International Journal of Multidisciplinary and Current Research

Research Article

ISSN: 2321-3124
Available at: http://ijmcr.com

HSApriori: High Speed Association Rule Mining using Apriori Based Algorithm for GPU

D.William Albert[1], Dr.K.Fayaz [2] and D.Veerabhadra Babu[3]

[1]Department of Computer Science & IT, Mahatma Gandhi University, Meghalaya, India
[2]Department of Computer Science & Technology, SK University, Anantapuram, India

Accepted 20 July2014, Available online 01 Aug 2014, **Vol.2 (July/Aug 2014 issue)**

Abstract

Apriori-Based algorithms are widely used for association rule mining. However, these algorithms cannot exploit the parallel processing power of modern GPU (Graphics Processing Unit). To make an algorithm to be compatible with GPU, it needs to be changed in representation of data, parallel processing and also in support count. In this paper we propose an Apriori-based algorithm ***HSApriori****for high speed association rule mining besides suitable data representation and support counting mechanisms. OpenCL is one of the best General Purpose Graphics Processing Unit (GPGPU) platforms used to implement the functionality of the algorithm. OpenCL with Java extensions are used for developing HSApriori. The datasets used include chess, pumsb, and accidents that are obtained from UCI machine learning repository. One more dataset used is synthetic in nature which is collected from IBM Almaden Quest Research Group. We built a prototype application to demonstrate the proof of concept. The experimental results are compared with BorgeltApriori. The results revealed that HSApriori outperforms BorgeltApriori.*

Keywords: *Data mining, high speed association rule mining, GPGPU platforms, Apriori*

1. Introduction

Association rule mining or frequent item set mining is done using many algorithms in data mining domain. The frequency is measured using support threshold which is computed as number of transactions containing a subset divided by the total number of transactions. The support threshold helps to extract user-interested rules that help in decision making. Apriori is one of the algorithms for association rule mining. Frequent item set mining has plethora of utilities in the real world such as sales data analysis, customer behavior analysis in banking, insurance, and other sectors. The frequent item set mining is also useful in computer vision, data stream analysis, and bioinformatics, information retrieval, and database management systems.

Graphics Processing Unit has special circuits that can help it processing graphics faster. Moreover they are equipped with parallel processing features. They are more efficient when compared with general purpose CPUs are being used traditionally. Modern GPUs are manufactured by companies like NVIDIA and AMD. The association rules mining or frequent item set generation can be done much faster using GPU. As the computing industry is all set to process big data which is characterized by features like velocity, variety and volume, the traditional data mining algorithms will not be able to use the power of GPU. Therefore it is required to rewrite a data mining algorithm so as to harness the parallel processing capacity of modern GPUs. As modern computers have processors that can involve in parallel processing it is good idea to have data mining algorithms to be modified to use the parallel processing nature of GPUs. Generally both CPU and GPU are used to have more flexibility. There should be proper communication between CPU and GPU. The communication mechanism also should be known to algorithm. In other words, it is essential that developers need to know the functioning of GPU and how it helps in working with traditional CPUs. Our contributions in this paper are as described here.

1. We have built an Apriori-based algorithm named “HSApriori” targeting GPGPU platform to leverage parallel processing of GPU and achieve high speed association rule mining.
2. We have premeditated and identified mechanisms for data representation and support counting to be compatible with GPU computing. These mechanisms are used by our algorithm HSApriori.
3. We have built a prototype application to demonstrate the functioning of the proposed algorithm by using synthetic and real datasets.

The remainder of this paper is structured as follows. Section II reviews literature on GPGPU platforms and data mining that targets these platforms. Section III provides information about the HSApriori implementation. Section IV presents experimental results while section V concludes the paper.

2. Related Works

This section reviews algorithms used for mining association rules besides GPGPU platforms. The best frequent pattern mining algorithms existed are FP-Growth [4], Eclat [5], Apriori [6] etc. The candidate generation process in Apriori and Eclat is similar except in the way they represent candidate and transaction data. FP-Growth operates differently. First of all it built FP-Tree and then mines frequent item sets. It does not use iterative process for generating candidates unlike the other two algorithms. Most of the frequent pattern mining algorithms that are existed use serial processing. It does mean that they work in traditional CPU. They are not written for exploiting the parallel processing power of modern GPU. With respect to single thread performance comparison, FP-Growth is better than the other two. However, Apriori outperforms FP-Growth when minimum support is high. Moreover Apriori algorithm is more suitable for parallel processing [7].

Recently many algorithms came into existence that is meant for discovering frequent patterns from large datasets. In [8] FerecBodon used candidate hashing and trie-based data structure for Apriori algorithm. Recursion pruning is used by Christian Borgelt in [9] for implementing Apriori. Based on Agrawal's algorithm [10] Bart Goethals implemented another Apriori algorithm. In this paper we proposed HSApriori which leverages the parallel power of GPU.

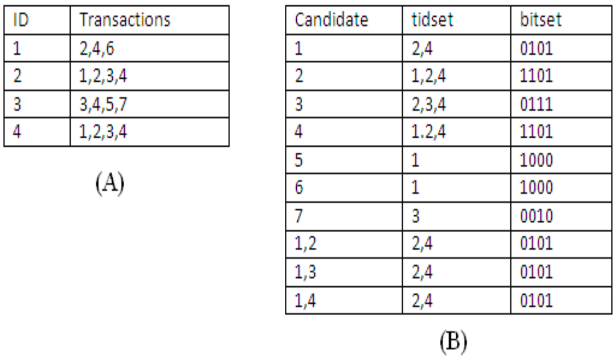
3. High Speed Association Rule Mining using Apriori with Gpu (HSApriori)

High speed association rule mining is possible when Apriori is able to run in GPGPU platforms like CUDA and OpenCL. In this section we describe the mechanisms that are different from traditional apriori and propose HSApriori which is targeted to GPGPU platforms. In this paper experiments are done using OpenCL with Java extensions. The following sub sections provide data structure suitability, support counting in OpenCL for building proposed algorithm.

A. Sample Dataset

Every transaction is given unique id which identifies a transaction. Each transaction has certain values. The representation of database transactions can have its convenience in generating frequent item sets as part of association rule mining.

Careful consideration of data structure is required while building Apriori for GPU. This is needed as GPU operates in parallel fashion. The algorithm should be able to utilize the simultaneous processing of multiple units supported by GPU. For this, vertical transaction list represented as shown in Figure 1(B) is considered. The tidsets are good for processing but they cause a problem while performing counting operation. This makes it unpredictable behavior and thus results in poor performance on GPU. This is because of the nature of GPU and its processing units need an approach that suits parallel processing. Figure 2 shows Tidset join and bitset join before understanding the best representation.



ID	Transactions
1	2,4,6
2	1,2,3,4
3	3,4,5,7
4	1,2,3,4

(A)

Candidate	tidset	bitset
1	2,4	0101
2	1,2,4	1101
3	2,3,4	0111
4	1.2,4	1101
5	1	1000
6	1	1000
7	3	0010
1,2	2,4	0101
1,3	2,4	0101
1,4	2,4	0101

(B)

Figure 1 Transactions in horizontal (A) and vertical (B) representations

B. Data Structure Suitability

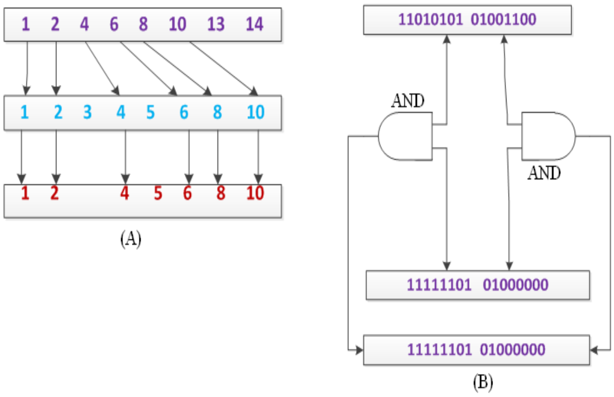


Figure 3 Tidset join (A) and Bitset join (B)

As can be seen in Figure 3 it is evident that there are two representations known as tidset and bitset. The tidset representation is good as it is compact in nature. However, this structure makes it difficult to parallelize the mining process. When it comes to bitset representation, it consumes more memory while it is best suited for GPU and parallel set join operation. Bitwise and operation between two bit vectors can be used to join two bitsets. This is more flexible way of representing data for parallel processing carried out on GPU.

C. Support Counting in Traditional Apriori vs. Apriori for GPU

In traditional Apriori algorithm scanning database transactions is required in order to compute support ratio. The support ratio is a measure used in frequent itemset mining to specify a threshold. In case of Aprioritrie traversal and binary search is required for support counting which causes irregular memory access while putting on GPU. It does mean that there is communication between the CPU and GPU and there are memory structures that are to be used carefully. Therefore the support counting functionality of the traditional Apriori has to be modified.

On GPU platform counting should be based on complete intersection where candidates are transferred from the traditional main memory to graphics memory through host code. Then the GPU is able to compute the support ratio using bitwise interactions on the given vertical transaction lists. Then the results of support values are sent back to main memory. The parallel processing of support counting on GPU is shown in Figure 4.

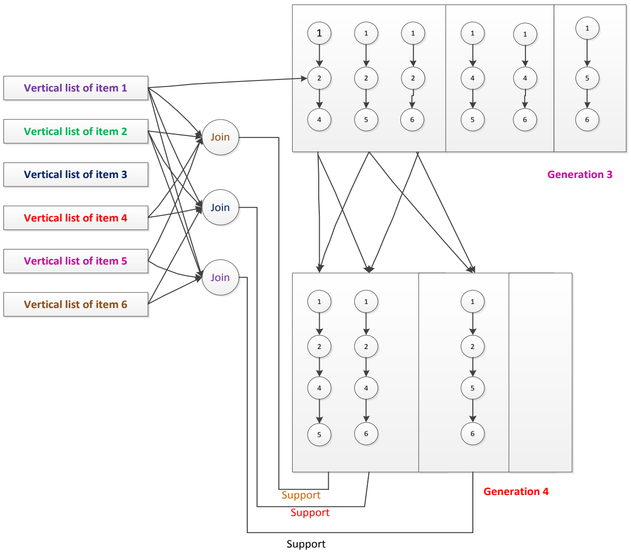


Figure 4 Support counting on GPU

As shown in Figure 4, it is evident that the processing of support counting is done simultaneously on GPU leveraging its parallel processing capabilities. This approach is known as complete intersection method. When compared to equivalent class clustering, the complete intersection approach is computationally complex but it reduces memory operations. The computational cost is negligible given the processing power of GPU.

D. Support Counting in OpenCL

As OpenCL is one of the GPGPU platforms that help in developing applications targeted for leveraging the parallel power of GPU, it is used for implementing modified Apriori. It is used along with its Java extensions. The computations in this platform are done using threads which are in turn organized into blocks with respect to GPU. Figure 5 shows how multiple threads organized into a block are involved in support counting.

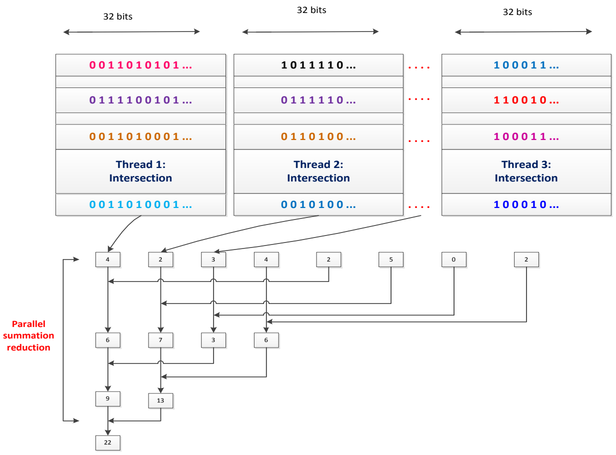


Figure 5 Support counting by threads

As can be seen in Figure 5 it is evident that the support counting is done simultaneously as expected in GPU. For the purpose of parallel summation reduction an algorithm proposed in [1] is modified and used. All the support values obtained from multiple threads are added into a single element. The result count is added to graphics memory and then sends back to the traditional memory. How programs targeting GPGPU platforms achieve coordination between main memory and graphics memory is an important consideration here. This has been explained in our previous paper. The robust coordination between two kinds of memory helps in leveraging parallel processing power of GPU.

4. Experimental Results

For making experiments the hardware environment used include Dell next-generation Intel Xeon processor with life cycle controller which can help in advanced system management. The system is connected to Tesla server to leverage its GPU processing power.

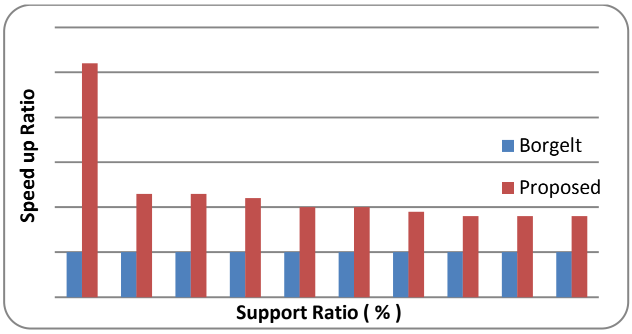


Figure 6 Performance comparison for synthetic dataset

As can be seen in Figure 6 it is evident that the speed up ratio of proposed Apriori is more when compared with that of BorgeltApriori.

Three datasets are taken from UCI machine learning repository [3] while one is synthetic dataset which has been taken from IBM Almaden Quest Research Group. The datasets obtained from UCI machine learning repository are accidents, chess and pumsb. The results show the performance comparison between the proposed Apriori for GPU and the BorgeltApriori [2] for CPU.

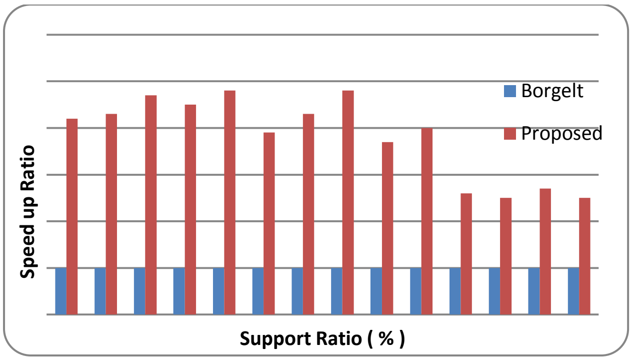


Figure 7 Performance comparison for synthetic dataset

As can be seen in Figure 7, it is evident that the speed up ratio of proposed Apriori is more when compared with that of BorgeltApriori.

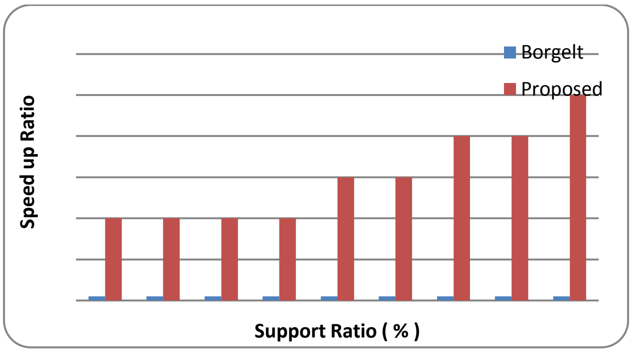


Figure 8 Performance comparison for synthetic dataset

As can be seen in Figure 8, it is evident that the speed up ratio of proposed Apriori is more when compared with that of BorgeltApriori.

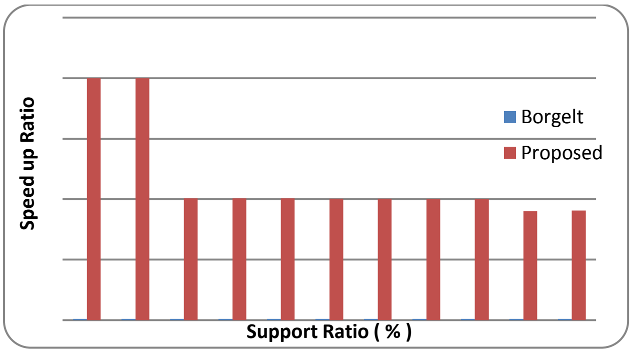


Figure 9 Performance comparison for synthetic dataset

As can be seen in Figure 9 it is evident that the speed up ratio of proposed Apriori is more when compared with that of BorgeltApriori.

Conclusions and Future Work

In this paper we studied GPGPU platforms for Apriori-based data mining algorithms. Based on Apriori, we proposed a new data mining algorithm named “HSApriori” that can exploit the parallel processing nature of modern GPU. We studied both tidset and bitset representations of dataset and identified the bitset representation is suitable for parallel processing. Another important aspect is support counting which is different for CPU and GPU based computing. We described the mechanism for support counting in GPU computing environment where multiple threads are used. Our algorithm is tested using a prototype application. We used datasets obtained from UCI machine learning repository along with a synthetic dataset which was taken from IBM Almaden Quest Research Group. The results revealed that the speed up ratio is substantially more with HSApriori which is compared with traditional HorgeltApriori. Our future work is to explore cloud computing, GPU and MapReduce programming for high speed association rule mining.

References

[1]. NVidia. Data Parallel Algorithm in CUDA SDK Available from: http://developer .download.nvidia. com.

[2]. C. Borgelt. Efficient Implementations of Apriori and Eclat. In Proc. FIMI. 2003.

[3]. UCI machine learning repository

[4]. J. Han, H. Pei, and Y. Yin. Mining Frequent Patterns without Candidate Generation. in SIGMOD. 2000. p. 1-12

[5]. M. J. Zaki and K. Gouda. Fast Vertical Mining Using Diffsets. in Proc. SIGKDD. 2003. p. 326-335

[6]. R. Agrawal and H. Mannila, Fast Discovery of Association Rules, in Advances in Knowledge Discovery and Data Mining. 1996. p. 307-328.

[7]. N. Govindaraju and M. Zaki, Advances in Frequent Itemset Mining Implementations, in FIMI. 2003.

[8]. F. Bodon, ATrie-based APRIORI Implementation for Mining Frequent Item Sequences, in OSDM. 2005. p. 56-65.

[9]. C. Borgelt. Efficient Implementations of Apriori and Eclat. In Proc. FIMI. 2003.

[10]. R. Agrawal and R. Srikant. Fast algorithm for mining association rules. in VLDB 1994. p. 487-499

D.William Albert, is a Research Scholar in Mahatma Gandhi University, Meghalaya in Data Mining specialization. Basically a B.Sc. Science Graduate and did his Post-graduation Masters in Computer Science in 2004 and Master of Technology in Computer Science & Engineering in 2006. He is a Life Member of profession body such as **I**ndian **S**ociety for **T**echnical **E**ducation, New Delhi. More than 15 National & International technical papers published and guided M.Tech, M.C.A. and B.Tech Students in their Project works. The area of interests includes Data Warehousing & Mining, Software Engineering, Software Testing Methodologies, Software Project Management, Distributed Operating System, Human Computer Interface, E-commerce, Enterprise Resource Planning, etc.

Dr.K.Fayaz is working as a faculty of system (System In-charge) in Sri Krishnadevaraya Institute of Management (SKIM), SriKrishnadevaraya University, Anantapuramu. He acquired B.Sc., from S.K.University, Post Graduate Diploma in Computer Application from JNTU, Anantapuramu and Master of Information Technology from Manipal Academy of Higher Education (MAHE) Deemed University. Manipal. Subsequently has done M.Phil and Ph.D from S.K.University in Computer Science & Technology. He has published more than 10 papers in Computer Science International Journals. He has attended and presented papers in several International, National Conferences, Seminars, Workshops.

D.Veerabhadra Babu graduated in Master of Computer Applications in 2005 and Master of Technology in Information Technology in 2011 from K.S.O.University, Karnataka and is a Research Scholar in Mahatma Gandhi University, Meghalaya in Data Mining specialization. The area of interests includes Data Warehousing & Mining, Software Engineering, Software Testing Methodologies, Software Project Management, Human Computer Interface, Management Information Systems, etc