DOI: https://doi.org/10.30837/csitic52021232890

A Study of the Influence of Computing Systems on the Text Vectorization Speed

Olesia Barkovska[1]

Vladyslav Kholiev[2]

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

Abstract. *The paper considers a relevant problem of accelerating the speed of text processing. The relevance of the development and modernization of electronic library systems results from the growing need for remote access to information objects. The purpose of the study is to reduce the execution time of one of the text processing methods at the stage of information accumulation, namely, vectorization. The results show that it is possible to achieve a greater acceleration of vectorization of small texts (874 words) on multiprocessor computing systems. Systems with massive parallelism produce good results (speedup up to 4.5 times) for large texts (8348 words).*

Keywords*: computing system, texts, vectorization, pre-processing, graphics processor, word2vec, bag of words.*

I. Introduction and Problem statement

The volume of information in the modern information field is growing so rapidly that the traditional methods of its processing, developed at the dawn of the computer era, can no longer complete the task in an acceptable time period. To remedy this, a set of approaches and methodologies was developed for efficient storage, processing, extraction of useful information from Big Data and was called Data Mining. A special case of Data Mining is Text Mining, designed to extract knowledge from huge sets of textual information. Text classification and text clustering are classic Text Mining methods, the quality of which greatly depends on the methods of preprocessing the input text, such as vectorization, stop word removal, lemmatization, stemming, etc. [1-2]

II. Research Task Rationale

The relevance of the development and modernization of electronic library systems (ELS) results from the growing need for remote access to information objects. The functioning and management of ELS is based on the information accumulation and access stages. In turn, the first stage is the most algorithmically and computationally laborious and time-consuming, including methods of normalizing the input text, converting it to a vector form, and the work of the classifier.

Thus, the purpose of the study is to reduce the execution time of one of the text processing methods at the stage of information accumulation, namely, vectorization.

III. Task Fulfillment

Reducing the execution time, which is required to transform the input text into vector form, is possible to achieve by adapting and modifying existing methods (Bag of Words; Word2Vec; TF-IDF) (figure 1) for parallel computing systems with shared memory and systems with massive parallelism.

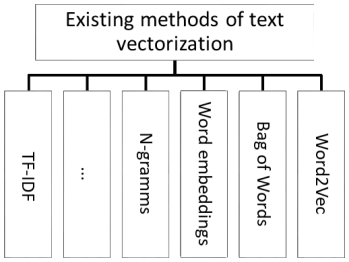

Figure 1. Existing methods of text vectorization

When studying the bag of words method (in which a dictionary-sized vector is created, after which the number of terms occurrences in the texts is counted and saved in the corresponding vector cell [3]) for an input text of 8348 words, an acceleration of 2.3 times was obtained for a computing system with shared memory AMD Ryzen 5 2600 3.4GHz (6 cores, 12 threads) and 4.5 times for a massively parallel system NVIDIA GeForce GTX 750 Ti with Maxwell architecture (5 multiprocessors, 650 CUDA cores). For a small amount of input data (874 words), the speedup was 3.6 times for a computing system with shared memory and 1.7 times for a system with massive parallelism.

The Word2Vec method (developed by a group of researchers at Google, the algorithm matches each term with a vector, giving the coordinates of the terms at the output [4]) also achieves the best result when using the computational resource of the GPU and CUDA technology for an input text of 8348 words 4.1 times and 1.8 times for a small amount of input data (874 words).

IV. Conclusions

In conclusion, the results show that it is possible to achieve a greater acceleration of vectorization of small texts (874 words) on multiprocessor computing systems. Systems with massive parallelism produce good results (speedup up to 4.5 times) for large texts (8348 words).

Further research involves the modification of the considered algorithms to achieve acceleration exceeding those obtained in this work.

References

[1] F.B.S.Prasdika, Dr. Bambang Sugiantoro, S.Si., M.T, “A review paper on big data and data mining”, IJID) International Journal on Informatics for Development Vol. 7, No. 1, 2018, Pp. 33-35. DOI: 10.14421/ijid.2018.07107

[2] V.Serdechnyi, O. Barkovska,D. Rosinskiy, N. Axak, M. Korablyov, “Model of the Internet Traffic Filtering System to Ensure Safe Web Surfing”, Advances in Intelligent Systems and Computing, 2020, 1020, стр. 133–147.

[3] Y. Goldberg, "Neural network methods for natural language processing", Synth. Lectures Hum. Lang. Technol., vol. 10, no. 1, pp. 1-309, Apr. 2017.

[4] T. Mikolov, K. Chen, G. Corrado, and J. Dean. Efficient estimation of word representations in vector space. In Proceedings of ICLR Workshops Track, 2013.