REPUBLIC OF AZERBAIJAN

On the rights of the manuscript

ABSTRACT

of the dissertation for the degree of Doctor of Philosophy

DEVELOPMENT OF AN INFORMATION RETRIEVAL SYSTEM IN THE DIRECTIVE DOCUMENTS DATABASE

Specialty:	3338.01 – System analysis, management and processing of information (by fields)
Field of science:	Technical Sciences

Applicant: **Sevinj Elman Pashayeva**

Baku - 2025

The dissertation was completed at Nakhchivan State University.

Dissertation advisor:	Doctor of Technical Sciences , Associate Professor **Fahrad Heydar Pashayev**
Official opponents:	Doctor of Technical Sciences, Professor **Salahaddin Imamali Yusifov**
	Doctor of Technical Sciences, Associate Professor **Elkhan Nariman Sabziev**
	Doctor of Technical Sciences, Associate Professor **Shahla Surkhay Huseynzade**

ED 1.20 Dissertation Council of the Higher Attestation Commission under the President of the Republic of Azerbaijan operating at the Institute of Control Systems of the Ministry of Science and Education of the Republic of Azerbaijan

Chair of the
Dissertation Council: Full Member of ANAS, Doctor of Technical Sciences, Professor
Ali Mahammad Abbasov

Academic Secretary
of the Dissertation Council: PhD in Technical Sciences,
Associate Professor
Tahir Ali Alizada

Chair of the
scientific seminar: Doctor of Technical Sciences,
Associate Professor
Gambar Agaverdi Guluyev

GENERAL DESCRIPTION OF THE DISSERTATION

The relevance and maturity of the topic.

It is known that the management of directives constitutes a significant part of electronic document management systems (EDMS). One of the main functions of EDMS is working with databases of electronic documents, determining the models of these databases, creating, organizing, recording their structures, forming various queries to the databases. Similarly, the main activity of information retrieval systems (IRS), built on directive document management systems, are carried out on the databases (DB) of such systems. Therefore, building an IRS on the basis of directives databases (DDB) remains relevant. For this purpose, electronic directives (ED) and databases in which these documents are stored must be in line with the requirements of the legislation of the Republic of Azerbaijan on electronic signature and electronic documents.[1] The regulation of work with these databases by the legislation of the Republic of Azerbaijan once again emphasizes the importance of a DDB-based IRS.

The concept of a directive document is given in various sources, sometimes with different, sometimes with similar definitions.

Originally, the term was used to refer to binding documents issued by higher authorities. These include laws, decrees, decisions, resolutions, etc., issued by higher organizations.

Currently, directives can also include standards, benchmarks and other normative documents created and used in technical, medical and other diagnostic systems.

Various information systems for working with directive documents and various methods of improving the operation of the created systems have been developed. One of the successful software applications is Norma CS. It was created to help specialists dealing with directives in various fields. It is designed to store and search for

[1] Law of the Republic of Azerbaijan on electronic signature and electronic document

various normative documents, standards and can be adapted to different conditions.

An illustrative example of an Electronic Document and Data System (EDDS) is the Unified Automated Management System (UAMS), which is currently employed by the Customs authorities of the Republic of Azerbaijan.

However, existing systems often fail to distinguish between ordinary electronic documents and directives.

The dissertation discusses aspects of the requirements for the storage of directives in electronic form that are different from the requirements for the storage of other electronic documents. The unacceptability of accidental or deliberate modification of EDs in the process of their storage imposes special requirements to the storage of these documents. ED databases and information retrieval systems working with such databases must comply with the requirements. Therefore, the dissertation investigates the development of a DB of a directives information retrieval system in accordance with the above-mentioned laws of the Republic of Azerbaijan. The advantages of using the relational database model (RDB) for creating successful search queries in the database are investigated. Rules and sequences of logical formulation of conditions for obtaining data from the RDB satisfying certain conditions are given. In accordance with the requirements of this base model, a query system was created based on the structure, attributes and conditions of the DB of the directives information retrieval system. The database contains information about documents, the organization that sent and stored the document, executors, and the current status of work on the document. When creating the ED database, considering the importance of these documents, some metric characteristics were defined and included in the database. Thus, the problem of detecting the facts of accidental or deliberate alteration of documents became easier to solve.

One method of optimizing IRS performance is to identify and automate those parts of these systems that can be automated. As a

result, manual operations are reduced and search operations are improved.[2]

DB creation, data model selection, database optimization methods and other issues are collected in one place in Novikov's textbook "Основы технологий баз данных" ("Basics of Database Technology").[3]

In many cases, data clustering, compression, and volume reduction are cited as optimization techniques. This is described in detail in the document Automatic Data Optimization with ORACLE Database.[4]

Solutions to problems such as parallelizing operations, improving data structure, etc. to optimize work with data are shown.[5]

In the above and other studies, various aspects of optimization problems were studied and significant results were obtained. However, some new and different approaches are presented in this dissertation.

a) A method of parallelizing file sharing using different polynomials to improve the efficiency of CRC algorithm implementation is presented. This method, unlike other systems, avoids delays, which is essential during peak usage hours.
b) In order to optimize the work of IRS in DDB, the principle of natural division of database files is developed, where database files are split into several parts according to the sources of their creation, and search operations are performed on smaller databases.

---

[2]В.О.Рахуба. Задачи автоматизированнго управления поисковой оптимизацией Интернет-ресурс (V.O. Rakhuba. Problems of automated management of search ortimization Internet resource)

[3] Новиков Б.А. Н73 Основы технологий баз данных: учебное пособие / Б. А.Новиков, Е.А.Горшкова, Н.Г.Графеева; под ред. Е.В.Рогова. — 2-е изд. — М.: ДМК Пресс, 2020. — 582 с. (Novikov B.A. N73 Fundamentals of database technologies: textbook / B.A. Novikov, Ye.A. Gorshkova, N.G. Grafeyeva; ed. by E.V. Rogov. - 2nd ed. - M.: DMK Press, 2020. - 582 p.)

[4]https://www.oracle.com/webfolder/s/delivery_production/docs/FY16h1/docs26/auto-data-opt-wp-12c-1896120-ru.pdf

[5] https://itentika.ru/news/kak-optimizirovat-rabotu-s-bazami-dannykh

It is known that EDMS databases store electronic documents created internally and received from the outside, as well as information on the attributes of these documents. Although documents are divided into two types, informational and directive, they can be stored in different file formats depending on the nature of their execution.

Each of the research areas to which the IRS in DDB refers is relevant and of interest to researchers in various research centers around the world. This applies, first of all, to the development of IRS. This field has been developing rapidly recently and is of great scientific and practical importance. Some of the scientists conducting research in this area are L.R. Fionova, S.P. Belov, M.V. Postnikova and others. The second main research area related to the object of research of the dissertation work is creation of DB. This area of research is also of major scientific and practical importance.

A distributed database is a number of databases distributed over a computer network and internally interconnected. A distributed CMS is similar to a distributed file system.

One of the areas of research to which IRS in DDB belongs is networks, in which such systems are used. Software tools, protocols, algorithms and technologies for creating such networks are given in the works of V.G. Olifer, N.A. Olifer, T.I. Aliyev, A.S. Tanenbaum, D.J. Wetherall and others.

The solution of some problems of mathematical support of systems and networks of electronic document search is given in the works of F.H. Pashayev, R.G. Alekperov, R.M. Aliguluyev, B.G. Ibragimov and A.Z. Malikov.

The aim of the dissertation is study and improvement of an information retrieval system in the directives database.

The **object of research** of the dissertation is electronic document management systems. The **subject of research** of the dissertation is the creation of an information retrieval system in EDMS.

In order to achieve the aim of the dissertation, the following issues are addressed:

- Study of characteristic features of ED;
- Development of algorithms for determining metric

characteristics and metric indicators of ED;

- Building stochastic and autoregressive models of IRS in DDB;
- Development of a natural partitioning principle for managing base files within the DDB;
- Creating operational algorithms for files of directives and their storage folders;
- Study of the topology of IRS in DDB and development of some software modules.
- Creation of the DB of IRS in DDB.

Research methods. Information theory, methods of probability theory, mathematical statistics, methods of calculation of cyclic codes, and the theory of database management were used as research methods in the dissertation. The correctness of the statements and results of the dissertation was verified on computer models and through a number of practical computational experiments.

Main statements put forward for defense:

- Development of algorithms for determining metric characteristics of electronic directives;
- Development of methods for working with files of directives in text format;
- A stochastic model of IRS in DDB and a method for determining experimental performance characteristics;
- Development of a natural partitioning principle for managing base files within the DDB;
- Development of algorithms for writing database files in same level file folders to simplify data writing, processing and reading;
- Creation of the structure of DB of IRS in DDB.

Scientific novelty. Scientific novelty of the results obtained in the dissertation work is as follows:

- EDS and their characteristic features were investigated, methods and algorithms for working with text files of directives have been developed;
- An algorithm for determining the metric characteristics of ED has been created;
- A stochastic model of IRS in DDB has been created and a method for determining experimental performance

characteristics has been developed;

- Autoregressive models of IRS DDB have been investigated and a method for determining adequate models for different life cycles of the system has been proposed;
- To enhance the performance of IRS in DDB the principle of natural division of database files has been developed. Here the database files are divided into several parts according to the sources of creation, and search operations are performed on databases of smaller size;
- To simplify data writing, processing and reading, algorithms have been developed to write database files into folders of the same level;
- The structure of the DB of IRS in DDB has been created.

Practical significance of the dissertation and implementation of the results:

The solution of the objectives set in the dissertation and the results obtained are of major theoretical and practical importance. The obtained results can be used in the creation, operation and ensuring the security of IRS and DB in DDB. The results of the dissertation can be successfully used in the creation of any IRS and corporate networks. Many of the obtained results can be used in the building of any EDMS. The scientific results obtained in the dissertation work were applied in the development and execution of laboratory works at the Informatics Department of Nakhchivan State University and as part of the monitoring, diagnostics and control systems complex operated at the facilities of Bibi-Heybat Oil Production Department.

Validation of the research. The main results of the dissertation were discussed at the following conferences and seminars:

- VII All-Ukrainian Scientific-Practical Conference "Computer Sciences and Systems Sciences", Poltava, March 10-12, 2016, pp. 223-225;
- Materials of the III Republican Scientific Conference on Applied Mathematics and New Information Technologies, SDU, Sumqayit, December 15-16, 2016, pp. 282-283.;
- VIII All-Ukrainian Scientific-Practical Conference

"Informatics and Systems Sciences", Poltava, March 16-18, 2017, pp. 211-213;

- Materials of the IV Republican Scientific Conference on Applied Mathematics and New Information Technologies, SDU, Sumqayit, December 9-10, 2021, pp.143-147

Publications. 14 scientific works have been published on the dissertation, 10 of which are articles.

Structure and volume of the dissertation. The dissertation consists of an introduction, 4 chapters, a conclusion, a list of references of 119 entries, and appendices. The main text consists of 146 pages, 17 figures, 6 tables, and 4 charts.

THE CONTENT OF THE DISSERTATION

The **introduction** substantiates the relevance of the topic of the dissertation, defines the aims and objectives of the research, and describes the scientific novelty and practical significance of the obtained results.

The first section of the first chapter states the objectives of the dissertation.

A conceptual model of IRS was created in the dissertation work (Fig.1).

In the conceptual model, users can organize communication with the query analysis and organization unit via a local area network. Depending on specific solutions, different types of communication can be used, including wireless communication. The query analysis and organization unit can communicate with various sources of information via LAN, Internet and other means of communication. The results of queries and any operation performed on the system are processed in the results analysis and decision-making unit. Here, the operational characteristics of the system are also calculated and refined after each query. Decisions can also be made to place remote data in the internal DB of the enterprise in order to ensure accessibility.

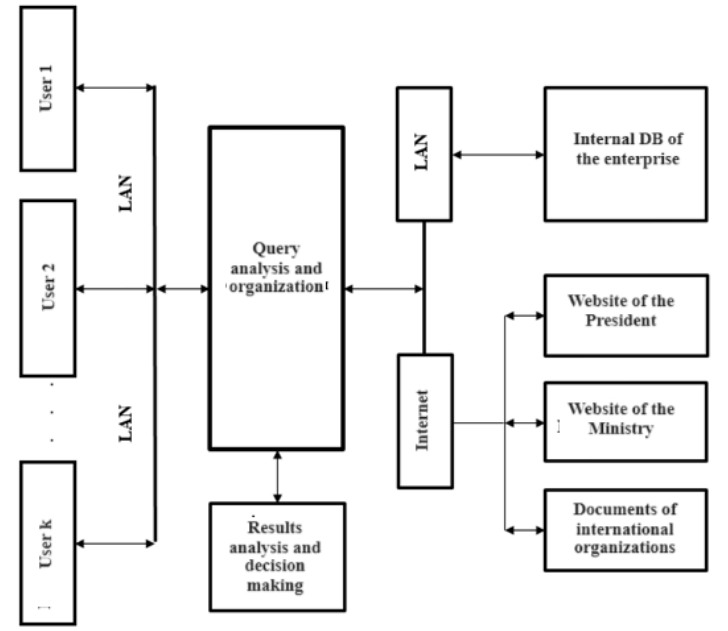

Fig. 1. A simplified conceptual model of the system

The second section of the first chapter focuses on the study of the topology of the information retrieval system in the directives database and the development of some software tools of the system. When considering topological models, it is emphasized that in many cases the use of wireless means of communication may be important in the creation of local networks. Here, existing topological models are reviewed and a unique topology of the DDB information retrieval system is proposed. (Fig. 2)

The **third section of the first chapter** of the dissertation contains program fragments reflecting the principle of operation and the operation of the timer of the information retrieval system.

An algorithm consisting of the following steps is created to organize a long waiting process without disrupting the operation of the system:

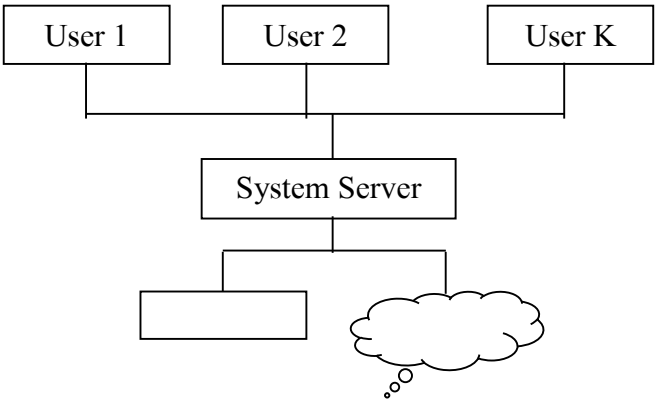

Fig. 2. Structure of links of IRS in DDB

Step 1. In the form.create subroutine, set a small interval value and assign the variable t1_step to the value 1.

timer1.Interval:=100; 0.2 seconds time specifies the intervals of the timer subroutine. It can be changed from the program.

t1_step:=1; this variable controls the operation of the Case statement and the timer subroutine. Control is given to the part of the Case statement corresponding to the number written in t1_step.

Timer1.Enable:=true; Gives permission for the timer to run.

As a result, the timer subroutine becomes active after 0.2 seconds and control is given to the first part of the Case statement.

Step 2. In the first part of the Case statement, a query to be sent is formed and is transmitted over the appropriate channel. The numbers Δt and p are set such that

$\Delta tp \geq T_{max} \geq \Delta t(p-1)$

timer1.Interval:= Δt;

Then t1_step:=2; is written in this part. As a result, control moves out of the first part of the Case statement and after time Δt control is given to the second part of the Case statement.

Step 3. When control is given to the second part of the Case statement, it is first checked to see if a response to the sent query has been received. If a response has been received, control is passed to the part of the case operator to receive, check and process the received response. To do this, it is simply necessary to write t1_step:=101 and exit the second part. As a result, since the other parameters have not

changed, control is passed to the 101st part of the case statement after the time Δt has elapsed. If there is no response, the operation P=p-1 is executed. The condition P=0 is checked. Fulfillment of the condition means that time has passed and no response has been received. Therefore, in this case, t1_step:=201 is written, control is passed to the 201st part to generate information about the incident and determine the need for a repeat query.

If the condition P=0 is not met, the waiting time T_{max} has not yet expired and control is returned to the second part of the Case statement.

Step 4. When control reaches the 101st part of the Case statement, the incoming response is accepted and the necessary operations are performed on it. The need to send a new query is determined and, if necessary, control is transferred to the first part of the Case statement. If there is no need to send a new query, the timer is temporarily stopped and control is transferred to the output.

Step 5. As is already known, when control reaches the 201st part of the Case statement, it is established that the waiting time has expired and no response to the sent query will arrive. Thus, information about the incident is generated, entered into the operational log and control is sent to the output.

The algorithm shown here is an abbreviated algorithm for an online timer subroutine. However, it is evident from this algorithm that this method can be used to organize and control the query process by staying inside the timer subroutine for short periods of time. In this case, the system can function normally without freezing.

The first section of the second chapter analyzes the characteristic features of IRS in DDB. It is known that modern directive information retrieval systems work mainly with electronic databases, which consist of electronically created documents. In general, directive information retrieval systems are a special type of EDMS systems with certain specific characteristics. The similarities and differences between electronic documents and electronic directives are discussed here.

It is noted that the EDMS DB stores electronic documents created internally and received from the outside of the enterprise, as

well as information reflecting the attributes of these documents. Although documents are divided into two categories depending on their functional purpose and nature: informational documents and directives, and they can be stored in different file formats.

The second chapter also defines the requirements that should be imposed on enterprises where electronic documents are stored and document management is carried out electronically. Necessary measures and obstacles for enterprises to switch to ED are identified.

This chapter analyzes policy documents and, in general, documents in electronic form, as well as the file formats in which the attributes of these documents are stored. The basic file formats used in modern EDMS are considered. The methods of using simple file formats when creating directives and EDMS software are given.

Algorithms and methods of using TEXT format files in software are given. Files of this type are used for software initialization, setting initial program modes and for other purposes.

It is recognized that there are special requirements for storing electronic documents and directives in particular. One such special requirement is to ensure that the directive is maintained without corrections or alterations. Obviously, electronic documents can be altered accidentally or deliberately. After each such correction, the files storing the document are overwritten with new control codes. Therefore, it is not easy to determine whether the original document has been altered. To solve this problem, the second chapter proposes to store information about incoming documents in a separate database and to identify and add their metric characteristics to it.

The dissertation sets and solves the problem of determining cyclic codes for the entire document, as well as the problem of computing cyclic codes for each non-text image. *Algorithms and methods* are given that allow data to be stored as a single variable in a database and operations to be performed backwards from that variable.

CRC algorithms. It is known that cyclic codes have been used to ensure the security of transmitted and received data since the time of Hamming [Hamming]. CRC algorithms using polynomials were developed to create more powerful error detection codes. When using these algorithms, the parties transmitting and receiving information

must establish the generator polynomial G(x). That is, the generator polynomial G(x) must be known to both parties before the operation. In the binary representation of this polynomial, the low and high bits must be one ("1"). When calculating CRC, if the binary representation of the transmitted polynomial M(x) has m bits, zeros are added to the division frame so that the resulting new polynomial is divisible by G(x).

Thus, the algorithm for calculating CRC is as follows:

1. Suppose that the degree of *G(x)* is *t*. Let us supplement the frame with r zeros, obtaining the polynomial $x^r M(x)$ in the frame. This new polynomial is divisible by the polynomial *G(x)*.

2. Divide the polynomial $x^r M(x)$ by the polynomial *G(x)* modulo 2.

3. Let us remove the remainder of division. It normally has r or less bits. Removing the remainder from the polynomial $x^r M(x)$ by subtraction modulo 2, we obtain the polynomial *T(x)* for transmission.

$$T(x) = x^r M(x) - \{x^r M(x)/G(x)\}.$$

From the above, it is clear that this new polynomial generated for transmission is divisible by G(x). As a result of this operation, the polynomial T(x) is transmitted by the transmitter to the receiver through communication channels. In a distortionless reception, that is, if the information is not distorted along the way, the receiving party will also receive the polynomial T(x). In our case, if the archived document is not distorted intentionally or accidentally, i.e., if it is not modified, it will remain the same when read again.

However, it is technically known that information can be accidentally added or lost during transmission. In both cases, if we compare the transmitted and received information, we will find that there is a difference equal to the polynomial E(x). Thus, the receiver receives information equal to *T(x)+E(x)*. The purpose of the operations performed is to determine the polynomial added during transmission, or at least to detect the fact of addition.

4. The receiver performs the following division:

$$[T(x) + E(x)]/G(x).$$

Here *E(x)* is the addition in the process of information transfer. In our case, it can be understood as changes caused by intentional or

accidental correction of the document. According to the above, if the information was received in a pure form, i.e., if the receiver received only *T(x)*, then *T(x)/G(x)* = 0.

This shows that the result of the check depends on the division of the polynomial *E(x)* added in the exchange. In general, if *E(x)* is added, the exponent of the degree is non-zero.

As mentioned above, the result of the division depends on *E(x)/G(x)*. This result is the first result of our operations. It should be noted that these operations have other results as well.

The second result is that the errors in the analogs of *G(x)* remain undetected. That is, the remainder of the division of errors equal to the analogs of *G(x)* is zero. All the other errors are detected by the algorithm.

The more important result is whether the information transmitted or stored in the file contains only single-bit errors or changes, i.e., $E(x) = x^i$, where *i* indicates which bit is erroneous.

That is, *i* here indicates which bit is erroneous. If *G(x)* consists of two or more components and *E(x)* is not divisible by *G(x)*, then all single-bit errors are determined.

The calculations used in implementing this algorithm are, in CRC arithmetic terms, simple division operations. All that will have to be done before the operation starts is to add zero bits to the information.

After determining all these parameters, the created model can be used to accurately describe the features of each CRC algorithm.

Table 1. Volumes of some directives

Title of document	Size of the file in which the document is stored (in bytes)	Note	File address
Director's instruction	187.512	Internal	Address

Table 1. Volumes of some directives (continued)

HAC Decision	538.609	For defense councils	Address
Technical sciences journals	538.609	Publication of the research results is recommended	Address
Regulation	152.304	On awarding academic degrees	Address
HAC Bulletin	593.257		Address
Guideline	2.229.789		Address
Rules	545.721		Address
Ordinance of the Ministry	86.221		Address
Order	556.930	Internal	Address
Rules for author's abstract	40.960	Internal	Address
...	...	...	...

An analysis of the directives used in educational institutions shows that the volumes of such directives vary as shown above (Table 1).

This table and an analysis of volume of many other documents show that documents are distributed approximately evenly in terms of volume. Therefore, when dealing with directives within an institution, simple methods can be used to improve the efficiency of CRC algorithms.

If we have a document in quantity I, then the volumes of documents can be denoted by v_i , $i \in [1, I]$.

To construct the distribution function of document volumes, denote the maximum value of the set by V_{max} and the minimum value by V_{min}. If we want to construct a distribution function for a segment M, we can calculate the width of the obtained equal intervals as follows:

$\Delta V = \frac{V_{max} - V_{min}}{M}$.

The values of the distribution function can be written as

y_m , $m \in [1, M]$,

and the values of the distribution density as

p_m , $m \in [1, M]$.

The values of the distribution function can be determined using the following algorithm:

Step 1. Clear the values of the distribution function for $\forall m \in [1, M]$=0, y_m=0

Step 2. If $v_i \in [(m-1)\Delta V,\ m\Delta V]$, then for each $\forall i \in [1, I]$ we can write $y_m = y_m + 1$. As a result, we obtain the values of the distribution function. The values of the distribution density function can be obtained as a result of the following operation. We can write as follows:

$$p_m = \frac{y_m}{\sum_{m=1}^{M} m}.$$

In the experiments, when the number of documents was about 500 and M=50, the chart of the distribution function had the following shape (Chart 1).

Plotted along the X axis of the chart shown in Fig. 1 are the values of $m \in [1.50]$ of the above quantity, the Y axis shows the number of documents whose volumes are in the corresponding intervals.

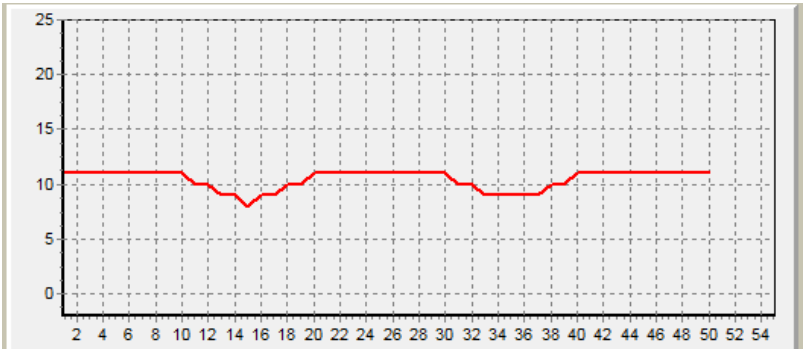

Chart 1. **Distribution function of the volume of directives**

CRC detection algorithms covering the whole document.

As previously mentioned, specific characteristics are calculated for directives and added to the general characteristics. CRC bytes (crc_low01 and crc_high01) are calculated based on the algorithms provided for the text parts of the document in the transmitting and receiving devices. The non-text parts of the document are then separated into BMP files, and CRC bytes are also calculated for these parts. As the name implies, these files are binary. These files store data in bytes. Each BMP file consists of four parts and has a special structure:

- The file header consists of 14 bytes and contains the following information:
  - file type BMP;
  - file size in bytes;
  - sequence number of the byte starting the image.
- The file information header consists of 40 bytes and contains the following information:
  - the number of rows and columns in the image;
  - the number of bits in a pixel;
  - compression law, if the image has been compressed;
  - other less significant and less common specifying information.
- Color table:
  - Not used for 24-bit pixels.
- Pixel storage:
  - information about the color of each pixel;
  - information in each branch below in each image;
  - each row is supplemented with enough bytes to ensure that the number of bytes in it is divisible by four. It is padded with zero bytes if necessary;
  - each line is written from left to right;
  - in 24-bit images, colors are formed from red, green, and blue.

A decimal number can be formed from the combination (arrangement) of bytes given in the table. If the number is given in four bytes, it can be calculated as follows.

Decimal number =

((4th byte*256+3rd byte)*256+2nd byte)*256+1st byte

Thus, a method is proposed to identify characteristics that can be special indicators of documents by applying cyclic codes.

This method, developed algorithms and program fragments can be used to solve similar problems.

The **fourth section** of this chapter contains the methods of working with BMP files described in the dissertation, as well as fragments of program modules corresponding to the structure of BMP files.

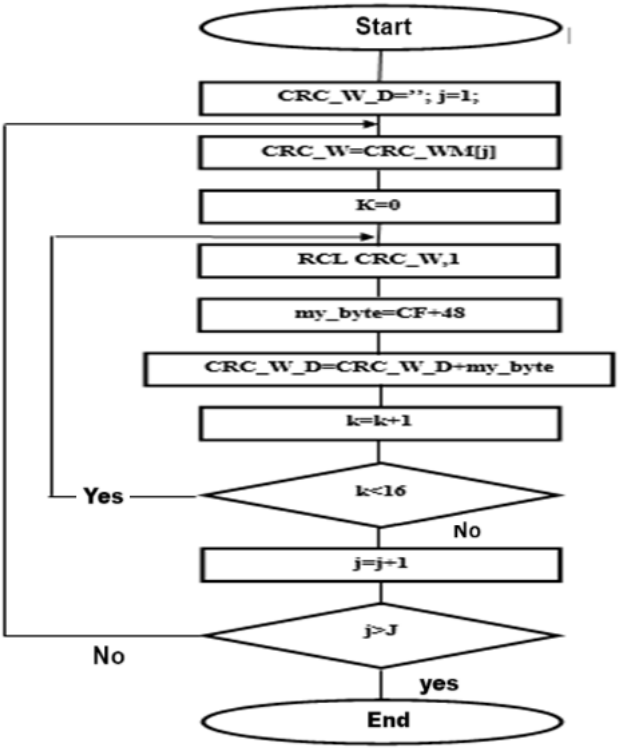

Fig. 3. Writing CRC bytes as a single string

To solve the above problem and calculate the CRC bytes for the whole document, after loading the image into the AR_BT array, the CRCV bytes of the image can be determined from this array. A new text variable CRC_W_D is created to be written to the database as a CRC_W variable. Each binary bit of each calculated CRC_W is written as a character in CRC_W_D. For this purpose, the cycle left shift command and the transition bit are used as follows:

Step 1: Cleanup operation CRC_W_D (CRC_W_D:='')

Step 2: Shift the next CRC_W cyclically one bit to the left.

Step 3: Write the transition bit my_byte.

Step 4: Add to convert to symbol 48 (my_byte:= my_byte+ 48).

Step 5: CRC_W_D:= CRC_W_D+ my_byte.

Step 6: Repeat steps 2-5 for each 16 bit.

Step 7: Repeat steps 1-6 for each CRC_W.

The block diagram of the above algorithm is given in Figure 3. In the block diagram, the set of CRC_Ws is denoted by CRC_WM. The total number of elements in the set is denoted by *J*. The current indices are denoted by *j* and *k*. The transition bit is denoted with CF.

As a result, a CRC_W_D of length 16*(number of CRC_W) bytes is obtained. If necessary, knowing CRC_W_D, it is possible to determine the number of images in the document:

Number of images=((length of CRC_W_D)-16):16 or

fig_count:=(length(CRC_W_D)-16)/16.

The **first section of the third chapter** of the dissertation presents a method for determining some of the performance characteristics of an information retrieval system in the directives database.

The life cycles of system operation are considered here, and the chart of system failure rate vs. time is investigated. It is shown that, although failures and malfunctions at the beginning of operation are high, they decrease rapidly over time, and failures become random during the main period of normal operation of the system. During the final period of wear and tear and breakdown of the system, the failure rate begins to increase again. (Chart 2)

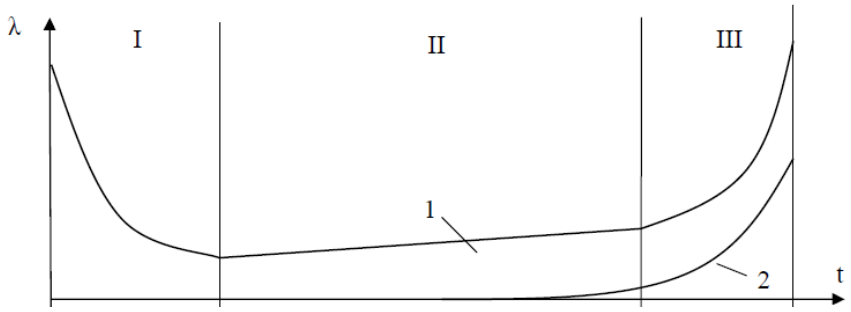

Chart 2. Intensity of failures over time: 1– curve of the intensity of failures λ(t); 2 – wear curve; I – initial period of operation; II – normal period of operation; III – wear period.

During monitoring, the total number of incoming queries from each user of the system, as well as the number of successful and unsuccessful queries are counted. Let N_i denote the number of incoming queries from each i-th user, $i \in [1, K]$, N_i^+ denotes the number of successful queries, and N_i^- denotes the number of unsuccessful queries. When any query from the first user arrives in the system, they can be calculated as follows:

$N_i = N_i + 1$,

if the query is successful

$N_i^+ = N_i^+ + 1$,

if the query is unsuccessful

$N_i^- = N_i^- + 1$.

It is clear that

$N_i^+ + N_i^- = N_i$

The main numerical measure of system reliability is the probability of failure-free operation during the operating period t. The probability of failure-free operation means the probability that the system will not fail within a certain period of time under given operating conditions. Since system failures are random events, the time of their occurrence t_0 should be considered a random event. Consequently, the probability of failure-free operation of the system can be

$$p(t) = p(t_0 > t).$$

Here t is a specified operating time. Since failure and failure-free operation are mutually exclusive events, the probability of a failure occurring in the system within a specified time t can be

$$q(t) = q(t_0 \leq t).$$

The sum of these two probabilities must satisfy the equality

$$p(t) + q(t) = 1.$$

The probability of IRS in DDB operating without failure up to a given time t and the probability of failure can be determined experimentally. We can denote these probabilities by $p^*(t)$ and $q^*(t)$, respectively. Then they can be calculated as follows:

$$p^*(t) = \frac{\sum_{i=1}^{K} N_i^+}{\sum_{i=1}^{N} N_i} \ , q^*(t) = \frac{\sum_{i=1}^{K} N_i^-}{\sum_{i=1}^{N} N_i}$$

It is clear here that

$$p^*(t) + q^*(t) = \frac{\sum_{i=1}^{K} N_i^+}{\sum_{i=1}^{K} N_i} + \frac{\sum_{i=1}^{K} N_i^-}{\sum_{i=1}^{K} N_i} = \frac{\sum_{i=1}^{K} N_i}{\sum_{i=1}^{K} N_i} = 1$$

Then, for simplicity of our subsequent calculations, let us take that

$$p^*(t) = p(t) \text{ and } q^*(t) = q(t).$$

The correctness of these assumptions can be seen once again from the fact that the equalities

$$p(t) = \lim_{N_i \to \infty} \frac{\sum_{i=1}^{K} N_i^+}{\sum_{i=1}^{K} N_i}$$

$$q(t) = \lim_{N_i \to \infty} \frac{\sum_{i=1}^{K} N_i^-}{\sum_{i=1}^{K} N_i}$$

are true. Of course, it is impossible to obtain $N_i \to \infty$ in practice. However, as time goes by, the number of queries received from each user during operation will take on increasingly large values. Over time, the number of failures in the system will start to increase. Consequently, the value of p will tend to decrease as it becomes greater than zero, and the value of q will tend to increase as it becomes less than 1.

One of the possible link structures of an information retrieval system is also given here. In practice, the basic link structure of a

medium-sized IRS in a database can be set as shown in the second chapter of the dissertation. Although the database is located in the local network, the system can have different sources of information, and these sources can be located both in the local network and in the global network. The monitoring of the IRS operations is organized on the system server. Based on the monitoring results, some operational characteristics of the system are calculated and refined over time.

It should be noted that various parameters can be selected to evaluate the system performance. Since this process is random, the change in the values of the recorded parameters cannot be smooth. Parameter values can change with random decreases and jumps. Therefore, reliable prediction of the values of the selected parameter is of great importance and relevance.

The second section of the third chapter shows that if we can write the performance characteristics that we calculate at different stages of the system life cycle as a time series, we can apply various autoregressive models to predict the future values of this series. Time series are indicators observed over a certain time interval. In practice, moving average or weighted moving average models are often used.

When applying the moving average model, the coefficients are taken to be equal to each other, assumed to be known, and the last few values of the time series are used. Here, if we use the last numerical value of the row as above, we can write as follows:

$$Z_t = \frac{1}{v} Z_{t-1} + \frac{1}{v} Z_{t-2} + \ldots + \frac{1}{v} Z_{t-v} + \varepsilon_t$$

Here, if the noise is white noise, it can be ignored. We can get the following approximate value.

$$\overline{Z_t} = \frac{1}{v} Z_{t-1} + \frac{1}{v} Z_{t-2} + \ldots + \frac{1}{v} Z_{t-v}$$

If we want to give priority to the last values of the time series we calculate in the model, then we get a weighted moving average model. Here, if we want to write the final approximate value, we get the following expression:

$$\overline{Z_t} = \frac{v}{v+(v-1)+\ldots+1} Z_{t-1} + \frac{v-1}{v+(v-1)+\ldots+1} Z_{t-2} + \ldots + \frac{1}{v+(v-1)+\ldots+1} Z_{t-v}$$

The issue of which of these models to prioritize for implementation at different stages of the life cycle of a IRS in the DBD is of practical importance. To solve this problem, it is necessary to record the performance characteristics obtained at different periods of the system's life in the form of separate time series. The results can be compared by applying moving average (MA) and weighted moving average (WMA) models to each obtained row.

The fragments (15 last values) taken from the time series, belonging to the first and second periods of the system life cycle, can be written as follows:

Values belonging to the first period: 5, 7, 6, 4, 5, 3, 3, 4, 2, 1, 2, 1, 2, 1, 1

Values belonging to the second period: 1, 0, 2, 3, 0, 1, 1, 0, 2, 1, 2, 1, 0, 1, 2

The moving average for the values taken from the first period will be called $MA1$, and the weighted moving average will be called $WMA1$. Consequently, the moving average for the values taken from the second period can be called $MA2$, and the weighted moving average $WMA2$.

Calculate the moving average from the values obtained from the first period:

$$MA1 = \frac{1}{15}(5 + 7 + 6 + 4 + 5 + 3 + 3 + 4 + 2 + 1 + 2 + 1 + 2 + 1 + 1) \approx 3.1$$

If we calculate the weighted moving average for this period, we get:

$$WMA1 = \frac{1}{120}(1 * 5 + 2 * 7 + 3 * 6 + 4 * 4 + 5 * 5 + 6 * 3 + 7 * 3 + 8 * 4 + 9 *$$

$$* 2 + 10 * 1 + 11 * 2 + 12 * 1 + 13 * 2 + 14 * 1 + 15 * 1) \approx 2.2$$

If we carry out similar operations for the second period, we get

$$MA2 = \frac{1}{15}(1 + 0 + 2 + 3 + 0 + 1 + 1 + 0 + 2 + 1 + 2 + 1 + 0 + 1 + 2) \approx 1.1$$

$$WMA2 = \frac{1}{120}(1 * 1 + 2 * 0 + 3 * 2 + 4 * 3 + 5 * 0 + 6 * 1 + 7 * 1 + 8 * 0 +$$

$$+9 * 2 + 10 * 1 + 11 * 2 + 12 * 1 + 13 * 0 + 14 * 1 + 15 * 2) \approx 1.2.$$

If we analyze our results, we will see that

$|MA1 - WMA1|$=0.9, and the last values of the first period are closer to WMA1. Thus, we can apply the weighted moving average model to predict new values in this period.

$|MA2 - WMA2|$=0.1, and the last values of this period are close to the predicted results obtained using both models. Thus, both models can be applied to predict new values in the second period.

If we were to make these comparisons for the third life period of the IRS in the DDB, we would have to favor the weighted moving average model here as well.

The third section of the third chapter notes that performance improvement of IRS in DDB can be done in different ways. The local or international significance of the documents, and whether they were created and executed by local or international organizations play an important role.

It has been found that there is an approximately linear relationship between the volume of documents and the execution of queries.

The linearity of this dependence allows simple methods to be applied to the division of documents.

According to Chart 3, we can say that the dependence is linear. However, for a more accurate synthesis of the model it is necessary to determine the parameters of different dependence models using the table, and then compare the obtained models with the experimental data.

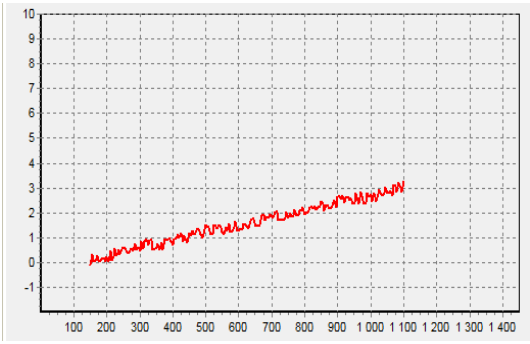

Chart 3. Query execution times vs. data volume

This division of documents and data according to the conceptual model (Fig.1) can be called natural. In general, in any serious organization, information and documents can be divided based on other principles. Let us consider some of the advantages that such a division provides. Suppose there are N documents or M data in sub-databases. The number of documents in these sub-databases, respectively, is $m_1, m_2, \dots, m_M$. Here, $\sum_{i=1}^{M} m_i = N$.

The set of documents stored in the i-th sub-database can be denoted as $S_i = \{s_{i1}, s_{i2}, \dots, s_{im_i}\}$. Here S_i indicates the i-th sub-database; and s_{im_i} indicates the document number m_i in the i-th sub-database.

Without taking into account the structure of the network in which the databases are located, we can assume that the search time is proportional to the number of elements stored in the database. The average value of the proportionality coefficient can be denoted by K. Thus, the search time in the i-th sub-database will be $T_i = Km_i$.

If no division has been performed, the search time over the entire whole database may be $T = KN$. Obviously, if the databases have been divided and sub-databases have been created, search time in the i-th sub-database can be saved to approximately $T - T_i$. The time spent will be fraction $\frac{T_i}{T}$ of the total time, and the saved time will be fraction $\frac{T-T_i}{T} = 1 - \frac{T_i}{T}$.

If
$\min\{m_1, m_2, \dots, m_M\} = m_{min}$, we denote the corresponding search time by
$T_{min} = Km_{min}$ and
for $\max\{m_1, m_2, \dots, m_M\} = m_{max}$, the corresponding search time by
$T_{max} = Km_{max}$,
then the minimum and maximum time saved resulting from the division will be
$T - Km_{min}$ and $T - Km_{max}$, respectively. Correspondingly, these are fractions
$1 - \frac{T_{min}}{T}$ and $1 - \frac{T_{max}}{T}$ of the total search time, respectively.

The first paragraph of the fourth chapter of the dissertation describes methods of storing information about large files when creating databases in various systems, as well as algorithms for working with such files. It is shown that the necessary information about such files, including information about the folders in which they are stored, is recorded in the database.

In the database of the information retrieval system in the directives database, algorithms for recording the files of the database into file folders of the same level were developed in order to simplify the recording, re-processing and reading of detailed information about the documents coming into the system and the organizations sending these documents. In particular, an algorithm for working with .dat files is given here:

At this stage, the following two problems need to be solved. The first problem is to read the data in the file and convert it to real number type; the second one is to sequentially read the names and creation dates of files in the folder and write them into the required arrays.

Solution of the first problem.

- Open the selected datatype file and write it to the AR_BT array. This array must have at least 16,000 bytes of available space. The first byte of each pair is the high byte and the second byte is the low byte. If the code is a positive number, the high byte will be less than 8, that is, the 12th sign bit will be zero. Otherwise, the high byte is greater than or equal to 8, i.e. the 12th sign bit is 1;
- The real numbers obtained after the conversion can be stored in the array of real numbers big_r. We will write the larger part of each pair from the AR_BT array into big_k bytes. Hence, if big_k< 8, then the number big_r is calculated as
  big_r[i]:=AR_BT[i*2-1]*256 + AR_BT[i*2].
  Otherwise, it can be calculated as
  big_r[i]:=-1*((255-(AR_BT[i*2-1]))*256+not(
  AR_BT[i*2])).

  The resulting array of real numbers big_r can be used for a variety of purposes, including analysis and plotting..

Solution of the second problem.

To save the names and creation dates of files in a folder, a record with the following structure must be created:

Fp : TSearchRec;

FAge: Integer;

FileParam: TDateTime;

Then the necessary information about the first file is written to the Fp record with the

FindFirst('..\BASA_S*.dat', faAnyFile, Fp) command

The FindNext(Fp) command writes information about other files in the folder to the Fp record.

Both commands return the file name in the Fp.Name parameter, the date the file was written in the FileAge(('..\BASA_S\'+ filename)) parameter, and the date and time of the write, respectively, via the FileParam:=FileDateToDateTime(filedate),

DateToStr(FileParam), TimeToStr(Fileparam) parameters.

By using these commands and parameters appropriately, the names and creation dates of all files in the BASA_S folder can be written to arrays. As a result, it becomes possible to perform the intended operations.

The proposed methods, algorithms and program fragments can be used in creating software for various systems.

The second section of the fourth chapter of the dissertation discusses the creation of a database of IRS in DDB at an enterprise. The Relational DB model is preferable for creating successful database search queries. Since there are higher requirements for the storage and execution of directives, some metric characteristics of these documents are defined and included in the database. Some example queries are given here.

In the dissertation, methods, technologies and algorithms have been developed to identify corrections made to documents after they have been written to the database.

The **results** section of the dissertation specifies the scientific results obtained in solving the objectives set in the dissertation: the main scientific-theoretical and scientific-practical results:

RESULTS

1. Electronic directives and their characteristics were considered, and methods for working with text files of directives have been developed. These methods have been applied to work with different file formats [7, 12];
2. An algorithm has been created for identifying the metric characteristics of electronic directives documents. These algorithms have simplified the detection of accidental or intentional corrections in directives documents [4, 6, 14];
3. A stochastic model of the information retrieval system in the directives database has been created, algorithms and methods for determining the experimental performance characteristics have been developed [1, 2, 11];
4. In order to improve the performance of the information retrieval system in the directives database, the principle of natural division of database files has been developed. Here, the database files are divided into several parts according to the sources of creation, and the search operations are performed on databases of smaller size [13];
5. A structure of the database has been created which contains various essential aspects of the information retrieval system in the directives database [3, 5];
6. To simplify the software writing, a solution has been proposed to store such folders on the same level as the software [10];
7. A general set of algorithms has been created for an information retrieval system in the directives database [8, 9].

THE FOLLOWING SCIENTIFIC WORKS HAVE BEEN PUBLISHED BASED ON THE DISSERTATION MATERIALS:

1. Musayeva N.F. , Paşayev İ.F. , Paşayeva S.E. Kompressor qurğusunun və ştanqlı, dərinlik nasoslu neft quyularının robast nəzarət və diaqnostika sistemində simsiz lokal şəbəkənin tətbiqi. AzMIU Elmi Əsərlər, 2015, 1, s. 47-52. https://azmiu.edu.az/upload/ckeditor/507207714.pdf

2. Musayeva N.F., Paşayev İ.F., Paşayeva S.E., Bayramov V.V., Cəfərov C.M., Süleymanli B.Ə. Neft sənayesi müəssisələrində simsiz lokal şəbəkələrin yaradılması prinsipləri. Azərbaycan Milli Elmlər Akademiyasının XƏBƏRLƏRİ, İnformasiya və İdarəetmə Problemləri, 2015, Cild XXXV, № 6, s. 95-103. https://icp.az/2015/6-10.php
3. Пашаев Ф.Г., Пашаев И.Ф., Пашаева С.Э., Алиева. Б.М. Локальный поиск документов в корпоративной среде. VII All-Ukrainian Scientific-Practical Conference «Computer Sciences and Systems Sciences», Poltava, 10-12 march 2016, pp. 223-225. http://dspace.puet.edu.ua
4. Пашаев Ф.Г., Пашаева С.Е. Определение некоторых метрических характеристик электронных директивных документов. Международный научный институт “Educatio”, Ежемесячный научный журнал, IV(22), 2016, с 64-67. (РИНЦ) https://edu-science.ru/
5. Paşayeva S.E. Direktiv sənəd axtarışı sisteminin verilənlər bazasının yaradılması. Naxçıvan Dövlət Universiteti. Elmi əsərlər, 2016, № 3 (77), s 59-67. https://ndu.edu.az/public/wp-content/uploads/Elmi%20Eserler/77%20deqiq%202016.pdf
6. Пашаева С.Э Метрические характеристики электронных директивных документов. Riyaziyyatin tətbiqi məsələləri və yeni informasiya texnologiyalari III Respublika Elmi Konfransı materialları, SDU, Sumqayıt 15-16 dekabr 2016-cı il, s 282-283.
7. Paşayeva. S.E. Mətn şəkilli fayllarla işləmək üsulları. Azərbaycan Milli Elmlər Akademiyasının XƏBƏRLƏRİ, İnformasiya və İdarəetmə Problemləri, 2016, Cild XXXVI, № 6. s. 104-111. (AAK) https://icp.az/2016/6-14.php
8. Пашаева С.Э. Параметры поиска в базе директивных документов. VIII All-Ukrainian Scientific-PracticalConference «Informatics and Systems Sciences» Poltava, 16-18 march 2017, pp. 211-213. http://dspace.puet.edu.ua
9. Paşayeva S.E. Direktiv sənəd bazasında informasiya axtarış sisteminin yaradılması. Azərbaycan Milli Elmlər

Akademiyasının XƏBƏRLƏRİ, İnformasiya və İdarəetmə Problemləri, 2018, Cild XXXVIII, № 6, s.110-119.(AAK) https://www.icp.az/2018/6-13.pdf

10. Pashayev F.H., Pashayeva S.E., Najafov H.T., Suleymanli B.A Operating algorithmes for folders and files. Информационные технологии, том 25, №3, 2019, с.152-156. http://novtex.ru/IT/it2019/number03.html
11. Pashayev F. H., Pashayeva S. E., Jafarov J. M., Pashayev I. F. Determination Method of Some Operational Characteristics of Information Search System in Directive Document Database. I.J. Mathematical Sciences and Computing, 2020, 3, 1-11 (GOOGLE Scholar) https://www.researchgate.net/journal/International-Journal-of-Mathematical-Sciences-and-Computing-2310-9025
12. Paşayeva S.E., Paşayev F.H., Bayramov V.V., Süleymanlı B.Ə. BMP tipli dinamoqram şəkillərinin rəqəmsal variantının yaradılması alqoritmləri. Riyaziyyatin tətbiqi məsələləri və yeni informasiya texnologiyalari. IV Respublika Elmi Konfransı materialları, SDU, Sumqayıt , 09-10 dekabr 2021, s.143-147 https://www.ssu-conferenceproceedings.edu.az/pdf/riyaziyyat2021.pdf
13. Пашаев Ф.Г., Пашаева С.Э., Максудова Н. А.. Оптимизация работы поисковой информационной системы в базе директивных документов. Поблемы информационной безопастности. Компютерные науки. 2024. № 4, с 152-161 (РИНЦ) DOI10.48612/jisp/8gta-g6v5-k2hh. https://jisp.ru/article/optimizatsiya-raboty-poiskovoj-informatsionnoj-sistemy-v-baze-direktivnyh-dokumentov/
14. Melent Y., Korol O., Shulha V., Milevskyi S., O. Voitko, Rzayev K., Husarova I., Kravchenko S., Pashayeva S.. Development of post-quantum cryptosystems based on the rao-nam scheme. Eastern-European JOURNAL of Enterprise Technologies. 1/9 (133) 2025, pp.35-48. SCOPUS DOI: 10.15587/1729-4061.2025.323195. https://journals.uran.ua/eejet/article/view/323195/314211

The defense of the dissertation will be held at the meeting of the ED 1.20 Dissertation Council operating at the Institute of Control Systems of the Ministry of Science and Education of the Republic of Azerbaijan on June 30, 2025, at 13:$^{00.}$

Address**:** 68 B. Vahabzade Str., Baku AZ1141, Republic of Azerbaijan

Dissertation is accessible at the library of the Institute of Control Systems of Ministry of Science and Education of the Republic of Azerbaijan.

Electronic versions of the dissertation and its abstract are available on the official website of the Institute of Control Systems of the Ministry of Science and Education of the Republic of Azerbaijan.

Abstract was sent to the required addresses on May 23 2025.

Signed for print: 22.05.2025
Paper format: (A5)
Volume: 37312 characters
Print run: 20 copies