Analysis of Accelerated Problem Solutions of Word Search in Texts

Zaiceva Sofia[1]

Barkovska Olesia[2]

[1,2]Kharkiv National University of Radio Electronics, 14 Nauky Ave, Kharkiv UA-61166, Ukraine, olesia.barkovska@nure.ua

Abstract. *The work is devoted to the topical problem of word search in texts, the implementation of which is essential in such tasks as electronic dictionaries formation, digital libraries creation, data compression, forecasting algorithms etc. The main problem faced by the majority of scientists who work on this issue is the speed of the existing algorithms implementation. In the work, the existing algorithms speed analysis was conducted and realization by means of a hybrid computer system was proposed.*

Keywords: *word pattern, parallelizing, high-performance computing system, Boyer-Moore algorithm.*

I. Introduction and Problem statement

Continual accumulation of information results in an increase of the information volume in electronic information resources (IR) warehouses. This does not complicate information processes connected with data mining, exchange, accumulation, storage, retrieval and transfer, however, this inhibits accomplishment and implementation time increment of such information processing operations as the given information search [2] and processing due to the growth of the number of computing operations necessary for the solution of the problem of information search in IR storages. Numerical, textual, graphical, audio and video data are distinguished according to the way of information representation. We will focus on the study of textual information being the kind of information represented in the form of written text, i.e. in the form of a predetermined sequence of symbols, because any kind of text analysis [1] (morphological, syntactical and semantic) is applied in a wide variety of applied fields such as marketing and market research, mass media and social networks monitoring, tonality analysis as well as opinion, feedback and complaints rating, for the search of answers to questions received by call-centers, for possible events forecast, in security systems, enabling to lock transfer of unwanted or sensitive information through the Internet etc.

All the above-mentioned types of text analytics and its areas of application explain the relevancy of developing methods of accelerated text search in large input text data arrays by means of the review and adaptation of the existing traditional methods of data mining for shared memory parallel computing systems and massively parallel systems.

II. Problem solution and Results

Among the existing algorithms of word pattern search in texts, the following algorithms are extensively used and have the given performance: linear search ($O((t-l)\times l)$), Aho-Corasick, Boyer Moore ($O(t/l)$), Knuth-Morris-Pratt ($O(t+l)$) and Rabin-Karp ($O((t-l)\times l)$) algorithms, in which t is the source text length, l is the search word length.

With regard to the given performance, Boyer Moore and Knuth-Morris-Pratt algorithms were adapted in the work for massively parallel systems.

Distribution of source text among shared memory computers and application of the fork-join program model supported by Boyer Moore algorithm provide for the rapid (where t=18589 and l=9, the search time for search word is 3,56ms) and almost error-free (0,018%) word pattern search in the text.

A large amount of computers available in massively parallel systems and GPGPU concept utilization with the application of Knuth-Morris-Pratt algorithm enable rapid (where t=18589 and l=9, the search time for search word is 1,85ms) almost error-free (0,084%) word pattern search in the text.

III. Conclusions

Analysis of the obtained results showed that an increase in the number of search words is followed by an increase in the search time as well as an increase in search time is caused by an increase in the source text size.

Application of shared memory multiprocessors enables to acceleration of up to 90 times while massively parallel systems provide for the acceleration of up to 135 times.

References

[1] Data Science & Big Data Analytics: Discovering, Analyzing, Visualizing and Presenting Data/ EMC Education Services. David Dietrich, Barry Heller, Beibei Yang. Published by John Wiley & Sons. Inc. USA, 2015. 435 p.

[2] 1. Barkovska O.Ju., Pyvovarova D.I., Serdechnyj V.S., Ljashova A.O. Pryskorenyj alghorytm poshuku sliv-obraziv u teksti z adaptyvnoju dekompozycijeju vykhidnykh danykh (Advanced Algorithm of Word Patterns Search in Texts with Adaptive Output Decomposition). // Systemy upravlinnja, navighaciji ta zv'jazku. – Poltava: PNTU, 2019. – Issue №. 4(56). – pp.28-34

[3] Najma Sultana, Sourabh Chandra, Smita Paira, Sk Safikul Alam. A Brief Study and Analysis of Different Searching Algorithms // 2017 SECOND IEEE INTERNATIONAL CONFERENCE ON ELECTRICAL, COMPUTER AND COMMUNICATION TECHNOLOGIES – 2017. – P. 944-948.