International Journal of Engineering and Advanced Technology (IJEAT)
ISSN: 2249-8958 (Online), Volume-8 Issue-5, June 2019

Implementing Frequent Itemset Mining By advanced Distributed Approach Using Matrix-Based Pruning

Srinivasa Rao Divvela, V Sucharita,

Abstract: ***Mining of frequent itemsets is a very crucial process in the mining of associate rules. Significant challenges are being encountered in this era where big data has drawn its own circle by shaping around space and time factors. A distributed procedure for the job is what that best be defined. So we increment our previous version with an enhanced version by the implementing a distributed approach which is better that FP growth. Differentiating of the existing and proposed algorithm is done using the practical valuable data that is available.***

Keywords: ***big data, FP growth, distributed approach, frequent itemset.***

I. INTRODUCTION

Frequent itemset mining is a crucial and important part of associate rule mining. Mining of frequent itemsets is a very crucial process in the mining of associate rules. Significant challenges are being encountered in this era where big data has drawn its own circle by shaping around space and time factors. A distributed procedure for the job is what that best be defined. The use of these techniques is more essential and crucial for the development of finding of the frequent and rare items for the procedure and are very useful in the mining aspects. The incorporation of bigdata is a major break through for the procedural techniques.

II. LITERATURE SURVEY

Marc Chong (14)	JC School of Public Health and Primary Care	China 2015	2015 IEEE International Congress on Big Data	Patient Flow Evaluation with System Dynamic Model in an Emergency Department (4)	Paper	Emergency Department process and activities	Data from: - Expertise review - Hong Kong Hospital Authority (HA) - Hong Kong Census and Statistics Department statistics - Accident & Emergency Medicine Academic Unit	NA	The creation of a simulation model to achieve greater efficiency taking into account waiting time, occupancy, admission volume, staff numbers in an emergency situation in ED.
Yong-Hong Kuo (15)	Stanley Ho Big Data Decision Analytics Research Centre	China 2015	2015 IEEE International Congress on Big Data	Embracing Big Data for Simulation Modelling of Emergency Department Processes and Activities (5)	Paper	Emergency Department process and activities	Data from: -Hospital data from The Prince of Wales Hospital (PWH), Hong Kong	NA	The use of RFI (radio frequency identification), to capture patient flow and activities in EDs through a simulation model which obtain all relevant treatment processes.
Eva Bruballa (16)	Tomàs Cerdà Computer Science School Universitat Autònoma de Barcelona (UAB) Barcelona, Spain	Barcelona 2014	2014 International Conference on Future Internet of Things and Cloud	Simulation and Big Data: A Way to Discover Unusual Knowledge in Emergency Departments (6)	Paper	Emergency Department process and activities	NA	NA	Obtain knowledge from data generated by simulation of the real system, integrated and transformed in a data warehouse and analyzed using data mining techniques to observe patterns and trends.
Shiming Yang (17)	University of Maryland School of Medicine, Baltimore	UK 2014	British Journal of Hospital Medicine	'Big data' approaches to trauma outcome prediction and autonomous resuscitation (7)	Review	Prediction of the outcome of Emergency Department's patients	NA	NA	With ambient non-invasive data sensors and appropriate collecting techniques, fractional information from different data sources can be gathered in real-time and be applied for specific studies.
R.AndrewTaylor (18)	Department of Emergency Medicine, Yale Univeristy	USA 2016	Academic Emergency Medicine	Prediction of in-hospital mortality in emergency department patients with sepsis: a local big data-driven, machine learning approach (8)	Article in a periodic journal	Prediction of the outcome of Emergency Department's patients	Data from: 1) urban, academic, level i trauma center 2) urban community-based, academic level ii trauma centre 3) community-based centre 4)sub-urban free-standing ed	Retrospective cohort study	The strength of using the large volume of data stored in electronic health records (EHRs) to improve the quality and the timing of sepsis mortality prediction of patients admitted in ED with sepsis.

Manuscript published on 30 June 2019.
***** Correspondence Author (s)
Srinivasa Rao Divvela, Research Scholar Department of CSE KL University Lakireddy Bali Reddy College of Engineering Vaddeswaram, India
Dr V Sucharita, Professor Department of CSE Narayana Engineering College Gudur, India

© The Authors. Published by Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP). This is an open access article under the CC-BY-NC-ND license http://creativecommons.org/licenses/by-nc-nd/4.0/

III. ALGORITHMS

We incorporate the matrix_based pruning algorithm procedure with the big data spark framework such that we can enhance the efficiency and scalability of the procedure.

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication (BEIESP)
© Copyright: All rights reserved.

The main factor is the acquiring of the Boolean vector for each item in the algorithm of a given database and calculation of the 2-*itemset* matrix, by using them the reduction of amount of candidate itemset is done.

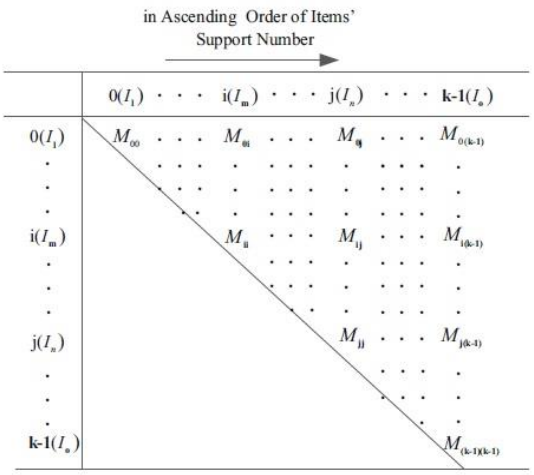


Fig: Matrix Prunning

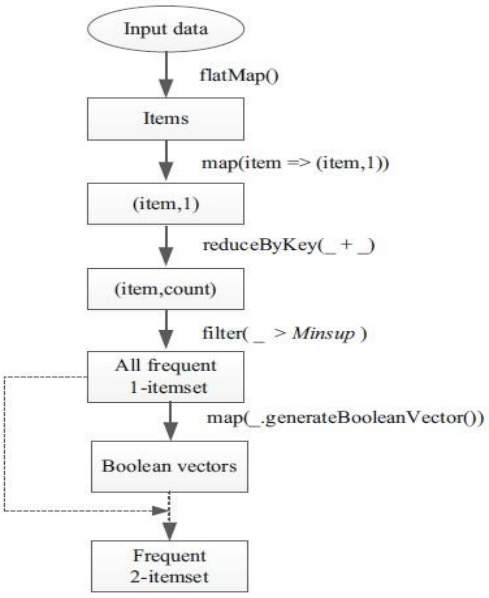
IV. IMPLEMENTATION

The implementation of the above algorithm for single system primarily based on spark is said in detail. In order to lessen thememory usage, hashmap is used in our program to save the boolean vector for every frequent 1 itemset and all of the common 2-itemset, in preference to saving the 2-itemset matrix m.

Step 1: achieve the boolean vectors for all common 1-itemset, then produce all the common 2-itemset based totally on these boolean vectors.

Step 2: get all common (k + 1)-itemset with the aid of the use of common ok-itemset(k ≥ 2).

With the help of these two steps we hence further improvise and improve the efficiency of the mining procedures on the framework of apache spark.



Step 1

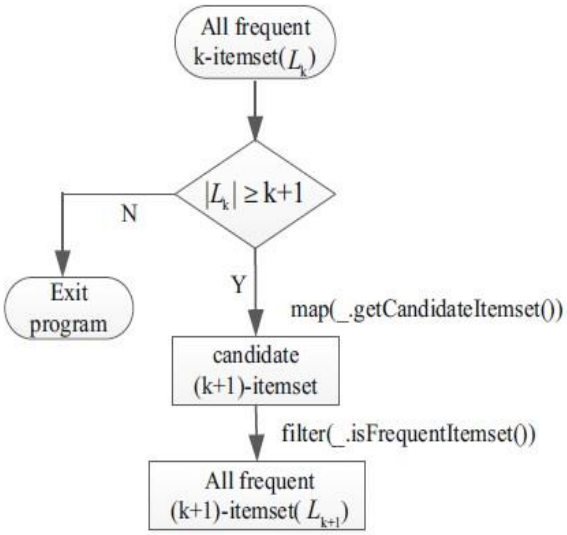


Fig. 3 The process of step 2 on Spark

Step 2

V. RESULTS

Here are the detailed reports of the attributes of the data sets that have been used in the process of the implementation of the distributed procedure we have proposed in this paper. Here the comparison of the previous PFP and the present implemented distributed procedure with respect to their runtime along with the detailed info of the datasets and data used in given in the tabular form for better and easy understanding of the process.

Table 2 The running time comparison on T10I4D100K and T40I10D100K

Dataset	Support degree	PFP (Spark 1.3)	DFIMA
T10I4D100K	0.01	64s	25s
	0.03	35s	17s
	0.05	37s	14s
T40I10D100K	0.01	321s	180s
	0.03	316s	81s
	0.05	250s	67s

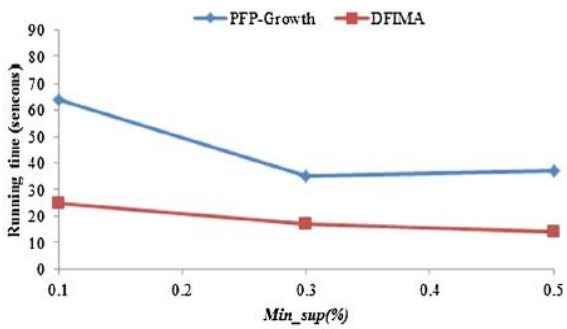


Fig. 4 The running time on T10I4D100K

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication (BEIESP)
© Copyright: All rights reserved.

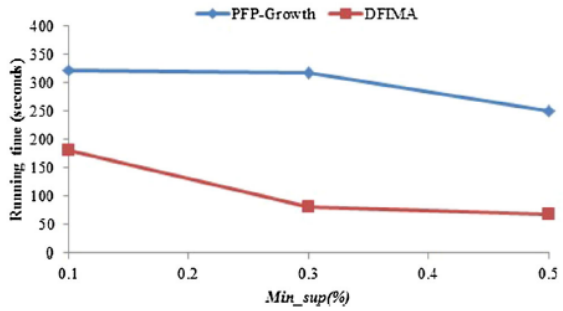


Fig. 5 The running time on T40I10D100K

Table 3 The running time comparison on "T40I10D100K" databases

Dataset	Support degree	PFP (Spark 1.3)	DFIMA
T40I10D100K*5 (73.8 MB)	0.5	46s	11s
	0.3	110s	15s
	0.1	262s	44s
	0.05	534s	126s
T40I10D100K*10 (147.6MB)	0.5	51s	16s
	0.3	58s	15s
	0.1	301s	105s
	0.05	749s	281s
T40I10D100K*20 (295.2 MB)	0.5	54s	24s
	0.3	58s	25s
	0.1	620s	284s
	0.05	1498s	598s
T40I10D100K*40 (590.4 MB)	0.5	72s	28s
	0.3	83s	26s
	0.1	1372s	578s
	0.05	2671s	611s
T40I10D100K*80 (1.2 GB)	0.5	103s	54s
	0.3	124s	51s
	0.1	2898s	365s
	0.05	5892s	583s

Activate W
Go to PC setti

Table 1 The properties of datasets used in experiment

Dataset	Size	Transactions	Items	Average length
T10I4D100K	3.84 MB	100,000	870	10
T40I10D100K	14.8 MB	100,000	1000	40

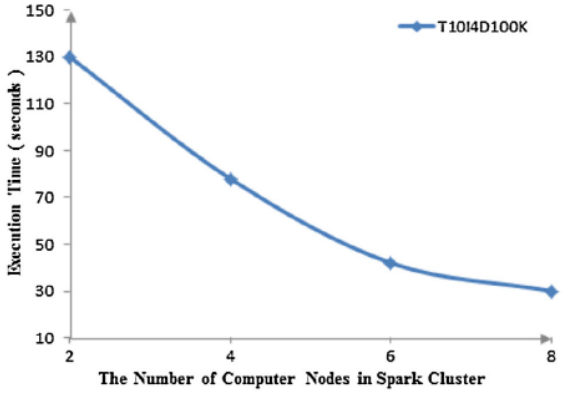


Fig. 6 The running time with different computer nodes

Table 4 The results of PFP-Growth and DFIMA on Car dataset

Dataset	Support number	PFP (Spark 1.3)	DFIMA
Car dataset	100	521s	201s
	150	435s	189s
	200	337s	158s

VI. CONCLUSION

The main hindrance and lacking in the map reducing approach is the repeated scanning but after overcoming the over scan problem we have the power and space to deal with.A distributed procedure for the job is what that best be defined. So we increment our previous version with an enhanced version by the implementing a distributed approach which is better that FP growth. Differentiating of the existing and proposed algorithm is done using the practical valuable data that is available. By using the matrix_based pruning and spark we overcome the problems and we have evaluated it on a standard dataset to ensure its genuinely on a whole. We have proved that the process proposed by us shows great amount of scalability and improves efficiency of the system. When there is a high support degree our proposed procedure shows most effective results. It can be improved further by incorporation of optimization techniques.

REFERENCES

1. A distributed frequent itemset mining algorithm using Spark for Big Data analytics, Feng Zhang1, Min Liu1, Feng Gui1, Weiming Shen2, AbdallahShami, Yunlong Ma1, Springer Science+Business Media New York 2015
2. Frequent itemset mining for big dataS Moens,E. Aksehirli,B.Goethals - Conference on Big Data, 2013 - ieeexplore.ieee.org
3. Neutrosophic_associationruleminingalgorithm for big data analysisM Abdel-Basset, M Mohamed, F Smarandache-Symmetry, 2018 - mdpi.com
4. Mining association rule in big data with NGEP Y Chen, F Li, J Fan - Cluster Computing, 2015

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication (BEIESP)
© Copyright: All rights reserved.