IJRSET 2018 Volume 5, Special-Issue-1 Pages: 30-33

ISSN 2394-739X

International Journal for Research in Science Engineering and Technology

A COMPARATIVE LEARNING ON REGULAR PATTERN MINING ALGORITHMS

[1] Dr. A. Mahendran, [2] Dr. C. Kavitha,
[1] Dept of CS, [2] Asst Prof of CS,
[1] Periyar University, Salem.11, Thiruvalluvar [1] Govt Arts College Rasipuram.

ABSTRACT: Frequent pattern mining has been an important subject matter in data mining from many years. Many efficient algorithms have been designed for finding frequent search patterns in transactional database .Discovering frequent itemsets is the computationally intensive step in the task of mining association rules. A large number of candidate itemsets generation is one of the main challenge in mining. The objective of frequent pattern mining is to find frequently appearing subsets in a given sequence of sets. Frequent pattern mining comes across as a sub-problem in various other fields of data mining such as association rules discovery, classification, market analysis, clustering, web mining, etc. Various methods and algorithms have been proposed for mining frequent pattern.This paper presents comparative study on frequent mining techniques – Apriori and FP-Growth. [2]

Keywords : [Aproiri,Fpgrowth,Hadoop,FrequentPattern Mining]

1. INTRODUCTION

I.1 Frequent itemset mining

Suppose that I = {I1, I2, . . . ,Im} is an itemset composed of m items. A database D consists of a series of transactions. Each transaction is a subset of I and has a unique label denoted by TID.Aset of items is referred to as an itemset. Anitemset that contains k items is a k-itemset. For instance, the set {beer, diaper} is a 2-itemset. The occurrence frequency of an itemset is the number of transactions that contain the itemset. Given an itemsetX, the support number of X is the number of transactions in D that contain. If the support number of X is greater than or equal to the specified minimum support threshold (abbreviated as MinSup), then the itemsetX is labelled as a frequent itemset. The purpose of frequent itemset mining is to find all frequent itemset in a given database. [3] A number of research works have been published that presenting new algorithm or improvements on existing algorithms to solve data mining problem efficiently. In that Apriori algorithm is the first algorithm proposed in this field. There are two categories of frequent pattern mining the algorithm, namely Apriori algorithm and Tree structure algorithm. The Apriori based algorithm uses generate and test strategy approach to find frequent pattern by constructing candidate items and checking their counts and frequency from transactional databases. The Tree structure algorithm uses a text only approach. There is no need to generate candidate item sets. Many tree based structures have been proposed to represent the

data for efficient pattern discovery including FP-Tree, CAT-Tree, CAN-Tree, CP-Tree, etc

1.2 APPLICATION OF FREQUENT PATTERN MINING

Frequent pattern mining were applied to inter-disciplinary domains beyond data mining. Frequent patterns reflecting strong associations among multiple items or object, capture the underlying semantics in data. Frequent pattern mining has huge application such as:

Indexing and similarity search of complex structured data Spatiotemporal and multimedia datamining Stream data mining Web mining Software Bug mining and page-fetch

1.3 APRIORI ALGORITHM

Apriori algorithm is introduced by Agrawal and Srikanth provide performance improvements over a naïve itemset search. Apriorialgorithm has been around almost as long as the concept of frequent itemsets and is very popular. The naive algorithm is a theoretical concept and is not used in practice. Aprioiri has become the classic implementation of frequent itemset mining. Aprioiri first builds a list of all singleton itemsets with sufficient support. Building on the monotonicity principle, the next set of candidate frequent itemsets is built of combinations of the singleton itemsets. This process continues until the maximum length specified for frequent itemsets is reached. The Apriori algorithm performs a breadth first search of the itemsets [4] By iteratively reducing the candidate itemsets, Apriori algorithm achieves good performance. The issue with Apriori however is that it entailsdata scans to find every frequent k-item set. Hence, the algorithm reaches a bottleneck easily when the length of the largest frequent itemset is relatively long as it then needs to generate huge candidate sets resulting in a dramatic performance decrease [1] Step 1:Get Frequent items The occurrence of the items which is greater than min support thresholds

Step 2: Get Frequent Itemsets The candidate itemset is been generated from the frequent items. Prune the results to find the frequent itemsets

Step 3: Applying Association rule Rules which satisfy the min support and min confidence threshold [6]

1.4. FP-GROWTH ALGORITHM

Frequent pattern growth also labelled as FP-growth is a tree based algorithm to mine frequent patterns in database, the idea of which was first presented by Han, Pei & Yin in the year 2000 .FP-growth takes a radically different approach to discovering frequent itemsets; the algorithm does not subscribe to the generate-and-test paradigm of Apriori. Instead, it encodes the data set using a compact data structure called an FP-tree and extracts frequent itemsets directly from this structure without generating candidate frequent item sets using divide and conquer method. FP-Tree, an extension of prefix tree structure, only stores the frequent items. Every node in the tree contains the label and the frequency of the item. The paths from the root to the leaves are set corresponding to the support value of the items such that the frequency of a parent is greater than or equal to the sum of the frequencies of its children[1]

Two data scans are necessary to construct a FP-Tree. Each item's support value is found in the initial scan. During the second scan, the previously calculated support values are used to sort the items within transactions in descending order. Now, if any of the two transactions have a same shared prefix, the common portion is fused and accordingly the frequencies of the nodes are incremented [1].It allows frequent item set discovery without candidate item set generation, it is a two-step approach

1. Build a compact data structure called FP-Tree.
2. Extracts frequent item sets from the FP-Tree FP-Tree is constructed by using two passes over data set.
3. Pass1 Scan the data and find support for each item.

Discard infrequent items Sort frequent items in decreasing order based on their support 4. Pass2 Construct the FP Tree by reading the

transactions. [2] Performance FP-Growth is generally the fastest and the most memory efficient algorithm. FP-Growth algorithm is efficient and scalable for mining both long and short frequent patterns. It is faster than the Apriori algorithm

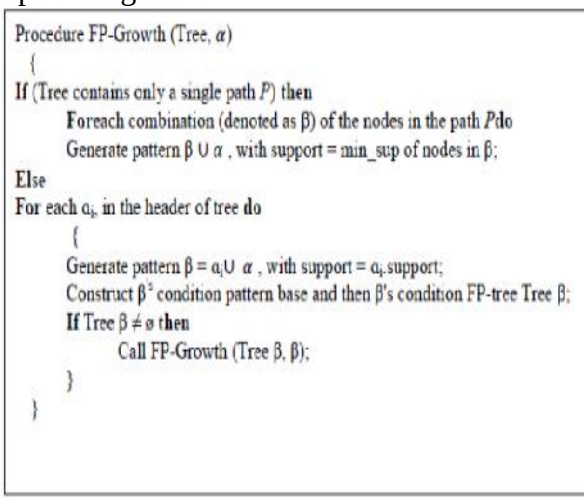

Figure1: Basic FP Growth Algorithm [5]

parameter	Apriori	FP Growth
Technique	Join and Prune	Constructing Fp Tree
Memory Utilization	Large memory space for candidate itemsets	Lesser memory due to compact structure
No of scans	Database is scanned multiple times	Database is scanned twice
Time	Larger execution time due to candidate itemset generation	Smaller Execution time

Table 1: Comparision of Apriori and FP Growth Algorithm [6]

2. HADOOP

Hadoop is a scalable, open source, fault tolerant Virtual Grid operating system architecture for data storage and processing. It runs on commodity hardware, it uses HDFS which is fault-tolerant high bandwidth clustered storage architecture. It runs MapReduce for distributed data processing Hadoop consists of distributed file system, data storage and analytics platforms and a layer that handles parallel computation, rate of flow (workflow) and configuration administration HDFS runs across the nodes in a Hadoop cluster and together connects the file systems on manyinput and output data nodes to make them into one big file system.and is works with structured and unstructured data

Author	Technique	Benefit
Dachuan Huang, Yang Song, RamaniRoutray, Feng Qin	Smart Cache – new Map Reduce technique with MRApriori Algorithm	Minimum support parameter is lowered
M.-Y. Lin, P.-Y. Lee, and S.-C. Hsueh	Apriori-based frequent itemset mining algorithms on MapReduce	Dynamically collects candidates of variable lengths for counting by mappers according to the number of candidates and the Execution time.
L. Zhou, Z. Zhong, J. Chang, J. Li, J. Huang, and S. Feng	Balanced parallel FP-growth with MapReduce	Improves performance of the original PFP algorithm by balancing load of the parallel FP-Growth phase.

Table 2: Represents a Comparative study of Algorithms on Hadoop

CONCLUSION

Numerous techniques have been put forward for mining frequent patterns. This paper presented an overview of two diverse techniques of frequent pattern mining that can be used in different ways to generate frequent itemsets. Methods proposedby various authors to extract frequent itemsets in a largedataset have also discussed. Every method has its own pros and cons. Performance of a specific technique is contingent on the available resources and input data

REFERENCES

[1] “Frequent Pattern Mining Algorithms: A Comparative Study,”MrSahilModak,MrSagarVikmam,Prof(Mrs) Lynette D’mello, International Journal of

Innovations & Advancement in Computer Science,Volume 4, Issue 9,2015
[2] "Frequent Pattern Mining Algorithms Analysis",RiteshGiri , Ananta Bhatt , Aadhya Bhatt , International Journal of Computer Applications, Volume 145 – No.9, July 2016
[3]" A distributed frequent itemsetmining algorithm using Spark for Big Data analytics", Feng Zhang, Min Liu, Weiming Shen, Article in Cluster Computing · October 2015
https://www.researchgate.net/publication/283299506,
[4]" Comparing Dataset Characteristics that Favour the Apriori, Eclat or FP-Growth Frequent Itemset Mining Algorithms", Jeff Heaton,arXiv:1701.09042v1 [cs.DB] 30 Jan 2017
[5]" An Enhanced Frequent Pattern Growth Based On MapReduce For Mining Association Rules", ArkanA. G. Al-Hamodi1 , Songfeng Lu, Yahya E. A. Al-Salhi, International Journal of Data Mining & Knowledge Management Process (IJDKP) Vol.6, No.2, March 2016
[6]" A comparative study of Frequent pattern mining Algorithms: Apriori and FP Growth on Apache Hadoop",Ahilandeeswari.G,Dr. R. ManickaChezian, International Journal of Innovations & Advancement in Computer Science, Volume 4, Special Issue ,March 2015, ISSN 2347 – 8616
[7]. M.-Y. Lin, P.-Y. Lee, and S.-C. Hsueh, "Apriori-based frequent itemset mining algorithms on mapreduce," in Proc. 6th Int. Conf. Ubiquitous Inform. Manag. Commun., 2012, pp. 76:1–76:8.
[8]. L. Zhou, Z. Zhong, J. Chang, J. Li, J. Huang, and S. Feng, "Balanced parallel FP-growth with mapreduce," in Proc. IEEE Youth Conf. Inform. Comput. Telecommun., 2010, pp. 243–246.