Journal of Fundamental and Applied Sciences

ISSN 1112-9867

Available online at http://www.jfas.info

Research Article

Special Issue

APPROXIMATE SEARCH IN THE SAMPLE ON THE BASIS MANBER-WU METHOD

S. V. Degtyarev[*], E. A. Titenko

Southwest State University 305040, Kursk, st.50 years of October, 94

Published online: 08 August 2017

ABSTRACT

In article approach to search in a sample for production systems is described. Search in a sample is applied to processing of symbolical information. For this purpose operands are presented in the form of a matrix in which processing of vectors is conducted by a conveyor way. Elements of two vectors are logically multiplied among themselves with shift by the beginning of bit of the previous vector of rather current vector. Such operation has received the name "shift" conjunction. She allows to find a sample as a part of the search text.

Keywords: binary vector of comparison, conveyor search, sample.

INTRODUCTION

Relevance of operation: Problem-search tasks of big dimensionality are to valid classes of tasks of the modern theoretical informatics and computer technology (CT). These tasks are guided by creation mathematical and the hardware-software intended for processing of knowledge [1]. In contrast to the computational-logical and combinatorial problems of large dimension, the problem-search problems are oriented to the processing of character information (OSI).OCI it is characterized by account not - factors of the description of a task and its decision (uncertainty of basic data, non-determination of process of the decision, ambiguity of result) [2]. This basic feature assumes use of effective schemes of the parallel calculations peculiar to natural intelligence. The scheme of decision-making "condition action" is one of significant schemes of parallel calculations.

Author Correspondence, e-mail: author@gmail.com

doi: http://dx.doi.org/10.4314/jfas.v9i2s.67

The system of rules (product system) is formal reflection of this scheme of parallel calculations. She has standard computing operations - search in a sample and modification of structure of data. The variability of a task of the left part of production (condition) allows to set naturally parallel transformations over incomplete or inexact basic data [1].

The modularity of structure of the product systemand standardization of basic operations allow to create the customized uniform computers and systems. They realize search and modification of symbolical data through elementary transformations over symbols - replacement, an insert, deletionof symbols taking into account replacement, addition and the admission of a symbol in the left part (sample) of production in any arrangement.Thus, the actual issue of modern CT is the organization and hardware support of parallel production calculations with implementation of approximate search [3].

Formulation of the problem

For the system of products, the problem of approximate search by model is formulated as follows.Let in the working alphabet of *A* the sample of *O* and the processed text of T as one-dimensional objects (words) of length n and m of symbols respectively be set ($n \leq m$). It is required to find all positions (addresses) of entries of a sample of *O* into the text of T taking into account possible wrong comparisons (the admission, replacement, addition of one symbol in structure of the word *T*) [4].

Mathematical formulation of the problem reducesto establishment of such smallest position (address) of *i* with which equality is fair

$$\widetilde{T}(i, i+n-1) = O(1, n)$$

where*i* – an initial position of a sample *O* in the text of *T* ($1 \leq i \leq k$, $k=m-n+1$), – a subline from *T* taking into account the possible admission or replacement or addition of one symbol in structure of the word *T*.

Consecutive methods and algorithms of search in a sample

The representative analytical review of methods and the algorithms of search in a sample based on them is shown in [5]. The majority of the existing algorithms is focused on a one-dimensional form of representation of symbolical operands. As a result algorithms have consecutive enumeration realization. Reduction of unproductive expenses of time in them contacts extraction of additional information from structure of a sample or a part of the analyzed text. Such approach restrictedly is subject to parallelization. This feature doesn't allow to use the best consecutive algorithms in uniform computers and systems.

Methods of conveyor comparison on a sample

Methods of conveyor search in a sample are known. They are based on receiving and serial processing of the binary vectors characterizing positions of occurrences of the current symbol of a sample *O* in the analyzed text of *T* - characteristic vectors (CT). Each *j* of bits of such vector accepts value logical "1" or "0" depending on whether there is *i* a sample symbol*O* at this position of the text of *T*.At such representation of symbolical operands operation of search in a sample comes down to serial irrevocable processing of *n* of binary vectors of the text of *T* word length in *m* of bits each vector. Algorithmization of this method comes down to cyclic processing current and previous *XV*on the basis of logical operation over two *XV*

$$XV_i^{t+1}(j) = XV_i^t(j) \& XV_{i-1}^t(j-1), \qquad (1)$$

where - resultant *j* of bits current *XV* on *i*to a sample symbol *O,* - initial *j* of bits current *XV* on *i* to a sample symbol *O*, - resultant (j-1) bit previous *XV* on (i-1) to *O* sample symbol. According to (1) elementary operation of processing of two binary vectors is hardware oriented. It represents bit-by-bit multiplication of binary vectors. Bits of the previous vector are taken with offset on one line item by the beginning in the current vector. In general, pattern-matching search consists in serial logical multiplication of *n* of characteristic vectors, since start *m*-bit value 1...11. The m-bit binary vector in which logical "1" units are specified entrance of a sample of *O* in the text of *T* is result of pipeline search.

Nevertheless, for problem-searching tasks of big dimensionality processing of binary vectors digit capacity in *m* of bits (*m* – tens, hundreds of thousands of bits) requires excess costs of memory of storage of the intermediate vectors or their parts. In this regard at the hardware level the original method of pipeline search of Manber-Wu is more preferable [5]. In this algorithm binary a vector are created on *O* sample length. They have digit capacity of *n* of bits (*n* - units, tens of bits). Their processing also begins with a start n-bit vector 1 … 11 and is carried out on (1).

Method of serial approximate pattern-matching search

Let in the working alphabet of *A*={ξ_1, ξ_2, ξ_3, ξ_S} the alphabetic $\xi \in A$variable be set and let the alphabetic $\mu_z$variables such be entered that is fair

$$\forall z(\mu_Z \in A_Z = \{A \setminus \xi_{1...Z}\}) \mid z = 1...s$$

Introduction of *s* of alphabets of $A_1 \div A_S$with an exception of the current letters of the alphabet of *A* allows to conduct approximate search at incomplete structural compliance of a sample and fragment of the text. The basis of approximate search is made by matrix (two-dimensional) representation of the next symbols of the text of *T* and current symbol of a sample *O*. In the local vicinity of the current ξ_{ij}symbol the next symbols of the text *T* can act as the passed, added or replaced elements of a sample of *O* (fig. 1).

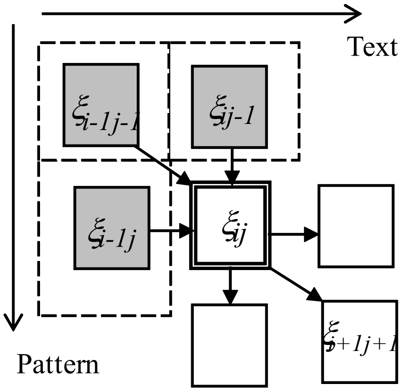

Fig.2. Area of the analysis of the vicinity of aξ_{ij}symbol in the direction of the text (Text) and a sample (Pattern)

The analysis of the vicinity of a binary matrix of comparisons of symbols of a sample and the text is carried out in the direction "at the left-to the right-down". It is conducted from an initial position of a starting vector of $XV_O = 1 \ldots 11$ and allows to consider errors of addition, replacement and the admission of one (next) symbol of a sample. A local word from a sample of $O = \ldots\xi_{i-1j-1}\xi_{ij}\xi_{i+1j+1}\ldots$ further it is modified to one of three types.

Accounting of positive coincidence of a symbol next at the left from ξ_{ij}it is equivalent to addition of one symbol in a sample and its modifications to a look

$O' = \ldots\xi_{i-1j-1}\xi_{i-1j}\xi_{i+1j+1}\ldots$ (2).

Accounting of positive coincidence of a symbol next from above from ξ_{ij}it is equivalent to addition of one symbol in the text and its modifications to a look

$O' = \xi_{i-1j-1}\xi_{ij-1}\ \xi_{i+1j+1}$ (3)

At last, accounting of positive coincidence of the next symbol at the left-above from ξ_{ij}it is equivalent to replacement of one symbol in a sample and its modification to a look

$O' = \ldots\xi_{i-1j-1}\mu_j\xi_{i+1j+1}\ldots$ (4).

As a result search in a sample is carried out:

- it agrees (1) – exact search;
- it agrees (2), (3), (4) – approximate search.

Thus, realization of approximate search is based on realization of three independent processing of adjacent symbols of rather current symbol. The end result on (2), (3), (4) is formed as operation of a disjunction of three positive comparisons (additions, the admission, replacement of one symbol) and changes of structure of a sample.

REFERENCES

1. Wa, B.U. The computer for processing of symbolical information / Louray M.B., Gotsze Li.//TIIER. 1989. t.77, N 4. Page 5-40.

2. Popov, E.V. Static and dynamic expert systems / E.V. Popov [etc.] M.: Finance and statistics, 1996. - 320 pages.

3. Wu S., Manber U. Fast text searching allowing errors, Commun. ACM. 1992. #35 (10) Pp.83-91.

4. Titenko, E.A. Metod and uniform computer of k-approximate search of occurrences in a sample / E.A. Titenko//Messenger of the Voronezh state technical university. t.7. No. 7. 2011 Pages 70-78.

5. Crochemore, M. Algorithms on Strings of [Text] / M. Crochemore, C. Hancart, T. Lecroq. – Cambridge University Press, 2007.

How to cite this article:
Degtyarev S V, Titenko E A. Approximate search in the sample on the basis manber-wu method. J. Fundam. Appl. Sci., 2017, *9(2S), 914-918.*