International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Index Copernicus Value (2013): 6.14 | Impact Factor (2015): 6.391

Improving Apriori Algorithm Using Shuffle Algorithm

Aasma Parveen[1], Shrikant Tiwari[2]

Department of Computer Science and Engineering, Faculty of Engineering and Technology, Shri Shankaracharya Technical Campus, Junwani, Bhilai, District-Durg, Chhattisgarh-490020, India

Abstract: ***Data mining is the method of extracting interesting (non-trivial, embedded, previously indefinite and potentially useful) in sequence or patterns from large information repositories . Association mining aims to extract frequent patterns, interesting correlation, association or untailored structures among the sets of objects in the transaction files or from the other data repositories. It plays a vital role in spawning frequent item sets from large transaction databases. The discovery of remarkable organization relationship among business transaction records in many commercial decision making method such as catalog decision, cross-marketing, and loss-leader analysis. It is also used to excerpt hidden information from huge data sets. The Association Rule Mining algorithms such as Apriori, FP-Growth wants repetitive scans over the entire file. All the input/output overheads that are being generated during the frequent perusing process, entire file decreases the performance of CPU, memory and I/O overheads. In this paper we have proposed An Cohesive tactic of Parallel Processing and ARM for mining Association Rules on Generalized data set that is basically altered from all the preceding algorithms in that it use database in transposed form and database rearrangement is done using Parallel rearrangement algorithm (Shuffle Transpose) so to generate all important association rules number of passes essential is abridged and equaled various classical Association Rule Mining algorithms and topical procedures. The proposed Apriori algorithm has decreased the time complexity, by reducing the the processing time of the Transposition of the data sets. The comparison is done between the sequential and shuffle transposition using apriori algorithm which indicates the time difference of 28 seconds when the 100 X 100 matrix is considered, which was a very important aspect of the work. In the future, the work can be extended by parallelizing the algorithm for a communal nothing multiprocessor machine.***

Keywords: Data Mining, Association Rule Mining (ARM), Association rule, Apriori algorithm, Frequent patterns

1. Introduction

Data mining is the technique to extract interesting (non-trivial, embedded, previously hazy and potentially useful) information or pattern from huge information repositories such as: relational database, data warehouses, XML repository, etc. Also data mining is recognized as one of the core processes of Knowledge Discovery in Database (KDD). Many people get data mining as a synonym for another popular term, Knowledge Discovery in Database (KDD).

Association basically deals with exploring to the association among data elements from the massive amount of data**.** Association rule mining is engaged to solve troubles in marketing place viz., market basket enquiry. This helps in understanding the buying strategy of the customers. **.** Association Rule of Mining was firstly introduced by Agrawal in the year 1993. According to the statement "Let I = {i 1,i 2,…..i n} be a set which has n binary attributes called as items. Let D = {t1,t2,…..tn} be a set which have n transactions called as database. Each and Every transaction in D has a different transaction ID and it contains a subset of all the items in I. An implication of the form X ® Y, where X, Y Ì I, and X ÇY = Ærepresents a rule. The sets of items (X and Y) on the left hand side of the rule are known as antecedent and items on the right hand side of the rule are known as consequent ".

a) Privacy Preserving

The Privacy preserving of the data must be the safeguard from the divulging sensitive of the data during the publication of individual of data. To the maintained to their privacy, a number of the techniques have been proposed by the modifying or transforming of the data. To avoid the data misuse, the data are anonymized. Many data mining techniques are modified to ensure their privacy. The techniques for the PPDM are based on the cryptography; data mining and information are hiding. In general on the statistics-based along with the crypto-based approach is used to tackling about PPDM.

The genetic algorithm, which is subclasses of the evolutionary of algorithms, is an optimization to technique inspired by the natural evolution, that is, mutation, inheritance, selection and crossover to be gain optimal solutions. Lately, it was enhanced by using a novel multi-parents crossover and a divers the operator instead of mutation in order to gain quick convergence, utilizing it in conjunction with several features are selection techniques, involving principle of components analysis, sequential floating, and correlation-based feature selection, using the controlled to elitism and dynamic crowding distance to present a general algorithms for the multi-objective optimization of the wind turbines, and utilizing a real encoded crossover and mutation operator to be gain the near global optimal solution of multimodal nonlinear of optimization problems.

b) Apriori Algorithm

A extraordinarily prominent association rule mining algorithm, Apriori, has been developed for rule mining in large transaction databases. Many additional algorithms developed are derivative and/or extensions of this algorithm. A major step forward for improving the performances of these algorithms was made by the introduction of a novel, compact of the data structure, referred to as frequent patterns tree, or FP-tree, and the associated mining algorithm, FP-growth. The main difference flanked by the two approach is

that the Apriori-like techniques are based on the bottom-up generation of frequent item set combination and the FP tree based ones are partition-based, divide-and-conquer methods.

c) Approaches for association rules

The numbers of transactions required for the item set are satisfy of minimum supports therefore referred to be as the minimum supports of a count. If an items set satisfies of minimum support, then it is a frequent item set (frequent pattern).The most common approaches to searching as association rules are break up the problem into two parts:

1) Find to all frequent items sets: By the definition, each and every of these item sets will be occurs at least as frequently a pre-determined minimum support of count.
2) Generate the strong association's rules from the frequent items sets: By the definition, these rules are must be satisfy the minimum support and minimum confidence. Additional interestingness measures can be applied it, if desired as the second step is the easier of the two. The overall performance of the mining association rules is indomitable by the initial step. As shown in, the performance, for largest databases, are as the most influenced by the combinatorial explosion of the number of the possible frequent item sets that must be considered and also by the number of the database scans that has to be performed.

2. Related Work

Agrawal et al. (1993) first proposed the issue of the mining association rule. They pointed out that some hidden relationships exist between purchased substance in transactional databases. Therefore, mining results can help decision-makers understand customers' purchasing performance. An association rule is in the form of X→Y, where X and Y represent Item set (I), or products, respectively and Item set includes all possible items{i1,i2, . . .,im}. The general transaction database (D= {T1, T2, . . .,Tk}) be able to represent the prospect that a customer will buy product Y after buying product X

Alatas et al. (2008) use Pareto based multiobjective differential evolution (DE) algorithm as a search strategy for mining accurate and graspable numeric association rules, which can be optimal in the wider sense that not her rules are superior to them when all objectives are concurrently considered. They formulated the association rule mining problem as a four-objective optimization difficulty, where sustain, confidence and comprehensibility of rules are maximized, while the amplitude of the intervals, which confirms the item set and rule is minimized.

Gupta et al (2012) used weighted particle swarm optimization (WPSO) for association rule mining for finding the suitable threshold values for minimum hold and minimum assurance. These parameters are used for extracting the valuable information. Nandhini et al. (2012) proposed association rule pulling out algorithm using PSO and domain ontology. They concluded that combining PSO with domain ontology interactively, decreased the number of rules generated without compromising the quality of rules.

Toivonen et al.(1996) presented sampling algorithm at1996. This algorithm is about finding association rules according to reduce database operations.Bender search algorithm (1998) by Lin et al. developed at 1998 can discover rules from most frequent item sets. Yang et al. (2012) offered an efficient hash based method named HMFS which combines DHP and Bender search algorithms results in reduction of database scan and filtering repeated item sets to find greatest repeated item set.

Gupta M. (2011) also offered a method for automatic finding of threshold value using weighted PSO. His results illustrate high effectiveness of PSO for associative rule mining. This approach also can gain better values of threshold in comparison with previous ones. Kuo et al. (2011) developed a PSO based methodfor automatic finding threshold value of minimal uphold. Their exertion show that basic PSO can find values faster and better than genetics algorithm

Gilan Atlas et al [23] proposed a multi-objective differential evolution algorithm for mining numeric association rules. Later, they proposed another numeric association rule mining method using rough particle swarm algorithm which had some improvements in performance and precision compared to the previous one. They also proposed another numeric association rule mining method chaos rough particle swarm algorithm

M. Kaya, et. Al. [24] The method uses an extension of elaborate encoding while relative confidence is the fitness function. A public search is performed based on genetic algorithm. As the method does not use minimum support, a system automation procedure is used instead. It can be extended for quantitative-valued association rule mining. In order to improve algorithm's efficiency, it uses a generalized FP-tree. Evaluation of the algorithm shows a considerable reduction in computational cost. Just interesting rules with constant length are discovered. In this method, the genes contain rank of fields. Final chromosome should be the best one and the process stops if it reaches the predefined number of iterations or the result is not improved. The fitness function is defined such a way that it stays in local optimum and causes many rules to be generated. Kaya proposed genetic clustering method.

B. J. Roberto.et. al. (2012) have worked in the DLG algorithm, logic AND operations are used to count the support. Both algorithms avoid generating candidate item sets, which is a very time consuming operation. Roberto proposed a Max-Miner algorithm [25] for efficiently identifying long frequent item sets which, in turn, can be used to generate other frequent item sets.

Park et al. proposed the DHP (standing for Direct Hashing and Pruning) algorithm [26] for efficient generation of frequent item sets. The DHP is a hash-based algorithm and is especially effective for the generation of frequent 2 − item sets. Based on the Apriori algorithm, the DHP algorithm uses an efficient approach to trim the number of candidate 2 − item sets. As a result, the number of candidate 2−item sets generated by the DHP algorithm is much smaller than those generated by the Apriori and the Apriori Tid algorithms.

3. Problem Identification

There were some problems which have been identified from the previous works while researching about the topic. Those are explained in brief.

a) Algorithmic issues

In the association rule for the field of mining, the most of the innovations were in the first place to enhance the algorithmic performance and in to the second place are diminish to the output set by allowing all the opportunity to express the constraints on the desired grades. Over the past decade, many algorithms that address all these problems through the refinement of the investigation strategies, pruning methods and the data structures have been urbanized. While many of the algorithms focus on the unambiguous discovery of all the regulations that satisfy nominal support and buoyancy constraints for a particular dataset, the increasing contemplation is being given to the particular algorithms trying to enhance the processing time or facilitate user analysis by reducing to the effect set size and by incorporating in the domain knowledge.

b) Data Processing issues

This issue addresses which are parts of the data streams are chosen to apply association rule mining. The data streams contain ordered series of items. Each and every set of items are normally called "transaction". The problem of the data processing model here is to find an approach to extract the transactions for the association rule mining process from the overall data streams. Because of these data streams coming continuously and unboundedly, the extracted transactions are getting altered from time to time. According to the researches, there are totally three kinds of stream data processing models, Damped Landmark, and Sliding Windows.

c) Issues regarding Memory Management

The next fundamental issue we needed to consider as how to optimize the memory space frenzied when running into the mining algorithm. This includes the way how to make a decision about the information i.e. we must also collect the data from the data streams and how to select a compact in-memory to the data structure that allows to the information to be stored, retrieved and updated efficiently. Fully addressing these challenges for mining algorithm can greatly enhance its performance.

d) Compact Data Structure

An efficient and compact the data structure are needed to be store, update and retrieving to the collected information. This is due to the bounded memory size and huge amounts of the data streams are coming continuously. Failure in to the developing such as the data structure will largely decrease the efficiency of the mining algorithm because, even if we store the information in to the disks, the additional I/O operations will increase in to the processing time. The data structure needs to be incrementally maintained since it is not possible to rescan entire whole input due to the huge amount of the data and requirement of the rapid online querying speed.

Different data stream application situation may have diverse to needs for an association rule mining algorithm. The Timeline Query rivulet the data, which come constantly over time. In some of the applications, user may be paying attention in triumph association rules, which are completely based on the data available throughout a fixed period of the time

4. Methodology

As we seen in our problem identification section topical algorithm is efficient then classical apriori algorithm but MESH Transpose distributed algorithm is not optimal which is used by topical algorithm. Mesh Transpose drawback can be overcome by Shuffle Transposition.

5. Proposed Algorithm

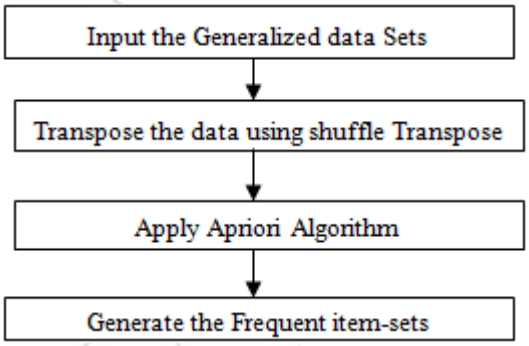

Figure 1: Flowchart of the methodology

The methodology which has been proposed for the solution of the problems identified in the project is as shown in the Figure 1.

Step 1:- Input the Generalized data Sets

- Consider Generalised data sets (GDS) as inputs in the proposed method.
- In the supervised kind of learning applications in machine learning and statistical learning theory, simplification mistake (also known as the out-of-sample error) is a measure of how accurately an algorithm is able to envisage outcome standards for previously unseen data. Because learning algorithms are evaluated on limited samples, the assessment of a learning algorithm may be sensitive to sampling error. As a result, measurements of prediction mistake on the present data may not provide much information about predictive ability on new data. The Generalization fault can be reduced by avoiding over fitting in the learning algorithm. The performance of a machine learning algorithm is measured by the plots of the generalization error values through the learning process and they are called learning curve.
- A data set (or dataset) is a collection of data. Most commonly a data set corresponds to the contents of a single database table, or a single statistical data matrix, where every column of the table represents a particular variable, and each row corresponds to a given member of the data set in question.

- Normally it's a collection of transaction of items and the entries consisting of attributes known as items in the form of rows and columns.

Step 2:- Transpose the data using shuffle Transpose

- Mesh Transpose distributed algorithm is not optimal which is used by topical algorithm. Mesh Transpose drawback can be overcome by Shuffle Transposition
- Consider a processor index k contains 2q bits. If k = 2q(i - 1) + (j - 1), then the q most significant bits of k represent i - I while the q least important bits represent] - 1. This is given in Figure 3(A). for q = 5, i = 5, andj = 12. After the processing of q shuffles (i.e., q cyclic shifts to the left), the element usually held by P_k will be in the processor whose index is

$$s = 2^q(j - 1) + (i - 1)$$

- As shown in Figure 4.2, a_{ij} has been moved to the position originally occupied by a_{ji}.

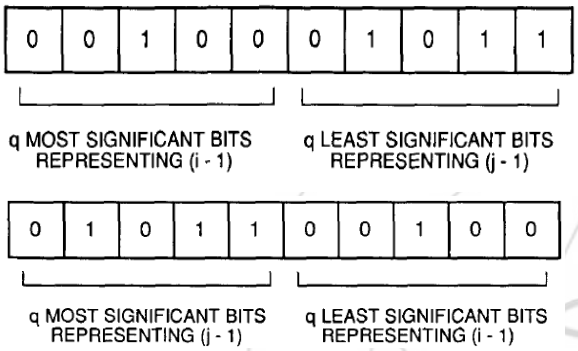

Figure 2: Derivation of number of shuffles required to transpose matrix

- The Shuffle Transpose method is a Parallel transposition method, which converts the generalized data sets to Boolean Data sets.
- Algorithm for Shuffle transpose

Procedure SHUFFLE TRANSPOSE (A)
for i= I **to** *q* **do**
for k = I **to** ***22q*** - 2 **do in parallel**
Pk sends the element of A it presently holds to
P2kmod(22q- 1)
end for
end for
end

Step 3:- Apply Apriori Algorithm

- The AprioriAlgorithm is an powerful algorithm for taking out the frequent item sets for the boolean union rules. Apriori uses a "bottom up" procedure, where the frequent subsets are extended on one article in a single time (a step known as candidate generation, and the groups of candidates are experienced against the data.
- The frequent item sets determined by Apriori can be used to determine association rules which highlight general trends in the database: this has applications in domains such as market basket analysis.
- Specify the Support count which is a threshold value and also notify the number of rows and columns

Procedure EPTA()
1.Shuffle Transpose(DataSet)//Transpose the transactional database
2. Read the database to count the support of C1 to determine L1 using sum of rows.
3. L_1= Frequent 1- itemsets and k:= 2
4. While (k-1 ≠ NULL set) do
Begin
C_k: = Call Gen_candidate_itemsets (L_k-1)
Call Prune (C_k)
for all itemsetsi□ I do
Calculate the support values using dot-multiplication of array;
L_k := All candidates in Ck with a minimum support;
k:=k+1End
5. End of step-4
End Procedure

Procedure SHUFFLE TRANSPOSE (A)
fori= I **to** *q* **do**
fork = I **to** ***22q*** - 2 **do in parallel**
Pk sends the element of A it currently holds to
P2kmod(22q- 1)
end for
end for
End

Procedure Gen_candidate_itemsets (Lk-1)
Ck= Φ
for all itemsets $I_1 \in L_{k-1}$*do*
for all itemsets $l_2 \in L_{k-1}$*do*
if $I_1[1] = I_2[1] \wedge I_1[2] = I_2[2] \wedge \ldots \wedge I_1[k-1] < I_2[k-1]$ then
c = $I_1[1], I_1[2] \ldots I_1[k-1], I_2[k-1]$
Ck = Ck□ □ {c}
End Procedure

Procedure Prune(Ck)
*for*all c ∈Ck
*for*all (k-1)-subsets d of c do
*if*d ∉Lk-1
*then*Ck= Ck– {c}
End Procedure
SHUFFLE TRANSPOSE

Consider a processor index k consisting of 2q bits. If k = 2q(i - 1) + (j - 1), then the q most important bits of k represent i - I while the q least significant bits represent] - 1. This is illustrated in Figure 3(A).for q = 5, i = 5, and j = 12. After q shuffle (i.e., q cyclic shifts to the left), the element originally held by P_k will be in the processor whose index is

$$s = 2^q(j - 1) + (i - 1)$$

Step 4:- Generate the Frequenct item-sets

- Frequent sets play an essential role in many Data Mining tasks that try to find interesting patterns from databases, such as association rules, correlations, sequences, episodes, classifiers and clusters.
- The mining of association rules is one of the most popular problems of all these. The identification of sets of items, products, symptoms and characteristics, which often occur together in the given database, can be seen as one of the most basic tasks in Data Mining.

6. Results

Based on the methodology discussed, we did got some extreme good and improved results while comparing with

the previous works, which have been shown with the help of table , screenshot and Graph.

Table 5.1: Comparison between proposed algorithm and typical Apriori algorithm

Data Size	Running Time of proposed Algorithm (in sec)	Running Time of Typical Apriori Algorithm (in sec)
5x5	0.002	0.0015
8x8	0.024	0.166
10x10	0.0957	0.429
20x20	0.1210	40.334

The Comparison of the previous method and the present method on the basis of the Parameters is shown with the help of Graph in the following figure 3.

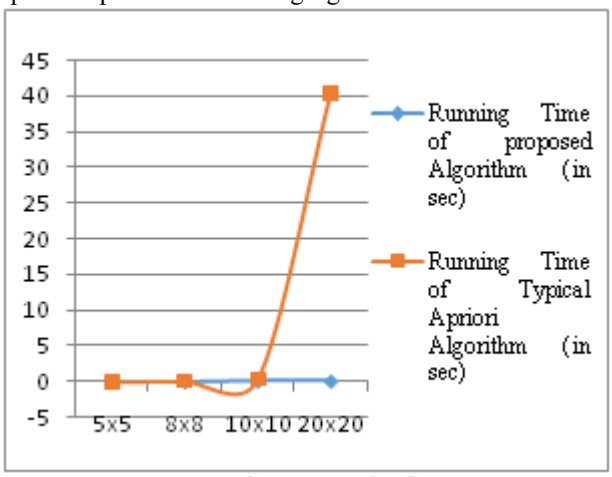

Figure 3: Comparison between typical and Proposed Algorithm

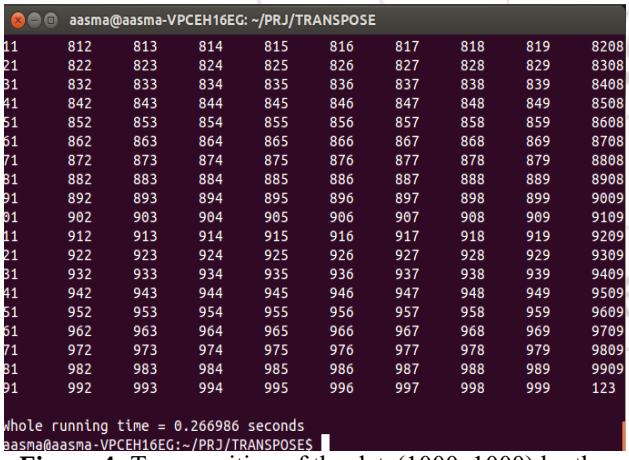

Figure 4: Transposition of the data(1000x1000) by the proposed method on the basis of the Shuffle Transpose.

7. Conclusion and Future Scope

The proposed Apriori algorithm has decreased the time complexity, by reducing the the processing time of the Transposition of the data sets. The comparison is done between the sequential and shuffle transposition using apriori algorithm which indicates the time difference of 28 seconds when the 100 X 100 matrix is considered, which was a very important aspect of the work. The CPU overhead has been reduced while comparing with the earlier works. The Efficiency and the accuracy of the apriori algorithm also has been increased. The work can be carried out with the help of EREW Transpose, which can decrease the time difficulty upto a large extent and even it will be cost efficient. With the accomplishment of our algorithm, it is interesting to revise and explore many related issues, extensions and application, like iceberg cube computation, classification and clustering. In the future, the work can be extended by parallelizing the algorithm for a communal nothing multiprocessor machine.

References

[1] Agrawal. R., and Srikant. R., "Fast Algorithms for Mining Association Rules", Proceedings of 20th International Conference of Very Large Data Bases. pp.487-499,1994.

[2] Jong Park, S., Ming-Syan, Chen, and Yu, P. S. "Using a Hash-Based Method with transaction Trimming for Mining Association Rules". IEEE Transactions on Knowledge and Data Engineering, 9(5), pp.813-825,1997.

[3] M.H.Margahny and A.A.Mitwaly, "Fast Algorithm for Mining Association Rules" in the conference proceedings of AIML, CICC, pp(36-40) Cairo, Egypt, 19-21 December 2005.

[4] Y.Fu., "Discovery of multiple-level rules from large databases", 1996.

[5] F.Bodon, "A Fast Apriori Implementation", in the Proc.1st IEEE ICDM Workshop on FrequentcItemset Mining Implementations (FIMI2003, Melbourne,FL).CEUR Workshop Proceedings 90, A acheme, Germany 2003.

[6] Akhilesh Tiwari, Rajendra K. Gupta, and Dev Prakash Agrawal, "Cluster Based Partition Approach for Mining Frequent Itemsets" in the International Journal of Computer Science and Network Security(IJCSNS), VOL.9 No.6,pp(191-199) June 2009.

[7] JiaWei Han Micheline Kamber."Data Mining:Concepts and Techniques"[M].Translated by Ming FAN, XaoFeng MENG etc. mechanical industrial publisher,BeiJing,2001,150-158.

[8] M.J. Zaki. "Scalable algorithms for association mining". IEEE Transactions on Knowledge and Data Engineering, 12 : 372 –390, 2000.

[9] JochenHipp, Ulrich G¨untzer, GholamrezaNakhaeizadeh. "Algorithms for Association Rule Mining – A General Survey and Comparison".ACM SIGKDD, July 2000, Vol-2, Issue 1, page 58-64.

[10] Sotiris Kotsiantis, Dimitris Kanellopoulos. "Association Rules Mining: A Recent Overview". GESTS International Transactions on Computer Science and Engineering, Vol.32 (1), 2006, pp. 71-82.

[11] S. Brin, R. Motwani, J. D. Ullman, AND S. Tsur, "Dynamic itemset counting and implication rules for market basket data", SIGMOD Record 26(2), pp. 255–276, 1997.Kim Man Lui, Keith C.C. Chan, and John TeofilNosek "The Effect of Pairs in Program Design Tasks" IEEE transactions on software engineering, VOL. 34, NO. 2, march/april 2008.

[12] Eui-Hong Han, George Karypis, and Kumar, V. Scalable "Parallel Data Mining for Association Rules".

IEEE Transaction on Knowledge and Data Engineering, 12(3), pp.728-737, 2000.

[13] Sanjeev Kumar Sharma &Ugrasen Suman "A Performance Based Transposition Algorithm for Frequent Itemsets Generation" International Journal of Data Engineering (IJDE), Volume (2) : Issue (2) : 2011

[14] Dr (Mrs).Sujni Paul "An Optimized Distributed Association Rule Mining Algorithm In Parallel and Distributed Data Mining With Xml Data For Improved Response Time".International Journal Of Computer Science And Information Technology, Volume 2, Number 2, April 2010

[15] ManojBahel, ChhayaDule "Analysis of frequent item set generation process in Apriori& RCS (Reduced Candidate Set) Algorithm" National Conference on Information and Communication Technology, Banglore April 2010

[16] Sedukhin, S.G.; Zekri, A.S.; Myiazaki, T."Orbital Algorithms and Unified Array Processor for Computing 2D Separable Transforms" Parallel Processing Workshops (ICPPW), 2010 39th International Conference Page(s): 127 – 134.