International Journal of Innovative Technology and Exploring Engineering (IJITEE)
ISSN: 2278-3075 (Online), Volume-9 Issue-4, February 2020

Reduction of Frequent Itemsets Mining in Big Data with the Help of FP Algorithm and Msegt-Tree

Srinivasa Rao Divvela, V Sucharita

Abstract: ***Frequent itemset mining is very crucial to minimize the cost and time of executions but when considering multiple distributed data streams in big data the frequent itemset mining has been a little cost consuming and taking more space and time complexity. In this paper we reduce the load and minimize the cost while minimizing the space and time complexities of the process by using reduction mechanism and indexing structures for preserving complexities. A 2-level architecture modal which will be helpful in handling the distributed data streams where the root node will be in level-0 and local nodes at level-1 is proposed. Each local node will evaluate the patterns in their specific data stream using the algorithm 'FP' which will help in lessening the burden on the root node and will be sent to root. With help of the patterns received from local nodes the root will generate a global pattern set.***

Key Words: ***Frequent Itemset Mining, Distributed Data Streams, Indexing Structures, Space and Time Complexity.***

I. INTRODUCTION

With technology upgrading day to day field that is used to analyze and extract information from too large or complex datasets by using data processing application software called big data is vastly used.Big data is characterized by 3V's:Volume, Variety and Velocity of data must be processed. By analyzing the past and the real time data can be used to improve their marketing strategies to satisfy the customer needs. Different branches of information found in the big data include: Comparative analysis which means compare the one company products with those of its competition. The uncertained data in an organization is accounted before it is used in bigdata analytics. Information technology also needs to ensure that they have enough accurate data available to produce correct results. Big data brings new and existing opportunities to companies who utilize the platforms available. Processing large datasets when coupled with hadoop distributed file systems can be used to handle big data is the programming methodology of the map reduction in big data. It has the capability to handle structured as well as unstructured data.
Knowledge identification through data mining technique is a kind of technique for learning from databases. At the end of the day Data Mining is extraction of fascinating examples or learning's from enormous quantify of information.

Revised Manuscript Received on February 28, 2020.
***** Correspondence Author
Srinivasa Rao Divvela*, Assistant Professor, Department of CSE, Lakireddy Bali Reddy College of Engineering, Mylavaram, srinumtechcse2007@gmail.com
Dr V Sucharita, Professor, Department of CSE, Narayana Engineering College Gudur, India. jesuchi78@yahoo.com

Fundamentally, it manages information recovery from vast databases relying upon the particular goal shown by every customer in a conceivable situation.
Information Mining is to find learning in huge social database that finds coordinating outcomes in enormous number of datasets.
Information mining basically tries to produce knowledge from a differed datasets and stored in a reasonable arrangement to fit for more purpose.

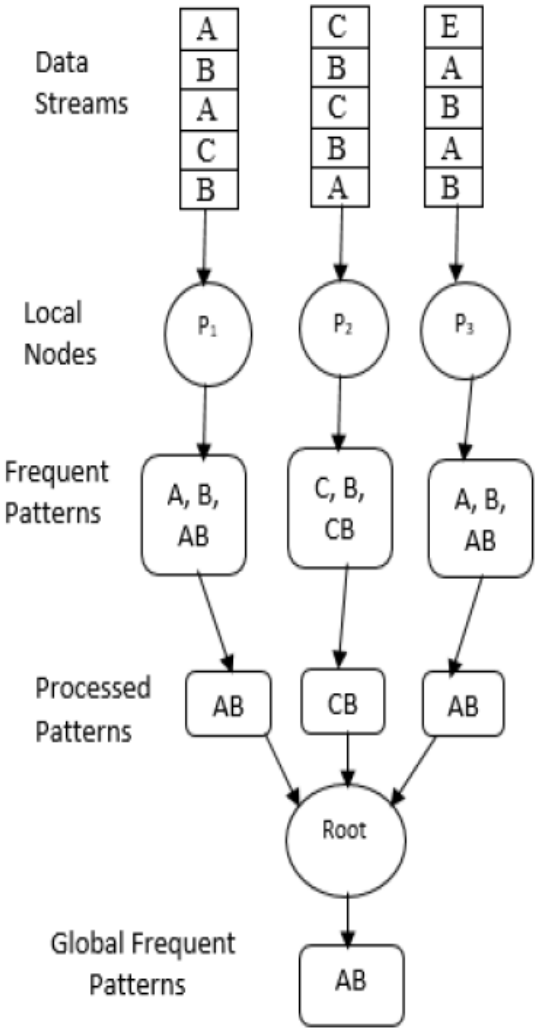
II. RELATED WORK

S.No	Algorithm Name	Author Name	Year	Disadvantage
1.	A-CLOSE (Apriori based closed frequent itemset)	Nicolas Pasquier Yves Bastide Rafik Taouil Lotfi Lakhal	1999	• Costly when mining long patterns or with low minimum support thresholds in large databases
2.	CHARM (Closed Association Rule Mining)	Mohammed Javeed Zaki, Ching-Jiu Hsiao	1999	• It does not work well in higher support of order of magnitude.
3.	CLOSET	Jian Pei, Jiawei Han, Runying Mao	2000	• It does not work well in lower support of order of magnitude.
4.	CLOSET+	Wang J, Han J, Pei J	2003	• Work well for datasets with small average row length,
5.	CARPENTER	Pan F, Cong G, Tung AKH, Yang J, Zaki M	2003	• Encounters problem for datasets that have large number of rows and features
6.	COBBLER (Combining Column and Row	Pan F, Tung AKH, Cong G, Xu X	2004	• It cannot make full use of the minimum support threshold to prune search space. As a result, experiments

III. PROBLEM & PROCEDURE

When we look upon a distributed system it has N number of data streams in it which are over looked by their corresponding local nodes and the overall performance is looked up by the root node. Let L1,L2,….Ln are the local nodes for D1,D2,….Dn data streams. The calculation of the frequent items is done at the root node which is the union of all the sets from the local nodes which is D1 U D2 U….Dn.
Let F = {f1,f2,….,fn} are the obtained sets of items from their respective nodes. The below figure depicts a 2-level architecture modal which will be helpful in handling the distributed data streams where the root node will be in level-0 and local nodes at level-1. Each local node will evaluate the patterns in their specific data stream using the algorithm 'FP' which will help in lessening the burden on the root node and will be sent to the root node.

The root node with help of the patterns received from local nodes will generate a global pattern set. If a change happens in any of the local node then immediately the updated set will be sent to the root node for updating. This will significantly reduce the load on root node and minimize the communication cost along with maintaining the root node updated all the times.



Architecture

IV. ALGORITHMS & IMPLEMENTATION

Processing at local node:

Each local node will evaluate the patterns in their specific data stream using the algorithm 'FP' which will help in lessening the burden on the root node and will be sent to the root node. If a change happens in any of the local node then immediately the updated set will be sent to the root node for updating.

Algorithm-'FP'

1. START 'FP'
2. In descending order arrange all the itemsets FIS (frequent itemsets) with respect to their sizes.
3. If FIS of size n then
4. RFIS ←FISn
5. Where n is the largest FIS & RFIS is the reduced set to forward to root
6. End if
7. If remaining FISi , i<=n-1 then
8. If FISi not in Subset (RFIS) then
9. RFIS ←FISi
10. End if
11. End if
12. Return RFIS
13. END 'FP'

Processing at Root node:

The root node with help of the patterns received from local nodes will generate a global pattern set. We implement the indexing algorithms for obtaining the global set of frequent items. Let's see two indexing algorithms best suited for the extraction. They are as follows:

I-List Algorithm:

This algorithm stores data in the form a linked list where the index value and frequent item value will be stored in the list.

```
1: procedure I_List (new local itemset p_i)
2:     H=H-list(p_1)
3:     mpre_i ← φ
4:     for n_p: H do
5:         mpre_p = mpre_p + n_p; count = 1; n_i = n_p;
6:         while count < p_i.length do
7:             flag=false; count++;
8:             for n_c : N_c do        ▷ N_c =child node of n_i
9:                 if n_c == P_i.next() then
10:                    mpre_p = mpre_p + n_c; n_i = n_c
11:                    flag=true;
12:                    break;
13:                end if
14:            end for
15:            if flag==false then
16:                break;
17:            end if
18:        end while
19:        if mpre_i < mpre_p.size then
20:            mpre_i = mpre_p
21:        end if
22:    end for
   return MsegT after p_i being inserted.
23: end procedure
```

Seg-Tree Modified Algorithm:

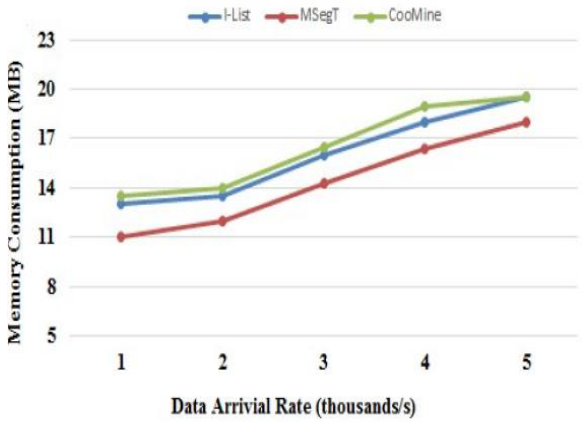
There are 3 steps for the calculation of the patterns required at the root node which are

1. Prefix matching – to identify local node itemsets sent to root.
2. Attribute value updation – for wvwry new set inserted the value has to be updated for processing.
3. Final global frequent itemset extraction – this again has 2 stages
   i. Global sets extraction
   ii. Global frequent sets extraction

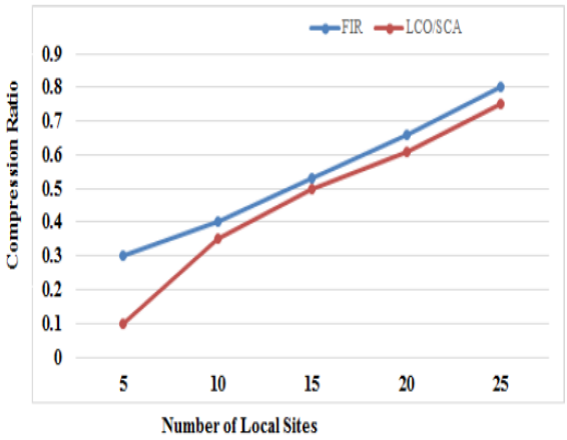
```
1: procedure MSEGT-GFI(H)        ▷ where H is pointer to H-list
2:     gfi ← φ, GFI ← φ          ▷ where gfi and GFI are the sets of global frequent items and itemsets respectively
3:     for each item i in H-list do
4:         Traverse the list by ref
5:         ζ_total_i ← φ
6:         for each node n in ref list of i do
7:             if | ζ_total_i | ≥ θ then
8:                 break;
9:             else
10:                ζ_total_i = ζ_total_i ∪ ζ_n
11:            end if
12:        end for
13:        if ζ_total_i ≥ θ then
14:            gfi ∪ {i}
15:        end if
16:    end for
17:    SubSeq← SUBSEQUENCE(gfi)
18:    Sort the SubSeq in decreasing order of size of their itemset.
19:    for each sq ∈ SubSeq do
20:        if sq ⊈ GFI_i for all GFI_i ∈ GFI then
21:            ζ_total_sq ← φ
22:            Traverse every branch of MsegT which are referred by sq_1 of H-list
               ▷ where sq = (sq_1, sq_2, ...., sq_n)
23:            if ζ_total_sq ≥ θ then
24:                break;
25:            else
26:                ζ_total_sq = ζ_total_sq ∪ ζ_n
27:            end if
28:            if ζ_total_sq ≥ θ then
29:                GFI ∪{sq}
30:            end if
31:        end if
32:    end for
   GFI = GFI ∪ {gfi}
   return GFI
33: end procedure
```

V. EVALUATIONS

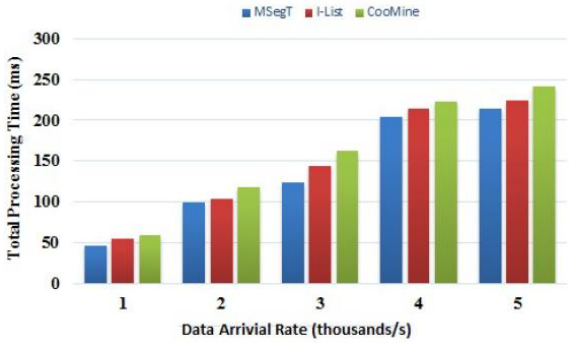
The evaluation is done on the basis of the performance exhibited by the proposed methodology and some of the existing models. The evaluation shows that the FP algorithm gives us relatively high compression ratio and minimizes the communication cost. The both space and time complexity is greately minimized with the help of Seg-Tree modified algorithm



Memory usage comparison



Compression ratio comparison



Processing time comparison

VI. CONCLUSION

In real-time distributed relational data streams data summarization is very crucial and impractical situation. So that for an outstanding data summarization on uncertain data streams an efficient system is needed to helps in continuously growing data environment. The proposed model solved the problem of global frequent itemset extraction when a union of data streams are involved. FP algorithm maximizes process speed as the load on root node is cut off and also it minimizes the communication cost for the nodes. Seg-Tree reduces the space and time complexities for the calculation of the itemsets. Further implications can be brought by using these approached in Utility mining and increasing the security factors.

REFERENCES

1. R. Aggarwal& R. Srikant, “Fast algorithms for mining association rules,” VLDB 1994, pp. 487–399.
2. M. Cannataro, A. Cuzzocrea, A. Pugliese, “A probabilistic approach to model adaptive hypermedia systems”, WebDyn 2001, pp. 50–60.
3. A. Cuzzocrea, G. Fortino, O. Rana, “Managing data and processes in cloud-enabled large-scale sensor networks: stateof-the-art and future research directions”, IEEE/ACM CCGrid 2013, pp. 583–588.
4. A. Cuzzocrea, F. Jiang, W. Lee, & C.K. Leung, “Efficient frequent itemset mining from dense data streams,” APWeb 2014, pp. 593–601.
5. J. Dean & S. Ghemawat, “MapReduce: simplfied data processing on large clusters,” CACM 51(1), Jan. 2008, pp. 107- 113.
6. P. Fournier-Viger, A. Gomariz, T. Gueniche, A. Soltani, C. Wu, & V. S. Tseng, “SPMF: a Java open-source pattern mining library,” Journal of Machine Learning Research 15(1), 2014, pp. 3389-3393.
7. J. Han, J. Pei, & Y. Yin, “Mining frequent patterns without candidate generation,” ACM SIGMOD 2000, pp. 1–12.
8. D. Jlailaty, D. Grigori, & K. Belhajjame, “Mining business process activities from email logs,” IEEE ICCC 2017, pp. 112-119.

9. Y. Lei, S. Chen, & P. S. Yu, "Heterogeneous service information mining based on parallel computing," IEEE ICCC 2017, pp. 120-123.
10. C. K. Leung, "Big data analysis and mining," Encyclopedia of Information Science and Technology, 4e, 2018, pp. 338-348.
11. C. K. Leung, F. Jiang, T. W. Poon, & P. Crevier, "Big data analytics of social network data: who cares most about you on Facebook?" Highlighting the Importance of Big Data Management and Analysis for Various Applications, 2018, pp. 1-15.
12. C. K. Leung & R. K. MacKinnon, "Balancing tree size and accuracy in fast mining of uncertain frequent patterns," DaWaK 2015, pp. 57-69.
13. H. Li, Y. Wang, D. Zhang, M. Zhang, & E. Y. Chang, "PFP: parallel FP-growth for query recommendation," ACMRecSys 2008, pp. 107-114.
14. Y. Li, M. Muthiah, A. Routh, & C. Dorai, "Cognitive computing in action to enhance invoice processing with customized language translation," IEEE ICCC 2017, pp. 136-139.
15. J. Liu, J. Li, S. Xu, & B. C. M. Fung, "Secure outsourced frequent pattern mining by fully homomorphic encryption," DaWaK 2015, pp. 70-81.
16. J. Liu, Y. Wu, Q. Zhou, B. C. M. Fung, F. Chen, & B. Yu, "Parallel Eclat for opportunistic mining of frequent itemsets," DEXA 2015, pp. 401-415.
17. S. Moens, E. Aksehirli, & B. Goethals, "Frequent itemset mining for big data," IEEE BigData 2013, pp. 111-118.
18. Z. Pan, Y. Ge, Y. C. Zhou, J. C. Huang, Y. L. Zheng, N. Zhang, X. X. Liang, P. Gao, G. Q. Zhang, Q. Wang, & S. Shi, "Cognitive acoustic analytics service for Internet of Things," IEEE ICCC 2017, pp. 96-103.
19. J. Pei, J. Han, H. Lu, S. Nishio, S. Tang, & D. Yang, "HMine: hyper-structure mining of frequent patterns in large databases," IEEE ICDM 2001, pp. 441–448.

AUTHORS PROFILE

Divvela Srinivasa Rao, is currently working as Sr.Assistant Professor, CSE Department in Lakireddy Bali Reddy College of Engineering, Mylavaram. He is a Research scholar in Koneru Lakshmaiah Educational Foundation, Guntur. His area of interest is Data Mining and Knowledge Engineering, Artificial Intelligence. He has total 13 years of teaching experienceExperience in various Engineering colleges in Andhra Pradesh. He has published 10 papers in various national and international Journals, conference proceedings and Conferences.

Dr.V.Sucharita, is currently working as Professor in the Department of Computer Science and Engineering in Narayana Engineering College,Gudur. She has obtained her Ph.D degree in the faculty of Computer Science with Data Mining and Neural Networks as specialization from Sri Padmavati Mahila Visvavidyalayam,Tirupathi. She has 18 years of Teaching Experience in various Engineering colleges in Andhra Pradesh. Prior to joining Narayana Engineering College, She worked as Professor in Koneru Lakshmaiah Educational Foundation, Guntur. She published 20 papers in various national and international Journals, conference proceedings and Conferences.