ABSTRACT

of dissertation work by Chinibayeva T.T. "Models and Methods of Management of Data with Heterogeneous Structures (Big Data)", submitted for the degree of Doctor of Philosophy (PhD) in the specialty 6D070400 - Computer Science and Software Engineering.

The development of modern society and technology is associated not only with the digitalization of new areas of human activity, but also with the widespread introduction of research and data analysis technologies for the development of management decisions.

Much attention is paid to the development of this issue throughout the world, in particular in Kazakhstan. An important document defining the main directions of the country's digital development is the state program "Digital Kazakhstan", adopted on December 12, 2017. The project passport states that in connection with a significant increase in the volume of data, the state will help create a large technological center for data analysis and ensure reliable operation, safety, integrity of national and state information resources, including the basis of existing initiatives.

The relevance of the research topic is determined by the presentation of models and methods for managing big data with a heterogeneous structure, used to monitor and analyze information describing the activities of scientific organizations.

The level of scientific character of the research topic. Over the past five to ten years, scientists from near and far abroad have contributed to the study of big data. In particular: A.F. Tuzovsky, L.V. Naykhanova, A.N. Soulless, V.A. Serebryakov I.S. Mikhailov, Yu.A. Zagorulko, Dietmar Bayer, K. Shahgeldyan, N. Guarino, N. Noy, M. Erig, A. Maedche, Yong Im Cho and domestic authors: A.A. Kuandykov, R.K. Uskenbaeva, T.G. Balova, A.A. Sharipbayev, I. T. Utepbergenov, R.R. Musabaev, U. A. Tukeyev, N.K. Mukazhanov. According to the results of scientific analysis of the problem of combining data with heterogeneous structures collected from different sources, knowledge in this subject area is unsystematic, and the lack of a unified approach to big data management determines the structure, purpose and objectives of the study.

The aim of the research is to develop models and methods for managing scientific information with a heterogeneous structure.

The object of the research is heterogeneous scientific data.

The subject of the research is models and methods of managing data with a heterogeneous structure in order to ensure the semantic compatibility of documents.

Research methods. The tasks posed in the course of the research were solved by methods of analysis of natural language texts, classification and software engineering. The results were presented by the apparatus of mathematical statistics and mathematical logic.

The scientific novelty of the dissertation is based on the borrowing of terms from the announcements of scientific conferences, as well as the development of a new algorithm for creating an ontology for a specific area of scientific knowledge using information obtained from search engines on the Internet. The evaluation of the computational complexity of its implementation is mathematically substantiated.

Distinctive features of the developed algorithm: automatic selection of terms in the field of study; the possibility of using other areas of scientific knowledge without changing the algorithm for creating ontologies; does not require specialist manual labor. Also, an algorithm was developed for identifying pairs of terms from a set of texts on a given topic. Unlike other classical algorithms, the term pair separation algorithm has shown high efficiency when comparing texts using classification and clustering. It is mathematically proved that the estimation of the complexity of the calculation and the main function of the weight of the term in the heading correspond to the requirements for it.

The following results are recommended for the defense of the thesis.

- Based on the results of studying the subject area when describing the results of a scientific organization, the development of algorithms used when working with big data with a heterogeneous structure using ontologies;

- a formal description of system queries created using the SPARQL language, which provides the necessary information in the ontology;

- It has been proven that the weights of headings in headings that are part of the algorithms for creating ontologies of certain areas of scientific knowledge and for extracting pairs of terms from text collections meet the requirements of the basic function;

- Analytical assessment of the complexity of the developed software.

Theoretical and practical significance of the work: scientific novelty and practical significance of the research are high. The research results are used to combine heterogeneous data and use them for further processing.

Approbation of work and publication.

The main provisions and scientific results of the work were reported and discussed at domestic and foreign international scientific conferences: The 14th International Conference on Control, Automation and Systems, ICCAS 2014 (South Korea, Busan, 2014); The 15th International Conference on Control, Automation and Systems, ICCAS 2014 (South Korea, Guangzhou, 2015).

The dissertation work was discussed at scientific seminars organized by the Department of Computer Engineering of the International Information Technologies University and in scientific seminars organized by the Faculty of Computer Engineering at Gachon University (South Korea, Seoul).

The main results obtained during the implementation of the dissertation work were published in 12 printed publications, of which 5 articles were published in editions recommended by the KKSON MES RK, 7 articles were published in collections of international conferences (Kazakhstan, South Korea, Latvia) 1 article was published in editions, indexed by the Scopus database (percentile 36%).

The structure and scope of the thesis. The structure of the thesis consists of an introduction, four chapters, a conclusion, a bibliography and an appendix. The total volume of work is 105 pages, including 38 figures, 11 tables, 77 list of used literature, 3 appendices.

The introduction provides an overview of the subject area and highlights key issues in this area. The significance of the dissertation is substantiated, the goal and requirements are formulated.

The first section is devoted to the current state and place in the market of big data technologies. To obtain reliable indicators of the effectiveness and focus of a scientist's research activities, a tool is needed that allows individual structural units to collect and analyze the scientist's works.

Big data technology plays a special role in the management of scientific information. Analysis of information systems used to solve similar tasks and presented on the Internet, allowed to distinguish several group systems, most of which are bibliographic and abstract databases of data, in the science portal Google, in the privacy of the Web. They combine these or other degrees in their functions, such as indexing and research. Part of the system, for example, M.V. The ISTINA MSU system. M.V. Lomonosov, information and analytical system of the Astrakhan State University "Results of scientific activity", the system of PURE company Elsevier carry out monitoring of scientific activity and results of the organization. A comparison of the characteristics of the largest web services used to manage scientific information in the world is given in Table 1.

Table 1 – Methods and tools of scientific information management

Methods of scientific information management	Advantages	Disadvantages
Quantitative conclusions based on research report information	counts manually and writes down the numeric index	quality of work is not considered
Expert opinion of the materials of the conference and the journal	The expert conducts high-quality and meaningful work	Articles are published in different languages
Analysis of main articles	Reduced stress through annotation	Information not included in the annotation will be excluded
Keyword search	The electronic version has a large amount of text information	actual demand is not covered

Big data technology plays a special role in the management of scientific information. Analysis of information systems used to solve similar tasks and presented on the Internet, allowed to distinguish several group systems, most of which are bibliographic and abstract databases of data, in the science portal Google, in the privacy of the Web. They combine these or other degrees in their functions, such as indexing and research. Part of the system, for example, M.V. The ISTINA MSU system. M.V. Lomonosov, information and analytical system of the Astrakhan State University "Results of scientific activity", the system of PURE company Elsevier carry out monitoring of scientific activity and results of the organization. A comparison of the characteristics of the largest web services used to manage scientific information in the world is given in Table 2.

At the end of the first chapter, the main shortcomings of the currently known systems for processing and analyzing scientific data are listed, which could be considered as possible solutions to the main problem. These disadvantages include: the complexity of data entry; the complexity and low ability to search for information; the use of rigid and uninformative models of the field of knowledge, lack of flexibility of systems; focus on processing information from the Internet, and not on semi-automatic input by the user; insufficient attention to the intellectualization of algorithms for loading, processing and retrieving information.

Table 2 - Comparison of the characteristics of the main web services

	№	Title	Authorized body	Advantages	Disadvantages	Data format
Large web service	1	Web of Science	Thomson Scientic	Articles in the system since 1900	The request is executed only by the keyword	.TXT
	2	Scopus	Elsever	Covers the entire subject area	The request is executed only by the keyword	.TXT
	3	Google Scholar	Google	Articles that are accepted but not yet published are taken into account	There are substandard and fake scientific publications	.TXT
Foreign projects	1	Bibster	University of Karlsruhe, University of Amsterdam, Bank of Dresden	Outputs data from the system in RDF format	Information is loaded into the system via a structured file	BibTeX
	2	JeromeDL	Gdansk (Poland) University of Technology, DERI Institute for Digital Research (Ireland)	The system can classify and contain electronic information in a database	Information is entered into the system in a structured manner or manually. complex queries are not performed, information is entered manually	BibTeX, Marc21, Dublin Core
	3	Flink	University of Amsterdam	Defines the area of the interface of scientists based on a keyword	Collects manually the ontology of the required subject area	FOAF, SWRC
	4	AIR	University of Wolverhampton (UK) and Alicante (Spain)	The system collects information from web pages in the DC structure	No complex ontology modeling the subject area	Dublin Core
SDB		Semantic database	Open Sourse	Available to any software developer	Providing a logical connection	RDF(s), OWL, SPARQL
R	1	«ISTINA»	Russia Moscow	Available to	Providing a logical	RDF(s),

				any software developer	connection	OWL, SPARQL
	2	"Results of scientific activity" of Astrakhan University	Russia, Astrakhan	Available to any software developer	Providing a logical connection	RDF(s), OWL, SPARQL

The second section presents architectural and technological solutions used in the system of automatic control of scientific information.

Assume that D is the area of scientific knowledge (for example, computer science). Let I be a set of descriptions of the unit of scientific and technical information in this area of knowledge (atomic measurement). Such blocks relate to: scientific articles; patents; reports; reports read at conferences; statements of accounting; monographs; textbooks and others. author's works (abstracts, translations). Each element of the multiplicity I contains a text description of the corresponding object.

The main purpose of the system is the execution of the search-analytical request. Indicate the number of typical queries with the symbol Q. The task is expressed by the expression q∈Q $r1: Q \rightarrow 2^I$ with the characteristics of the block of scientific and technical information $I_q \subseteq I$.

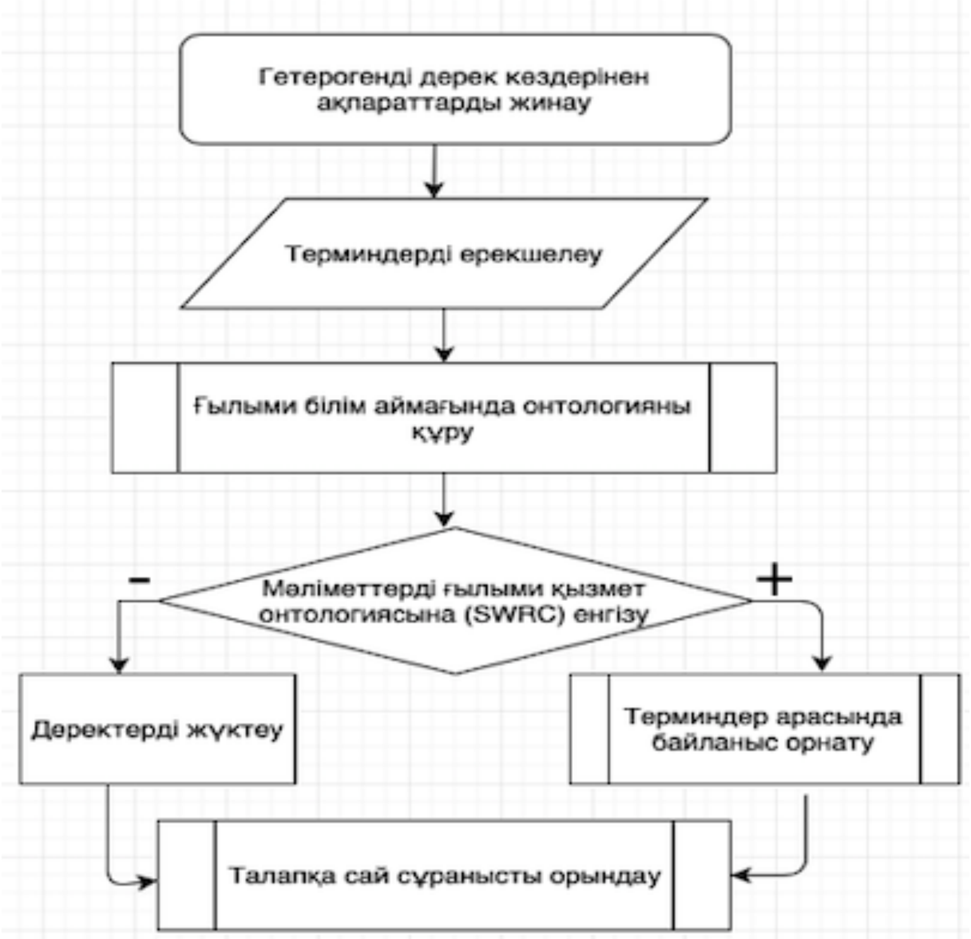

Гетерогенді дерек кездерінен ақпараттарды жинау
Терминдерді ерекшелеу
Ғылыми білім аймағында онтологияны құру
Мәліметтерді ғылыми қызмет онтологиясына (SWRC) енгізу
-
+
Деректерді жүктеу
Терминдер арасында байланыс орнату
Талапқа сай сұранысты орындау

Figure 1 - General architecture of a scientific information management system

The general scheme of the system is presented in Figure 1 and consists of the following models:

- to distinguish the terms describing the area of scientific knowledge D, from the text description of the scientific and technical conference, dedicated to the area of knowledge;
- D creation of the considered ontology in the field of scientific knowledge;
- download data on the results of the scientific database of employees;
- to establish a link between the instance, collected in the field of education, and the information downloaded from the results of scientific research;
- Execution of the analytical request on the received information;
- The general scheme consists of the following stages:
- D to distinguish terms describing the area of scientific knowledge (key word);
- Development of ontology areas of scientific knowledge D;
- Downloading of information differs depending on the area of science;
- to establish communication between the concept of the developed ontology and scientific conclusions of users;
- A sample that responds to queries contains a summary of the information received.

The next step is to describe each step. Was obtained with the help of semantic, in particular, linguistic and statistical methods for the separation of terms describing the field of knowledge. During the development of the algorithm for distinguishing terms, the following definitions were formed.

In accordance with the requirements for the system, the prototype developed by the author assumes the following three possible data entry methods:

- analysis of bibliographic references;
- parsing BibTeX records (BibTeX, MathML, LaTeX, FinXML);
- fill in the fields manually.

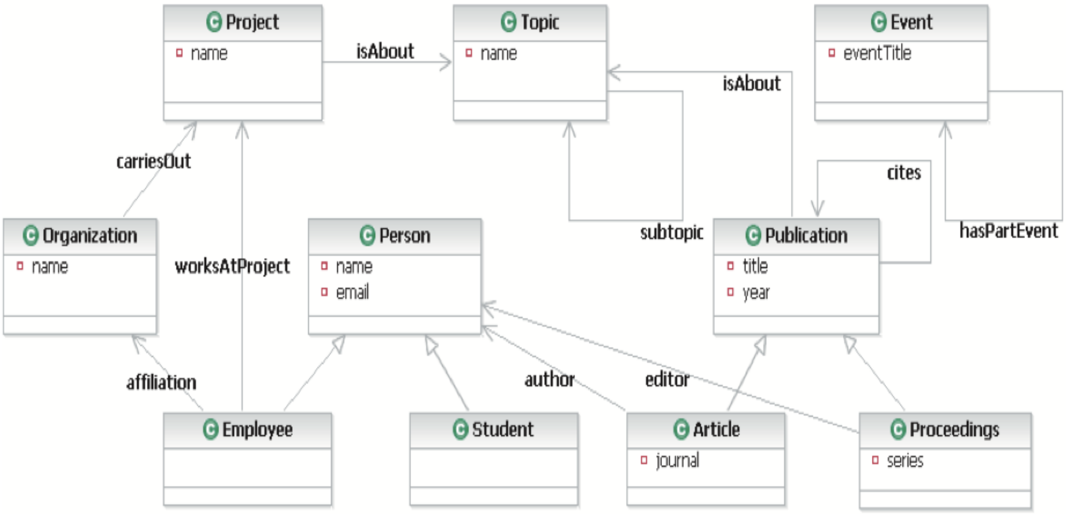
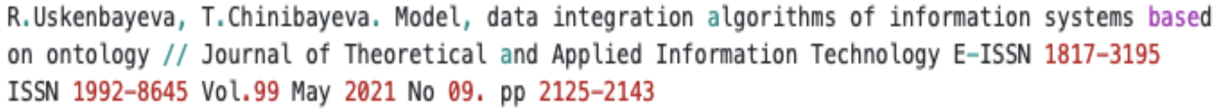
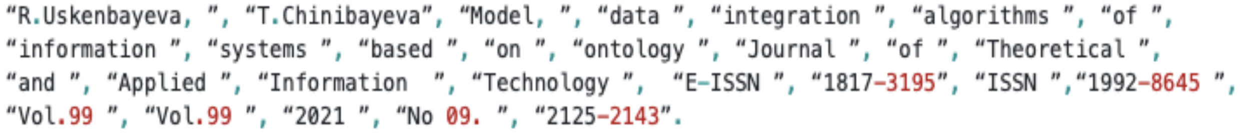


Figure 2 - Ontology SWRC (*fragment)*

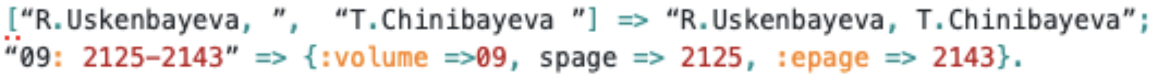
Rules of bibliographic links. Obtaining information from bibliographic links is the task of obtaining information from unstructured text. Algorithm of conditional field fields (CRF), which showed the greatest influence on the results of testing the method of adjustment of referenced bibliographic links. The FreeCite software package, developed at Brown University in the United States, was used in the CRF ++ library, which implements this algorithm.



```
R.Uskenbayeva, T.Chinibayeva. Model, data integration algorithms of information systems based
on ontology // Journal of Theoretical and Applied Information Technology E-ISSN 1817-3195
ISSN 1992-8645 Vol.99 May 2021 No 09. pp 2125-2143
```



```
“R.Uskenbayeva, ”, “T.Chinibayeva”, “Model, ”, “data ”, “integration ”, “algorithms ”, “of ”,
“information ”, “systems ”, “based ”, “on ”, “ontology ”, “Journal ”, “of ”, “Theoretical ”,
“and ”, “Applied ”, “Information  ”, “Technology ”,  “E-ISSN ”, “1817-3195”, “ISSN ”,“1992-8645 ”,
“Vol.99 ”, “Vol.99 ”, “2021 ”, “No 09. ”, “2125-2143”.
```

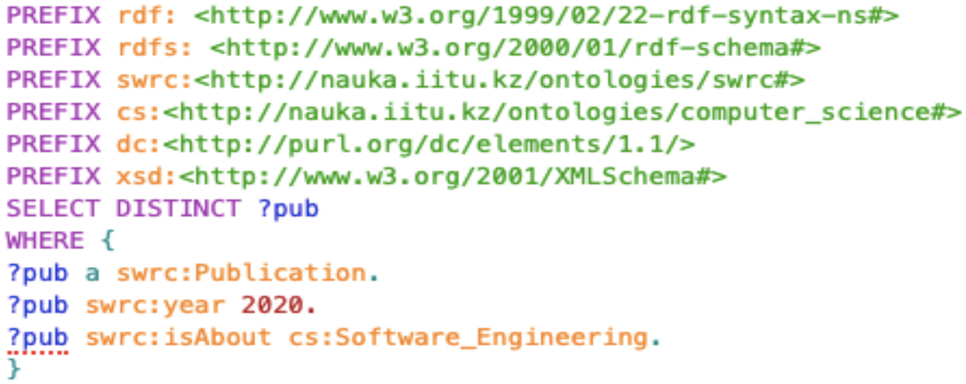


```
[“R.Uskenbayeva, ”,  “T.Chinibayeva ”] => “R.Uskenbayeva, T.Chinibayeva”;
“09: 2125-2143” => {:volume =>09, spage => 2125, :epage => 2143}.
```

Establishment of a relationship between the existing model in the field of education and information obtained from the downloaded texts, consisting of the results of scientific work of students, is necessary for the implementation of analytical requirements. From the document used up to this level, only the amount of information about the scientific work of the employee differs.

Ontological action, related to knowledge, allows the use of current and past approbation of algorithms performing analytical queries. In particular, rewriting a query using an ontology can be performed automatically using a logical output mechanism.

Here is an example of a query that allows you to get the publication 2020 for the development of software security ("Software Engineering") to demonstrate the syntax of the language SPARQL.



```
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX swrc:<http://nauka.iitu.kz/ontologies/swrc#>
PREFIX cs:<http://nauka.iitu.kz/ontologies/computer_science#>
PREFIX dc:<http://purl.org/dc/elements/1.1/>
PREFIX xsd:<http://www.w3.org/2001/XMLSchema#>
SELECT DISTINCT ?pub
WHERE {
?pub a swrc:Publication.
?pub swrc:year 2020.
?pub swrc:isAbout cs:Software_Engineering.
}
```

Description of the algorithm for selecting terms from a set of texts with given thematic sections. Suppose that W - ε is the majority of all words found in all documents, including the empty word Doc, and PW is a pair of all ordered words, that is PW = W * W. The document d represents d: N → W, for each natural number of n words in n-m direction in this set of documents. Figures without words (after the end of the document). Correspondingly, the new line p is assigned in the specified paragraph as p: N → W, which corresponds to each positive whole number of words n in n-th position. The number of the place where the word is not written is marked as an empty word. All new lines in the set are marked with the letter P. r only the documents that form the title, and more precisely - r∈2 ^ Doc. The capacity of the title is the same as the size of the documents in it. will be determined. Let's define the majority of data titles R.

We also highlight a number of additional functions:

- $\tau_1: PW \to W, \tau_2: PW \to W$ – a pair for many other words, in which a pair of words corresponds to the first word (corresponds to the second);
- $Freq: PW * Doc \to \mathbb{N} \cup \{0\}$- d ∈Doc a function that determines the number of pairs pϖ∈PW entered into the document;
- $Freq: W * Doc \to \mathbb{N} \cup \{0\}$ - $d \in$ introducing a pair w∈W into a document a function that determines a number
- $L(d) = |\{n \in \mathbb{N} | d(n) \neq \varepsilon\}|$ - d - document length;

- $id(a) = a$ – similar image;

$Av(f, A) = \frac{\sum_{a \in A} f(a)}{|A|}$ - A - the average value of the function f in the last.

For example, Av (|·|, R) is the average number of documents in the header, $Av(L, Doc)$ is the average length of the document, $Av(id, A)$ is the arithmetic average of the set $A = \{a_1, \dots, a_k\}$.

The algorithm consists of four stages. In each of them, using some rules, the set M_i and M_{i-1} obtained in the previous step is selected. At the first stage, a selection is made from a set of PWs (all pairs of words), i.e. $M_0 = PW$. The M_4 set is a pair of terms that satisfies all four dimensions.

An algorithm for constructing an ontology of the field of scientific knowledge based on a collection of scientific conference announcements, divided into headings, as well as information from search engines on the Internet. Conference announcements, called call for papers (CFP), are used as the main data source for constructing an ontology. This approach has important advantages. In particular, it allows one to obtain sufficiently reliable, relevant and complete information about the field of scientific knowledge.

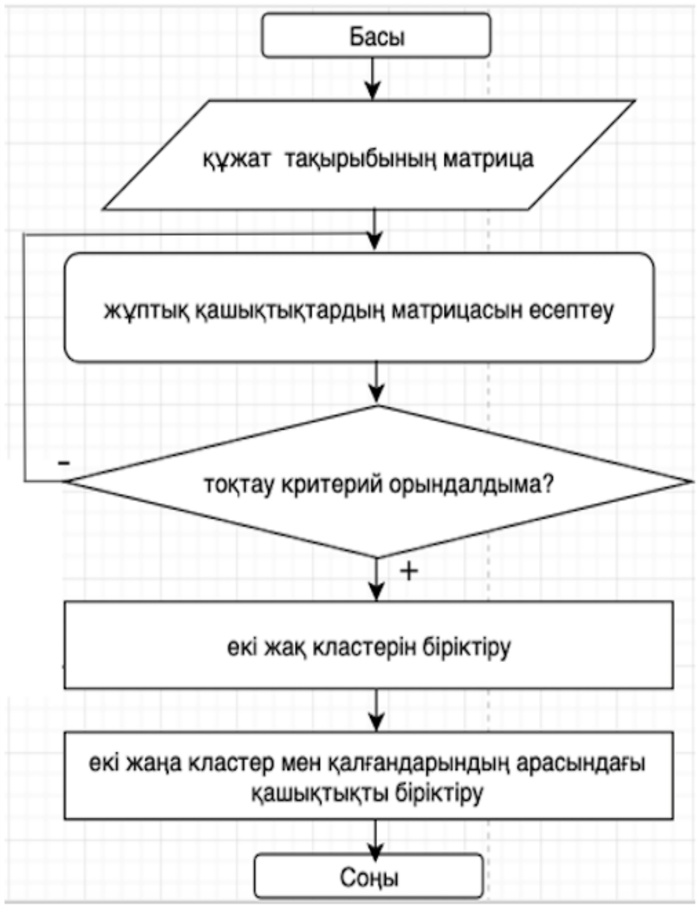

Басы
құжат тақырыбының матрица
жұптық қашықтықтардың матрицасын есептеу
-
тоқтау критерий орындалдыма?
+
екі жақ кластерін біріктіру
екі жаңа кластер мен қалғандарындың арасындағы қашықтықты біріктіру
Соңы

Figure 3 - Algorithm for constructing an ontology in the field of scientific knowledge

Google Normalized Distance (NGD) is a generic term used to define the level of semantic similarity between two terms. Let A and B be terms and N be the total number of pages indexed by the search engine. Then the level of semantic similarity NGD between A and B is determined by the following formula:

$$NGD(A,B) = \frac{\max\{\log hits(A), \log hits(B)\} - \log hits("A\ AND\ B")}{\log N - \min\{\log hits(A), \log(B)\}}$$

The next step in the algorithm is to build a hierarchy of terms. The classical algorithm for constructing a hierarchy of concepts using linguistic templates, developed by Hirst, turns out to be ineffective for building a hierarchy of scientific directions. Within the framework of this work, linguistic templates have been developed specifically to solve this problem. The main template looks like

$$A\ is * \ keyword * prep(aux)?\, B$$

Conclusion. The thesis describes methods and techniques for building a scientific information management system. The theoretical basis for action is ontology. The version proposed by the author of the system includes such templates as query execution, ontology and results of scientific work of scientists, a template for loading information, a template that forms a formal model in the field of education.

Approbation of work:

1. R. Uskenbayeva, Y. Chinibayev, A. Kassymova, T. Temirbolatova, K. Mukhanov. Technology of integration of diverse databases on the example of medical records//Proceedings of the 14th International Conference on Control, Automation and Systems (ICCAS 2014) - Gyeonggi -do, Korea, 2014. P 282-285. ISSN: 2093- 7121.
2. R.Uskenbayeva, T.Temirbolatova, Young Im Cho, Z.Uskenbayeva, G.Bektemyssova, A. Kassymova. Recursive decomposition as a method for integrating heterogeneous data sources//Proceedings of the 15th International Conference on Control, Automation and Systems (ICCAS 2015). – Busan, South Korea. October 13-16, 2015 – P.2076-2079. ISSN: 2093 - 7121
3. Р.К. Ускенбаев, Т.Т.Темірболатова, А.Б. Касымова. Бұлттық есептеуде mapreduce технологиясымен үлкен деректерді өңдеу - // Вестник КазНТУ имени К.Сатпаева No5 (111). – 2015. С.50 - 53. ISSN 1680- 9211
4. Р.Ускенбаева, Г. Бектемысова, Т.Темірболатова. Интеграция больших неоднородных данных с использованием языка R и HADOOP - Вестник КазАТК - №4 2015-11-01
5. Ускенбаева Р.К., Аманжолова С.Т., Темірболатова Т.Т. Анализ и локализация инцидентов снижения работоспособности распределенных вычислительных систем. Труды международного форума «инженерное образование и наука в XXI веке: проблемы и перспективы», посвященного 80-летию Каз НТУ им. К.И. Сатпаева
6. T. Temirbolatova, D. Beisenov Automatic asynchronous exchange of business object between heterogeneous systems - The 12th ICIT&M 2014. 2014 April 16-17, 2014, Information Systems Management Institute, Riga, Latvia
7. T. Temirbolatova, A.Khamitov, A. Keldybay, T.Sembayeva Manage different-structured Big Data - The 12th ICIT&M 2014. 2014 April 16-17, 2014, Information Systems Management Institute, Riga, Latvia
8. Temirbolatova T. Jarmukhambetov Y., Temirbolatova U. The method of extracting semantic meta descriptions from databases//2nd International scientific conference «Information Technologies in Science &Industry» International IT University, May 19, 2016 Almaty, Kazakhstan. ISBN 978-601-7407-33-9
9. T. Chinibayeva Security semantic database problems // Herald of the Kazakh-british technical university ISSN1998-6688. V INTERNATIONAL CONFERENCE "DIGITAL TECHNOLOGY IN SCIENCE AND INDUSTRY - 2019» (DTSI-2019), 10th Anniversary INFORMATION TECHNOLOGY INTERNATIONAL UNIVERSITY Vol.16, No.3 (2019), pp. 168-174
10. R.Uskenbayeva, T.Chinibayeva. Algorithm for the construction of an ontology in the field of scientific knowledge//The Bulletin of Kazakh Academy of Transport and Communications named after M. Tynyshpayev ISSN 1609-1817. Vol. 107, No.4 (2018), pp. 259-266
11. R.Uskenbayeva, T.Chinibayeva. Method of extracting meta description from databases//Herald of the Kazakh-british technical university ISSN1998-6688. Vol.15, No.4 (2018), pp. 116-123
12. R.Uskenbayeva, T.Chinibayeva. Model, data integration algorithms of information systems based on ontology // Journal of Theoretical and Applied

Information Technology E-ISSN 1817-3195 ISSN 1992-8645 Vol.99 May 2021 No 09. pp 2125-2143