A SURVEY OF COMPARATIVE ANALYSIS OF NAIVE BAYES ALGORITHM THROUGH GPU AND CPU

[1]Deep Shah, [2]Yash Barot, [3]Isha Kazi, [4]Mr. Swapnil Gharat
[1]Dept of I.T, Rajiv Gandhi Institute of Technology, Mumbai, India
[2]Dept of I.T, Rajiv Gandhi Institute of Technology, Mumbai, India
[3]Dept of I.T, Rajiv Gandhi Institute of Technology, Mumbai, India
[4]Associate Professor, Dept of I.T, Rajiv Gandhi Institute of Technology, Mumbai, India
[1]Rajiv Gandhi Institute of Technology
Off Juhu Versova Link Road, Versova, Andheri (W), Mumbai, India

***Abstract*:** Recent advances in computing have led to an explosion in the amount of data being generated. Processing the ever-growing data in a timely manner has made computing throughput an important aspect for emerging applications. In the past few years there have been many studies claiming GPUs deliver substantial speedups (between 10X and 1000X) over multi-core CPUs on these kernels. To understand where such large performance difference comes from, a rigorous performance analysis for both CPUs and GPUs will be performed. The Naïve Bayes algorithm, which is one of the most widely used document classification algorithms, will be executed through both, a normal CPU and the GPU. Query processing i.e. firing queries with certain conditions processed by the GPU will result in an output dataset in short response time than in CPU. The analysis of the classification time taken by the Naïve Bayes algorithm helps to analyse the performance difference between CPU and GPU computers. The results show that processing through GPU can speed up the classification process when compared to a sequential CPU-based implementation, with basically the same effectiveness in most cases.

***Index terms*: GPU, CPU, Naïve Bayes Algorithm.**

I.INTRODUCTION

Big data is a term for data sets that are so large or complex that traditional data processing application software is inadequate to deal with them. Big data challenges include capturing data, data storage, data analysis, search, sharing, transfer, visualization, querying, updating and information privacy. The term "big data" tends to refer to the use of predictive analytics, user behavior analytics, or certain other advanced data analytics methods that extract value from data, and seldom to a particular size of data set. There is little doubt that the quantities of data now available are indeed large, but that’s not the most relevant characteristic of this new data ecosystem. Analysis of data sets can find new correlations to spot business trends, prevent diseases, combat crime and so on. Scientists, business executives, practitioners of medicine, advertising and governments alike regularly meet difficulties with large data-sets in areas including Internet search, fintech, urban informatics, and business informatics.

The Naïve Bayes algorithm, which is one of the most widely used document classification algorithms, and run the algorithm through both, a normal CPU and the GPU .For GPU computing, the computer powered by NVidia graphics processor while for CPU computing a computer powered by Intel Core i7 processor is used. Evaluation of performance of the Naïve Bayes algorithm on the general purpose CPU and the NVidia powered GPU. The performance analysis of CPU and GPU is based on the time taken by the Naïve Bayes Algorithm on both the machines. The results show that processing through GPU can speed up the classification process when compared to a sequential CPU-based implementation, with basically the same effectiveness in most cases.

II.SIMULATIONS AND EXPERIMENTAL RESULTS

Platform/Parameter	**Time(seconds)**	**Accuracy (%)**
Central Processing Unit	**14.412**	**50.15**
Graphics Processing Unit	**1.241**	**50.00**
Difference(Percentage)	**91.38**	**0.3**

Table No.: 1 Comparision of CPU and GPU for Naïve Bayes Classification.

III.CONCLUSION

By executing and performing a rigorous analysis of the Naïve Bayes Algorithm through a general purpose CPU and GPU, we arrive at the conclusion that processing the algorithm through a GPU is much more time efficient and accurate. Thus, it will be very helpful in processing data of large size as well as small size data through a GPU rather than a general purpose CPU.

REFERENCES

[1] D. Kumarihamy and L. Arundhati, "Implementing data mining algorithms using NVIDIA CUDA," 2009.

[2] G. Andrade, G. Ramos, D. Madeira, R. S. Oliveira, R. Ferreira, and L. C. da Rocha, "G-dbscan: A GPU accelerated algorithm for density based clustering," in ICCS, 2013.

[3] A. McCallum and K. Nigam, "A comparison of event models for naive bayes text classification," in Workshop on Learning for Text Categorization at the 15th Conf. of the American Association for Artificial Intelligence. Madison, Wisconsin: AAAI Press. 1998.

[4] A. Freitas, "A survey of parallel data mining," in Proc 2nd Int Conf on the Practical Applications of Knowledge Discovery and Data Mining, H. Arner and N. Mackin, Eds. London: The Practical Application Company, 1998.

[5] C. Z. Jiang, L. and D. Wang, "Improving naive bayes for classification," International Journal of Computers and Applications, 2010.