Bulletin of TUIT: Management and Communication Technologies

Volume 4 *2021-third number* Article 4

April 2021

Intellect models used to search for data in stochastic information environment in Data Systems

Bahodir Muminov
"Bulletin of TUIT: Management and Communication Technologies", mbbahodir@gmail.com

Ulug'bek Bekmurodov
Bulletin of TUIT: Management and Communication Technologies, ulugbek_bekmurodov@mail.ru

Zarif Qodirov
qodirov-zarif_1992@mail.ru

Sunnat Hojiyev
Bulletin of TUIT: Management and Communication Technologies, Hojiyevsunnatullo_1991@mail.ru

Intellect models used to search for data in stochastic information environment in Data Systems

Muminov B. B.,
DSc., chief of department TUIT named after Muhammad al Khwarizmi Tashkent, Uzbekistane-mail: mbbahodir@gmail.com
Bekmurodov U.B.,
PhD student TUIT named after Muhammad al Khwarizmi Tashkent, Uzbekistan e-mail: ulugbek_bekmurodov@mail.ru
Qodirov Z.Z.,
PhD student TUIT named after Muhammad al Khwarizmi Tashkent, Uzbekistan e-mail: qodirov-zarif_1992@mail.ru
Hojiyev S.N.,
assistent TUIT named after Muhammad al Khwarizmi Tashkent, Uzbekistan e-mail: hojiyev-sunnat_1991@mail.ru

Abstract: IDEF models of the method of forming a semantic connection through calculations of the relevance on the basis of the area with the zone and attributes, and the calculation of the most cost-effective value of g, the method of calculating the weight of the data frequency and the frequency with inverse data, based on the model of intellectual semantic search, the creation of the ISIDR in the SIS and for software.

Keywords: *information retrieval, data, scholastic IE, IDEF model*

INTRODUCTION

Search and study of data associated with science and education play an important role in the development stages of society [7]. Requests, in models of intellectual search in information environments, are considered a series of terms. And this will be based on the models of storing the requested information with clear rules and their processing [7]. Also, the question of the connection with the queries given in the Stochastics information environments (SIS) and metadata (metadata) in the data on a number of terms is important. These methods designate the area of the

methodical, usually they are called data attributes. The types and values of these attributes are limited in advance.

The implementation of consumer requests through the adjustment to these attributes provides the probability of inter-request and query. Denote the logical query - q, and the data - d. (Q, d) pair, the resampling in the interval [0,1] in the weight attribute method from the relevance. Let the data zone take on logical values.

Consider the data collection in which there are data with l zones. Then let $g_1, g_2, \dots, g_n \in$ [0,1]_n∈ [0,1] and satisfy the following (1) expression.

$$\sum_{i=1}^{l} q_i = 1 \qquad (1)$$

Here, in $1 \le i \le l$ q- the query and the ratio *i* of the zone, s_s_i- logical dimension. If all query terms belong to one zone, exactly 1 if there is exactly 0. And in truth, any logical function

based on terms in the {0,1} zone can satisfy it. Such a weighting zone can be designated as follows:

$$V = \sum_{i=1}^{l} g_i s \qquad (2)$$

One can say a probable search for weighted areas of relevance on the basis of relevance or an innovative system of intelligent data retrieval (ISIDR) [5].

We give the most effective way of calculating the value of g for this system. Define the logical functions $s_T(d,q)$ and $s_B(d,q)$ for the query q and the data d. Let the logical functions determine whether the query q refers to the title of the data d. To calculate the relevance value in [0,1] on the basis of each data-request pair, we can use the values $s_T(d,q)$ and $s_B(d,q)$, and the constant g∈ [0,1]. Namely:

$$Result(d,q) = g \cdot s_T(d,q) + (1-g) \cdot s_B(d,q) \qquad (3)$$

Define the constant g based on the training example defined on the basis of $\Phi_j = \left(d_j, q_j, r\left(d_i, q_j\right)\right)$. Each training example is defined based on d_i data, q_i query and $r\left(d_i, q_j\right)$ relevance value. If the relevance is determined by (3), the expression will be exactly *1-g*. Expression (3) can also be written for other combinations $s_T\left(d_j, q_j\right)$ and $s_B\left(d_j, q_j\right)$ (look to table 1).

TABLE 1. COMBINATION OF FUNCTIONS S_T AND S_B

$s_T\left(d_j, q_j\right)$	$s_B\left(d_j, q_j\right)$	relevance
0	0	0
0	1	1-g
1	0	g
1	1	1

For $s_T\left(d_j, q_j\right) = 0$ and $s_B\left(d_j, q_j\right) = 1$, the number of training examples when the human factor is actual (i.e. 1) (hence irrelevant (i.e. 0)) is denoted n_{01r} (hence n_{01n}) . Then the expression (3) for $s_T\left(d_j, q_j\right) = 0$ and $s_B\left(d_j, q_j\right) = 1$ will be as follows:

$$R = [1-(1-g)]^2 n_{01r} + [0-(1-g)]^2 n_{01n}$$

Similarly, we can write the sum defined by expression (3) for other combinations $s_T\left(d_j, q_j\right)$ and $s_B\left(d_j, q_j\right)$..

$$R = (n_{01r} + n_{10n})g^2 + (n_{10r} + n_{01n})(1-g)^2 + + n_{00r} + n_{11n} \qquad (4)$$

If we simplify (4) The expression for g and the equality to 0, we determine the most effective value of g.

$$R = \frac{n_{10r} + n_{01n}}{n_{10r} + n_{10n} + n_{01r} + n_{01n}} \quad (5)$$

The relevance of the data will be related to the zones and the most effective value of g [1]. Now let's look at another story, i.e. We will take into account the relevance of the query for the zone or the data of the frequent term in the queries and give it a big value. This is convenient for entering sample queries, i.e. Meaning free entry, without the auxiliary means of querying in Search Engines (PDS). This approach is based on web technology and in it queries are created in natural language, and the query is considered as a collection of terms. To calculate the data values, it is sufficient to summarize the values of these terms in the queries.

The combination of each term meaning is related to the frequency of the term's involvement in the data and leads to the definition of the term weight. From this, it is possible to determine, through the weight of the term t in the data d, the relationship between the query q and the data d. This weight can be designated as the number of the term t in the data d. This algorithm is also called the frequency of the term and is denoted by $tf_{t,d}$. Here, t and d in the indices are the term and the data [5].

To improve the weight of the data frequency, we denote the number of data in the collection by the letter N and determine the frequency of the term t with the inverse data:

$$idf_t = \log \frac{N}{df_t} \quad (6)$$

So, the frequency of the inverse data of the little-configured terms exceeds the multifigure terms (table 2). In the second table, the number of data is N = 32021177 and it is computed using expression (6).

TABLE 2. EXAMPLE OF REVERSE DATA RATE

Terms	cf_t	idf_t
One	1980	4.209
Two	1981	4.21
Three	2004	4.20355
four	2005	4.2033

Thus, the weight $tf - idf_{t,d}$ of the term t in the data d has the following features:

1. If the term t will appear in a lot of data, it will have a maximum value (only this condition shows how it differs from other data).

2. If you meet several times in some kind of data, or will occur in many data, it will decrease.

3. If the term appears in all data, it will have a minimum value.

4. In the methods of calculating the relevance based on vector collections in the IRS, the definitions of the vectors $\bar{V}(q)$and $\bar{V}(d)$differ (Table 3) [2].

In the methods of calculating the relevance based on vector collections in the IRS, the definitions of the vectors $\bar{V}(q)$and $\bar{V}(d)$differ (Table 3) [2].

TABLE 3. TF-IDF OPTIONS FOR BASED SMART SYSTEMS

Frequency of terms	**Frequency Data**	**Normalization**
$tf_{t,d}$	1	1

$1+\log(tf_{t,d})$	$log\frac{N}{dt}$	$\frac{1}{\sqrt{w_1{}^2+w_2{}^2+\cdots}}$
$0.5 + \frac{0.5tf_{t,d}}{max_\tau(tf_{t,d})}$	$max\left[0, log\frac{N}{\cdots}\right.$	1/u ; uniqie
$\begin{cases}1, & \text{агар } t \\ 0, & \text{қолган ҳо}\end{cases}$		1/charLength^a, a<1 size)
$\frac{1+log(tf_{t,d})}{1+\log(ave_{red}(tf_{t,}}$		

In this table, charLength is the number of characters in the data. Table 3 lists the methods for calculating weights based on the basic schemes for computing the weights of the vectors $\bar{V}(q)$ and $\bar{V}(d)$, various storing them in memory.

We believe that these criteria will be paramount in the evaluation or use of ISIDR. Only ISIDR should justify the following criteria.

In the practice of users is not satisfied with the intellectual search of the above data. Because, to find clear and important data it takes time, and this in turn is related to the data elements [3].

1. The number of data;

2. Technical means (high-speed and external memory);

3. The way of information security;

4. Network infrastructure;

We present the model of the logical semantic search for decreasing the time for ISIDR in the SIS.

Suppose that there are collections of queries $q = \{q_i\}$ and $R^i = \{r_j^i\}, j = 1 \dots M, i = 1 \dots N$ … N corresponding to q_i. M - logical number, number of data in the system, N - number of requests in the system. Let's write the logical semantic search model for one pair:

$$f(q) = \{\langle q_i | q_j \rangle\} = \frac{|R^j \cup R^i|}{|R^j|} \quad (7)$$

Here, q is a collection of queries, {a|b} – | - an action that signifies the connection and similarity of a and b, which is determined by the action $\frac{|R^j \cup R^i|}{|R^j|}$. In general, in the logic of semantic search model, based on the grouping of semantic connection and similarity, expression (7) can be written as follows:

$$f(q) = \frac{|\bigcup R^i|}{|R^i|} \quad (8)$$

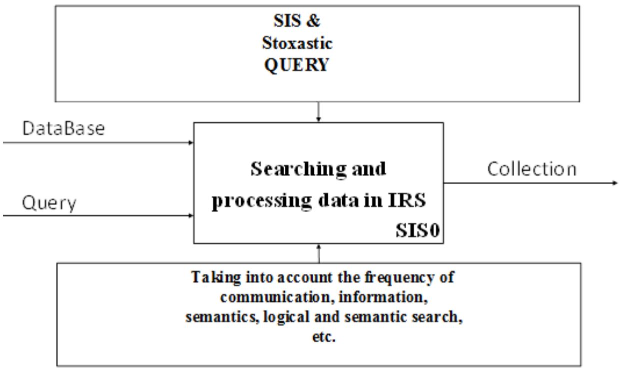
If the user request q can be generated with the help of requests in the SIS, i.e.

$$qNew = \bigcup qOld \quad (9)$$

In this image, the speed of data retrieval will change noticeably. Because, the logical semantic connection and similarity of data for queries in the system are defined by expressions (7) and (8) [3].

We will develop methods of basic elements and modeling factors of ISIDR in SIS, based on data frequency and vector models, software design models based on the calculation of relevance and IDEF technologies.

The general functionality of the ISIDR in the SIS is depicted with the help of the IDEF0 model (look to Pic 1).



Pic. 1. General functional IDEF0 model ISIDR in SIS

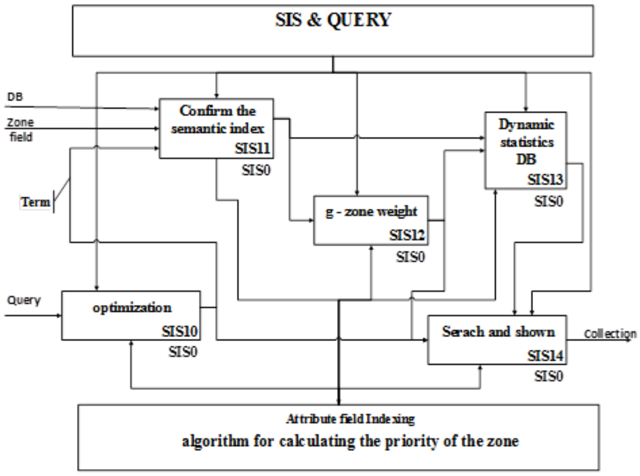
In the general functional IDEF0 model, the input parameters "query" and "DB" and the output parameters "Collection of sources". To create a collection of sources corresponding to the query in the SIS, it is necessary to develop IDEF0 models in which the functionality of the tools consisting of 3 consecutive stages will be displayed [6]. It:

stage 1. Indexing of the semantic connection, g - calculation of the weight zone and the formation of dynamic statistics (look to pic. 2.)

stage 2. Reading of the number of terms, normalization with respect to tf, and the formation of dynamic statistics based on the calculation of the frequency of the term (look to pic. 3.)

stage 3. Verification of communication and similarity based on the model of logical semantic search, indexing of queries and corresponding sources (look to pic. 4.)

These stages are based on the ways and models of generating statistical data for the search for queries (terms) in the SIS.

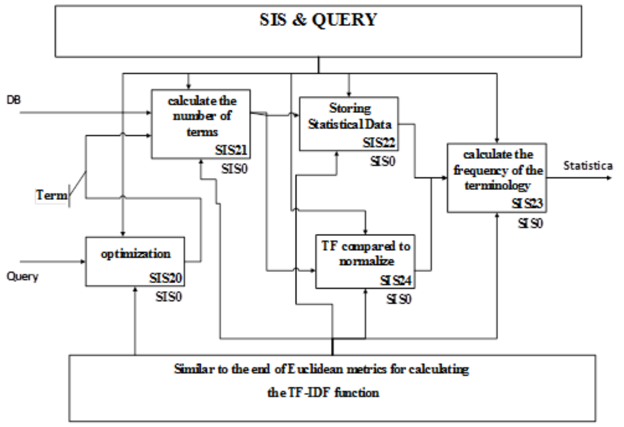


Pic. 2. Model IDEF0 of the stage 1 of ISIDR in the SIS

In the IDEF0 model of the stage 1 of the ISIDR in the SIS, the incoming parameters "query", "database", "zone area" and output parameters "data collection".

For "Area of a zone" the data can be formed in 2 special extended types, i.e. In a clear form and on the basis of preliminary templates.

After the optimization, this query is divided into terms. Based on the terms, the area of the zone is determined, and after determining the compliance with the "Area zone" templates are taken as data. Based on the data from the corresponding "Area of the Zone", sources in the database are compiled a semantic link between terms and sources and indexed. The value g of the area of the zone corresponding to the sources is calculated and dynamic statistics are written in the database. Dynamic statistics in the database should update the semantic link and recalculate the value of g when the term or source is added. The first stage serves to calculate the relevance of the source corresponding to this request through processing based on the area of the zones.

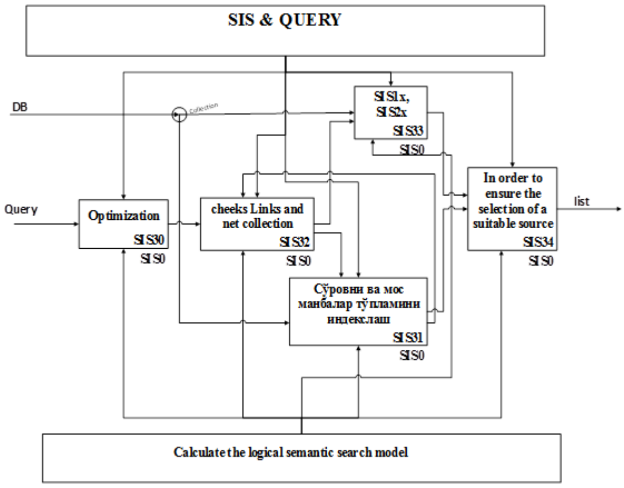


Pic. 3. IDEF0 model of the stage 2 of ISIDR in the SIS.

In the IDEF0 model, the steps 2 of the ISIDR in the SIS are the incoming parameters "query", "DB", and the output parameters "List".

In the stage 2 of the IDEF0 model, zones without area or zones that do not coincide with templates are considered effective. Here, queries are given as a sequence of terms and after optimization only collections of terms are formed. For these terms, the relevance of the sources is determined with the calculation of the frequency of terms. For this purpose, the number of terms in the source (SIS21) is calculated and the statistical data is stored. When the corresponding sources are presented, the frequency of the term is calculated and sorted in the decreasing order. The frequency of the term is considered to be non-permanent, so it is always calculated.

Computations with dynamic data in the first and second stages affect the speed of data retrieval in the ISIDR in the SIS. To do this, you can use the IDEF0 model of the 3 stage of the ISIDR in the SIS, checking the connection and similarity based on the software tools, i.e .n the logic model of semantic search.



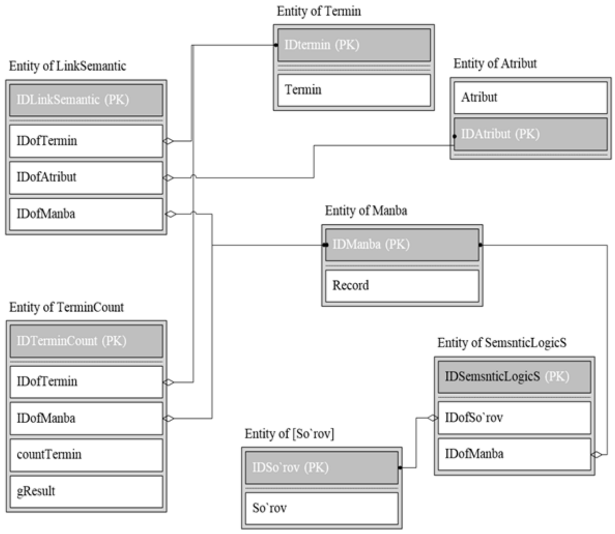
Pic. 4. IDEF0 model of the stages 3 of the ISIDR in the SIS

In the IDEF0 model, the 3 steps of the ISIDR in the SIS include the incoming parameters "query", "database", and the output parameters "List".

After optimization, the query is indexed. If the query can be formed from combinations of previous requests, the relationship between the requests is determined, the relevant sources are indexed and presented. If the request can not be formed from combinations of previous requests, the software tools from the first and second stages resort to tools and on the basis of them collections of sources are formed, and after the indexation are presented. The third stage significantly accelerates the search for information in the MIT at the expense of actions in the previous stages.

This developed stage of 3 model for ISIDR in SIS makes it possible to create instrumental software tools. To integrate, localize and work with other systems with IDEF0 models, IDEF1x models should be developed and integrated. Because, when IDEF0 models are integrated into other systems, data structures from IDEF1x models are needed.

IDEF1x model is a database model for IDEF0 adapted to store static and dynamic data based on relational data structures.



Pic. 5. IDEF1x ISIDR model in SIS

The IDEF1x model shows 7 objects and their characteristics:

1. "Entity of Termin" - created for collections of terms and has 2 characteristics of "IDTermin" and "Termin".

2. "Entity of Atribut" - created for collections of the area of the term zone and has 2 characteristics of "IDAtribut" and "Atribut"

3. "Entity of Manba" - created for collections of sources and has 2 characteristics of "ID Manba" and "Manba".

4. "Entity of SemanticLogicS" - created for collections of sources corresponding to requests and has 3 characteristics, "IDSemanticLogicS", "IDofSo`rov" and "IDofManba".

5. "Entity of [So`rov]" is created for collections of requests and has 2 characteristics of "IDSo`rov" and "So`rov".

6. "Entity of TerminCount" is designed to store the number of terms in the source and the value of g, and has 5 characteristics, "IDTerminCount", "IDofTermin", "IDofManba", "countTermin" and "gResult".

7. "Entity of SemanticLink" - is designed to collect semantic links of terms, area zones and sources, and has 4 characteristics, "IDLinkSemantic", "IDofTermin", "IDofManba" and "IDofAtribut".

Between the objects in the IDEF1x model there is only one-to-many relation (1: N).

5-TABLE. LINKS IN THE IDEF1X MODEL FOR THE IDEF0 MODELFOR THE 1-STAGE IDEF0 MODEL

No.	PK Field name	FK Field name	Communica-tion
IDEF0 model for the stage 1			
1	IDTermin	IDofTermin	(1:N)
2	IDManba	IDofManba	(1:N)
3	IDAtribut	IDofAtribut	(1:N)
IDEF0 model for the stage 2			
1	IDTermin	IDofTermin	(1:N)
2	IDManba	IDofManba	(1:N)
IDEF0 model for the stage 3			
1	IDSo`rov	IDofSo`rov	(1:N)
2	IDManba	IDofManba	(1:N)

To integrate the IDEF0 and IDEF1x models created for ISIDR in the SIS, the software package should be based on the principle of the first based on MVC technologies, or should be able to process models. This makes it possible to process the IDEF0 and IDEF1x models, and create their objects. The created objects are used to manage and exchange information between the software complex and objects.

CONCLUSION

Conclusion and recommendations based on the characteristics and models of ISIDR:

1. You can define and index the semantic relationship between the query and the data.

2. It is possible to identify and optimize important modeling elements for queries that have the characteristics of calculating the relevance of terms and area areas.

3. It is possible to use SMART systems for weight schemes of sources and queries in ISIDR, and for their normalization.

4. To accelerate the search for data in the ISIDR and other information systems, mathematical models of logical semantic search can be used.

5. You can use mathematical models of logical semantic search to create different semantic links.

6. You can create new IDEF0 models based on the IDEF0 model created for modeling the functionality of the software tools.

7. You can explore indexing, compression, information security and the choice of hardware in the SIS based on the IDEF0 and IDEF1x models.

REFERENCES

[1] Anagnostopoulos A., Andrei Z., Kunal P. Effective and efficient classification on a search-engine model. 2006. In Proc. CIKM, -P. -208-217.

[2] Arlitsch K., Obrien P., Rossmann B. Managing Search Engine Optimization: An Introduction for Library Administrators. Journal of Library Administration. Feb2013, Vol. 53 Issue 2-3, -p. -177-188.

[3] Azevedo C., Iacob M., Almeida J., Van S., Pires L., Guizzardi G. Modeling resources and capabilities in enterprise architecture: A well-founded ontology-based proposal for ArchiMate. Information Systems. Dec2015, Vol. 54, -P. -235-262.

[4] Bhat N. A., Ganaie Sh. A. E-resources: use and search strategies adopted by users of Dr Y.S. Parmar University of Horticulture and Forestry. Collection Building. 2016, Vol. 35 Issue 1, -P. -16-21.

[5] Cai F., Rijke M. Learning from homologous queries and semantically related terms for query auto completion. Information Processing & Management. Jul2016, Vol. 52 Issue 4, -P. -628-643.

[6] Cheang B., Chu S. OR/MS journals evaluation based on a refined PageRank method: an updated and more comprehensive review. Scientometrics. Aug2014, Vol. 100 Issue 2, -P. -339-361.

[7] Muminov B.B. Methods of information retrieval and cash algorithms. Transactions of the international scientific conference "Perspectives for the development of information technologies ITPA – 2014" 4-5 November, 2014. 194-196 p.

[8] Muminov B.B. The model of logical semantic searching. Proceedings of the uzbek-japan symposium on ecotechnologies "Innovation for sustainability-harmonizing science technology and economic development with human and natural environment" 14 may Tashkent – 2016, 72-77 p.

[9]_____Boltaevich, M.B., Ilxomovich, E.K.(2019) *Constructing a model of the*

process of receiving and storing oilseeds in oil and fat enterprises through the Petri net. International Conference on Information Science and Communications Technologies: Applications, Trends and Opportunities, ICISCT 2019

[10]Boltaevich, M.B., Ogli, N.R.H., Qizi, G.N.S., Ogli, M.S.S. (2019) *Estimation affects of formats and resizing process to the accuracy of convolutional neural network.* International Conference on Information Science and Communications Technologies: Applications, Trends and Opportunities, ICISCT 2019.

[11]. Bouville, M. (2008) Plagiarism: Words and ideas. Science and Engineering Ethics. Science and Engineering Ethics. Downloaded from https://arxiv.org/pdf/0803.1526.pdf

[12]. Doreen, S., Grace G. (2017) ACT BSSS What is plagiarism? (pp. 12-19) South Australia: ACT Board of Senior Secondary Studies. Downloaded from https://www.um.edu.mt/data/assets/pdf_file/avdplagiarism.pdf

[13]. Dyagilev, V., Say, A., Butakov, S. (2017) The architecture of service of detecting plagiarism. Moscow: Vestnik NGU.

[14]. Muminov B.B. Methods of information retrieval and cash algorithms. Transactions of the international scientific conference “Perspectives for the development of information technologies ITPA – 2014” 4-5 November, 2014. 194-196 p.

[15]. Boltaevich, M.B., Ilxomovich, E.K.(2019) Constructing a model of the process of receiving and storing oilseeds in oil and fat enterprises through the Petri net. International Conference on Information Science and Communications Technologies: Applications, Trends and Opportunities, ICISCT 2019 http://sersc.org/journals/index.php/IJAST/article/download/21872/10895/

[16]. Boltaevich, M.B., Ogli, N.R.H., Qizi, G.N.S., Ogli, M.S.S. (2019) Estimation affects of formats and resizing process to the accuracy of convolutional neural network. International Conference on Information Science and Communications Technologies: Applications, Trends and Opportunities, ICISCT 2019 https://ieeexplore.ieee.org/document/9011858

[17]. Bahodir, M., Bahrombek, S. (2019) Information, analysis ximmingning four neurons chemical name of the information in intellectual evidence. International Journal of Innovative Technology and Exploring Engineering. https://www.ijitee.org/download/volume-8-issue-11/

[18]. Mo’minov, B.B., Eshankulov, K. (2019) Modelling asynchronous parallel process with Petri net. International Journal of Engineering and Advanced Technology. https://www.ijeat.org/download/volume-8-issue-5s3/