ISSN 2348-1196 (print)
International Journal of Computer Science and Information Technology Research ISSN 2348-120X (online)
Vol. 3, Issue 2, pp: (114-119), Month: April - June 2015, Available at: www.researchpublish.com

MULTI DIM SEQ-DATA MINING WITH A PARALLEL APPROACH

[1]Prof. Manjitsing Valvi, [2]Rinit Lathia, [3]Ronak Shah, [4]Tejas Sampat

[1,2,3,4]Department of Information Technology, K.J. Somaiya College of Engineering, Vidyavihar, Mumbai, India

Abstract: **Algorithm PTPSPM (a parallel algorithm based on prefix tree for sequence pattern mining) is proposed in order to deal with the speed limited and effectiveness problem of the sequence pattern mining in massive data. In this paper, a new prefix-tree structure and an improved prefix-span algorithm are introduced to mine the local sequence, the global sequence are obtained by merging all the local sequences. A new prefix tree pruning technique is presented to delete the global k-sequence which cannot be attended. PTPSPM algorithm applies project database identifier index table of dynamic scheduling to avoid the processor idle waiting. Additionally, it cites selective sampling techniques to balance the loads between processors. The experiment results demonstrate that PTPSPM algorithm has better execution performance and speedup.**

Keywords: **sequence mining; parallel mining; prefix-tree; local sequence; global sequence.**

I. INTRODUCTION

Sequence pattern mining, plays an important role in analysis of shopping sequences, discovery of DNA sequences pattern, analysis of network access mode, and soon.[1] Parallel mining algorithms have high performance in massive data mining.[2] Sequences are an important type of data which occur frequently in many scientific, medical, security, business and other applications.[1] For example sequences can be used to capture how individual humans behave through various temporal activity histories such as weblogs and customer purchase histories.[1] Sequences can also be used to describe how organizations behave through sales histories such as the total sales of various items over time for a supermarket, etc.[1]

Huge amounts of sequence data have been and continue to be collected in genomic and medical studies, in security applications, in business applications, etc.[1] In these applications, the analysis of the data needs to be carried out in different ways to satisfy different application requirements, and it needs to be carried out in an efficient manner. [1]

The mined sequence patterns will be given as output. We assume the system to provide analysis of purchase patterns for better sales forecasting. Accordingly based on the mining result decisions can be made based on how customers purchase their products.

The rest of this paper is organized as follows: we describe definition of the parallel sequence pattern mining; the parallel sequence pattern mining algorithm based on prefix-tree[2]; Experiment result is given; and finally the conclusion.

Parallel System used for mining:

Assume that there is a parallel system which has n-network interconnects sites: S1, S2,...,Sn, and each site are a stand-alone computer, S represents the sets of all sites, S= {S1, S2… Sm}. The data sequences on site Si (i=1,2, ..., m) which is expressed with dbi(i=1,2,…,m), where DB represents all the data sequences, dbi ⊆ DB ， db1_db2_…_dbm=DB and db1∩db2∩…∩ dbm=NULL. The data sequences on the site are expressed with (ID Sequence), ID represents sequence identity and Sequence is a raw sequence data. On the site Si, the minimum support count is mincounti=|dbi|×minsup, where minsup is the user-defined minimum support. Global minimum support countmin count =|DB| × minsup =

ISSN 2348-1196 (print)
International Journal of Computer Science and Information Technology Research ISSN 2348-120X (online)
Vol. 3, Issue 2, pp: (114-119), Month: April - June 2015, Available at: www.researchpublish.com

(|db1|+|db2|+|dbm|) ×minsup. The number of sequence which contains on site Si known as the local count of s on site Si, denoted as counti(s), if counti(s) ≥mincounti, then sequence s is a local

sequence pattern on site Si. The total counts for sequence s is count (s), if count(s) ≥mincount, then s is the global sequence. Obviously, count(s) =Σ=ni 1 counti(s) . For sequence s and α, S_α means S connects with α,containing item extension and sequence extension,expressed as _ i and _ s.

Sequence pattern of length k is called k-sequence pattern. Suppose F (k) is the global set of k-sequence patterns, for any s∈F (k), count(s) ≥ minsup. For the same reason, suppose Fi (k) is the local set of k-sequence patterns on siteSi, for any s∈Fi (k), counti(s) ≥mincounti.

Definition 1 the structure of Prefix tree:

The prefix tree is consisted of all the global sequence patterns that meets the minimum support. Each node in the tree maintains a triple (item, count, and branch), *item* denotes the last item of a global sequence, *count* is the support of sequence and *branch* is a branch of the tree pointing to its child. In the prefix tree, there are two kind of *branch*, the branch with a dashed line means that the child node is the item extension of its parent, it is the sequence extension of its parent if the branch with a solid line. Each branch (from the root node to a leaf node) represents a candidate sequence.

Here is the sequence database that shows in Table 1:

Table 1

ID	Sequence
1	<(af)(d)(e)(a)>
2	<(e)(a)(b)>
3	<(e)(abf)(bde)>

The prefix-tree is constructed in Figure 1. For example, the dashed line and f2 in level 2 represents 2-sequence pattern item extension <(af)>, where the count is 2; the solid line and d2 in level 3 represents 3-sequence pattern sequence extension <(af)(d)>, the corresponding count is 2.

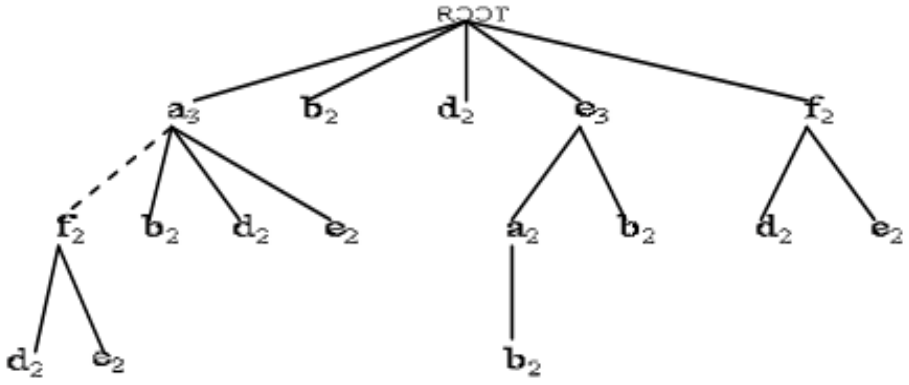

Figure 1. prefix-tree

Definition 2 (Lk sub tree):

In the prefix tree, the root is in level 0, the first level of the tree are L1 sequence and Li sequence is stored in the ith level of the tree(i=2,3,..,n). The sequence tree under the L1 sequence can be divided into multiple sub-trees, called L1 sub-tree, accordingly, the sub tree which can be divided from sequence of length k are called Lk sub tree. The sub trees formed by the local sequence patterns of every site are called local sub tree, and global sub tree is consist of all global sequence patterns. For instance, in figure 1, the second level of the tree represents five L1 sequence pattern(a3,b2,d2,e3,f2), it can

be divided into five sub-tree{(a3),(b2),(d2),(e3),(f2)}. A Parallel Algorithm Based on prefix tree for Sequence Pattern Mining In the parallel system, the implementation of the algorithm mainly composes of two processes on all sites, the main process is responsible for generating global sequence patterns and controlling the mining progress. And the sub-process is responsible for generation of local sequence pattern, data transmission and further implementation of the global sequence pattern mining. The difference between the host node and other nodes is that the host needs to collect the entire local sequence patterns which mined by other nodes, and to output the final result.

II. RELATED WORK

Agrawal proposed GSP which presented time constraints, sliding time window and user-defined taxonomies to decrease the number of sequences and to reduce the overhead [3][4]. Shintani proposed three parallel tactics based on GSP(Generalise System of Preferences): NPSPM(Non Partitioned Sequential Pattern mining), SPSPM(Simple Partitioned Sequential Pattern mining) and HPSPM(Hash Partitioned Sequential Pattern mining), as the hash mechanism was used in HPSPM, it has the best performance, and is better than the first two algorithms[5]. But they all need to scan the database for many times and to exchange remote database partition which result in greater communication overhead and I/O costs.In order to address the above problems, Zaki presented pSPADE , which was based on a serial algorithm SPADE and a shared memory parallel structure, lattice theory was used to minimize I/O costs, due to the limited bandwidth of the shared memory parallel structure, the scalability may be inhibited at some point[6].In reference, Wang made a comprehensive survey on parallel frequent pattern mining technology, he pointed out the efficiency, scalability of parallel technique in massive data mining and the transformation of the parallel mining platform from distributed systems to multi-core system; Wu proposed EDMA algorithm to mine association rules, it minimized the number of candidate sets, exchanged messages by local and global pruning and reduced the scan time by decreasing the size of average transactions and datasets[8]; The PartSpan algorithm was proposed in, data parallel and task parallel were used to divide and distribute the projection database, but it was lack of the necessary load-balancing mechanism.So Zhou makes an improvement on the load imbalance in parallel sequence mining algorithm DPA, a new approach was proposed to generate a balanced workload among processors and to reduce processor idle time[8][9]. On the basis of the tree projection, Valerie presented a new parallel algorithm based on distributed storage: STPF, it has good scalability by using breadth-first approach in the static load balancing mechanism. Han proposed Par-CSP algorithm for parallel sequence pattern mining, dynamic load balancing and divide and conquer strategy were introduced to minimize the costs and to obtain better speedup. At the basic of Par-CSP, Niagara proposed an improved Par-ClosP algorithm to solve the problem of parallel closed sequence mining, it introduced a new pruning method and pseudo projection technique to minimize the use of time and space.

III. PROPOSED SYSTEM

MultiDimSeq (Multidimensional Sequential pattern mining using prefix span algorithm on parallel systems) is proposed in order to deal with the underutilization and effectiveness problem of the sequence pattern mining in massive data. In this system, an improved prefix-span algorithm is introduced to mine the local sequence, the global sequence are obtained by merging all the local sequences.[2] The prefix tree pruning technique is presented to delete the global k-sequence which cannot be attended.[2] MULTIDIMSEQ will be applied on shopping purchase histories to extract frequent purchase records. The subordinate billing systems are used to mine the local k-sequences. Additionally, it cites selective sampling techniques to balance the loads between processors. After login in the system we can mine all the frequent sequential patterns from the past purchase records. This helps us to analyze in what sequence the products are being purchased. Thus, the sales company can plan for product combos and discounts so as to benefit the customers as well as themselves creating a win-win situation for both the parties. Other functionalities from user point of view are that it provides other features of appending records to existing datasets, viewing database relations and generating reports for analysis purpose.

Objective of the system is to develop a standalone data mining application which mines the frequent sequential patterns on parallel system, for analyzing and predicting future sales of products for a shopping mart.

The application would provide features like:

• Search based on minimum support of the item in the dataset. The result varies with different minimum support values. Customized mined reports can be dynamically generated. Comparison of sequential patterns based on time, price, name and sales.

ISSN 2348-1196 (print)
International Journal of Computer Science and Information Technology Research ISSN 2348-120X (online)
Vol. 3, Issue 2, pp: (114-119), Month: April - June 2015, Available at: www.researchpublish.com

- **ADDING RECORDS TO EXISTING DATASET:**

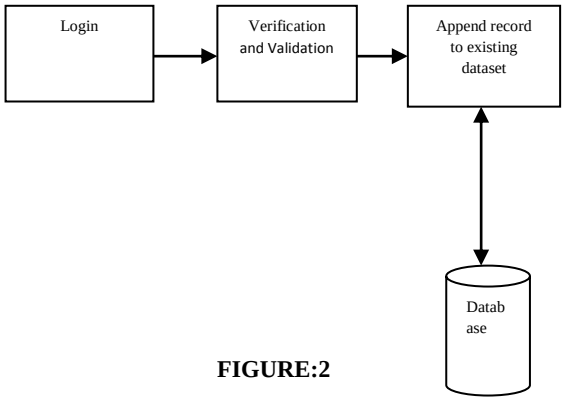

FIGURE:2

- **LOCAL SEQUENCE MINING:**

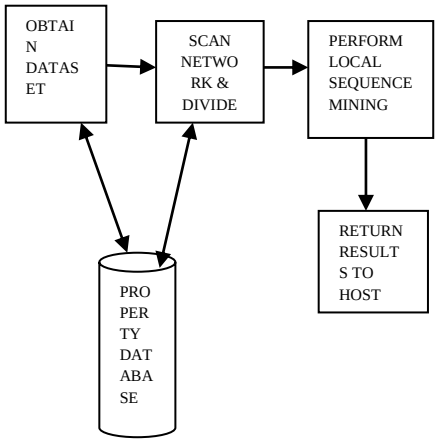

FIGURE:3

- **MERGING PHASE:**

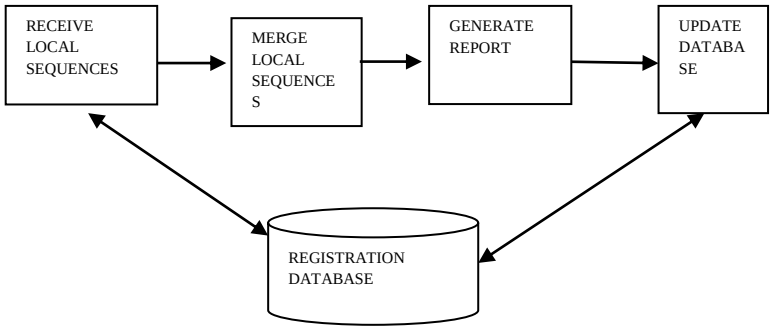

FIGURE:4

IV. FUTURE SCOPE

The main aim is to develop a standalone data mining application which mines the frequent sequential patterns on parallel system, for analyzing and predicting future sales of products for a shopping mart. The analysis will help to boost the sales of a particular product and also to attract the target audience. This will ultimately help to grow the revenue of the concerned organization.

ISSN 2348-1196 (print)
International Journal of Computer Science and Information Technology Research ISSN 2348-120X (online)
Vol. 3, Issue 2, pp: (114-119), Month: April - June 2015, Available at: www.researchpublish.com

V. EXPERIMENT RESULTS

To test the performance of PTPSPM algorithm, our experiments were performed on four computers in the local area network with the speed of 10MPS. Each computer has a 3.2GHZ Pentium _ processor and 1GB memory, with the Windows XP Professional operating system. We use the dataset from the IBM Almaden datasets center. This paper choose HPSPM as comparison since it was the most representative algorithm in parallel sequence mining, further more, to prove the effectiveness of the proposed scheme. It also used PartSpan algorithm to compare with the proposed algorithm. The experiment results are shown in Figure4 and Figure5. In Figure 4 all the execution time decreases as the number of processors increases. Whereas, PartSpan and PTPSPM decreases much than HPSPM.That is because the dynamic scheduling technique which can avoid the processor idle waiting, In particular, our proposed HPSPM is the winner in any case of distinct processors, since it introduces an improved algorithm and pruning technique and adds the selective sampling techniques to balance the loads between processors on the basis of PartSpan. As we can see from Figure 5, PTPSPM and PartSpan have good speedup than HPSPM, especially as the processors number increased for the same reasons that mentioned above.

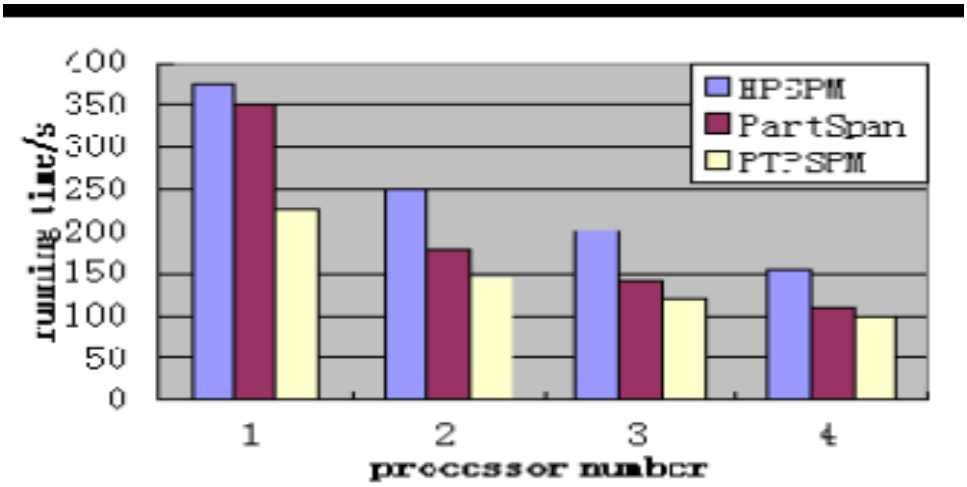

Figure4. execution time comparison of three algorithms for varying processors

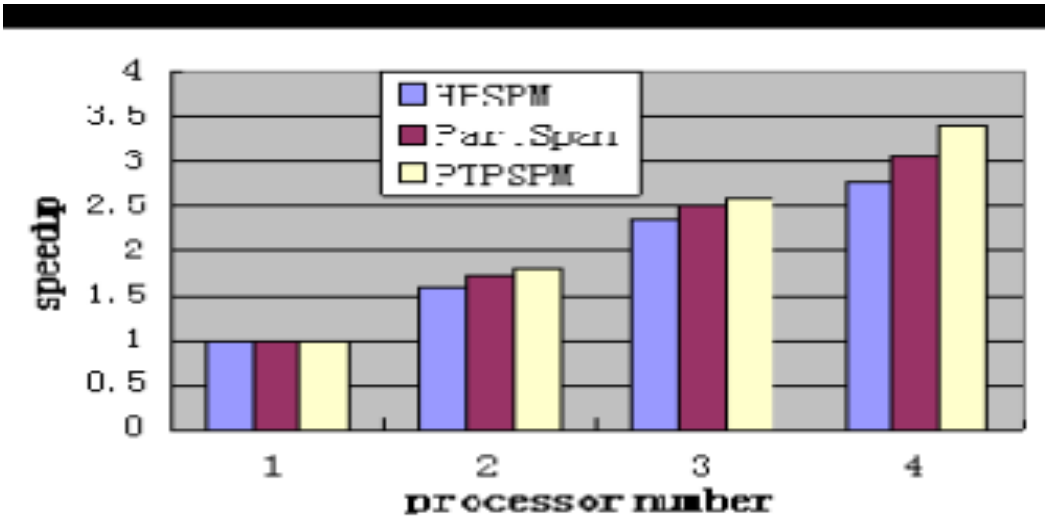

Figure5. speedup comparison of three algorithms for varying processors

VI. CONCLUSION

Thus we hope to implement the MULTIDIMSEQ algorithm to reduce the execution time of mining the required data. We also expect our algorithm to be better than the previously executed algorithms. Our algorithms can be used or implemented in shopping marts to analyze sales.

REFERENCES

[1] Sequence Data Mining, Gouzhu Dong & Jian Pei, Springer Publications, ISBN-13: 978-0-387-69936-3.

[2] A Parallel Algorithm Based on prefix tree for Sequence Pattern Mining, Jia-dong Ren, 2010 First ACIS International Symposium on Cryptography, and Network Security, Data Mining and Knowledge Discovery, E-Commerce and Its Applications, and Embedded Systems.

[3] R. Agrawal, and R. Srikant, "Mining sequence patterns," proceedings of the 11th International Conference on Data Engineering. Taipei, 1995, pp3-14.

[4] R. Agrawal, and R. Srikant, "Mining sequence patterns:Generalizations and Performance improvements," proceedings of the 11th International Conference on Extending Database Technology. Heidelberg, Springer-Verlag, 1996, pp13-20.

[5] T. Shintani, and K. Kitsuregawa, "Mining Algorithms for sequence Patterns in Parallel: Hash Based Approach," in Research and Development in Knowledge Discovery and Data Mining: Second Pacific Asia Conference (PAKDD98). Australia: Melbourne, 1998, pp283-294.

[6] Zaki, "Parallel sequence mining on share-memory machines", Journal of Parallel and Distributed Computing. vol. 61, pp401-426, 2001.

[7] W. Jianyong, Z. Lizhu, and Z.Yuzhou, "Parallel Frequent Pattern Discovery: Challenges and Methodology," Tsinghua Science and Technology. vol.12, pp719-728, 2007.

[8] W. Jian, and L. Xingming, "An efficient association rule mining algorithm in distributed database," the first International Workshop on Knowledge Discovery and Data Mining. 2008, pp108-113.

[9] Q. Shaojie, T. Changjie, D. Shucheng, Z. Mingfang, P. Jing, L.Hongjun, and K. Yungchang,, "PartSpan: Parallel Sequence Mining of Trajectory Patterns," the 1.