978-1-5090-3525-0/16/$31.00 ©2016 IEEE

Accelerating FCM-Based Text Classification Algorithm Using GPUs

Moahmmed Shehab, Qussai Yaseen, Mahmoud Al-Ayyoub, Firas Albalas and Yaser Jararweh
Jordan University of Science and Technology, Jordan

***Abstract*—Text classification is one of the fundamental tasks in information retrieval and text mining. A recent approach for classification is to employ a clustering algorithm to separate textual data to clusters. A very common algorithm for this purpose is the Fuzzy C-Means (FCM) algorithm. However, such algorithms face a serious problem when dealing with big data due to the long processing time needed. To handle this problem, this paper proposes using GPUs to accelerate the processing of large scale text classification data sets. The initial results are promising and with a decent performance enhancement.**

I. Introduction

Data classification is an important problem in many fields including Data Mining (DM), Machine Learning (ML), Information Retrieval (IR) and Natural Language Processing (NLP). It is concerned with separating new data elements into a predefined set of classes/categories based on what can be learned from an already classified/categorized dataset (known as the training set). Generally speaking, classification algorithms look for similar/common patterns in the data elements of each class (in what is known as the training phase) in order to determine to which class a new element belongs. Text Classification (TC) methods are used to classify textual data. There are many examples of TC including determining the domain or the author of a document, spam detection, Sentiment Analysis (SA), etc. Clustering is the problem of separating data into groups such that the elements of a single group are very "similar" and any two elements belonging to different groups are very "dissimilar." Unlike classification algorithms, clustering algorithms requires no training data. However, they might need certain inputs such as the number of clusters, quality measures for the clustering, whether there is any hierarchy in the generated clusters or not, etc.

The emerging graphics processing unit (GPU) are providing an opportunity for accelerating the processing of large scale problems. It is used in many applications such as image processing, weather forecasting etc [1]. This paper is using GPUs to accelerate large scale text classification problems.

II. Methodology and Experimental Evaluation

This research focuses on improving the performance of classification algorithm that is proposed in [2]. The proposed algorithm has been designed to work correctly and efficiently on big data, which is a main issue nowadays. The paper used a database with In-Memory dataset to build the lexicon. The training phase was implemented using CPUs and GPUs. The first step in the training phase was executed using parallel computing on the CPU. This operation was performed using two main technologies to improve the performance. The first technology used is in-memory dataset.The second technology used is the pipeline method.

TABLE I
The improvement of performance using Advanced CPU and parallel version implementation

Step name	Sequential version	Parallel version	Improvement
Build term documents matrix+Calculate the mean and stander deviation	0h 40m 57s	0h 17m 04s	2.3X
Adapted Fuzzy C-Mean	0h 00m 02s	0h 00m 00.63s	3.2X
Total of training step	0h 40m 59s	0h 17m 04s	2.4X
Testing step	0h 17m 27s	00h 06m 33s	2.6X
Total time	0h 58m 25s	00h 23m 37s	2.5X

Adapted Fuzzy C-Mean is executed using the GPU. The upper and lower memberships are transferred to the GPU memory, which is 5X faster than the CPU. When the AFCM step ends, the data is transferred from the GPU side to the CPU side. Finally, the testing phase is executed in parallel on CPU side using the same two technologies, which are in-memory data and pipelines processing.

Table 1 shows the final results between parallel and the best version of sequential implementation. Obviously, the parallel version is 2.5X faster than sequential implementation.

III. Conclusion

Text classification is one of the most common techniques in data mining. Many algorithms were proposed for text classification. However, the existing algorithm have serious issues when dealing with big data. Therefore, the paper has proposed and tested the fuzzy C-Mean algorithm, which is a very common algorithm, to solve this problem. The paper used Fuzzy C-Mean as clustering algorithm in training phase. The results have shown that this version is 2.5 faster than the sequential version.

References

[1] M. A. Shehab, M. Al-Ayyoub, and Y. Jararweh, "Improving fcm and t2fcm algorithms performance using gpus for medical images segmentation," in *Information and Communication Systems (ICICS), 2015 6th International Conference on*. IEEE, 2015, pp. 130–135.

[2] B. Harish, B. Prasad, and B. Udayasri, "Classification of text documents using adaptive fuzzy c-means clustering," in *Recent Advances in Intelligent Informatics*. Springer, 2014, pp. 205–214.