnaires.

– Interviews — similar to questionnaires, but with more flexibility and personal contact with the person being interviewed.

– Participatory design and evaluation — methods that allow different types of participants to participate in the evaluation or design of systems.

– Thought aloud method — users continuously say out loud all their thoughts, beliefs, expectations, doubts, discoveries while using the system under test.

– Creative methods — methods that involve identifying properties of new products and systems, usually as a result of interactions among group members, often with users as members of such groups.

2) Methods that involve indirect user participation, which are used when it is not possible to collect usage data due to the absence of users, or in cases where they provide additional data and information:

– Model-based approaches — the use of models, which are an abstract representation of the product being

evaluated, that allow prediction of user actions.

– Document review methods — the study of existing documents by a usability specialist to provide a professional evaluation of the system.

– Automated evaluation — algorithms based on ergonomic knowledge that identify product defects by comparing them with specified data.

– Expert evaluation — an evaluation based on a usability specialist's knowledge, professionalism, and practical experience in the field of ergonomics.

Let's take a closer look at the automatic scoring methods:

1. Entropy of the RGB profile. The visual complexity of the system is estimated.

2. Information productivity. The ratio of the minimum amount of information needed to complete a task to the amount of information the user has to input.

3. Determination of the average time required by the user according to the GOMS, KLM methodology. Based on the averages, the average time spent by the user on the main tasks is calculated. User scenarios are determined individually for each project.

4. XML tree analysis. The complexity of the structure of the provided page is checked. This method requires specialization in web client development and principles of site optimization.

5. Number of classes into which interface objects can be divided.

4. Analyzing Methods Used to Assess Element Heterogeneity

The variety of web elements is one of the important criteria that make up the satisfaction score. This criterion affects the ease of assimilation of information, the perception of a web page and the ease of management of the system.

It is important to note that interfaces usually serve two main purposes [3]:

1. To provide information to the user.

2. Providing interaction with the system.

The process of creating interfaces is divided into two stages:

1. User Experience (UX) — shaping the interaction.

2. User interface (UI) — visualizes or materializes the interaction.

Experimental studies

The paper [4] presents tables describing such important attributes as understandability, well presented and organized information, interactivity, navigation (the ability to easily navigate between different pages of a resource), ease of use, which in varying proportions create a measure of heterogeneity. Their brief decoding is provided in this paper.

Learn more about the automatic scoring methods.

1. Understandability — the clarity and completeness of information on web pages.

2. Well-presented — the quality of information published on websites.

3. Ease of use — shows how easy it is for users to use the website's features.

4. Well organized — controlled (i.e. intuitive organization) and structured web environment.

These criteria were used to create a version of the questionnaire to assess the heterogeneity of the interface of neural network training sites based on user questionnaires. Based on these questionnaires, a neural network was constructed. The results and the weight of each parameter in the neural network are shown in Figures 1 and 2.

	System quality	Information quality	Security-privacy
Understandability	0.133	**0.854**	0.192
Reliability	0.266	**0.692**	0.253
Usefulness	0.216	**0.853**	0.137
Access	**0.807**	0.150	0.247
Friendliness	**0.828**	0.158	0.174
Navigation	**0.770**	0.282	0.065
Interactivity	**0.653**	0.201	0.480
Privacy	0.293	0.255	**0.795**
Security	0.156	0.206	**0.882**

Note: Bold values indicate the highest influence weight.

Fig. 1. Neural Network Weights for General Site Criteria

	System quality	Information quality	Security-privacy
Easy to comprehend	0.099	**0.809**	0.165
Well-presented	0.117	**0.773**	0.172
Accurate	0.263	**0.738**	0.169
Up-to-date	0.271	**0.667**	0.203
Relevant	0.141	**0.817**	0.134
Detailed	0.138	**0.796**	0.192
Speed of access	**0.729**	0.175	0.169
Availability	**0.778**	0.088	0.096
Ease of use	**0.801**	0.174	0.147
Well-organised	**0.794**	0.168	0.192
Page-loading	**0.702**	0.168	0.140
Hyperlinks	**0.749**	0.178	0.091
Two-way communication	**0.655**	0.192	0.425
Active control	**0.606**	0.109	0.456
Confidentiality	0.316	0.267	**0.701**
Authorisation	0.174	0.222	**0.800**
Integrity	0.225	0.218	**0.784**
Protection	0.130	0.185	**0.830**

Note: Bold values indicate the highest influence weight.

Fig. 2. Neural Network Weights for Private Site Criteria

As you can see, robustness is rated the highest by the neural network in the information quality section.

Similarly, well presented, well organized, and ease of use are highly rated by the trained neural network.

In another paper [3], similar metrics were obtained to create user interface quality evaluation systems. Figure 3 shows the ontology diagram.

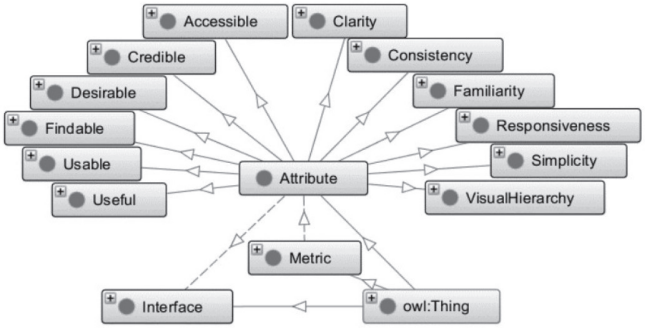

Fig. 3. User Interface Ontology Diagram

Based on the task, the directions of the ontology were determined to create a tool for evaluating attributes such as readability, comfort, cleanliness, and simplicity [3].

5. Standards and specifications

We managed to formalize these aspects using ISO standards related to human interaction with displays and interfaces. It is regulated by standards in the field of usability and human-machine interaction:

1. The User-Centered Interactive Systems Design Process provides guidance on how to organize the interface design process and integrate it seamlessly into the overall software production process. It describes the usability methods necessary for: determining the context of product use, identifying user and customer requirements for the system, prototyping and usability testing the product.

2. Ergonomics of human-computer interaction, description of the process of designing user-oriented interfaces. It describes in detail the maturity model of the organization in terms of the level of use of the UCD process. Recommendations for moving to higher maturity levels are given.

3. Ergonomics of multimedia user interface software. Recommendations are given for designing controls for multimedia products.

4. Human-system interaction ergonomics. Guidelines for access to human-machine interfaces.

5. Human-system interaction ergonomics. Usability-based methods for ensuring human-centered design.

Since ISO provide only technical requirements for the implementation of web pages, the visual and variable part of the design is outside the scope of the above standards. It is important to note that the standards define the color palette and normalize its contrast. This is an important criterion that can affect the heterogeneity of different elements together. They also standardize input/output methods, the basics of element behavior (for example, the principles of interaction with the "button" element), and their variability.

Expert research

Nielsen & Norman Group is a large expert firm that provides services for evaluating and improving UX/UI design.

Their publications have also been used to develop expert evaluation methods. In particular, an article that provides more than a hundred tips for creating a high-quality Web site [12]. Some recommendations related to the selected metrics were selected for the expert system.

Also, the works of NN Group employees describe patterns, best practices, and obvious mistakes in the design of web page elements [13-21].

Based on the knowledge gained, it is possible to develop the basis for a method to assess the heterogeneity of web page elements:

Summarizing the data from the different sources described above, we can assume that to assess heterogeneity, the following criteria should be taken into account: accessibility, quality of the information provided, ease of use, organization of components, comfort, cleanliness, simplicity.

The example of the implementation of an expert system using a neural network capable of assessing the quality of a web site shows that there is practical evidence for the quality metrics proposed. The resulting weights for the criteria give a good idea of the value of each evaluation criterion. Based on the given task, we can understand that there are all theoretical prerequisites for creating a system based on neural networks to assess the heterogeneity of elements.

The result of the research part is the developed concept of interface heterogeneity. Interface heterogeneity is the number of objects and their classes.

6. Analyzing and comparing analogues

In the course of studying analogs, we examined web applications for user interface evaluation that use screenshots as input data [12].

UsabilityHub is a web application that allows you to determine the quality of usability based on an uploaded screenshot. The analysis provides information on how easy it is for users to navigate a website page, identifies the elements that attract the most attention, and creates a heat map of clicks. The evaluation is based on a user survey.

UserPlus is a web application that allows you to determine the quality of usability based on an uploaded screenshot. Each screenshot is independently marked by the user of the service and then, after the survey, the result of the usability analysis is published for each marked interface element.

Usabilla is a web application that conducts user surveys based on uploaded screenshots and pre-prepared questions and generates analysis based on the results.

ConceptFeedback is an online resource where you can get a user interface evaluation from professional designers.

Based on the results of the comparative analysis, it can be concluded that most web applications use questionnaires and surveys of users and testers as a method of interface evaluation. A number of programs also monitor user activity on a web page. None of the existing analogues uses an automatic user interface evaluation system.

User interface evaluation according to the international standard is performed only in the UserPlus application. However, this evaluation method mainly refers to the individual elements of the interface rather than to the overall assessment of the interface usability.

Based on the results of the review of modern methods of evaluating web interfaces, as well as on the identified shortcomings among the studied analogues of web

applications for evaluating interfaces, there is a need to develop our own method of evaluating the hierarchical interface of web pages and web applications, which provides work with screenshots based on the developed method.

Characteristics of Analog Selection

There are currently no finished public products available to users, nor are there any implementations using the approach presented in this paper. Therefore, it was decided to study prototypes, the results of theoretical studies, and related solutions.

The research paper [4] describes theories such as the two-factor theory, the expectation of refutation theory, and the three-factor theory. These theories argue that the impact of a website attribute on satisfaction can have different weights for different characteristics, which means that their importance depends on their effectiveness. This fact leads to non-linear and asymmetric relationships that are difficult to evaluate using traditional methods. Therefore, successful results are obtained using neural networks, which are presented below.

The following analogues implement a part of the task at hand, so the description of each of them is a confirmed variant of the system component to be implemented, taking into account the specifics of the task at hand.

Evaluation of an expert system based on a questionnaire

Paper [3] attempts to determine the relationship between overall user satisfaction and website attributes. The paper uses the experimental results of a large questionnaire-based survey. The input data are also questionnaires. The purpose of the survey is to determine the overall satisfaction of website users by answering questions related to specific website attributes. The survey asked a set of 370 Internet users to rate the effectiveness of 18 specific and 9 general attributes, and to indicate their overall satisfaction on a nine-point scale ranging from "very dissatisfied" to "very satisfied". The results were tested and validated using reliability and validity procedures, showing that there is a relationship structure as certain general and specific website attributes create a link to user satisfaction. In this article, we try to find out the relationship between overall satisfaction and specific website attributes using neural networks to approximate the functions.

The result of the research is an expert neural network that shows the results of the value of each criterion on user satisfaction depending on the quality of that criterion. The results are shown in Figures 4-7.

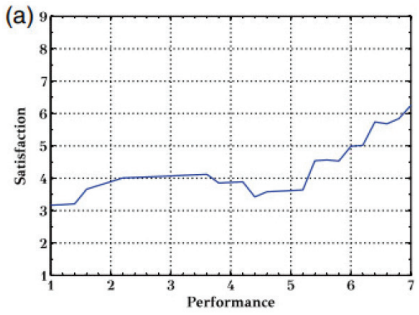

Fig. 4. Ratio of Performance to Satisfaction for the Understandability Metric

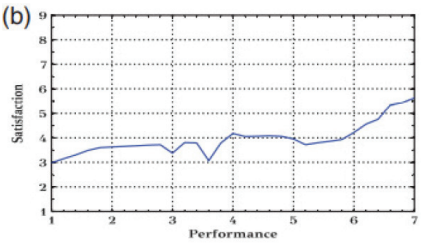

Fig. 5. Ratio of Performance to Satisfaction of the Well-Described Metric

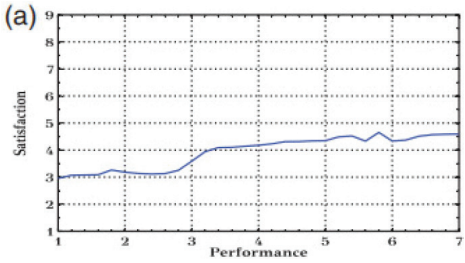

Fig. 6. Ratio of Productivity to Satisfaction of the Well-Organized Metric

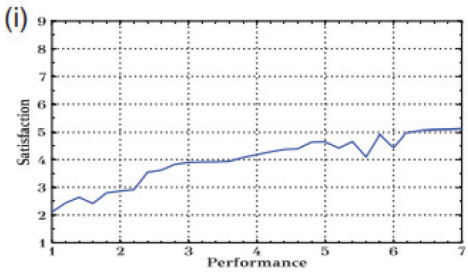

Fig. 7. Performance vs. Satisfaction Ratio for the Usability Metric

These figures allow us to understand the value of each of the criteria in the process of comparing them with their results.

It can be seen that these criteria together give an increase in satisfaction in relation to each other. It turns out that the proposed evaluation method should be productive for any image quality and the evaluation result should be intuitive for the system user.

Unfortunately, the system described in this paper requires an expert who is able to reliably convey the initial evaluation of a web page. In general, the neural network and its results are for research and exploratory purposes only and are not suitable for automated image processing.

Overview of the implementation of UI element retrieval using neural networks

Paper [22] provides comprehensive information on the performance of neural network models. The comparison includes Faster RCNN, Cascade RCNN, and YOLOV4 in Figure 8.

Run ID	Model	Overall precision	mAP@IoU 0.5	recall@IoU 0.5
67413	**baseline** Faster RCNN	0.94789	57.2	40.3
67833	Cascade RCNN	0.95035	68.16	53.3
67710	Cascade RCNN	0.94909	64.92	50.5
67722	Cascade RCNN	0.93463	72.33	58.5
67829	YOLOv4	0.93300	73.82	55.6
67707	YOLOv4	0.93125	79.24	59.4
67831	YOLOv4	0.92987	79.11	60
67972	Cascade RCNN	**0.95044**	71.53	55.6
67706	YOLOv4	0.93437	**79.36**	**59.8**

Run 67972: 10000 iterations; Run 67706: 7000 iterations

Fig. 8. Neural Network Performance Results

As you can see, YOLOv4 shows a good performance result. It should be clarified that when choosing the implementation tools and the technical experiment, it was decided to use YOLOv5s, since it returns the answer faster, which is extremely important for server applications.

We used 2950 images to train the neural network to search for UI elements: 2363 images were used to train the neural network, and 587 images were used for testing.

The knowledge gained from the analysis of related programs allowed us to choose the means of implementation, the future architecture of the system, the amount of necessary expert, training, verification and test data.

Selection of the means of realization

The choice of implementation tools is based on the knowledge of neural networks gained during training, as well as on the results of project and course work.

We chose the YOLOv5 neural network model to classify web page elements because it is quite compact and efficient and meets the performance requirements of server applications. Accordingly, the framework for running the neural network is PyTorch, the programming language is Python, and the framework is FastApi. Since some components of the system were developed during the internship and coursework, the second neural network, which plays the role of an expert, is implemented using the Keras framework, which was proposed for work during the training. Heterogeneity is one of the criteria for the quality of the user interface, so it was decided not to use frameworks to create SPA applications, as it is obvious that at this stage of client development it is unnecessary.

For the same reason, the search for tools and technologies for databases is not taken into account due to the small functionality of the program — the user simply has no reason to save images for further work.

The user is given the opportunity to download the current result, which is quite enough.

Technical Experiment

Two frameworks were chosen to implement the server API: Flask and FastApi. Both frameworks are positioned as easy to understand and easy to use.

To select the optimal framework, a technical experiment was performed: a prototype program was implemented using Flask and FastApi, and standard prediction models for YOLOv5s and YOLOv3 were used as a neural network to classify objects. As a result, the FastApi framework was chosen because of its implementation of asynchrony, the ability to easily create parallel threads, and simple and flexible tools for configuring server endpoints, such as simple and transparent validation implementation.

YOLOv5 was chosen as a preliminary model because it is a more productive version and requires less memory to run. An important factor was the speed of training, since the only place to train a neural network is provided by Google Colab, which has technical limitations in terms of resources and time. You can see the difference in the required resources in Figures 9, 10.

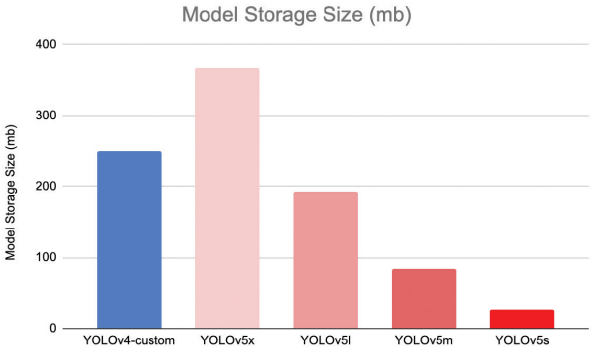

Fig. 9. Megabyte Model File Size Chart

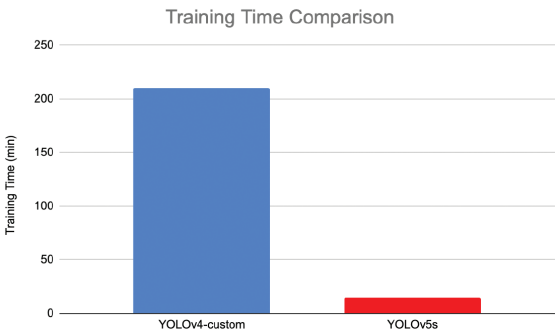

Fig. 10. Diagram of Neural Network Training Speed in Minutes

The choice of a simplified model had a positive effect on the quality of the trained model: more epochs improved the classification accuracy. As part of the task, it is first necessary to find all the classes represented in the image.

The accuracy of YOLOv5s is lower than that of YOLOv3, so it was decided to allow the user to choose a model if speed is not so important. The choice of YOLOv5s is also due to the fact that the latency is acceptable for mass processing, e.g. object classification of 850 Full HD images (1920 x 1080 pixels) in multi-threaded, asynchronous mode takes no more than two minutes using AMD FX-6300 CPU 6 6 threads instead of 4-5 minutes using YOLOv3 and full CPU utilization. See the System

Performance Testing section for more information. At the time of selecting system development tools, the current information is sufficient.

Conclusions

The quality of the user interface is a difficult concept to evaluate. The variety of web elements is one of the important criteria that make up the satisfaction score.

This criterion affects the ease of assimilation of information, the perception of a web page and the ease of managing the system. The relevance of the topic is confirmed by many works.

As part of this work, we have developed a web application for evaluating the hierarchical nature of web page interfaces based on the analysis of screenshots.

In particular, the following tasks were solved:

1) A review of existing methods for evaluating user interface usability and existing analogs.

2) Analyzed the requirements, developed a method for evaluating the hierarchy of web interfaces, and designed the architecture of the web application.

3) The subject area is studied and a comparative analysis of methods for evaluating interface heterogeneity is made.

4) Developed a methodology for assessing the heterogeneity of web page interface.

5) Designed the architecture of the system for assessing the heterogeneity of web pages using neural network technologies.

Further research is planned to solve the following tasks:

1) Extend the list of features of the hierarchy evaluation.

2) Investigate the relationship between expert opinion and the value of the metric to be calculated.

3) Combining the modules for hierarchicality and heterogeneity evaluation of the web page interface into a single system.

In addition to practical and research experience, the work provided invaluable experience in integrating trained neural network models and a client-server program.

The tasks of neural network integration and practical application were completed.

The work itself has further development potential: extending the evaluation methods, improving the current solution, publishing the service and providing access to it.

Conflict of Interest

The authors declare no conflict of interest.

REFERENCES

[1] Stewart T. 2012. Websites – Quality and Usability // Behaviour & Information Technology, 2016. – Pp. 645-646.

[2] How long do users stay on web-pages. URL: https://www.nngroup.com/ articles/how-long-do-users-stay-on-web-pages

[3] Asil Oztekin, Dursun Delen, Ali Turkyilmaz, Selim Zaim A machine learning-based usability evaluation method for eLearning systems // Decision Support Systems, 2013. – Volume 56. – Pp. 63-73.

[4] Amanatiadisa A., Mitsinisb N., Maditinos D. A neural network-based approach for user experience assessment // Behaviour & Information Technology, 2014. – Pp. 321-333.

[5] Márcio José Moutinho da Ponte, Antonio Morais da Silveira. A Methodology for Evaluation the Usability of Software for Industrial Automation Using Artificial Neural Networks: Case Study—Eletrobrás // 2008 International Conference on Computational Intelligence for Modelling Control & Automation, Vienna, Austria, 2008. – Pp. 430-435.

[6] DeLone W., McLean E. Information Systems Success: The Quest for the Dependent Variable // Information Systems Research, 2012. – Pp. 60-95.

[7] Noriaki K., Seraku N., Takahashi F., Tsuji S. Attractive Quality and Must-Be Quality. // The Journal of the Japanese Society for Quality Control, 1984. – Pp. 39-48.

[8] Courage C., Baxter K. Understanding Your Users: A Practical Guide to User Requirements Methods, Tools, and Techniques. – Morgan Kaufmann, The Morgan Kaufmann Series in Interactive Technologies, 2005. – 214 p.

[9] Hassenzahl M., Tractinsky N. User Experience – A Research Agenda. // Behaviour & Information Technology, 2006. – Pp. 91-97.

[10] Kincl T., Štrach P. Measuring Website Quality: Asymmetric Effect of User Satisfaction. // Behaviour & Information Technology, 2012. – Pp. 647-657.

[11] Nielsen Norman. Group Articles & Videos. URL: https://www.nngroup.com/articles

[12] Palmer J.W. Web site usability, design, and performance metrics. //Information systems research, 2002. – Pp. 151-167.

[13] Kara Pernice. QuickLinks label intranet. URL: https://www.nngroup.com/articles/quicklinks-label-intranet

[14] Jen Cardello. Navigation IA tests. URL: https://www.nngroup.com/ articles/navigation-ia-tests

[15] Anna Kaley. Popup problems. URL: https://www.nngroup.com/ videos/popup-problems

[16] Jakob Nielsen. Rules UX. URL: https://www.nngroup.com/videos/rules-ux

[17] Therese Fessenden. Effective online advertising. URL: https://www.nngroup.com/videos/effective-online-advertising

[18] Therese Fessenden. Footers. URL: https://www.nngroup.com/articles/footers

[19] Kate Moran. Designing search suggestions. URL: https://www.nngroup.com/videos/designing-search-suggestions

[20] Kathryn Whitenton Better forms visual organization. URL: https://www.nngroup.com/videos/better-forms-visual-organization

[21] Therese Fessenden. Grid layouts. URL: https://www.nngroup.com/videos/grid-layouts

[22] Narayanan N., Balaji N., Jaganathan K. Deep Learning for UI Element Detection: DrawnUI 2020. – Sri Sivasubramaniya Nadar College Of Engineering, 2020. – 157 p.

The article was delivered to editorial stuff on the 27.07.2023

ПРАВИЛА
оформлення рукописів для авторів науково-технічного журналу «БІОНІКА ІНТЕЛЕКТУ»

Науково-технічний журнал «Біоніка інтелекту» приймає до друку написані спеціально для нього оригінальні рукописи, які раніше ніде не друкувались. Структура рукопису повинна бути такою: індекс УДК, відомості про авторів, заголовок, анотації (на трьох мовах), ключові слова, вступ, основний текст статті, висновки, список використаної літератури, резюме.

Відповідно до Постанови ВАК України від 15.01.2003 №7-05/1 (Бюлетень ВАК, №1, 2003, с. 2), стаття повинна мати такі необхідні елементи: постановка проблеми у загальному вигляді та її зв'язок із важливими науковими чи практичними завданнями; аналіз останніх досліджень і публікацій і виділення не вирішених раніше частин загальної проблеми в даній області; формулювання цілей та завдань дослідження; виклад основного матеріалу досліджень з повним обґрунтуванням отриманих наукових результатів; висновки з даного дослідження та перспективи подальших досліджень у даному напрямку.

Статті мають бути виконані в редакторі Microsoft Word. Формат сторінки – А4 (210×297 мм), поля: верхнє – 25 мм, нижнє – 20 мм, ліве, праве – 17 мм. Кількість колонок – 2, з інтервалом між ними 5 мм, основний шрифт Times New Roman, кегль основного тексту – 10 пунктів, міжрядковий інтервал – множник (1,1), абзацний відступ – 6 мм. Обсяг рукопису – від 6 до 12 сторінок (мови: українська, англійська, російська та мовою оригінала).

УДК друкується з першого рядка, без відступів, вирівнювання по лівому краю.

ПІБ автора (-ів), назва статті, назва та адреса учбового закладу необхідно надати повністю російською, українською та англійською мовами.

Назва статті друкується прописними літерами; шрифт прямий, напівжирний, кегль 12.

Назви розділів нумерують арабськими цифрами, виділяють жирним шрифтом. Відступи для назви статті, ініціалів та прізвищ авторів, відомостей про авторів, назв розділів, вступу та висновків, списку літератури: зверху – 6 пт, знизу – 3 пт.

Анотації (мовою статті, абзац 6–12 рядків, кегль 9) розміщують на початку статті, в ній має бути розміщена інформація про очікувані результати описаних досліджень (на трьох мовах).

Ключові слова (4–10 слів з тексту статті, які з точки зору інформаційного пошуку несуть змістовне навантаження) наводять мовою рукопису, через кому в називному відмінку, кегль 9.

Рисунки та *таблиці* (чорно-білі, контрастні) розміщуються у тексті після першого посилання у вигляді окремих об'єктів і нумерують арабськими цифрами наскрізною нумерацією за наявності більше ніж одного об'єкта. Невеликі схеми, що складаються з 3–4 елементів виконують, використовуючи вставку об'єкта Рисунок Microsoft Word. Більш складні виконують у графічних редакторах у вигляді чорно-білих графічних файлів форматів .tif, .jpg, .wmf, .cdr із розділенням 300 dpi. Рисунки мають міститися у текстовому файлі й обов'язково подаватися окремими файлами з відповідними назвами (наприклад, рис1.jpg).

Усі елементи рисунка, включаючи написи, повинні бути згруповані. Усі написи в рисунках і таблицях мають бути виконані шрифтом Times New Roman, кегль у рисунках – 10, у таблицях – 9.

Рисунок повинен мати центрований підпис (поза рисунком), шрифт 9, відступи зверху і знизу по 6 пт. Ширина рисунка має відповідати ширині колонки (або ширині сторінки).

Формули, символи, змінні повинні бути набрані в **редакторі формул MathType**. Формули розміщують посередині рядка й нумерують за наявності посилань на них у рукописі. Шрифт – Times New Roman. Висота змінної – 10 пунктів, великих і малих індексів – 8 пт, основний математичний символ – 12 (10) пт. Змінні, позначені латинськими літерами, набирають курсивом, грецькі літери, скорочення російських слів і цифри – прямим написанням. Змінні, які є в тексті, також набирають у редакторі формул.

Список літератури вміщує опубліковані джерела, на які є посилання в тексті, укладені у квадратні дужки, друкують без абзацного відступу, кегль 9 пт, відступ зверху – 6 пт.

Після списку літератури з відступом зверху 6 пт зазначають *дату подання статті до редколегії*. Число та місяць задають двозначними числами через крапку. Розмір шрифта – 9 пт, курсив, вирівнювання по правому краю.

Резюме (Times New Roman, кегль — 10 пунктів,) подають англійською мовою: обсяг резюме до 2000 знаків (бажаний переклад). *Структура резюме:* **Background, Materials and methods, Results, Conclusion.**

Разом із рукописом (на аркушах білого паперу формату А4 щільністю 80-90 г/м^2, надрукований на лазерному принтері) необхідно подати такі документи:

1. Заяву, яку повинні підписати всі автори.
2. Акт експертизи про можливість опублікування матеріалів у відкритому друці (якщо потрібно).
3. Рецензію, підписану доктором чи кандидатом наук.
4. Відомості про авторів.
5. Електронний варіант рукопису, резюме та відомостей про авторів.
6. Зробити оплату публікації.

Необхідно також зазначити один з наступних тематичних розділів, якому відповідає рукопис:

1. Теоретичні основи інформатики та кібернетики. Теорія інтелекту.
2. Математичне моделювання. Системний аналіз. Прийняття рішень.
3. Інтелектуальна обробка інформації. Розпізнавання образів.
4. Інформаційні технології та програмно-технічні комплекси.
5. Структурна, прикладна та математична лінгвістика.
6. Дискусійні повідомлення.

ЗМІСТ

ІНФОРМАЦІЙНІ ТЕХНОЛОГІЇ. МАШИННЕ НАВЧАННЯ. БАЗИ ДАНИХ

ОБ'ЄКТНЕ МОДЕЛЮВАННЯ. НЕЙРОННІ МЕРЕЖІ ТА НЕЙРОМАТЕМАТИКА

ІНТЕЛЕКТУАЛЬНІ СИСТЕМИ. НЕЙРОННІ ТА ЛОГІЧНІ МЕРЕЖІ

СТРУКТУРНА, ПРИКЛАДНА ТА МАТЕМАТИЧНА ЛІНГВІСТИКА

ІНФОРМАЦІЙНІ ТЕХНОЛОГІЇ. ОПТИМІЗАЦІЯ ТА ГРАФІЧНІ РІШЕННЯ

Наукове видання

БІОНІКА ІНТЕЛЕКТУ

інформація, мова, інтелект

Науково-технічний журнал

№ 1 (99)

2023

Головний редактор — *Г. Г. Четвериков*

Відповідальний редактор — *І.В. Кириченко*

Комп'ютерна верстка — *О. Б. Ісаєва*

Рекомендовано Вченою Радою
Харківського національного університету радіоелектроніки
(протокол № 14 от 26.12.2023)

Адреса редакції:

Україна, 61166, Харків-166, просп. Науки, 14,
Харківський національний університет радіоелектроніки, к. 127
тел. 702-14-77, факс 702-10-13,
e-mail: bionics@nure.ua

Підписано до друку 29.12.2023. Формат $60 \times 84\ ^1/_8$. Друк ризографічний.
Папір офсетний. Гарнітура Newton. Умов. друк. арк. 13,5. Обл.-вид. арк. 13,0.
Тираж 20 прим.

Віддруковано в редакційно-видавничому відділі ХНУРЕ
61166, Харків, просп. Науки, 14.