Design and Implementation of an Improved Routing Algorithm using Frequent Item Set Mining

Sanjeev Bansal, Ph.D
Department of CSE
Amity University, Noida, India

Sovers Singh Bisht
Department of CSE
Amity University, Noida, India

ABSTRACT
This paper presents how routers can manage large amount of data packets as routers exchange information with each other to find the best possible route to a destination thus graphs represent a more general class of structures than sets, sequences, lattices and trees. Since on web heavy range of information is represented by a graph as in social networks so modelling a sophisticated network by statistical analysis requires efficiency in routing packets within an autonomous system hence this approach here focuses on how routers can upgrade their configuration to exactly route the frequently occuring data packets within a homogeneous network.

General Terms
RTB-FP–Tree graph representation of a homogeneous network, improved algorithm and network statistics based forwarding table.

Keywords
Frequent item set mining, statistical modeling of networks, routing table, autonomous system and graph representation.

1. INTRODUCTION
A network consist of a set of nodes each corresponding to an object associated with a set of properties and a set of edges connecting those nodes, representing relationship between objects thus a homogeneous network is one in which nodes and edges of the same type are present such as a social network, a coauthor network or a web page network else it is called a heterogeneous network such as books in a library or an event management network. Thus with the advent of increasing traffic within an autonomous system this principle of improved frequent item set mining works within a homogeneous network where variety of hosts are accessing large amount of information that causes the network to be increasingly dense and shrinkage in diameter of the network. As it is clear that routers play a major part in distributing the packets within a virtual circuit environment they are times when routing these packets within a homogeneous network become cumbersome as the traffic density increases and the router configuration limits the size of traffic by limiting the forwarding table attributes. Mining of frequent item sets was proposed by Agarwal et al. [1] who proposed algorithm apriori. An apriori is a breath first search bottom up approach algorithm. It uses hash trees to store frequent item sets and generate candidate sets but the FP-tree based frequent item set algorithm is a depth first search preorder based algorithm which works without generation of candidate keys. Since apriori lacks space complexity so as FP-tree lacks small amount of time complexity. As it is clear that in router based configuration system generally rely on data packets which are basically representation of IP addresses and with their next hop interfaces within their routing table thus in a homogeneous network router constructs here a new frequent item set based forwarding table which will limit the use of routing algorithms which are constantly generating next hop interfaces for the packets arriving from a remote host. Thus correlation is a statistical approach in data mining where correlation between the attributes with respect to their arrival rate within a router is done and if their arrival rate increases hence a packet can be directly forwarded to the next hop based on its previous hop pattern. Router maintains a queue based on FIFO mechanism where the system tends to outflow the data packets constantly to minimize the data traffic.

2. LITERATURE STUDY
Router configuration basically causes two types of delivery of packets one within a physical network and other to a destination host within another autonomous system using multiple intermediate routers thus every router works on the principle of store and forward mechanism where a forwarding table is maintained by every router to manage its next hop delivery using the link interfaces. Router mechanism basically works on link state routing and distance vector routing algorithms by finding the best possible shortest path between host and destination thus the datagram network approach in connectionless environment is first implemented with the use of routing algorithms and then a virtual circuit is created between the source and destination using a virtual circuit sequence number which in routing table identifies the next hop on the edge (link) within the network topological graph. Now when a data traffic is large and enormous there is a increase in time complexity with the routing algorithms as well as space complexity causes the traffic to be slow and working with congestion, to avoid this parameter the system heavily rely on the routing table which are static and dynamic in nature. Static table says that the routes on the outgoing network are already destined and no updating is required while dynamic routing table heavily rely on datagram approach where the system have to constantly update the ongoing traffic within a router.

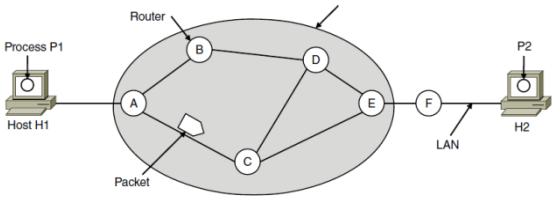

Fig 1: A Store and forward packet switched environment.

The above network topology shows how different routers choose the best possible path to forward the data packet thus this complex mechanism is based on the routing table and this algorithm approach is thus constrained in managing this table.

The above network comprises of an autonomous system represented by a single arrow and nodes as routers within this autonomous system which shows packet movement from node A to C to reach destination host H2 with receiving process P2.

3. ACCOMPLISHED WORK

Work accomplished is the store and forward packet switching where a routing table maintains the details of every outgoing and incoming packet delivered to the host. In the work accomplished below shows how a router works in a connection less environment and maintains a routing table which is updated frequently using the various algorithms proposed earlier. This connection less service given by router is later on upgraded with a static table using a virtual circuit connection oriented service thus the system now have a dedicated path to serve the source and the destination.

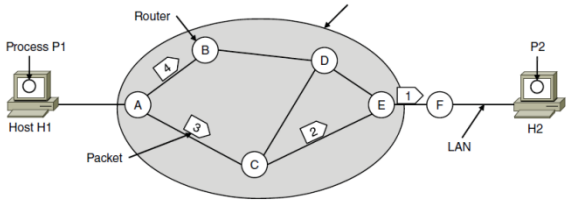

Fig 2: A connection less datagram network environment.

A's table (initially) A's table (later) C's Table E's Table

Dest.	Line
A	⊠
B	B
C	C
D	B
E	C
F	C

A	⊠
B	B
C	C
D	B
E	D
F	D

A	A
B	A
C	⊠
D	E
E	E
F	E

A	C
B	D
C	C
D	D
E	⊠
F	F

Thus the routing table maintains the forwarding table as the longest prefix matching technique where the network address, next hop and link interfaces is stored for every packet to be delivered though the system know that the network is distributed in a geographical area so there is very little scope of low traffic and congestion. The above network shows how routers maintain the outgoing and incoming of packets 3, 4, 2 & 1 where only packet 2's forwarding table is depicted as an example in tables above.

4. PROBLEM STATEMENT

Congestion in a network may occur when the incoming traffic within a router is larger than the capacity of the network to manage the routing table to forwarding table. This congestion within a network topology is impossible to manage and remove completely but it is possible to detect and minimize the problem by studying its flow mechanism. This congestion within a network can increase delay uninterrupted, packet loss and cut the channel throughput. This network traffic is monitored by the router in the network layer thus it maintains the routing table for every packet that enters and leaves the link by store and forward mechanism. Hence the problem of congestion is caused in network by the following reasons:-

- Bandwidth constraint
- Limited buffer capacity of intermediate nodes
- Non availability of multipath routing procedure
- Distributing the load efficiently by load balancing means
- Constant route changes due to node mobility as well as the unreliable medium result in unsteady packet delivery delays and packet losses.

5. PROPOSED CONCEPT

The concept proposed here is basically dealt with the store and forward mechanism of the routing table where the system generates frequent item sets based on apriori property of data mining. Here generally the system is dealing with large number of data packets which are increasing data traffic in the network and are stored for delivery within a router which maintains the routing table. These routing packets are generally based on the network IP addresses and the link interfaces hence whenever these packets are generated within a homogeneous network where the system is repeatedly getting a request from the same IP, the system is again routing the packets by implementing and updating the routing table thus system needs to implement the routing algorithms to update the table but if it works on the apriori property than it does not need to update the table again and again thus all frequent network addresses within a homogeneous network will be updated once and whenever it will get the same request it can directly route the packet based upon the tree generated by the algorithm with frequent item sets. The algorithm on which this basic approach is implemented is the FP-tree algorithms which is based on divide and conquer approach and compresses the item sets representation into a frequent pattern tree which on the contrary reduces the size of the data base searched.

5.1 Correlation Concept

As the system is aware of the fact that mining frequent item sets rely on the concept of correlation thus here according to the probability distribution theory two item sets can be correlated according to the entropy of information contained within the item sets. Entropy is the resultant of the probability distribution of elements which signifies how much information an element contains within a network.

According to the entropy elements is generally represented by the method as follows:

$$\text{Entropy} = -\sum_{i=1}^{N} Pi * \log(Pi) \text{ where,}$$

P_i of an element i is given as frequency of element i/total number of elementshence grouping the elements according to their occurrence in a particular router configuration to get the reduced tree based on frequent item set approach.
Two items packet P1 and P2 can be correlated if they satisfy the following conditions:

$$P(P1P2) > P(P1)\,P(P2)$$

Here
P (P1P2) = Probability of packets P1 and P2 occurring simultaneously in a particular network i.e. the probability of occurrence of both P1 and P2 together/total number of occurrences, similarly measure probability of P1 and P2. This shows that if two packets occur together within a network than their information gained is more as compared to packets occurring alone thus in a homogeneous network similar packets can be grouped to reduce the space of the forwarding table and above concept represents that

Observed Probability P (P1P2) > Expected P (P1) P (P2)

5.2 Proposed Algorithm

The idea behind this proposed algorithm is generally based upon the forwarding table within the router where the system

have data packets within a homogeneous network such as a social networks i.e. the data packets containing IP may be similar coming to a router for delivery. Here data packets are represented as P packets which are maintained in a matrix format within a router.

Algorithm: RTB-FP growth.
Mine frequent item sets using an RTB-FP (Routing tree based frequent packets) growth algorithm.
Input:
Incoming Packets, a network database within a router;
Min sup, the minimum support count threshold in terms of frequency of occurrence of packets. i.e. required to group packets for delivery.

Output:

The complete set of frequent packets in a homogeneous network as correlated.

Method:

1. The RTB-FP -tree is constructed in the following steps:

(a) Scan the network database D within a router once. Collect *P*, the set of frequent packets, and their frequency of occurrence. Sort *P* in support count descending order as *L*, the *list* of frequent packets.

(b) Create the root of an RTB-FP tree, and label it as "null." For each transaction *Trans* in *D* do the following.

Select and sort the frequent packets in network according to the order of *L*. Let the sorted frequent item list in *network* be [*N*], where *n* is the first element and *N* is the remaining list. Call insert tree ([*n in N*], which is performed as follows. If *T* has a child *X* such that X.*item-name=n.item-name*, then increment *packets*' count by 1; else create a new node *A*, and let its count be 1, its parent link be linked to *T*, and its node-link to the nodes with the same *item-name* via the node-link structure. If *A* is nonempty, call insert tree (N, *A*) recursively.

2. The RTB-FP -tree is mined by calling RTB-FP growth (*FP tree*, *null*), which is implemented as a following procedure to generate RTB-FP growth tree:-

1. If tree contains a single path p then
2. For each combination generate nodes(b) in the path p and generate entropy by probability of occurrence packets of each group in a bottom up fashion
3. pattern b[a with *support count = minimum support count o f packets in* b;
4. else till each header of the tree {
5. continue this process for every combination of group of packet in the tree with correlation
6. generate every conditional pattern base for each node in correlation such as
7. P (P1P2) > P (P1) P (P2)
8. Thus if observed probability > expected probability and it satisfies *support count = minimum support count o f packets in* b} then
9. Call RTB-FP growth tree

Introducing the correlation among existing packets will generate the correlation coefficient and reduce the size of tree and remove redundancy also at a greater extent where frequent packets will be generated and stored within a table arising time to time to the router thus packets are grouped together and stored in the routing table and whenever it is expected to have the same packet again it can directly get routed to the stored link interfaces and the system do not need to update the table again for a homogeneous network environment. Here all correlated patterns are satisfying the conditional pattern for the observed probability to be minimum and every packet will at once be repeated to the same router from the host of that router in that autonomous system so it is possible that packets will be repeated and grouped according to their destination next interface link. This will make low less amount of generating of candidate sets and the network database can be reduced to a far much extent as the RTB-FP tree is highly compressed. Frequency of occurrence of packets can be easily projected in the main memory where traversing is very easy since each packet is stored in form of packet Id and next node or link interface it is attached with.

5.3 Example

To prove this above algorithm follow a simple example to demonstrate the above procedure as follows:

Table 1. Network database with incoming packets

Network database
P1,P2,P5
P2,P4
P2,P3
P1,P2,P4
P1,P3
P2,P3
P1,P3
P1,P2,P3,P5
P1,P2,P3

The following database within shows how a routing matrix is formed with arriving packets at different intervals from different hosts here the packets are basically represented in form of IP addressed.
Now set the frequent packets in order of their decreasing frequency i.e. support count.

Table 2. Packets with frequency of occurrence

Packets	Frequency	Probability(f/n)
P2	7	.30
P1	6	.26
P3	6	.26
P4	2	.086
P5	2	.086

Implementing the above proposed algorithm to construct a tree based on conditional packets within a router. This shows construction of an RTB-FP tree based upon the network matrix is constructed as follows. Here the root consists of null value and every node is represented as a depth first search pre order traversal list.

Table 3. Stored frequent items projection in memory.

Header list/count list							
P2	0	0	0	0	0	0	0
P1	0	0	0	0	0	0	
P3	0	0	0	0	0	0	
P4	0	0					
P5	0	0					

Here the frequent items are projected in the memory a compressed representations of an element thus the tree represented below consist of the link interfaces of its compressed representation.

Now according to the condition of conditional pattern base it can generate the table for observed probability and expected probability depending upon the tree structure. The tree below is traversed in a bottom up approach every node is traversed up to its root node and for every packet it generate a conditional pattern where it can group patterns into their observed probability correlation concept and store it in the forwarding table of the router.

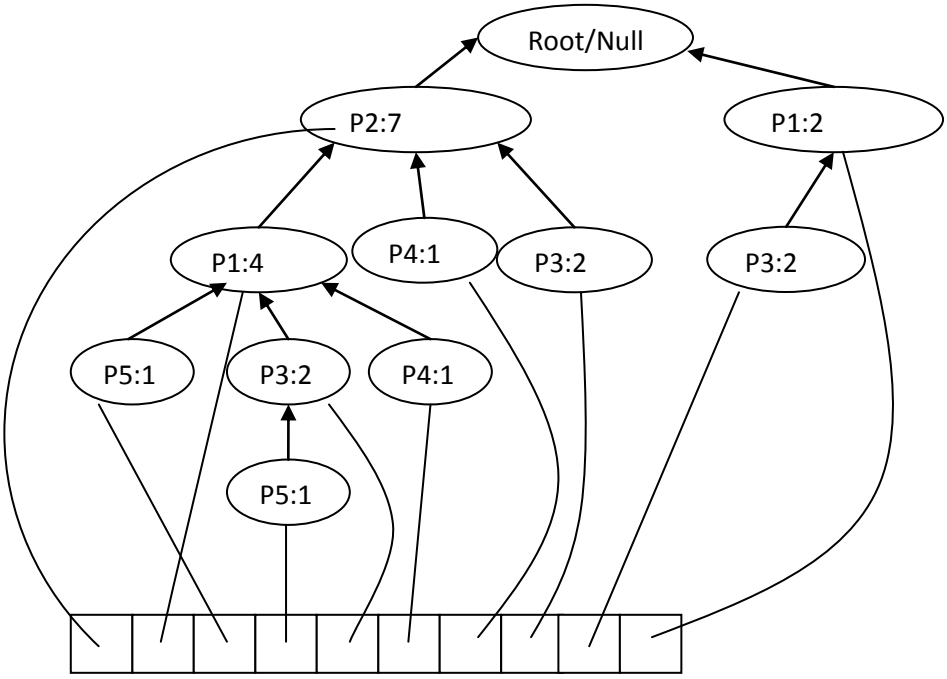

Fig 3: A routing tree based on frequent packets

5.4 Generation of observed probability

The observed probability for the above routing tree can be generated as follows which shows the relation between every node with highest occurrence of packets in this regard observed probability > expected probability of the data packets.

Table 4. Creating conditional pattern for packets

Packets	Conditional Patten(Bottom up approach)
P5	**{{P2,P1:1},{P2,P1,P3:1}}**
P4	**{{P2,P1:1},{P2:1}}**
P3	**{{P2,P1:2},{P2:2},{P1:1}}**
P1	**{P2:4}**

The above table shows how a conditional pattern can be generated via a tree based structure hence in the table below the system can calculate the frequent pattern generated.

Table 5. Frequent pattern for packets

Packets	Frequent pattern Generated
P5	**{{P2,P5:2},{P1,P5:2},{P2,P1,P5:2}}**
P4	**{P2,P4:2}**
P3	**{{P2,P3:4},{P1,P3:4},{P1,P2,P3:2}}**
P1	**{P2,P1:4}**

So the maximum possible combination generated with the data packets is a combination of three packets altogether in the network router configuration. Here the observed pattern shows a minimum of three patterns generated with respect to terminal node packet in a bottom up approach this shows that least support count for probability function should be two packets as then only observed probability can be encountered.

Table 6. Observed pattern for packets

Packets	Frequent Pattern Observed Probability
P5	**{P2,P1,P5: 0.006708 }**
P4	**{P2,P4: 0.0258}**
P3	**{P2,P1,P3: 0.02028}**
P1	**{P2,P1: 0.078}**

5.5 Result Analysis.

The above analysis shows how a pattern can be generated and compressed with the probability of occurrence of elements since every incoming packet is related to its frequency and of arrival from a different host. The result obtained from above algorithm is a tree based frequent item set approach and after obtaining the frequent and compressed pattern it can be stored in a routing table based upon its next hops delivery and incoming interface.

Table 7. Creating new pattern forwarding table.

Packets	Frequency	Probability	Frequent Pattern	Next Hop Interface
P1	6	.30	P1,P2,P5	Link #
P2	7	.26	-	-
P3	6	.26	P2,P1,P3	Link #
P4	2	.086	-	-
P5	2	.086	-	-

The frequent pattern is responsible for its IP addressing mechanism and the mask of the address while the next hop interface defines the next link which the packet will take whenever it encounters a similar pattern from a sending client host which is functioning within a homogeneous network.
So this can come to a conclusion that packets which are frequent within a network router can be identified based upon the above algorithm compressed, recorded and then forwarded.

5.6 Conclusion

This paper hereby presents a mechanism how the system can generate the frequent occurring patterns of any router configuration and achieve high quality information for analysis as the packets are masked and frequent packets can be directed to their next hop sequence directly. Hence the advantages of this algorithm are given as follows:

- As much less is the probability that much less is the probability of occurrence of elements i.e. elements will be frequent but their probability of occurrence in a router based configuration will be less.
- Since the support count of the elements is kept minimum and is generating a frequent pattern it proves this result.
- The observed probability of occurrence of elements is less than probability of individually occurring elements hence redundancy is reduced for packets in a routing table and since router is connected to hosts therefore maximum possible combinational probability can be generated within a routing table to get frequent patterns.
- Whenever the system finds probabilities of large datasets containing packets it becomes easier to generate fast moving data streams such as real time traffic and network monitoring thus clustering ahead becomes easier and compact.
- The tree generated is highly compressed and the redundancies of datasets have been highly reduced due to the grouping of data packets.
- This approach saves time and cost of scanning highly large incoming data packets within a router. This algorithm transforms the problem of finding long frequent patterns into searching for shorter ones in much smaller conditional databases recursively and then concatenation the suffix.
- Therefore this approach reduces the size of pattern to be searched for and generates a routing table that reduces the time complexity on large streams of data.

6. REFERENCES

[1] R.Agrawal and R.Srikant. Fast algorithms for mining association rules.In VLDB'94, pp.487 {499.}

[2] Data Mining: Concepts and Techniques: Concepts and Techniques -Jiawei Han, Micheline Kamber, Jian Pei.2012

[3] Improved algorithm for frequent itemset mining based on apriori and FPtree-Sujata Dandu,B.L.Deekshatulu & Preei Chandra Global Journals Inc. (USA) 2013

[4] An improved frequent pattern tree based association rule mining technique.ICISA-2011

[5] Data communication and networking –Behrouz A forouzan -2006.

[6] An improved association tree mining with FP tree using positive and negative integration-Rashmi & Nitin Shukla-JGRCS-2012

[7] An effective approach in data mining to reduce redundancies in large databases-Sovers Singh & Dr.Sanjeev Bansal-IJETAE-2012

[8] J.Han,J.Pei and Y.Yin," Mining frequent patterns without candidate generation", Proceedings of the ACM SIGMOND ,May 2000.

[9] Contrasting Correlations by an efficient double clique Condition Aixiang Li,Makoto Haraguchi,and Yoshiaki Okubo.

[10] J.Pei, J.Han and H.lu.Hmine: Hyper-structure mining of frequent patterns in large databases.ICDM 2001.pp441-448.