International Journal of Advanced Technology in Engineering and Science
Vol. No.5, Issue No. 03, March 2017
www.ijates.com
ijates
ISSN 2348 - 7550

MINING OF DATA BY MEANS OF PARALLEL COMPUTING FOR E-GOVERNACE BY KNOWLEDGE RETRIEVAL BY COMPUTE UNIFIED DEVICE ARCHITECTURE

G.Janardhan[1] , Mr.Banda Srikanth Reddy[2], T.Madhu [3]

[1]*Department of Computer Science and Engineering JNTUH, (India)*
[2]*Department of Computer Science and Engineering JNTUH, (India)*
[3]*Department of Science and technology, SRIS, Nalgonda, (India)*

ABSTRACT

In this paper, a scalable framework aimed at providing a platform for developing and using high performance data mining applications on heterogeneous platforms. The framework incorporates a software infrastructure and a library of high performance kernels. The exponential increase in the generation and collection of data has led us in a new era of data analysis and information extraction and focuses on the security of the information. In which the three methods is being used first K-means algorithm, decision tree and the artificial neural network.

After using the data mining concept there is use of more mature association model for information system. The information security is more concerning concept for various large scale organizations. As information system provides many facilities to us but the information security is another concern of any computer system. The information system is "Double Edge Sword" because it provides many facilities but if there is no security there will be a great loss and inconvenience. The security of any information system is interrupted by virus, hacker, and leakage of secrets, system failure and many other things.

Conventional systems based on general-purpose processors are unable to keep pace with the heavy computational requirements of data mining techniques. High performance coprocessors like GPUs and FPGAs have the potential to handle large computational workloads.

Data mining is widely used in various domains and has significant applications. But the data mining is mostly used in many fields differently like E-governance, information security, spatial mining, web mining and also with parallel computing. This paper will present the combination of three concepts which are E-government and information security by using parallel computing plate form. Using Parallel Computing concept will makes the E-Governance system much faster and the Information Security will make it more reliable.

In this paper the CUDA, (the Compute Unified Device Architecture), is a parallel computing platform and programming model created by NVIDIA and implemented by the graphics processing units (GPUS) that they produce. CUDA gives program developers direct access to the virtual instruction set and memory of the parallel computational elements in CUDA GPUs. Using CUDA, the GPUs can be used for general purpose processing

(i.e., not exclusively graphics); this approach is known as GPGPU. Unlike CPUs, however, GPUs have a parallel throughput architecture that emphasizes executing many concurrent threads slowly, rather than executing a single thread very quickly.

Keywords: ***CUDA, GPU, Double Edge Sword, Parallel Computing, K-Mean Algorithm, Apriori, Decision Tree, Security, Information Extraction***.

I. INTRODUCTION

1.1 Data Mining in Information Security for E Government

E-government security is considered one of the crucial factors for achieving an advanced stage of e-government. As the number of e-government services introduced to the user increases, a higher level of e-government security is required. Since an underdeveloped country whose development can be rapid through proper E-Government implementation. Presently, it is in infnt ancy stage. One of the major failure factors identified at this stage is the improper security consideration. This dissertation also contributes in proposing a cost effective security framework for underdeveloped country.

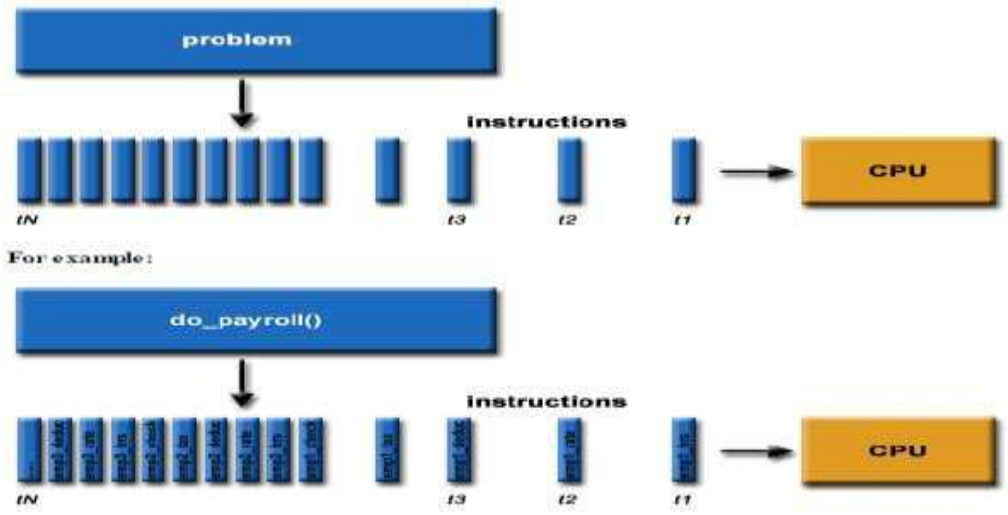
In twenty-first Century, information technology has rapidly permeated into every field of human society. Form the Large national or international fields to the small families or individuals, more and more people use of information technology to provide convenient, fast and efficient work and business. The expanding demands of All walks of life prompted the investment and the scale of informatization construction from the underlying physical layer to the top of the application system constantly strengthen intensify. According to the United States of FBI survey, economic losses caused by the network security are more than $170 billion in USA per year. 75% companies reported that the financial losses were caused by results of computer system security problems. From CN CERT and the China Internet Network Information Center's annual reports, in the first half of 2010, CN CERT had received 4780 network security incident reports, increase of 105%. In the past year, the service fee expenditure for processing safety events totaling up to 153 billion Yuan.

1.2 Parallel Computing

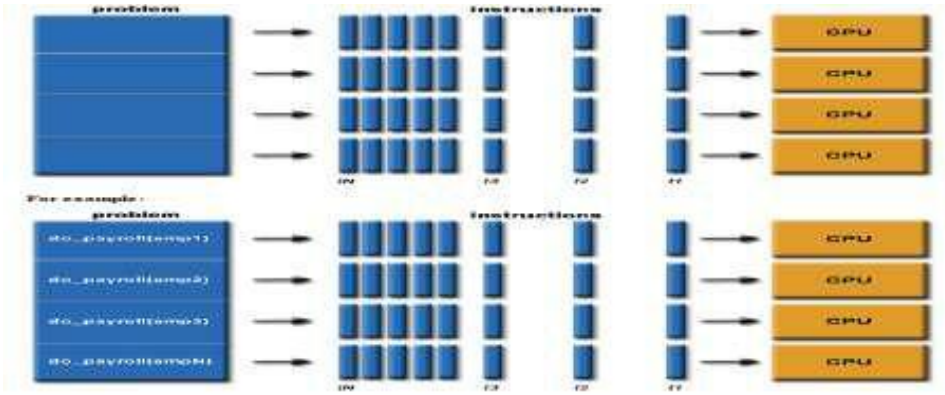
HISTORY OF GPU COMPUTING:- Graphics chips started a fixed-function graphics pipelines. Over the years, these graphics chips became increasingly programmable, which led NVIDIA to introduce the first GPU or Graphics Processing Unit. In the 1999-2000 timeframe, computer scientists in particular, along with researchers in fields such as medical imaging and electromagnetic started using GPUs for accelerating a range of scientific applications. This was the advent of the movement called GPGPU or General Purpose computing on GPUs. The challenge was that GPGPU required using graphics programming languages like OpenGL and Cg to program the GPU. Developers had to make their scientific applications look like graphics applications and map them into problems that drew triangles and polygons. This limited the accessibility of tremendous performance of GPUs for science. NVIDIA realized the potential to bring this performance to the larger scientific community and decided to invest in modifying the GPU to make it fully programmable for scientific applications and added

support for high-level languages like C, C++, and Fortran. This led to the CUDA parallel computing platform for the GPU.

To be run on a single computer having a single Central Processing Unit (CPU). A problem is broken into a discrete series of instructions. Instructions are executed one after another. Only one instruction may execute at any moment in time.



In the simplest sense, *parallel computing is* the simultaneous use of multiple compute resources to solve a computational problem. To be run using multiple CPUs. A problem is broken into discrete parts that can be solved concurrently. Each part is further broken down to a series of instructions. Instructions from each part execute simultaneously on different CPUs.



II. K-MEANS ALGORITHM

K-means algorithm (Lloyd, 1982) is a simple and effective statistical clustering technology, it gives specific classes number K, and put N objects into the K classes, to make the maximum similarity within objects in any class, and to make the minimum similarity among class. The algorithm first need to make an initial judgment that is to select the initial class number and the initial cluster center, and then, each sample is placed in the similar class. Similarity measure can be defined in many different ways. The most commonly used similarity

metric is simple Euclidean distance. After all samples are placed into the appropriate class, the class center was update through the calculation of each new class of average. The process is repeated until a certain iteration of generating class center no longer change.

III. DECISION TREE

The decision tree algorithm was first proposed as ID3 algorithm by Quinlan, Later there reappeared many kinds of decision tree algorithm such as ID4, ID5, C4.5, CART, CLOUDS, PUBLIC, SLIQ, RAINFOREST, SPRINT, and ScalParC and so on. It is a common structure to supervise learning, mainly used for data classification. First, we should select portion of the samples to create a decision tree from the training set, and the remainder of the training samples are used to inspect the accuracy of the tree established. If the decision tree can correctly classify the remaining samples, the process will be end. If some sample's classification is error, this sample is added to the training set and create a new tree. In this way, we design a tree which can classify all training samples correctly.

IV. ARTIFICIAL NEURAL NETWORK

Artificial neural network was born in 1950, and Rosenblatt put the single-layer perception application in pattern classification. Its principle is the human brain thinking system's simple structure simulation. It's a multilayered network that is made up of a number of neuronal connections, and can imitate the human brain function of neuron. It is also the adaptive function estimator that does not rely on the model. It does not need any model to realize arbitrary function relation. Its advantage is capable of parallel processing, and has the learning ability, adaptability and strong fault tolerant ability.

V. THE FORMAL DESCRIPTION FOR CORRELATION ANALYSIS MODEL

Correlation analysis model is adapted to find out the meaningful connection hiding in the large data sets. The connection found can be represented by association rules or frequent item sets form. We need to deal with two critical issues in the data correlation analysis. First, finding in computing mode from the large object data concentration may be costly, second, some models found may be false, because they may have occurred by chance.

5.1 Data's Two Elements Expression

The hypothesis that certain things and the Item set included in these things in the set of relations as shown in

Table 1

TID	i1	i2	i3	i4	i5	i6
1	1	1	0	0	0	0
2	1	0	1	1	1	0
3	0	1	1	1	0	1
4	1	1	1	1	0	0
5	1	1	1	0	0	1

If I={i1,i2,…,id} is the collection of all items, and T={t1,t2,…,tn} is the collection of all things. Each thing of ti contains item set is the subset of I, that is ti being contained in I. Item set's an important property is the support count, that is the number of things containing specific item sets, With σ(X) expressed as:

$$\sigma(X) = |\{t_i | X \subseteq t_i, t_i \in T\}|$$

5.2 The Support Degree and Confidence Degree

Support degree is an important measure, because the rules which support degree is very low may Sometimes occur. From the data analysis point of view, most of low support degree of rules will be meaningless, because it was more important to improve the research and find out countermeasures to focus on more important things than on the things which does not often happen. Therefore, when we carry out data analysis, we can use the support threshold interval to delete those meaningless rules. In addition, support degree also has desired properties, can be used for association rules effectively discover. The confidence degree measures the reasoning reliability by rules or mathematical model. For a given rule X→Y, the higher the confidence degree, the more likely that Y will appear in the things which contain X. The confidence degree can also estimate the Y defined in X under the conditional probability. If the association rules are expressed as shaped like X→Y implication expression, where X and Y are disjoint sets, namely X∩Y=ø. The strength of association rule can be measured by support degree and confidence degree. Support degree determines the rules that can be used for a given data set frequency, while the confidence degree determines the frequent degree that Y in X contains things appeared. Support degree (s) and confidence degree (c) these two metric representation just as below:

$$s(X \rightarrow Y) = \frac{\sigma(X \cup Y)}{N}$$

$$c(X \rightarrow Y) = \frac{\sigma(X \cup Y)}{\sigma(X)}$$

Support degree measures the importance (or range) of association rule, confidence degree measures the accuracy of association rules. At the same time satisfying the minimum support threshold (min-support) and minimum confidence threshold (min-confidence) rules called the strong rule. The problem of association rule mining is the strong rule that is satisfied for min-support and min-confidence at the same time when mining in transaction database.

5.3 Algorithm optimization

Mining association rules in a primitive method is to calculate for each possible rules support degree and confidence degree. But this method is costly, and step back, because of the number of rules extracted from the data set up to index level. For example, the total number of rules extracted from a data set containing d is, R=3d-2d+1+1. If there are 6 items of data set, it needs to calculate the 602 rules of the support degree and confidence degree. The optimization strategy of mining algorithm of association rules is to decompose the association rules mining task into two processes. The first is the frequent item sets generation process that is the process finding to meet the minimum support threshold. The second process is the generation rules that are to extract all the high confidence rules from frequent item sets found from the previous step. There are many optimization algorithm in

frequent item sets generation process, such as the transcendental principle, apriori algorithm, candidate item sets generation and pruning algorithm.

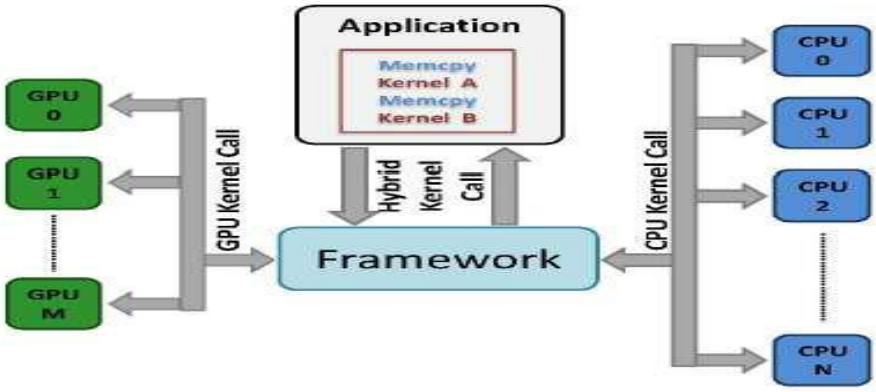
5.4 Association Analysis Application in the Analysis Of Information Security

In this part, we give an IT environment of the government with hundreds of branches as a case. We give an information security risk assessment analysis on years of Information Security Survey and information security events and other research data according to the proposed model relational analysis, and give a step by step sample. This paper describes the data up to the end of 2010 census data completed, and streamlines a part of the data to be example. Exclusion of different industry, the information security census data is up to thousands items, users wanted to investigate the relationship amount some projects according to some empirical, and analyzed

Some valuable information related with information security from the artificial.

5.5 Hybrid Implementation

Hybrid implementation refers to harnessing the capabilities of both

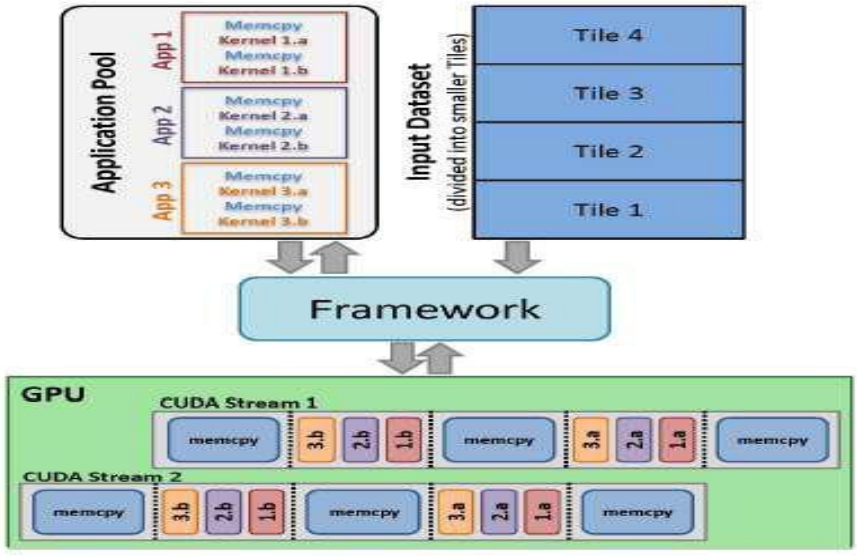


GPUs and CPUs simultaneously. Consider a situation when we have a GPU and a multi-core CPU in the system. It would be desirable to distribute the tasks between the GPU and the CPU cores. Since the computational power of GPU is significantly higher than that of the CPUs, the data need to be distributed such that the work remains balanced. Our framework provides the functionality to run an application in the hybrid mode. In this mode, the data will be distributed among the nodes and the corresponding CPU or GPU kernels will be launched. Figure 5 shows how a hybrid kernel call gets broken down into architecture specific kernel calls using the framework.

5.6 Multiple Kernel Optimizations

This optimization is specific to the CUDA implementation. We notice that data mining kernels process huge amounts of data and it is not always possible to fit the entire data in the GPU device memory. As mentioned in Section III-B2, this will require the usage of CUDA Streams to lower the overhead caused by copying data from host memory to the device memory. We further notice in our experiments that kernel execution time is smaller than the time it takes to copy smaller tiles of data to the GPU device memory. This presents us a unique opportunity to leverage the time difference. In practical situations, a number of different data mining algorithms

are used on a given dataset. We propose to run kernels from different applications on the dataset while it is in the device memory so as to reduce the overhead of memory copy as much as possible. Figure 6 shows the idea behind this optimization. As an example, three different applications App1, App2, and App3 are shown in the figure. For each memory transfer call put on a CUDA Stream, one kernel call from each of the applications (1.a, 2.a, and 3.a) is allocated on that particular stream as shown. This can be viewed as a single kernel whose execution time is close to the combined execution time of the same kernels running separately. The kernel execution and host-device memory copy times can be used to predict the number of applications which can be interleaved in the above fashion.



VI. CONCLUSION

As explained earlier the information technology is the double edge sword that's why the security is main issue for it. The developed country are also analyze that information issue is big issue for any country development in modern competitive era. The governing bodies of country or the state will also moves towards the online application which makes the people to easy and quick transaction with government and each other. E-governance is electronic government means everyone in the area will access it so it must be secure that's why no one can suffer from the any security problems. As at the time many people use same server access so the parallel computing will give fast access to the system without any interference and less cost will also beneficial for the governing bodies.

REFERENCES

[1]. High Performance Data Mining Using Data Cubes On Parallel Computers _Sanjay Goil Alok Choudhary ECE Department and CPDC, Northwestern University Technological Institute, 2145 Sheridan Road, Evanston IL-60208.

[2]. High Performance Data Mining Using R on Heterogeneous Platforms, Prabhat Kumar, Berkin Ozisikyilmaz, Wei-Keng Liao, Gokhan Memik.

[3]. Alok Choudhary Department of Electrical Engineering and Computer Science, Northwestern University, Evanston, IL, USA, fpku649, boz283, wkliao, memik, choudharg@ece.northwestern.edu.

[4]. Parallel Data Mining Techniques on Graphics Processing Unit, with CUDA, Liheng Jian Ying Liu(yingliu@gucas.ac.cn) Shenshen Liang, School of Information Science and Engineering, Graduate University of Chinese Academy of Sciences, Beijing, China

[5]. Data Mining Applications in E-Government Information Security Tongwei yuan*, Peng Chen.

[6]. http://www.infosecurity-us.com/view/11292/. Cyber-crimes cost organizations 38 million per year. 27 July 2010.

[7]. http://www.edu.cn. Network black industry challenge network security 04 Jan 2011.

[8]. Litao, Zhang; Security management model research based on system boundary analysis information. (2005)

[9]. Chen. Jinbo; The study on the application of data mining in telecommunication CRM. (2006)

[10]. PangNing Tan, Michael Steinbach, Vipin Kumar; Introduction to Data Mining; Pearson Addison Wesley (2006) 200-238

[11]. J.A.Berry Michael, S.Linoff Gordon; Data Mining Techniques; Wiley Publishing (2011) 245-290.

	G.JanardhanReceived the B.Tech degree in Computer Science and Engineering and M.Tech degree in Computer Science and Engineering from J N T U Hyderabad University. He is Currently working as a Assistant Professor in Swami RamanandaTirtha Institute of Science and technology, Nalgonda,Telangana, India. He has having 6 years of teaching Experience. His current research interests include Web Database Integration Systems and Data Mining
	Mr. Banda SrikanthReddy received the B.Tech degree in Computer Science and Engineering and M.Tech degree in Computer Science and Engineering from J N T U Hyderabad University. He is Currently working as a Assistant Professor in Swami RamanandaTirtha Institute of Science and technology, Nalgonda, Telangana, India. He has having 8 years of teaching experience. . His current research interests include Data Mining and Cloud Computing
	T.Madhu is working as a Associate Professor in Swami RamanandaTirtha Institute of Science and technology, Nalgonda,Telangana, India