IJCSN International Journal of Computer Science and Network, Volume 5, Issue 2, April 2016
ISSN (Online) : 2277-5420 www.IJCSN.org
Impact Factor: 1.02

Enhancing Speed of XML Data Mining on GPU

[1] Deepak V. Mangela, [2] Manoj A. Kale, [3] Sandip M. Walunj

[1, 2, 3] B.E, Computer Department
Sandip Institute of Technology and Research Center, SPPU University
Nashik, India

Abstract - We propose a system for data mining dynamic XML document using a GPU. As a GPU is capable of processing in parallel in real time, it promises on increasing the speed of data discovery process. Nvidia has provided its CUDA platform to utilize its graphics processor for implementing non-graphical algorithms on it. The WWW is structured over the Extensive Markup Language (XML) and hence provides access to the data structured over the WWW. The main objective is to find association rules by extracting items on XML nodes into the XML documents. The use of GPU comes in picture in the pre-processing phase and sorting phase when the item set is being shortened over the occurrences of the items, as this process takes a lot of time the use of a GPU promises a very significant decrease in the pre-processing time.

Keywords - *GPU, Data Mining, XML, CUDA, Parallel Processing.*

1. Introduction

The Internet or the World Wide Web (WWW) has developed exponentially over the past years since its evolution from ARPANET to the more commercial world. With so much information available on WWW, the caliber of exploring new knowledge over this has brought many researchers into this field and their research in enhancing good quality of the data retrieved. Now E-commerce becoming a part of our day to day life, the ease of Buying and Selling online has attracted the end users. This again contributes to the amount of data available on web. With the availability of such enormous amount of data stored in databases and other repositories, there is a need to develop powerful means to acquire knowledge from this structured data. This knowledge can be used for decision making in complex business problems and other areas of research.

XML (Extensible Markup Language) was first presented by The World Wide Web Consortium to categorize the data exchange format on the web and also at the same time achieve operability between dissimilar technologies and implements involved. As a result extracting knowledge from XML data repositories turned into a very important and necessary characteristic. More efficient ways to traverse the data needs to be developed to get the desired information. Many algorithms are developed for traversing but need to be equipped to handle scalable data. Most of the algorithms to mine XML data proposed till date relies on customary relational database with an XML interface.XML mining includes mining both the edifice and the matters from XML documents. Mining of structure, which is essentially mining the XML representation, includes intra-structure mining (mining the inner structure of XML document) and inter-structure mining (mining the structures between XML documents). Mining of content involves content examination and edifice explanation. Content examination is concerned with analyzing texts within the XML document. Edifice explanation is concerned with determining the related documents based on their content.

The Static XML document does not tend to alter their content and edifice with time. For example, an XML document containing details of sales transactions is a static document. Dynamic, or multi-versioned, XML documents are open to alteration to their assembly or content with time. For example, if the content of a bidding event were to be represented in XML format, it would alter daily based on the bids. Association rules is one of the important algorithm used in data mining where there is a role of transactions and a need to find relationship among data. They have been found to be useful especially in the area of discovering interesting relations in very large data sets. From a large data set correlation among different are found. Applying XML mining using association was first introduced by Braga et al. From research it is evident that the approaches used so far rely on an Apriori-like candidate set generation-and test approach. However, this approach is found to be time consuming and costly, especially when the dataset is very large. This paper proposes to introduce parallelism in pre-processing of XML documents during the deserialization procedure. Henceforth the de-serialized data is sorted in parallel using selection sort. The parallelism is incorporated in pre-

processing stage so as to make the dataset favorable for mining. This considerably speeds up the mining process and hence takes care of the scalability issue. The paper further describes the background information related to data mine XML documents on a Graphics Processor and the algorithms associated with its mining.

CUDA is the Parallel Programming architecture provided by NVIDIA to utilize their GPU's for implementing Non-Graphical Algorithms on it. As parallel processing provides a massive decrease in processing time the CUDA architecture is being utilized by researchers to implement time consuming Non-Graphical Algorithms.

2. Related Work

In the paper "Accelerating XML Mining using Graphic Processors" By S. Rathi, C. Dhote, V. Bangera they have successfully accelerated static XML documents over a graphics processor by deserialization of XML using XPATH and then Pre-processing the XML data set for parallelizing it over a GPU. The transactions used were synthetically random generated. The algorithm is tested on 100, 1K, 10K, 50K, 100K and 500K transactions respectively. A clear increase in performance can be observed with the increase in the number of transactions. However, the GPU implementation is not efficient for small data sets compared to the CPU. The time span also includes the calling overhead for invoking the GPU kernel. Because of the memory bound GPU- based implementation; there are more memory accesses than floating point operations. Their system analyzed performances and quality results of our algorithm on different size of datasets ranging from 100 transactions to 5,00,000 transactions. The experiments were performed both on CPU and GPU-CPU based setup. It was found that as the number of transactions goes on increasing the time complexity of CPU increases and GPU performance surpasses the CPU remarkably [1].

In the paper Extracting Association Rules from XML documents using XQuery by J.W. Wan and G. Dobbie, they have demonstrated the use of XQuery to find association rules from XML data and examine the performance of the XQuery execution of the Apriori algorithm. The results show that use of XQuery was not efficient and was time consuming in case of I/O accesses rather than the C++ version of the algorithm enactment. Though their research is a good lead in mining XML data through the Apriori approach, it still has some problems considering that their approach can mine any data for which an expression can be written. The structure of the XML document can be more complex and it has to be taken into considerations too. Another problem of proceeding with the Apriori approach is it needs too many accesses to the database for calculation of the support and confidence, Hence an alternative to it can be the FP-growth algorithm which is way faster and only needs 2 accesses to the data [2].

In the paper Implementation of the Web Mining Based on XML Technology by C. Zheng, Y. Fang and Y. Shen, This paper implements a framework for Web mining based on XML technology and proposes the improved strategy for VTD. The HTML pages that have no rules in format are converted into well formatted XML documents. Then we pre-process the data obtained by XML and store them by VTD in order to mine useful knowledge. With the continuous development of XML technologies, Web Mining will also achieve greater efficiency improvements [3].

The entire framework consists of three modules: data acquisition module, data pre-processing module and data mining module. In data acquisition module, the Web pages are identified through the meta-search engine. In the data pre-processing module, the data can be converted into their Corresponding HTML format through the URL in Java, and then converted into XHTML format by JTidy and converted into the corresponding XML documents through XSLT. Then these converted XML documents are integrated. Finally the data extracted from the integrated XML documents are stored to the database through the technology of VTD. In data mining module, the data in the database are pre-processed again to standardized data sets. Then data mining is carried out on them in order to extract useful knowledge.

3. Methodology

3.1 Proposed System

The proposed system utilizes the GPU's Parallel processing strength to pre-process the data extracted using XPath over the XML document. Later on the arranged data is then sorted based on their occurrences in parallel, while the items which do not occur more often are rejected. The process of data mining is very simple yet very complex as the simple processes take a very long time when considering massive amounts of data altogether. Each step is powered by parallel execution over the GPU and hence the process tends to take less time as compared to serial execution.

The first phase consists of extracting the specific class of data from the XML documents, for this the XPath language is very useful just by creating an expression in the language and applying it on the data can extract only the proper data required. All the data collected over a specific class is then arranged for a more meaningful way

and for further processing and searching for the frequently occurring items. All this searching a sorting of data is done in parallel to achieve a greater decrease in time for the frequent item discovery process.

Now considering how the implementation varies over a GPU as it is different from executing instructions on a CPU the data has to be assigned to threads, bundles and grids before executing them in parallel. The assigning and management of the data such that the GPU executes efficiently is dependent on the memory size and CUDA cores [13]. Fig.1 represents the flow of events in the system. The documents are processed using XPath to extract the items which appeared frequent enough to fulfill the threshold or minimum support value. This data is then arranged and used to create association rules. The Apriori approach is very space consuming and needs to access the data many times during its implementation. The data access is very costly in terms of time and hence Apriori approach is not used by our system. The proposed system works and uses the FP-Growth algorithm to find frequent item set and association rules.

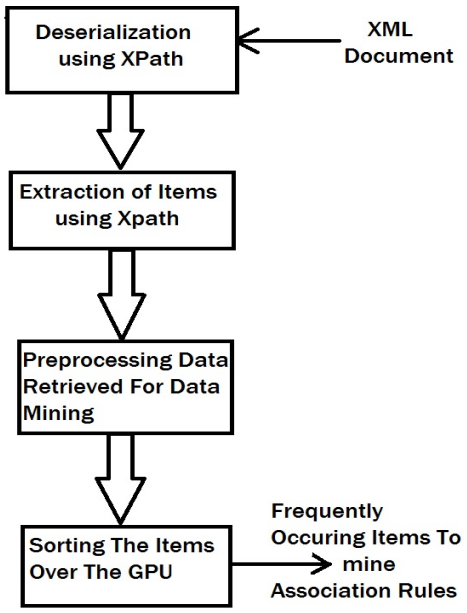

Fig. 1. Proposed System Architecture

3.2 Mathematical Model

The proposed system can be defined as a set of tuples shown below:

$$S= \{I, O, \delta, \lambda, D, G\} \quad (1)$$

Where,

I = Input XML document

O = Output

δ = function to extract items from the XML code

λ = function to count occurrences of items in the item set

G = GPU Cores

Let the No. of Nodes in the Given XML Document I be N

Say when we pass the input consisting of N nodes to the function δ we get O as an output with n items each related to the N nodes respectively.

$$\delta (I) = IS\{o_1, o_2, o_3 \dots o_n\} \quad (2)$$

Now,
We have a set of items IS for which we have to find the occurrences to be able to find some association rules. Hence we pass the set IS to the function λ to find the count of each item in the whole set IS.

$$\lambda (IS) = O\{(o_1, c_1), (o_2, c_2), (o_3, c_3) \dots (o_n, c_n)\} \quad (3)$$

Here O is the set of elements frequently occurring with their count.

Hence with the elements in the O set the association rules can be built to give a more understandable result of the mined data and can be analyzed to build a strong knowledge foundation.

4. Software and Hardware Requirement

4.1 Software Requirement

i. CUDA Toolkit 6.5
ii. g++ 4.4
iii. gcc 4.4
iv. XPath library XQuilla or libxml2
v. NVCC Cuda Compiler
vi. Operating System- Ubuntu 14.04 or Windows 7

4.2 Hardware Requirement

i. CPU: Intel i5 2nd gen 64-bit processor
ii. GPU: Nvidia Geforce GT525M 96 CUDA Cores

5. Conclusion

The findings of this paper suggests that the only use of graphics processors is not only just implementing graphical algorithms but their parallel processing architecture can be made use to enhance the speeds of other non-graphical algorithms too. In the near future the graphics processors would not be known for just increasing the visual experience and imaging but would also be known in every field where there is a need of fast processing and computational power.

6. Limitations

This system shows the power of GPU computing over a basic level mobile GPU with least CUDA cores and graphics memory of 1GB. Hence the speed up factor and scalability is limited. Yet it shows significant increase in processing speed and time efficiency.

7. Future Scope

This system can be scaled to an array or GPU's or implemented over computing clusters of GPU's. The industry grade GPU's such as the Tesla series promises far more speedup in discovering knowledge than the basic GPUs present in daily use computers.

Acknowledgment

We would like to appreciate all the authors for contributing in the field of data mining and parallel processing, their works encouraged and also helped us in understanding the working of these technologies and also how to use them to enhance the computational experience in everyday problems. This research paper was pursued with the help of our mentor and co-author Prof. Sandip M. Walunj, Sandip Institute of Technology and Research Center, Nashik, Maharashtra.

References

[1] S.Rathi, C.Dhote, V. Bangera, Accelerating XML Mining using Graphic Processors International Conference on Control, Instrumentation, Communication and Computational Technologies (ICCICCT) 2014

[2] J.W. Wan and G. Dobbie, Extracting Association Rules from XML documents using XQuery, 2003

[3] C. Zheng, Y. Fang and Y. Shen, Web Mining Based on XML Technology, Computational Intelligence and Security, 2009. CIS '09. International Conference on (Volume:1)

[4] R.Porkodi, V.Bhuvaneswari, R.Rajesh, T.Amudha, \emph{An Improved Association Rule Mining Technique for Xml Data Using Xquery and Apriori Algorithm}, 2009 IEEE International Advance Computing Conference (IACC 2009) Patiala, India, 978-1-4244-1888-6/08 2008 IEEE.

[5] Liu, Md. Sumon Shahriar and Jixue. \emph{On Mining Association Rules with Semantic Constraints in XML}, ICDIM, pp.1-5, 978-1-4577-1539-6/11 IEEE, 2011.

[6] Zhi-gang Wang, Chi-she Wang. \emph{A parallel association-rule mining algorithm}. In Proceedings of the 2012 International conference on Web Information Systems and Mining, WISM'12, Springer, pp. 125-129.

[7] Han J, Pei J, Yin Y, \emph{Mining frequent patterns without candidate generation}, In: Proc. of the ACM SIGM0D Conference on Management of Data. Dallas, TX, 2000.2

[8] Fang R., He B., Lu M., Yang K., Govindaraju N. K., Luo Q., Sander P. V.: \emph{GPUQP:query co-processing using graphics processors}. In ACM SIGMOD International Conference on Managament of Data, pp. 10611063, New York, NY, USA, 2007. ACM.

[9] Braga, A. Campi, S. Ceri, M. Klemettinen, and P. L. Lanzi. \emph{Mining association rules from xml data}. In Proc.Of 4th InternationalConference on DataWarehousing and Knowledge Discovery(DaWaK'02), volume 2454 of LNCS. Springer, pp.21-30,2002.

[10] S. Che, M. Boyer, J. Meng, D. Tarjan, J. W. Sheaffer, and K. Skadron. \emph{A performance study of general-purpose applications on graphics processors using cuda}. J. Parallel Distrib. Comput.,68(10):1370-1380, 2008.

[11] J. Dean and S. Ghemawat. Mapreduce: \emph{Simplified data processing on large clusters}. Commun. ACM, 51(1):107-113, 2008

[12] W. Fang, K. K. Lau, M. Lu, X. Xiao, C. K. Lam, P. Y. Yang, B. Hel, Q. Luo, P. V. Sander, and K. Yang.\emph{ Parallel data mining on graphics processors.} Technical report, Hong Kong University of Science and Technology,2008

[13] CUDA C/C++ Basics – Nvidia www.nvidia.com/docs/IO/116711/sc11-cuda-c-basics.pdf