ISSN 2518-170X (Online),
ISSN 2224-5278 (Print)

ҚАЗАҚСТАН РЕСПУБЛИКАСЫ
ҰЛТТЫҚ ҒЫЛЫМ АКАДЕМИЯСЫНЫҢ
Қ. И. Сәтпаев атындағы Қазақ ұлттық техникалық зерттеу университеті

Х А Б А Р Л А Р Ы

ИЗВЕСТИЯ

НАЦИОНАЛЬНОЙ АКАДЕМИИ НАУК
РЕСПУБЛИКИ КАЗАХСТАН
Казахский национальный исследовательский
технический университет им. К. И. Сатпаева

N E W S

OF THE ACADEMY OF SCIENCES
OF THE REPUBLIC OF KAZAKHSTAN
Kazakh national research technical university
named after K. I. Satpayev

SERIES
OF GEOLOGY AND TECHNICAL SCIENCES

3 (435)

MAY – JUNE 2019

THE JOURNAL WAS FOUNDED IN 1940

PUBLISHED 6 TIMES A YEAR

ALMATY, NAS RK

NAS RK is pleased to announce that News of NAS RK. Series of geology and technical sciences scientific journal has been accepted for indexing in the Emerging Sources Citation Index, a new edition of Web of Science. Content in this index is under consideration by Clarivate Analytics to be accepted in the Science Citation Index Expanded, the Social Sciences Citation Index, and the Arts & Humanities Citation Index. The quality and depth of content Web of Science offers to researchers, authors, publishers, and institutions sets it apart from other research databases. The inclusion of News of NAS RK. Series of geology and technical sciences in the Emerging Sources Citation Index demonstrates our dedication to providing the most relevant and influential content of geology and engineering sciences to our community.

Қазақстан Республикасы Ұлттық ғылым академиясы "ҚР ҰҒА Хабарлары. Геология және техникалық ғылымдар сериясы" ғылыми журналының Web of Science-тің жаңаланған нұсқасы Emerging Sources Citation Index-те индекстелуге қабылданғанын хабарлайды. Бұл индекстелу барысында Clarivate Analytics компаниясы журналды одан әрі the Science Citation Index Expanded, the Social Sciences Citation Index және the Arts & Humanities Citation Index-ке қабылдау мәселесін қарастыруда. Webof Science зерттеушілер, авторлар, баспашылар мен мекемелерге контент тереңдігі мен сапасын ұсынады. ҚР ҰҒА Хабарлары. Геология және техникалық ғылымдар сериясы Emerging Sources Citation Index-ке енуі біздің қоғамдастық үшін ең өзекті және беделді геология және техникалық ғылымдар бойынша контентке адалдығымызды білдіреді.

НАН РК сообщает, что научный журнал «Известия НАН РК. Серия геологии и технических наук» был принят для индексирования в Emerging Sources Citation Index, обновленной версии Web of Science. Содержание в этом индексировании находится в стадии рассмотрения компанией Clarivate Analytics для дальнейшего принятия журнала в the Science Citation Index Expanded, the Social Sciences Citation Index и the Arts & Humanities Citation Index. Web of Science предлагает качество и глубину контента для исследователей, авторов, издателей и учреждений. Включение Известия НАН РК. Серия геологии и технических наук в Emerging Sources Citation Index демонстрирует нашу приверженность к наиболее актуальному и влиятельному контенту по геологии и техническим наукам для нашего сообщества.

«ҚР ҰҒА Хабарлары. Геология мен техникалық ғылымдар сериясы».
ISSN 2518-170X (Online),
ISSN 2224-5278 (Print)
Меншіктенуші: «Қазақстан Республикасының Ұлттық ғылым академиясы» РҚБ (Алматы қ.).
Қазақстан республикасының Мәдениет пен ақпарат министрлігінің Ақпарат және мұрағат комитетінде 30.04.2010 ж. берілген **№10892-Ж** мерзімдік басылым тіркеуіне қойылу туралы куәлік.

Мерзімділігі: жылына 6 рет.
Тиражы: 300 дана.

Редакцияның мекенжайы: 050010, Алматы қ., Шевченко көш., 28, 219 бөл., 220, тел.: 272-13-19, 272-13-18,
http://www.geolog-technical.kz/index.php/en/

Редакцияның мекенжайы: Қазақстан, 050010, Алматы қ., Қабанбай батыра көш., 69а.
Қ. И. Сәтбаев атындағы геология ғылымдар институты, 334 бөлме. Тел.: 291-59-38.

Типографияның мекенжайы: «Аруна» ЖК, Алматы қ., Муратбаева көш., 75.

«Известия НАН РК. Серия геологии и технических наук».
ISSN 2518-170X (Online),
ISSN 2224-5278 (Print)
Собственник: Республиканское общественное объединение «Национальная академия наук Республики Казахстан (г. Алматы)
Свидетельство о постановке на учет периодического печатного издания в Комитете информации и архивов Министерства культуры и информации Республики Казахстан **№10892-Ж,** выданное 30.04.2010 г.

Периодичность: 6 раз в год
Тираж: 300 экземпляров

Адрес редакции: 050010, г. Алматы, ул. Шевченко, 28, ком. 219, 220, тел.: 272-13-19, 272-13-18,
http://nauka-nanrk.kz /geology-technical.kz

Адрес редакции: Казахстан, 050010, г. Алматы, ул. Кабанбай батыра, 69а.
Институт геологических наук им. К. И. Сатпаева, комната 334. Тел.: 291-59-38.

Адрес типографии: ИП «Аруна», г. Алматы, ул. Муратбаева, 75

News of the National Academy of Sciences of the Republic of Kazakhstan. Series of geology and technology sciences.
ISSN 2518-170X (Online),
ISSN 2224-5278 (Print)
Owner: RPA "National Academy of Sciences of the Republic of Kazakhstan" (Almaty)
The certificate of registration of a periodic printed publication in the Committee of information and archives of the Ministry of culture and information of the Republic of Kazakhstan N 10892-Ж, issued 30.04.2010

Periodicity: 6 times a year
Circulation: 300 copies

Editorial address: 28, Shevchenko str., of. 219, 220, Almaty, 050010, tel. 272-13-19, 272-13-18,
http://nauka-nanrk.kz/geology-technical.kz

Editorial address: Institute of Geological Sciences named after K.I. Satpayev
69a, Kabanbai batyr str., of. 334, Almaty, 050010, Kazakhstan, tel.: 291-59-38.

Address of printing house: ST "Aruna", 75, Muratbayev str, Almaty

N E W S
OF THE NATIONAL ACADEMY OF SCIENCES OF THE REPUBLIC OF KAZAKHSTAN
SERIES OF GEOLOGY AND TECHNICAL SCIENCES
ISSN 2224-5278
Volume 3, Number 435 (2019), 240 – 246 **https://doi.org/10.32014/2019.2518-170X.91**

UDC 615.035.4

B. Sinchev[1], A.B. Sinchev[2], J. Akzhanova[3], A.M. Mukhanova[4]

[1]International University of Information Technologies, Almaty, Kazakhstan,
[2]JSC "National Information Technologies", Astana, Kazakhstan,
[3]Astana LRT, Astana, Kazakhstan,
[4]Almaty Technological University, Almaty, Kazakhstan.
E-mail: sinchev@mail.ru, askar.sinchev@gmail.com, zyekudayeva@gmail.com, nuraksulu72@mail.ru

NEW METHODS OF INFORMATION SEARCH. I

Abstract. The paper discusses new methods used to solve the problem of information retrieval of unstructured (text) data. The search for documents is carried out by keywords in natural language used in search engines. The proposed search methods are fundamentally different from the existing methods in time and memory used, as well as in the simplicity of implementing software products based on the developed algorithms. The theorems of sampling a subset satisfying the sum (certificate) *S*, the sum of subset problems, lemmas and algorithms for solving the problem of searching for unstructured data based on a search query with several (two or three) keywords are given. The time and memory required for a search query with two keywords are proportional to $O(n)$. The task of information retrieval with three keywords is reduced to the task of searching for information with two keywords or to the problem of computational geometry. These scientific results are fully based on the materials cited in the USPTO USA, filed on 17.12. 2018 year.

Key words: search, method, algorithm, unstructured information.

Introduction. One and the first fundamental reviews of the tasks of information retrieval and search engines was presented in [1]. To perform a search, many search engines build on the basis of the initial information logical and physical data structures, which are a search index that allows you to implement some given information retrieval model. We define the main types of search queries. It is enough to list them: boolean search, search for relevance, search by pattern (mask), etc. There is no need to decipher each type of search, as it is given in [2]. Currently, several methods of sequential and binary searches are used. The systems that use these keyword search methods include the most common search engines, including such Web search systems as Yandex, Google, AstaVista, Yahoo, etc.

Basic concepts of information retrieval. The task of information retrieval: to find one or more elements in the set, and the desired elements must have a certain property. This property can be absolute or relative. Relative property characterizes in relation to others: for example, the minimum element in the set of numbers.

Definition 1. Let us call the alphabet, a finite set of symbols $A = \{\tau, \alpha_1, \dots, \alpha_k\}$, where τ is a space character, | A | = k, and k> 0 is the number of characters in the alphabet.

Definition 2. A word is a finite sequence of characters from the alphabet, not including the whitespace character τ. We assume that the set of words *W* is always finite.

*Definition*3. A string of length n, we call the sequence of words $D = \{d_1, \dots, d_n, \$\}$, where ∀i, $d_i \in$ W and *$* is a special character that does not belong to the alphabet and denotes the end of the line.

Definition 4. The search query $P = \{p_1, \dots, p_m\}$ is the string consisting of a finite set of words separated by a space character τ. In this case, $| P | = m$ is the length of the query in words, $p_i \in W$, where *i* is the number of the word in the pattern, and *W* is the set of all words. Words in search queries and documents will be called terms (keywords).

Comment. The length of the search query P will always be denoted by the symbol m, and the total length of the source data D, for which the search problem will be solved, is denoted by n.

The main provisions of tabular search methods. Mathematical substantiation of new methods and algorithms for searching unstructured text information by keywords is carried out using a vector-spatial model (each element of the search digital index acts as a coordinate of the vector space), excluding the probabilistic and boolean models.

Thus, the task of information retrieval is reduced to the sum of the subset belonging to the NP class to the complete problem, and the proposed approach allows us to apply the existing criteria for the relevance of documents in vector spaces.

Subset sum problem. Given a set of n numbers and a number S. It is required to determine whether there exists at least one subset whose sum of elements is equal to S.

Currently, information retrieval tasks have been solved for two, three, and four keywords based on tabular m-sums (table m-sum problem, mxn dimension table is specified).

There is a table 2xn and a given number S. You need to find 2 numbers from different lines (one from each line), giving a total of S.

Algorithm A. Brute force. The running time is $O(n^2)$.

Algorithm B. Brute force sorting. Sort the first line, for each element from the second line subtract it from S and look for this difference in the first line. Runtime -O ($n$log$n$). Memory requirement O (n).

There is a table 3xn and a given number S. You need to find 3 numbers from different lines, giving a total of S.

Algorithm A. Brute force. The running time is $O(n^3)$.

Algorithm B. For a single line, find all the differences from S, and for the other two, go through all the options. The running time is $O(n^2)$.

In the future, this material will allow a comparative analysis with the following proposed scientific results.

Open problems finding unstructured textual information. The works [3, 4] are devoted to solving the problem (subset sum problem), in which the search time $T=O(2^{n/2})$ and the required memory $M=O(2^{n/4})$ do not allow the results to be applied in practice. The main disadvantage of tabular methods is the construction of each row of the table by property defined by each keyword. This means that we are obliged to carry out preliminary work on some structuring of input data. In turn, there is an additional problem of splitting a vector space into subspaces according to each keyword.

Therefore, we will change the problem definitions of tabular sums in a more generalized form suitable for the practical search of any information, using the sum of the subset problem. In the future, algorithms for solving these problems can be directly applied to the search for arbitrary unstructured information based on a vector-spatial model and a searchable digital index.

Let us reformulate the formulation of the main problem directly related to the length m of the search query.

The main practical task. Given a set of n numbers and the number S. It is required to find out if there is one or several subsets, each of which consists of m elements, and the sum of these elements is equal to S.

1. A set of n numbers and a number S are given. It is required to find out if there is one or several subsets of two numbers, the sum of whose elements is S and with a running time shorter than $O(nlogn)$.

2. A set of n numbers and a number S are given. It is required to find out if there is one or several subsets of three numbers, the sum of whose elements is S and with a running time shorter than $O(n^2)$.

Now we can move on to the mathematical formulations of the search problems and their solutions.

The main practical task. Given a set of integer (natural) numbers

$(x_1, x_2, \ldots, x_n) \in X^n$ of dimension n. It is required to find out whether there exists a subset X_m of dimension m such that the following conditions are fulfilled:

$$X_m = \{x_i + x_j + \cdots + x_g + x_h = S, i \neq j \neq \cdots \neq g \neq h, x_i, x_j, \cdots, x_g, x_h \in X^n,$$

$$(i, j, \cdots, g, h) \in N = (1,2, \ldots, n), \ \ m \leq n\} \quad (1)$$

Here $x_i, x_j, \cdots, x_g, x_h \in X_m$ with the number of elements $x_i, x_j, \cdots, x_g, x_h$ equals m.

We introduce the following notation: C_n^m -complex, S_n^m-sum of elements of one subset from the set of subsets X_m of the set X^n. In this case, the variable *m* can vary from 0, 1, 2, ..., n. The set of these subsets X_m is determined on the basis of the combination

$$C_n^m = \frac{n!}{m!(n-m)!}. \quad (2)$$

We sort the given vector *x* from the set X^n in descending order and get the sorted set Z^n - the set of vectors *x* whose values of the elements are sorted in descending order, and find the values

$$S_{min}^m = \sum_0^m z_{n-m}, \quad (3)$$

$$S_{max}^m = \sum_0^m z_m . \quad (4)$$

It should be noted that

$$S_{min}^0 = S_{max}^0 = 0, S_{min}^n = S_{max}^n = \sum_1^n x_i = \sum_1^n z_i . \quad (5)$$

We compose the possible ranges of the certificate *S* belonging to a subset of the set of subsets X_m, $S \in [S_{min}^m, S_{max}^m]$. (6)

The solution of the problem of the sum of subsets is based on the following theorems.

Theorem 1. Let certificate S belong to the range $[S_{min}^m, S_{max}^m]$. Then there is a subset X_m,whose sum of elements is equal to *S*.

Proof. The fulfillment of the condition of the theorem (or condition (6)) means that it is necessary to generate all the subsets X_m based on the formula (2) of the set X^n, the sum of the elements of each of them changes from the minimum value S_{min}^m to the maximum value S_{max}^m. This is equivalent to generating all *n*-dimensional vectors from zeros and ones ($e \in E^n$). The above condition allows the enumeration of subsets in order of minimal change in the binary code of the vector *e*. If the *i*-th index of the vector *e* is 1 (one), this means that this element is included in this subset and must be taken into account when calculating the sum of the elements. The definition of *m* indices on which there are units uniquely determines the vector *e* corresponding to one subset of the set of subsets X_m. Then there is a vector *e* such that the sum *S* is calculated on the basis of the scalar product: *S* = (e, x).

Remark1. The dimension of the set X_m easily extends to n if other elements of this set are considered zeros, except for elements with indices $(i, j, \cdots, g, h) \in N$. On the other hand, we can use one of the properties of the combination (2) to reduce the parameter n: $C_n^m = C_{n-1}^{m-1} + C_{n-1}^{m-1}$.

The result obtained can be extended to a set of subsets $X_{n-m}$from the set X^n and introduce the ranges $[S_{min}^{n-m}, S_{max}^{n-m}]$.

Theorem 2. Let certificate *S* belong to the range $[S_{min}^{n-m}, S_{max}^{n-m}]$. Then there exists a subset X_{n-m} whose sum of elements is equal to *S*.

Proof. Based on the equality of combinations $C_n^m = C_n^{n-m}$, we can replace the variable m with the variable *n-m* and the vector *e* from theorem 1 with the vector $\bar{e}$, in which the zeros of the vector *e* are replaced by ones and the ones with zeros. Then there is a vector $\bar{e}$ such that the certificate is calculated on the basis of the scalar product: $S = (\bar{e}, x)$ or $S = S_{min}^n - (e, x)$.

It is easy to find the running time of the algorithm based on theorem 1 using the sorted vector x and the merge method:

$$T = O(C_n^m) < O(2^n). \quad (7)$$

The required memory $M = O(n)$ is necessary to save the vector *e*. Generation of vectors *e* can be made on the basis of the Gray code.

Example 1 of [4]. Consider a vector *x* = (7,3,9,6,2), *S*=11, $C_5^2 = C_5^3 = 10$, $S \in [S_{min}^2, S_{max}^2] = [5,16]$ or $S \in [S_{min}^3, S_{max}^3] = [11,22]$. Then the solutions of the problem of the sum of subsets are the vectors $e = (00101)$, $\bar{e} = (01011)$ and sum $S = (e, x) = 11$ or $S = (\bar{e}, x) = 11$, at $S = 10$, $e = (11000)$.

We turn to solving practical problems.

Task1. It is required to find out if a subset exists

$$X_2 = \{x_i + x_j = S;\ i \neq j;\ x_i, x_j \in X^n;\ i,j \in N\} \quad (8)$$

where $X^n = (x_1, x_2, \dots, x_n)$ is the set of integer (or natural) numbers, N = (1,2, ..., n) is the set of natural numbers.

Task2. It is required to find out if a subset exists

$$X_3=\{x_i+x_j+x_k=S;\ i \neq j \neq k;\ x_i, x_j, x_k \in X^n\ ;\ i,j,k \in N\}. \quad (9)$$

To solve these problems, we introduce the mapping of the set X^n into the set Y^n:

$$y = \tau(S,x) = (S-x)x,\ \forall x \in X^n. \quad (10)$$

Based on the mapping (10), we have that

$$Y^n = \{y_1, y_2, \dots, y_n \leftrightarrow\ \tau(S,x_i) = y_i, x_i \in X^n, i = 1,2, \dots, n\}. \quad (11)$$

Suppose that among the set Y^n there exist elements such that the identity holds:

$$y_i = y_j,\ i \neq j; i,j \in N. \quad (12)$$

Certificate S allows you to find a set of subsets $X_2 = \{x_i, x_j\}$, consisting of pairs of elements of the original set X^n, based on formulas (3) and (4).

Lemma 1. Let certificate S belong to the range $[S^2_{min},\ S^2_{max}]$ and the identity (12) holds for set (11). Then problem1 is solvable.

Proof. The first condition shows the existence of a subset X_2 from theorem1 satisfying the certificate S. To construct vectors e from identity (12), we have $y_i = \tau(S,x_i) = (S-x_i)x_i = x_j x_i$, assuming that $x_j = S - x_i$. On the other hand, $y_j = \tau(S,x_j) = (S-x_j)x_j = x_i x_j$, likewise assuming that $x_i = S - x_j$. In fact, the quantities x_i, x_j are the roots of the quadratic equation $x^2 - Sx + c = 0$. According to the Viet's theorem $c = x_i x_j$. Thus, we get $y_i = y_j = x_i x_j$. The latter means that the fulfillment of identity (12). Then there are elements x_i and x_j such that $x_i+x_j=S$.

We introduce the value

$$S(x_k) = S - x_k,\ \forall x_k \in X^n. \quad (13)$$

Lemma 2a. Let certificate S belong to the range $[S^3_{min},\ S^3_{max}]$ and for some element $x_k \in X^n$ and taking into account formula (13), identity (12) holds for $i \neq j \neq k$; $i,j,k \in N$. Then problem 2 is solvable.

Proof. The first condition shows the existence of a subset X_3 from theorem1 satisfying the certificate S. To construct vectors e from identity (12) with formula (10), we have $\tau(S(x_k),x_i) = (S(x_k)-x_i)x_i = = x_j x_i$, assuming that $S(x_k) - x_i = x_j$. On the other hand, $\tau(S(x_k),x_j) = (S(x_k)-x_j)x_j = x_i x_j$, likewise assuming that $S(x_k) - x_j = x_i$. The latter means that the conditions of Lemma 1 are satisfied when formula (13) is taken into account. Then we have that $x_i+x_j+x_k=S$.

The following lemma is based on computational geometry and is of independent scientific interest. The well-known fact that problem2 was reduced to an equivalent problem of the belonging of three points of one straight line on a plane. However, problem 2 in this formulation has not been solved to date.

In this case, the mapping (10) for the search query with three keywords will be rewritten as:

$$y = \tau(S,x) = (S-x)xx,\ \forall x \in X^n \quad (14)$$

and enter a 3x3 matrix

$$H = \begin{pmatrix} x_i & y_i & 1 \\ x_j & y_j & 1 \\ x_k & y_k & 1 \end{pmatrix}. \quad (15)$$

The coordinates (x_k, y_k) on the plane are calculated by the formulas

$$x_k = (S-(x_i+x_j)), y_k = \left(S-(x_i+x_j)\right)^2 (x_i+x_j),\ x_i, x_j \in X^n. \quad (16)$$

Lemma 2b. Let the certificate S belong to the range $[S^3_{min},\ S^3_{max}]$ and the determinant Δ of the matrix (15) is zero when considering formulas (16) and some element x_k, defined by the first formula of expression (16), belongs to the set $X^n,\ i \neq j \neq k,\ i,j,k \in N$. Then problem 2 is solvable.

Proof. The first condition shows the existence of a subset X_3 of theorem1 satisfying the certificate *S*. The second condition ensures the construction of vectors *e* based on the application of the well-known result [5] of the belonging of three points $(x_i, y_i), (x_j, y_j), (x_k, y_k)$ of one line lying on the plane (*x*, *y*). Replacing the third coordinate (x_k, y_k) with variables (16) completes the proof of the lemma.

Algorithms search. On the basis of these lemmas, we formulate the search algorithms.

Search algorithm1 for the first task.

Step 1. Input of the initial data: setX^n, *n*, *S*.

Step 2. Formation of the set Y^n on the basis of the map (4).

Step 3. Verification of the identity (5) and the formation of a subset

X_2={ $\tau(S, x_i) - \tau(S, x_j) = 0: i \neq j; x_i, x_j \in X^n$; $i,j \in N$} }.

Step 4. Subset output X_2.

The time of the search algorithm $T=O(n)$, the required memory $M = O(n)$ for the formation of the set Y^n.

For the tabular 2-sum, the exhaustive search time is $T=O(n^2)$ and in the case of sorting $T=O(nlogn)$.

Note2. The maximum number of pairs in the set Y^n is $m=[n/2]$. Thus, the number of pairs in Y^n can vary from 1 to *n/2*.

It should be noted that this algorithm allows you to find all the subsets of X_2 with a small modification.

Example 2. Given a set X^7={2,1,6,4,3,5,3} dimension *n*=7. It is required to find out whether there exists a subset X_2 ={x_i, x_j}, the sum of these elements from the set X^7 is equal to *S* = 6. Here $S \in [S^2_{min}, S^2_{max}] = [3,11]$. Initially, on the basis of the mapping $\tau(S, x)$ (map (8)) we translate the set X^7 into the set Y^7={8,5,0,8,9,5,9}. Further, to find the subset X_2, we use the identity (12): $y_i=y_j$, $y_i, y_j \in Y^7$, $i,j \in N$=(1,2,3,4,5,6,7). Then we get X_2=(2,4), X_2=(1,5), X_2=(3,3).

The search algorithm 2a for the second task.

Step 1. Input of the initial data: set X^n, *n*, *S*.

Step2. Calculation of $S(x_k) = S - x_k$ for some element $x_k \in X^n$.

Step3. Formation of the set Y^non the basis of the map (4) with regard to $S(x_k)$.

Step4. The formation of the subset X_3={ $\tau(S(x_k), x_i) - \tau(S(x_k), x_j) = 0$ for $i \neq j \neq k$; $x_i, x_j, x_k \in X^n$; $i, j, k \in N$}.

Step5. Output subsets of X_3.

Remark 2. Given in the search algorithm1 allows us to determine the search time $T=O((n-2m)2m)$. Here, *m* is the number of pairs in the set Y^n, *n*-2*m* is the number of remaining indices without taking into account the used index *k*. It is easy to show that as $m \to \left[\frac{n-1}{2}\right]$ tends, the search time varies with in $O(n) \leq T \leq O((n-2m)2m)$. So, the running time of the algorithm is $T = O\left(\frac{n^2}{2}\right)$.

Search time for tabular 3-sum $T=O(n^2)$.

Example 3. Given the set X^9 ={17,43,38,14,20,10,36,47} of dimension *n* = 9. It is required to find out whether there exists a subset X_3 ={x_i,x_j,x_k}, the sum of these elements from the set X^9 is equal to *S* = 100. Here $S \in [S^3_{min}, S^3_{max}] = [51,126]$. First, choose an arbitrary element $x_k = x_6$=10. Find $S(x_k)$ based on the formula (13) $S(x_6)= S - x_6 = 100 - 10 = 90$. Now we use the mapping (10) ($\tau(S, x)$) from the first part of the work and define the set Y^9 for the value $S(x_k)$. Next, apply identity (12): $y_2=y_9$, $S=x_2+x_9+x_6$=43+47+10=100.

Search algorithm 2b for the second task.

Step 1. Input of the initial data: set X^n, *n*, *S*.

Step 2. The formation of the matrix *H*.

Step 3. Check condition Δ = | *H* | = 0.

Step 4. Checking the ownership of the calculated item

$x_k = (S - (x_i + x_j))$ множеству X^n.

Step 5. Output of subset X_3.

The running time of the algorithm varies within $O(n) \leq T < O(n^2)$, the required memory is $M = O(n)$.

Remark 3. From all combinations of n^2 determinants $|H| \neq 0$ and $|H| = 0$, those for which $\Delta = 0$ are selected and all the elements x_k, belonging to the initial set X^n, are selected from them, the number of such elements is at most n-2 .

Example 4. Given the set X^9 ={2,1,6,4,3,5,3,9,7} of dimension n = 9. It is required to find out whether there exists a subset X_3 ={x_i, x_j, x_k}, the sum of these elements from the set X^9 is equal to S = 15. Here $S \in [S^3_{min}, S^3_{max}] = [6,22]$. If x_2=1, x_6=5, and x_k is determined by the formula (16), x_8 = 9. Then, for these elements, the determinant $\Delta = |H| = 0$, $x_8 \in X^9$, the subset X_3={ x_2, x_6, x_8} is satisfied. It is easy to obtain other subsets, in particular, X_3={ x_1, x_3, x_9}.

The discussion of the results. There are a lot of information retrieval algorithms in the scientific literature based on exponential algorithms from [3,4]. The search time and the required memory are $O(2^{n/2})$ and $O(2^{n/4})$ respectively. The use of these algorithms is difficult due to the finding of $2^{n/2}$ subsets. Tabular search methods are based on the construction of tables. The proposed theorems are virtually independent of the length of the search query and require finding only one subset of the sum of subsets task. Lemmas and examples 2-4 show the solution of the tasks set independently of the combination (7). The developed search algorithms with two and three keywords are the most effective compared to tabular methods. In particular, when m = 1 and m = n-1, the theorems follow the traditional search methods: sequential search and pattern matching (mask search). Theorems 1 and 2 allow us to construct a whole family of algorithms for sampling unstructured data for a “short” search query with m keywords and a “long” search query with n-m keywords.

Conclusion. The analysis shows that new methods of information retrieval based on a search query with several keywords significantly reduce the search time for unstructured data, as well as reduce the hardware requirements for the power of computers, servers and other computing devices used. The developed mathematical theory of information retrieval of unstructured data eliminates the need to use arrays, trees, index arrays, index trees, and other well-known information retrieval algorithms that do not drastically improve the search time.

Б. Синчев[1], А. Б. Синчев[2], Ж. Ақжанова[3], А. М. Мұқанова[4]

[1]Халықаралық ақпараттық технологиялар университеті, Алматы, Қазақстан,
[2]«Ұлттық ақпараттық технологиялар» АҚ, Астана, Қазақстан,
[3]Астана LRT, Астана, Қазақстан,
[4]Алматы технологиялық университеті, Алматы, Қазақстан

АҚПАРАТТЫҚ ІЗДЕУДІҢ ЖАҢА ӘДІСТЕРІ. I

Аннотация. Мақалада құрылымдық емес (мәтіндік) деректерді іздеудің проблемасын шешу үшін қолданылатын жаңа әдістер талқыланды. Құжаттарды іздестіру іздеу жүйелерінде қолданылатын табиғи тілдегі негізгі сөздермен жүзеге асырылады. Ұсынылған іздеу әдістері қолданыстағы әдістерден уақыт пен жадыдан, сондай-ақ дамыған алгоритмдер негізінде бағдарламалық өнімдерді енгізудің қарапайымдылығымен түбегейлі ерекшеленеді. Бірнеше (екі немесе үш) кілт сөзбен іздеу сұранысы негізінде құрылымдық емес деректерді іздестіру мәселесін шешу үшін S (сомасы) сомасын (сертификатын) қанағаттандыратын шағын жиынтықтаудың теоремалары келтірілген. Екі кілтік сөзбен іздеу сұрауы үшін қажетті уақыт пен жад O (n) үшін пропорционалды. Ақпаратты үш кілт сөзбен іздеу қызметі екі кілт сөзбен немесе есептік геометрия мәселесіне ақпарат іздеу тапсырмасына дейін азаяды. Бұл ғылыми нәтижелер 17.12-де жарияланған АҚШ патенттік өтінімінде келтірілген материалдарға негізделген. 2018 жылы.

Түйін сөздер: іздеу, әдіс, алгоритм, құрылымдық емес ақпарат.

Б. Синчев[1], А. Б. Синчев[2], Ж. Акжанова[3], А. М. Муханова[4]

[1]Международный университет информационных технологий, Алматы, Казахстан,
[2] АО «Национальные информационные технологии», Астана, Казахстан,
[3]ТОО «Астана LRT», Астана, Казахстан,
[4]Алматинский технологический университет, Алматы, Казахстан

НОВЫЕ МЕТОДЫ ИНФОРМАЦИОННОГО ПОИСКА. I

Абстракт. В работе рассмотрены новые методы, применяемые для решения задачи информационного поиска неструктурированных (текстовых) данных. Поиск документов осуществляется по ключевым словам

на естественном языке, применяемых в поисковых машинах. Предлагаемые методы поиска принципиально отличаются от существующих методов по времени и используемой памяти, а также - простоте реализации программных продуктов на основе разработанных алгоритмов. Приведены теоремы выборки подмножества, удовлетворяющего сумме (сертификату) S, задачи о сумме подмножеств, леммы и алгоритмы решения задачи поиска неструктурированных данных на основе поискового запроса с несколькими (двумя либо тремя) ключевыми словами. Время и требуемая память для поискового запроса с двумя ключевыми словами пропорциональны $O(n)$. Задача информационного поиска с тремя ключевыми словами сведена к задаче поиска информации с двумя ключевыми словами либо к задаче вычислительной геометрии. Эти научные результаты полностью опираются на материалы, приведенные в заявке на патент USPTO США, поданной 17.12. 2018 года.

Ключевые слова: поиск, метод, алгоритм, неструктурированная информация.

Information about authors:

Sinchev B., International University of Information Technologies, Almaty, Kazakhstan; sinchev@mail.ru; https//orcid.org/0000-0001-8557-8458

Sinchev A. B., National Information Technologies JSC, Astana, Kazakhstan; askar.sinchev@gmail.com; https//orcid.org/0000-0002-7333-2255

Akzhanova J., Astana LRT LLP, Astana, Kazakhstan; zyekudayeva@gmail.com; https//orcid.org/0000-0003-1250-8744

Mukhanova A. M., Almaty Technological University, Almaty, Kazakhstan; nuraksulu72@mail.ru; https//orcid.org/0000-0001-6781-5501

REFERENCES

[1] Van Rijsbergen C.J. "Information Retrieval". Dept. of Computer Science. University of Glasgow, 1979 (in Eng.).

[2] Adamansky A. Overview of methods and algorithms for full-text search. Novosibirsk: Novosibirsk State University, 2018. 26 p. (in Rus.).

[3] Horowitz E., Sanni S. Computing Partitions with the Application to the Knapsack Problem // Journal of the ACM (JACM), 1974, T21. P. 277-292 (in Eng.).

[4] Schroeppel R., Shamir A. A $T=O(2^{n/2})$, $S = O(2^{n/4})$ Algorithm for Certain NP-Complete Problem // SIAM Journal on Computing. 1981. Vol. 10, N 3. P. 456-464 (in Eng.).

[5] Korn A., Korn M. Mathematical Handbook. New York: McGraw-Hill Company, 1968. 832 p.

[6] Lifshiz Y. Exact algorithms and open problems // yura@logic.pdmi.ras.ru (in Rus.).

[7] Akhtanova S.S. Algorithms of Data Search // Modern Technologies. 2007. N 3. P. 11-17.

[8] Simakov V.S., Tolkachev D.M. Methods and algorithms for finding information on the Internet. M.: Globus, 2017. 332 p. (in Rus.).

[9] Knut D. The art of programming. Sort and search. Vol. 3. M.: Williams, 2000. 844 p. (in Rus.).

[10] McConnell J. Analysis of Algorithms. M.: Tekhnosfera, 2002. 304 p. (in Eng.).

[11] Urvacheva V.A. Review of information retrieval methods // Bulletin TI them. A. P. Chekhov. 2016. P. 1-7 (in Rus.).

[12] Sinchev B., Sinchev A.B., Akzhanova Z.A. Search for unstructured information // Application for a patent in the USPTO USA dated December 17, 2018. 70 p. (in Eng.).

[13] Sinchev B., Mukhanova A.M. **(2018)** The design of unique mechanisms and machines. II //

News of the National academy of sciences of the Republic of Kazakhstan. Series of geology and technical sciences. 2018. Vol. 5, N 431. P. 210-217. https://doi.org/10.32014/2018.2518-170X.27 ISSN 2518-170X(Online), ISSN 2224-5278(Print).

Publication Ethics and Publication Malpractice
in the journals of the National Academy of Sciences of the Republic of Kazakhstan

For information on Ethics in publishing and Ethical guidelines for journal publication see http://www.elsevier.com/publishingethics and http://www.elsevier.com/journal-authors/ethics.

Submission of an article to the National Academy of Sciences of the Republic of Kazakhstan implies that the described work has not been published previously (except in the form of an abstract or as part of a published lecture or academic thesis or as an electronic preprint, see http://www.elsevier.com/postingpolicy), that it is not under consideration for publication elsewhere, that its publication is approved by all authors and tacitly or explicitly by the responsible authorities where the work was carried out, and that, if accepted, it will not be published elsewhere in the same form, in English or in any other language, including electronically without the written consent of the copyright-holder. In particular, translations into English of papers already published in another language are not accepted.

No other forms of scientific misconduct are allowed, such as plagiarism, falsification, fraudulent data, incorrect interpretation of other works, incorrect citations, etc. The National Academy of Sciences of the Republic of Kazakhstan follows the Code of Conduct of the Committee on Publication Ethics (COPE), and follows the COPE Flowcharts for Resolving Cases of Suspected Misconduct (http://publicationethics.org/files/u2/New_Code.pdf). To verify originality, your article may be checked by the Cross Check originality detection service http://www.elsevier.com/editors/plagdetect.

The authors are obliged to participate in peer review process and be ready to provide corrections, clarifications, retractions and apologies when needed. All authors of a paper should have significantly contributed to the research.

The reviewers should provide objective judgments and should point out relevant published works which are not yet cited. Reviewed articles should be treated confidentially. The reviewers will be chosen in such a way that there is no conflict of interests with respect to the research, the authors and/or the research funders.

The editors have complete responsibility and authority to reject or accept a paper, and they will only accept a paper when reasonably certain. They will preserve anonymity of reviewers and promote publication of corrections, clarifications, retractions and apologies when needed. The acceptance of a paper automatically implies the copyright transfer to the National Academy of Sciences of the Republic of Kazakhstan.

The Editorial Board of the National Academy of Sciences of the Republic of Kazakhstan will monitor and safeguard publishing ethics.

Правила оформления статьи для публикации в журнале смотреть на сайте:

www:nauka-nanrk.kz

ISSN 2518-170X (Online), ISSN 2224-5278 (Print)

http://www.geolog-technical.kz/index.php/en/

Верстка *Д. Н. Калкабековой*

Подписано в печать 11.06.2019.
Формат 70x881/8. Бумага офсетная. Печать – ризограф.
15,7 п.л. Тираж 300. Заказ 3.

Национальная академия наук РК
050010, Алматы, ул. Шевченко 28, т. 272-13-19, 272-13-18