PROJECT REPROT

AKSHAY KULKARNI

Data Structures Used

- Adjacency List is used for storing Projected Database. In Adjacency List we store mapping of transaction-ids with item-ids. Items corresponding to particular transaction are stored in sorted order. Below is sample example to show how projected DB is stored.

  Adjacency List

Transaction – 1	Item - 1	Item - 2	Item - 4	Item - 5
Transaction – 2	Item - 2	Item - 3	Item - 5	
Transaction – 3	Item - 3	Item – 4	Item - 5	
Transaction - 4	Item - 1	Item - 3	Item - 5	

- Two Hash-Maps are used for efficient processing. One Hash-Map is used to store item frequency. This enables us to compute support count for items with single Transaction DB scan. Second Hash-Map is used to store frequent-items along with its support count. This enables in faster generation of association rules.

ALGORITHM

1) Support count for each item is determined in single Transaction DB scan.

2) Now, all infrequent items i.e. item whose support count is less than **minsup** is removed. This pruned Transaction DB is used for generating frequent item-set.

3) During pruning process, vector **(sc)** that stores all frequent items along with its support count is generated. Depending upon **option** parameter, that vector is sorted.

4) Each item in **sc** is processed one by one. For each item**(e)** we remove all transactions that did not contain **e**. This item **e** is added to vector which maintains frequent items in current recursion.

5) In the remaining transactions, we remove all items that are lexicographically smaller than **e.** Since, items in transactions are sorted we can use binary search for removing the items that are lexicographically smaller than **e.**

6) Now we have obtained Projected DB for item **e**. Steps 1 to 5 are now implemented on smaller Transaction DB. The recursion terminates when we have generated all frequent item-sets

ANALYSIS

Using **option 2,** i.e. by processing elements in increasing order of their support count, performed significantly better than option 1 and option3.

Let's assume initial size (number of transactions) of Transaction DB be **S** and average transaction width be **w**. Then let after processing of first element according to option 1,2,3 the sizes be S1, S2, S3 respectively. S2 will be smallest among S1, S2, S3 because item for which projected DB was created had minimum support due to which most to transaction will be eliminated. Thus recursion tree won't grow deep and will die out quickly. On the other hand, if we use option 3, the size of projected DB will reduce slowly due to which recursion will grow deep and time taken to generate frequent item-set will increase.

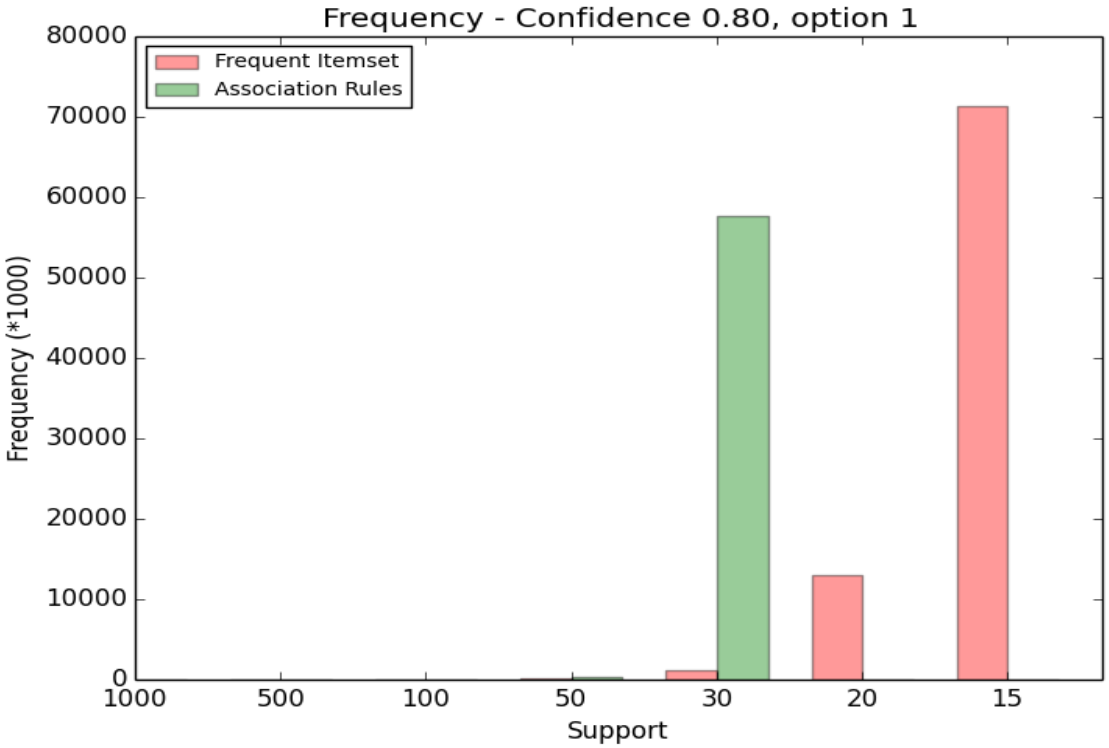
SUMMARY OF RESULTS

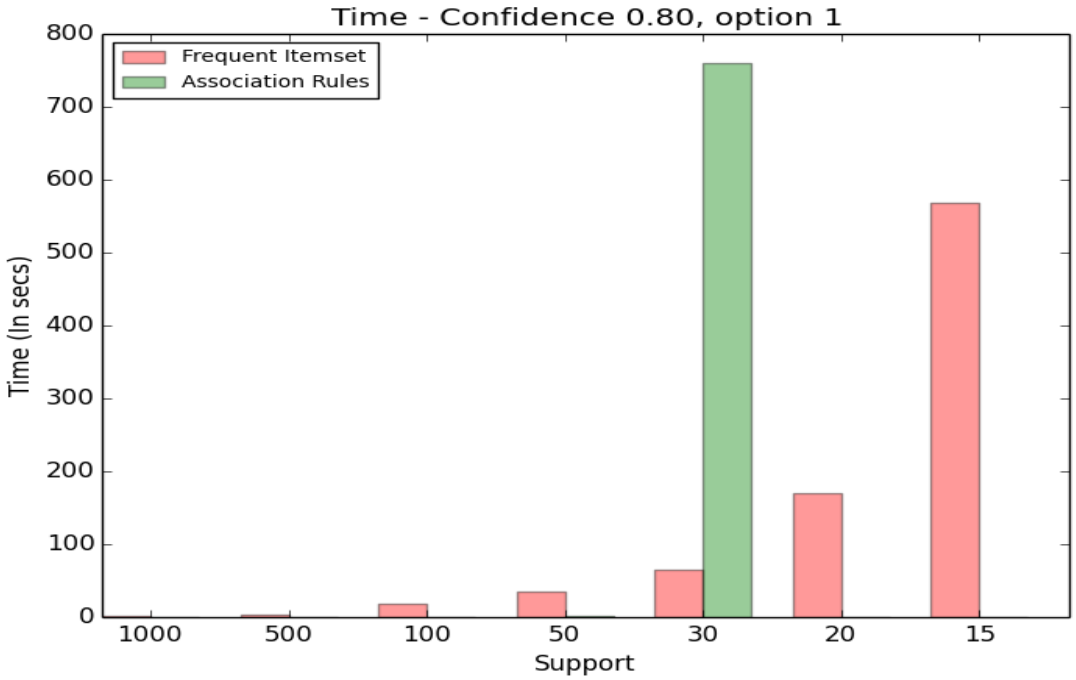
Tests were conducted on one of the CSE Lab machines

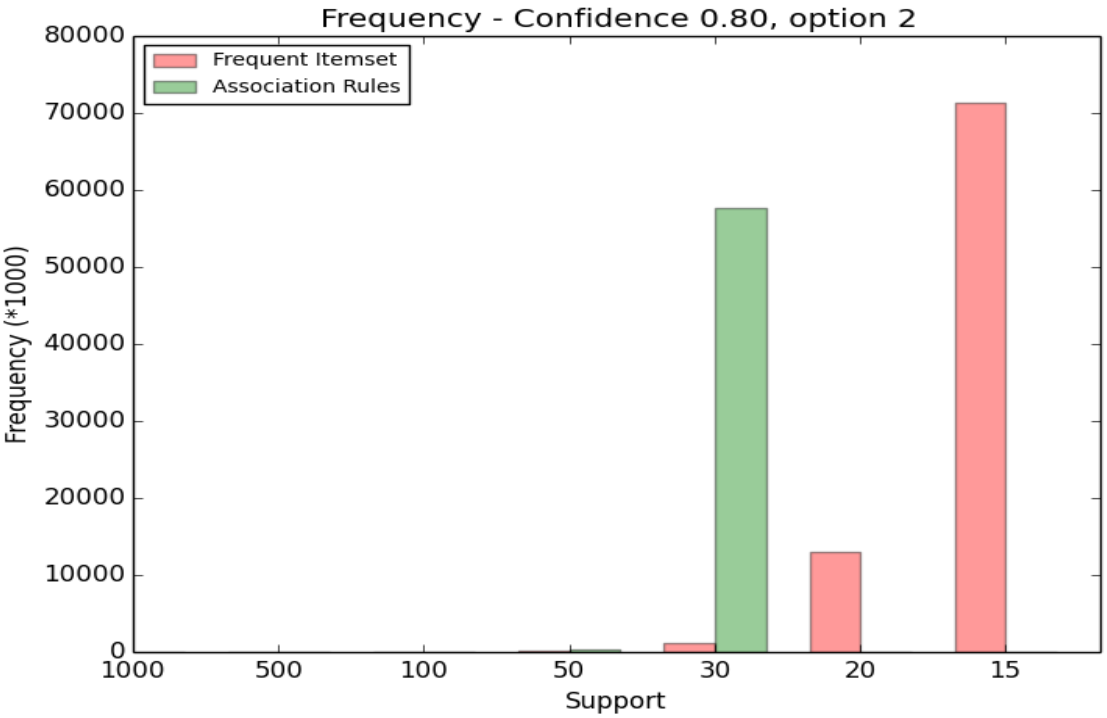
Support	Confidence	Option	# Frequent Item-Sets	# Association Rules	Time for Frequent Item Set (Secs)	Time for Association Rules(Secs)
1000	0.80	1	635	37	0.97585	0.00049
1000	0.80	2	635	37	0.77137	0.00047
1000	0.80	3	635	37	0.97848	0.00049
1000	0.90	1	635	3	0.98335	0.00047
1000	0.90	2	635	3	0.78259	0.00045
1000	0.90	3	635	3	0.97169	0.00046
1000	0.95	1	635	0	0.97482	0.00046
1000	0.95	2	635	0	0.77937	0.00044
1000	0.95	3	635	0	0.97729	0.00045
500	0.80	1	2303	168	2.78148	0.00303
500	0.80	2	2303	168	2.05063	0.00209
500	0.80	3	2303	168	2.78147	0.00225
500	0.90	1	2303	15	2.78217	0.00206
500	0.90	2	2303	15	2.04909	0.00186
500	0.90	3	2303	15	2.78134	0.00206
500	0.95	1	2303	1	2.78253	0.00209
500	0.95	2	2303	1	2.05001	0.00186
500	0.95	3	2303	1	2.77195	0.00203
100	0.80	1	39278	5403	18.48263	0.07940
100	0.80	2	39278	5403	10.68831	0.07660
100	0.80	3	39278	5403	18.47515	0.07725
100	0.90	1	39278	768	18.48352	0.06910
100	0.90	2	39278	768	11.03521	0.07194
100	0.90	3	39278	768	18.56809	0.06885
100	0.95	1	39278	197	18.52267	0.07017
100	0.95	2	39278	197	10.79977	0.07115
100	0.95	3	39278	197	18.53351	0.06994
50	0.80	1	157915	329304	35.44647	2.21037
50	0.80	2	157915	329304	17.39766	2.25917
50	0.80	3	157915	329304	32.90120	2.12954
50	0.90	1	157915	160077	35.38391	1.36499

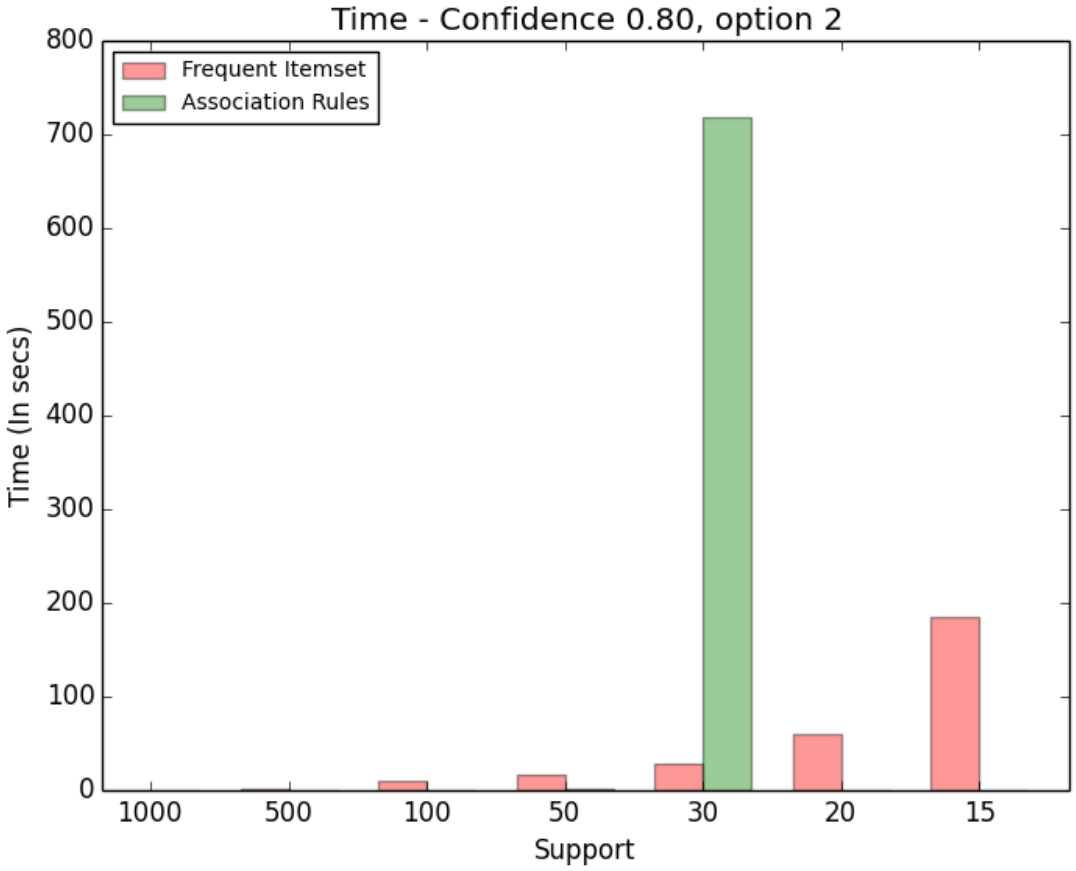
50	0.90	2	157915	160077	17.36397	1.42803
50	0.90	3	157915	160077	35.26747	1.37116
50	0.95	1	157915	65679	35.31544	0.80421
50	0.95	2	157915	65679	17.46071	0.86978
50	0.95	3	157915	65679	36.33309	0.85020
30	0.80	1	1197364	57726221	64.60285	759.67109
30	0.80	2	1197364	57726221	28.40344	718.31389
30	0.80	3	1197364	57726221	61.46600	774.28957
30	0.90	1	1197364	26168852	60.52567	350.90224
30	0.90	2	1197364	26168852	28.05649	316.41924
30	0.90	3	1197364	26168852	60.44548	351.43600
30	0.95	1	1197364	11569806	60.21311	166.14495
30	0.95	2	1197364	11569806	28.25273	148.90896
30	0.95	3	1197364	11569806	64.57120	171.21523
20	0.80	1	12917208	0	170.14184	0.00000
20	0.80	2	12917208	0	60.43505	0.00000
20	0.80	3	12917208	0	175.87408	0.00000
20	0.90	1	12917208	0	171.16314	0.00000
20	0.90	2	12917208	0	60.53357	0.00000
20	0.90	3	12917208	0	176.03557	0.00000
20	0.95	1	12917208	0	171.18156	0.00000
20	0.95	2	12917208	0	60.32637	0.00000
20	0.95	3	12917208	0	175.81910	0.00000
15	0.80	1	71274569	0	569.05801	0.00000
15	0.80	2	71274569	0	184.70767	0.00000
15	0.80	3	71274569	0	565.92290	0.00000
15	0.90	1	71274569	0	570.59370	0.00000
15	0.90	2	71274569	0	184.60955	0.00000
15	0.90	3	71274569	0	567.60210	0.00000
15	0.95	1	71274569	0	570.21841	0.00000
15	0.95	2	71274569	0	184.63225	0.00000
15	0.95	3	71274569	0	565.18789	0.00000

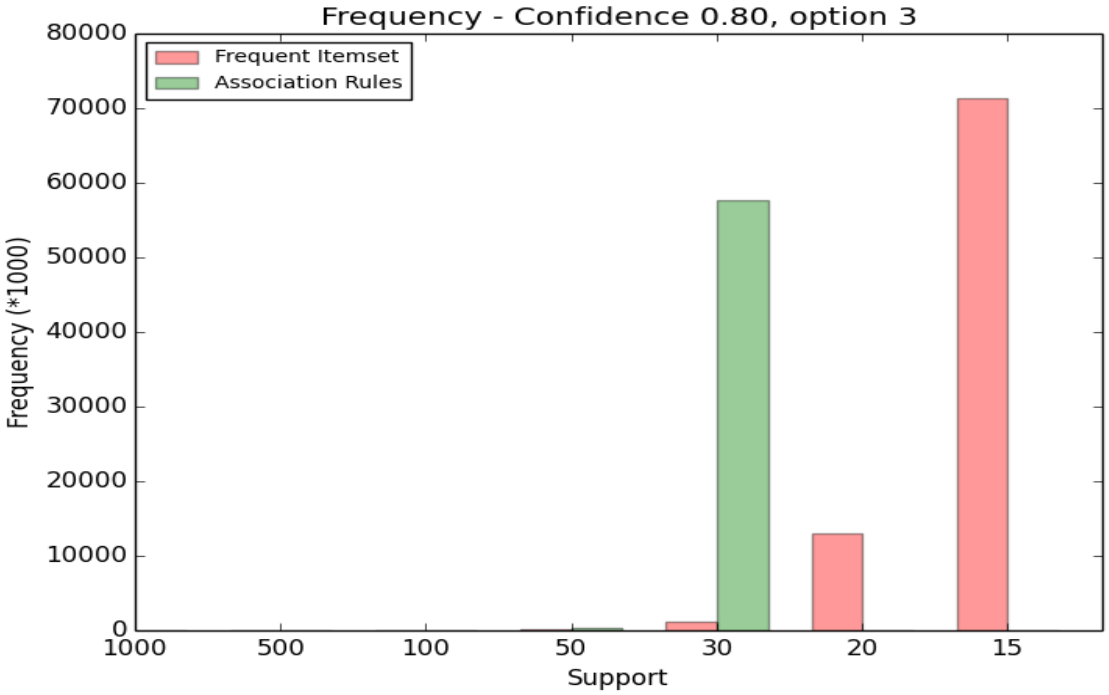
BAR GRAPH FOR LARGE DATA-SETS

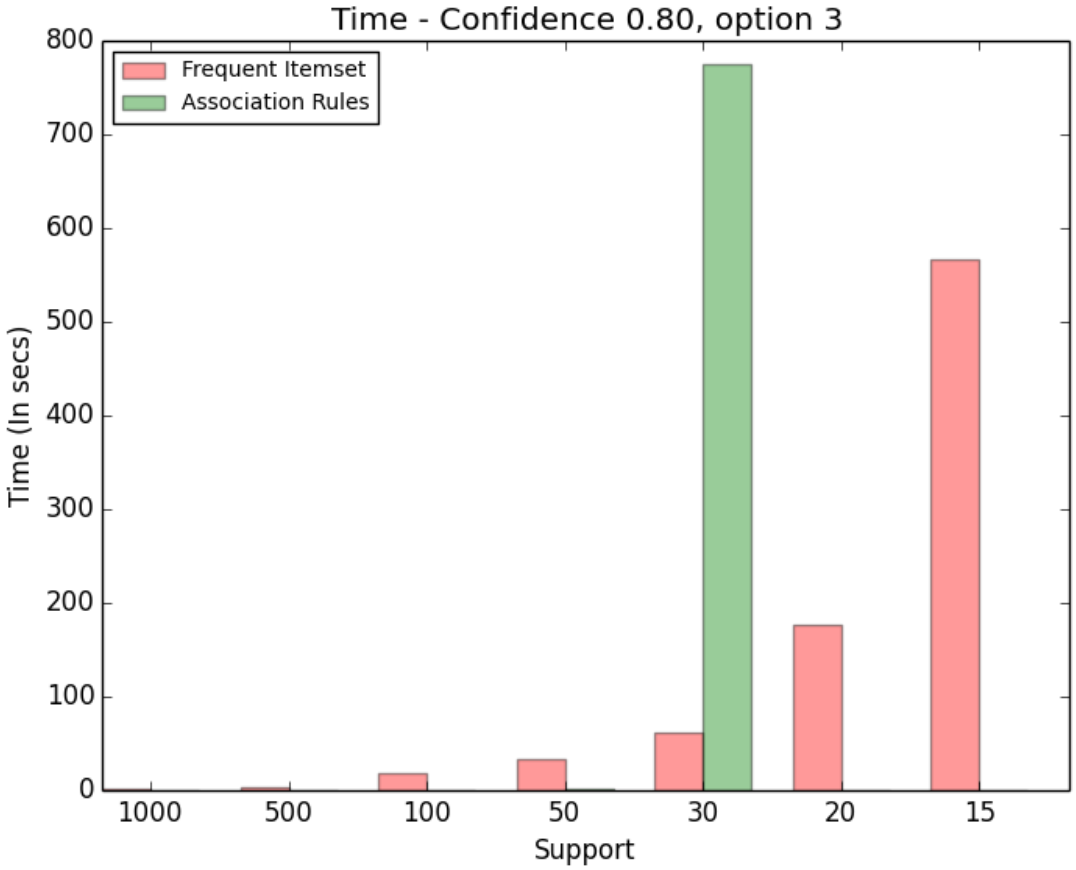


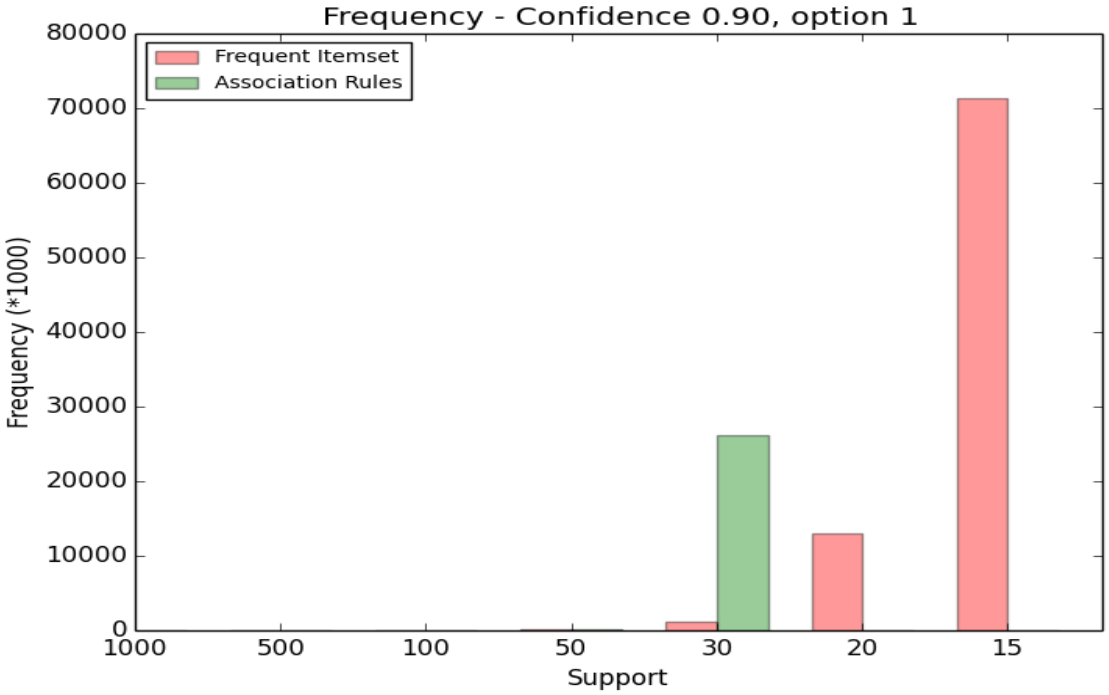


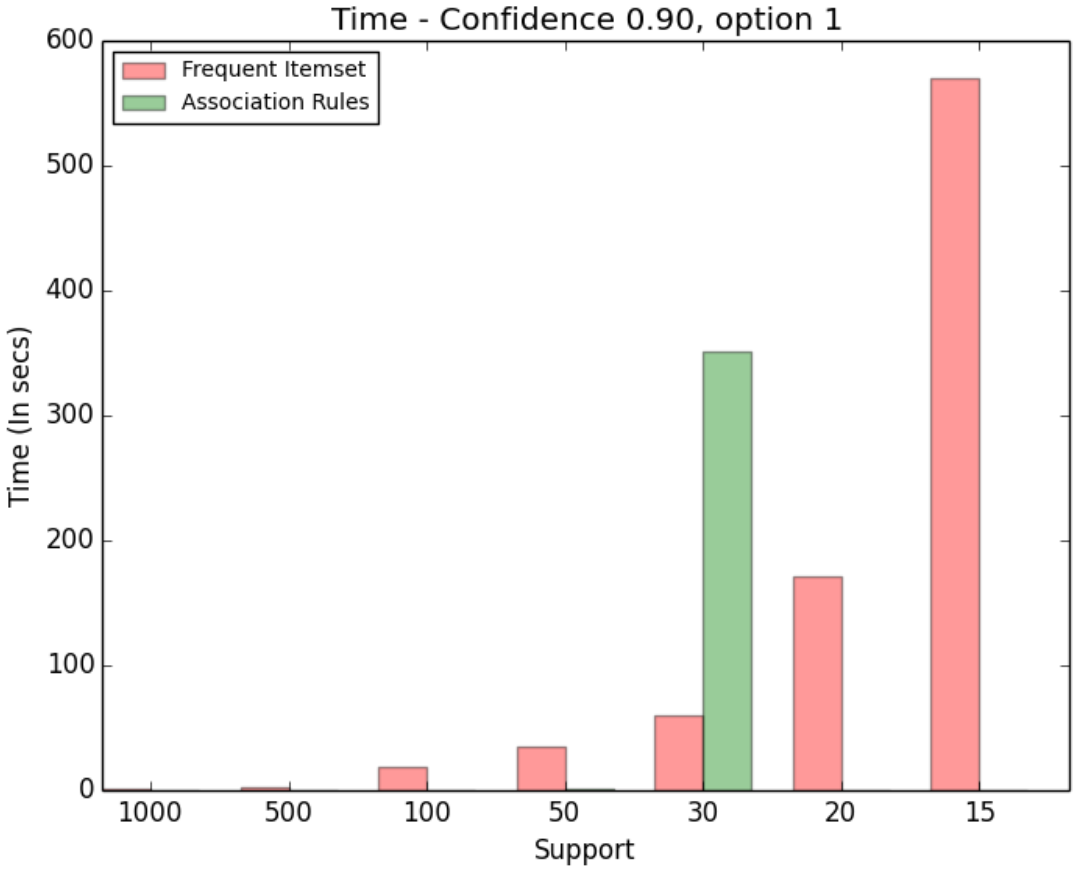

Frequency - Confidence 0.80, option 2
Frequent Itemset
Association Rules
Frequency (*1000)
80000
70000
60000
50000
40000
30000
20000
10000
0
1000
500
100
50
30
20
15
Support

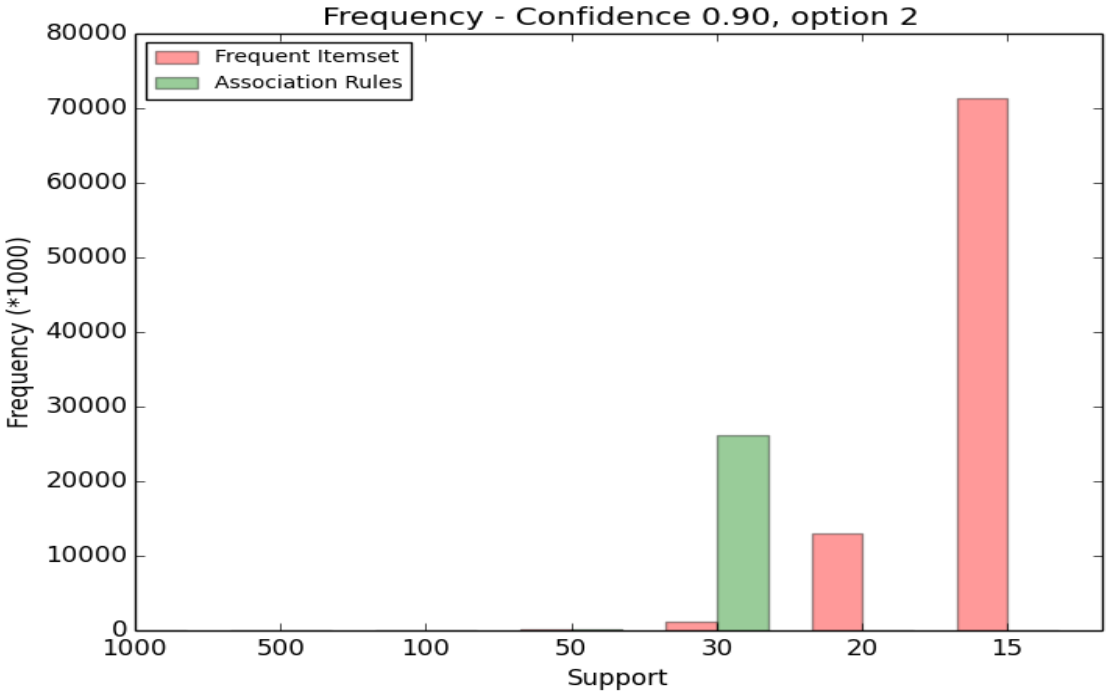

Time - Confidence 0.80, option 2
Frequent Itemset
Association Rules
Time (ln secs)
800
700
600
500
400
300
200
100
0
1000
500
100
50
30
20
15
Support

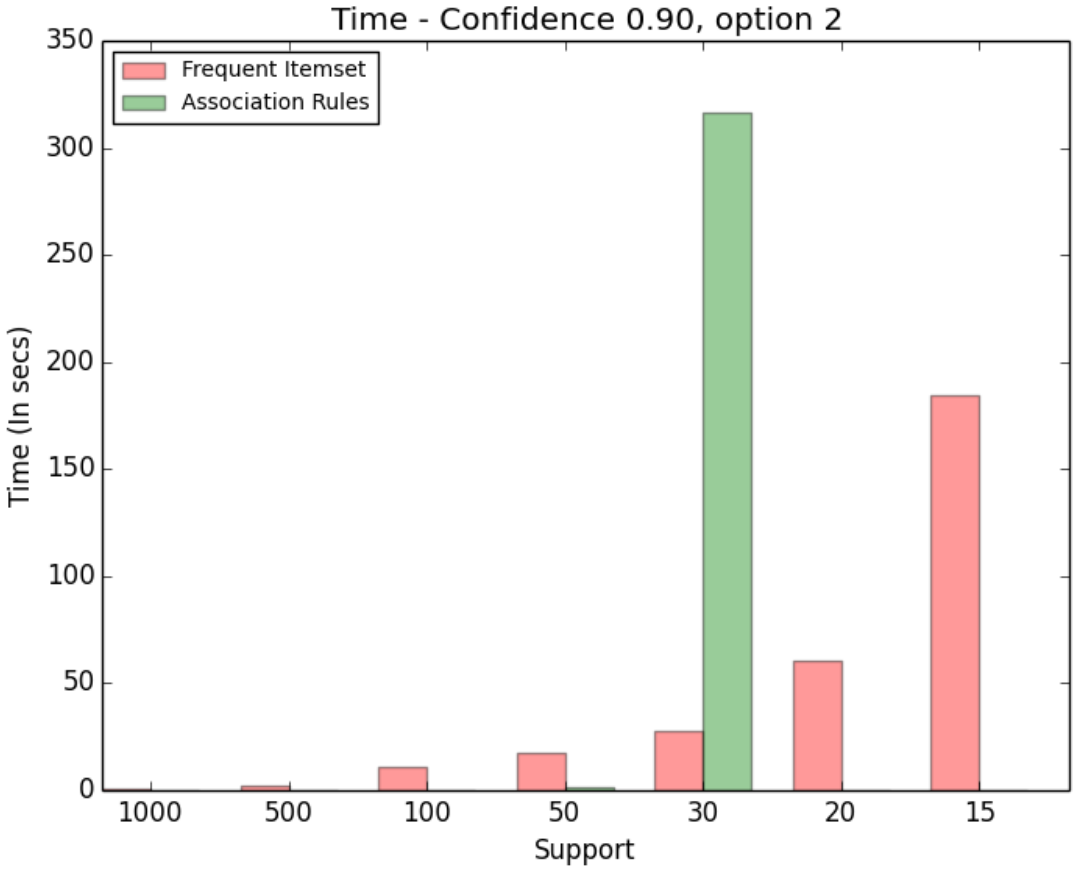

Frequency - Confidence 0.80, option 3
Frequent Itemset
Association Rules
Frequency (*1000)
80000
70000
60000
50000
40000
30000
20000
10000
0
1000
500
100
50
30
20
15
Support

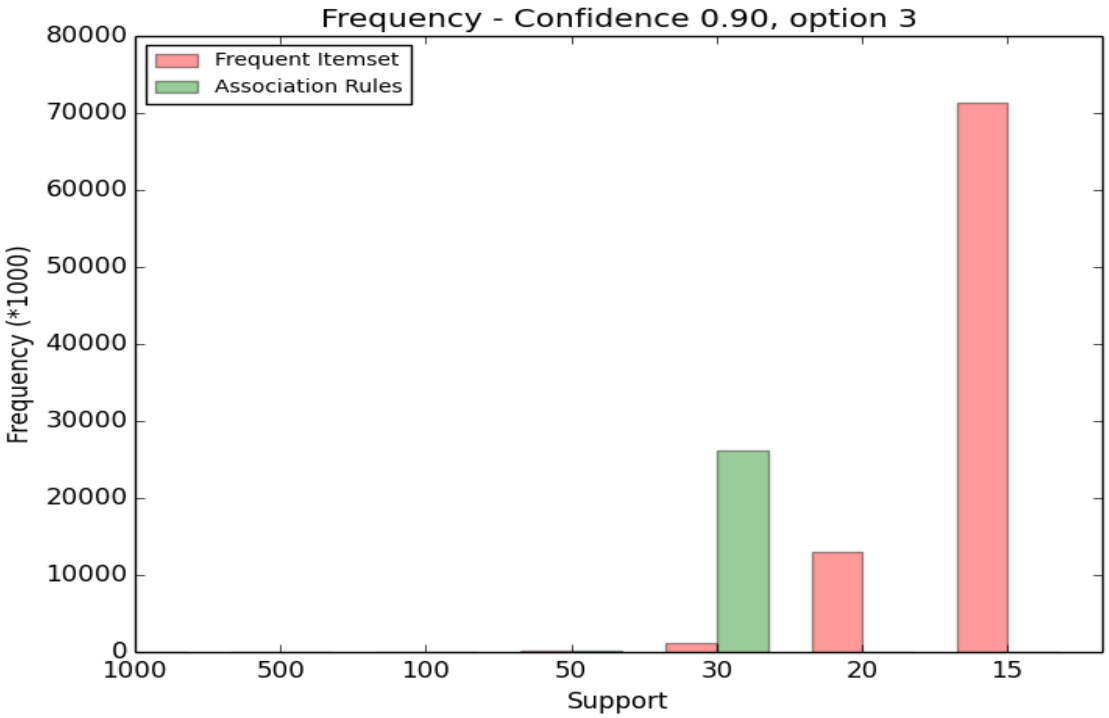

Time - Confidence 0.80, option 3
Frequent Itemset
Association Rules
Time (In secs)
800
700
600
500
400
300
200
100
0
1000
500
100
50
30
20
15
Support

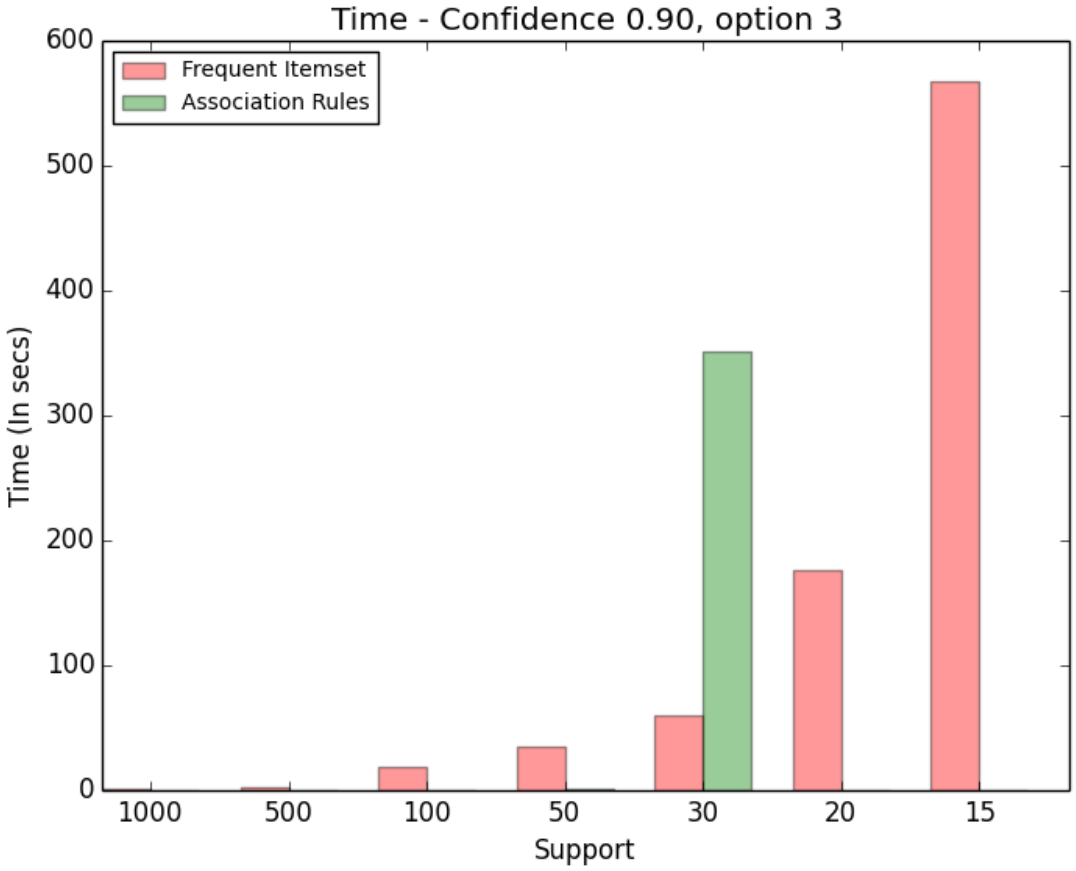

Frequency - Confidence 0.90, option 1
Frequent Itemset
Association Rules
Frequency (*1000)
80000
70000
60000
50000
40000
30000
20000
10000
0
1000
500
100
50
30
20
15
Support

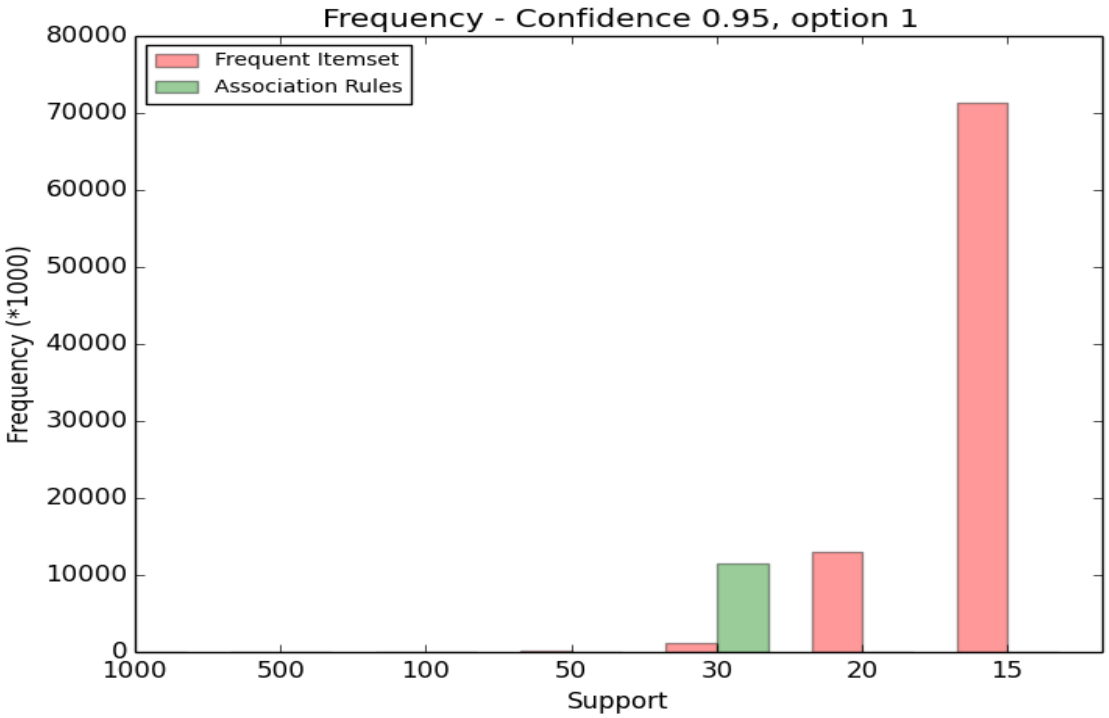

Time - Confidence 0.90, option 1
Frequent Itemset
Association Rules
Time (In secs)
600
500
400
300
200
100
0
1000
500
100
50
30
20
15
Support

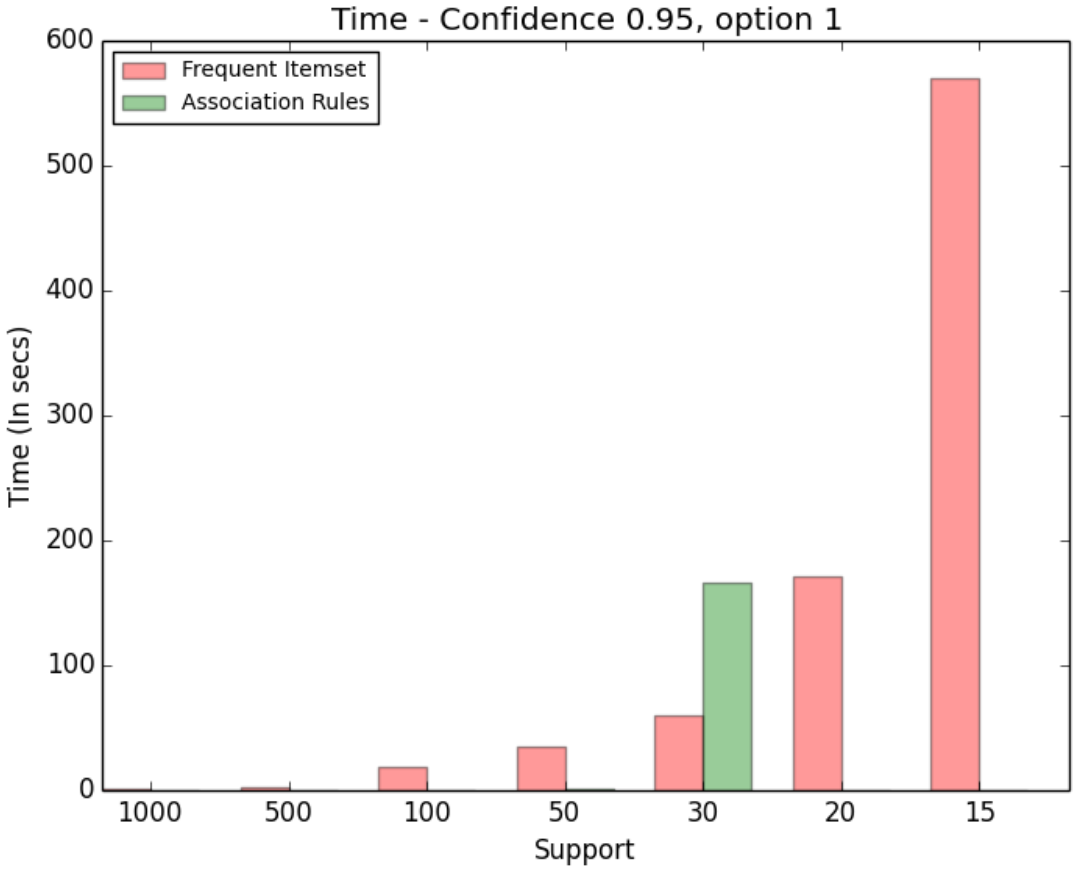

Frequency - Confidence 0.90, option 2
Frequent Itemset
Association Rules
Frequency (*1000)
80000
70000
60000
50000
40000
30000
20000
10000
0
1000
500
100
50
30
20
15
Support

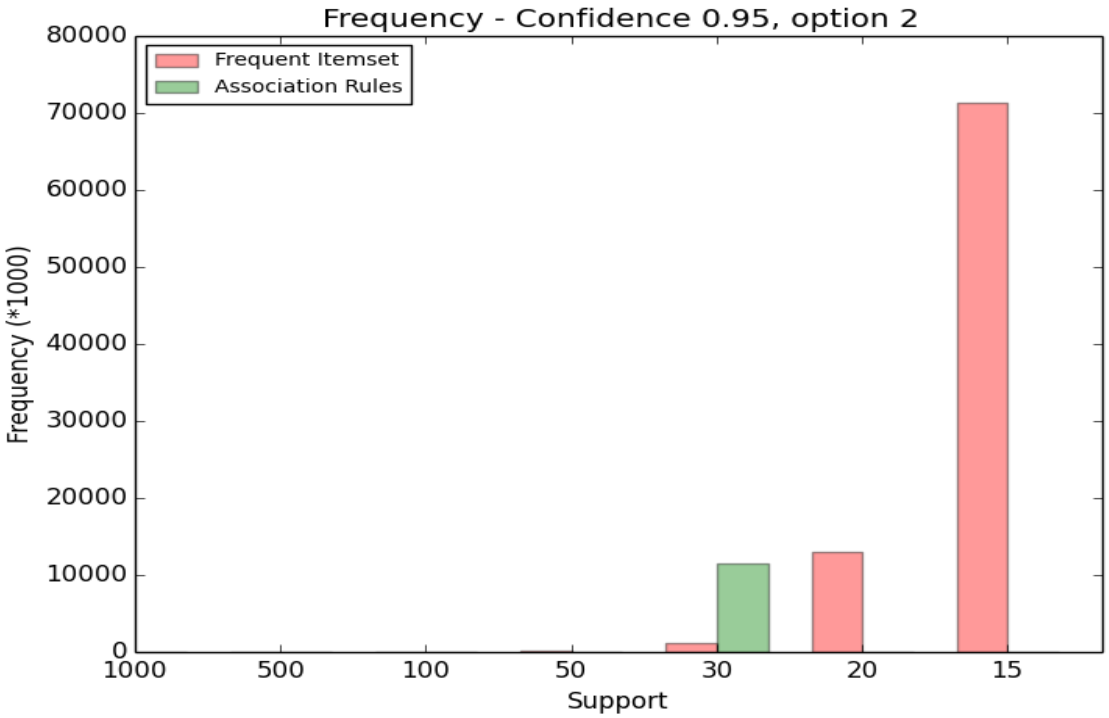

Time - Confidence 0.90, option 2
Frequent Itemset
Association Rules
Time (In secs)
350
300
250
200
150
100
50
0
1000
500
100
50
30
20
15
Support

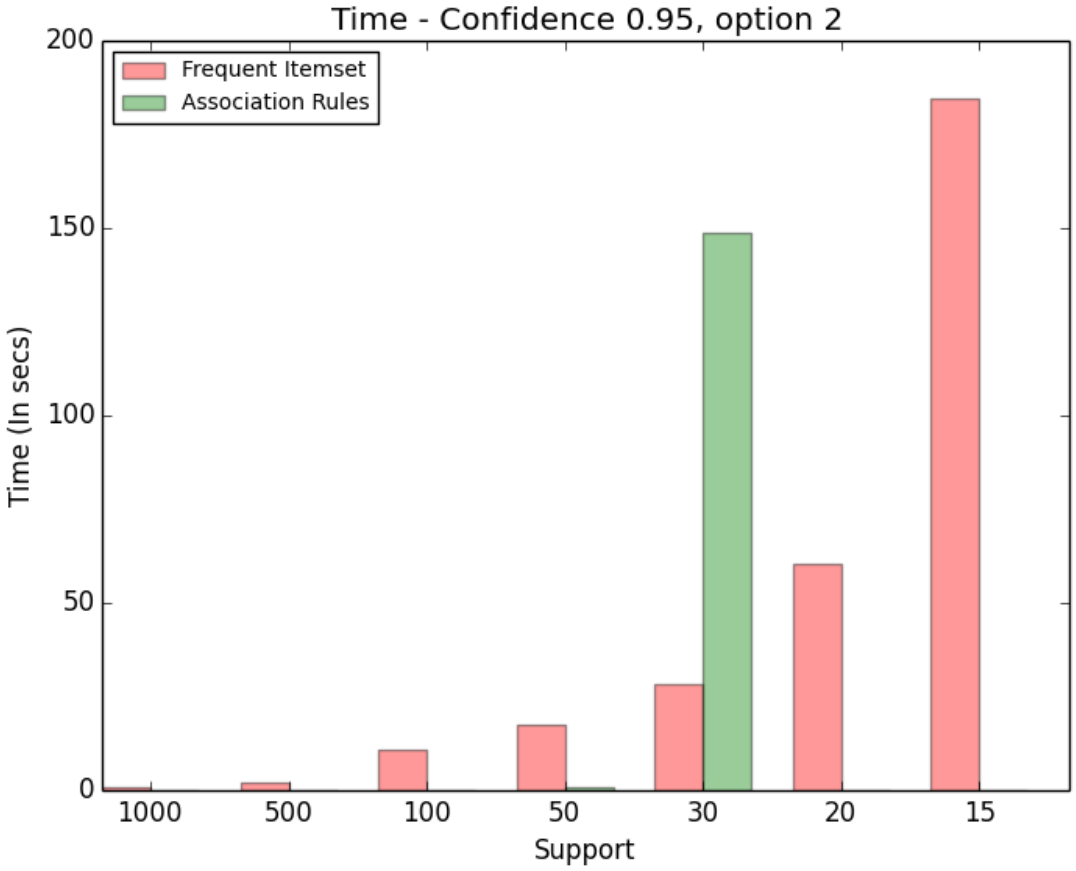

Frequency - Confidence 0.90, option 3
Frequent Itemset
Association Rules
Frequency (*1000)
80000
70000
60000
50000
40000
30000
20000
10000
0
1000
500
100
50
30
20
15
Support

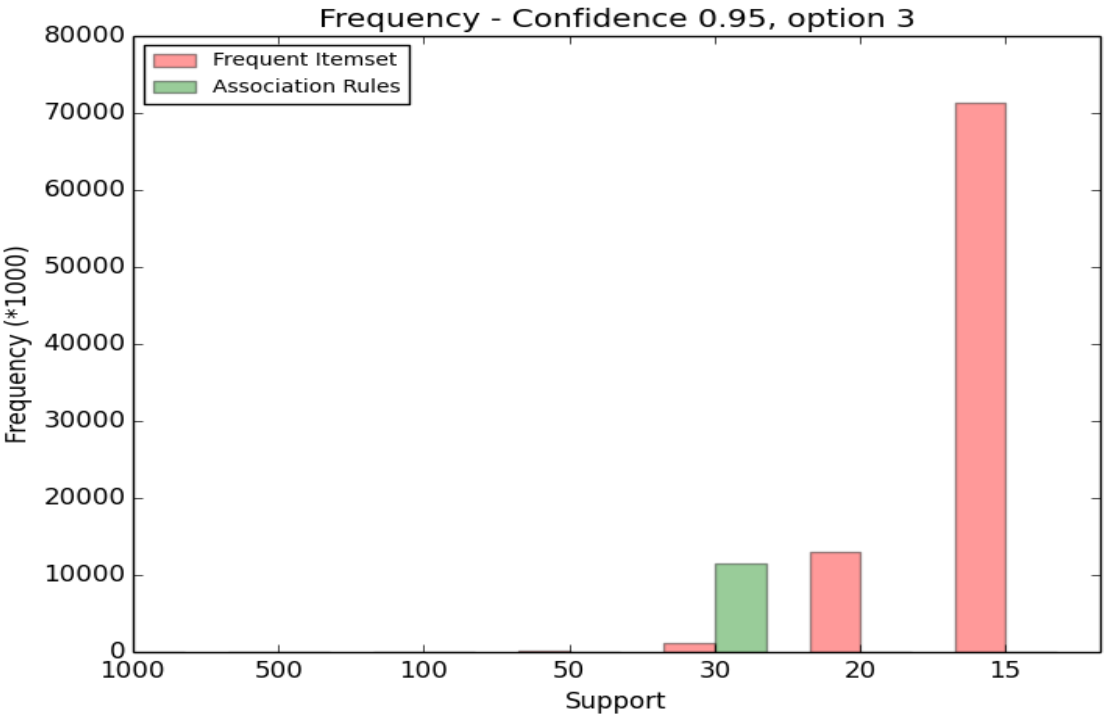

Time - Confidence 0.90, option 3
Frequent Itemset
Association Rules
Time (In secs)
600
500
400
300
200
100
0
1000
500
100
50
30
20
15
Support

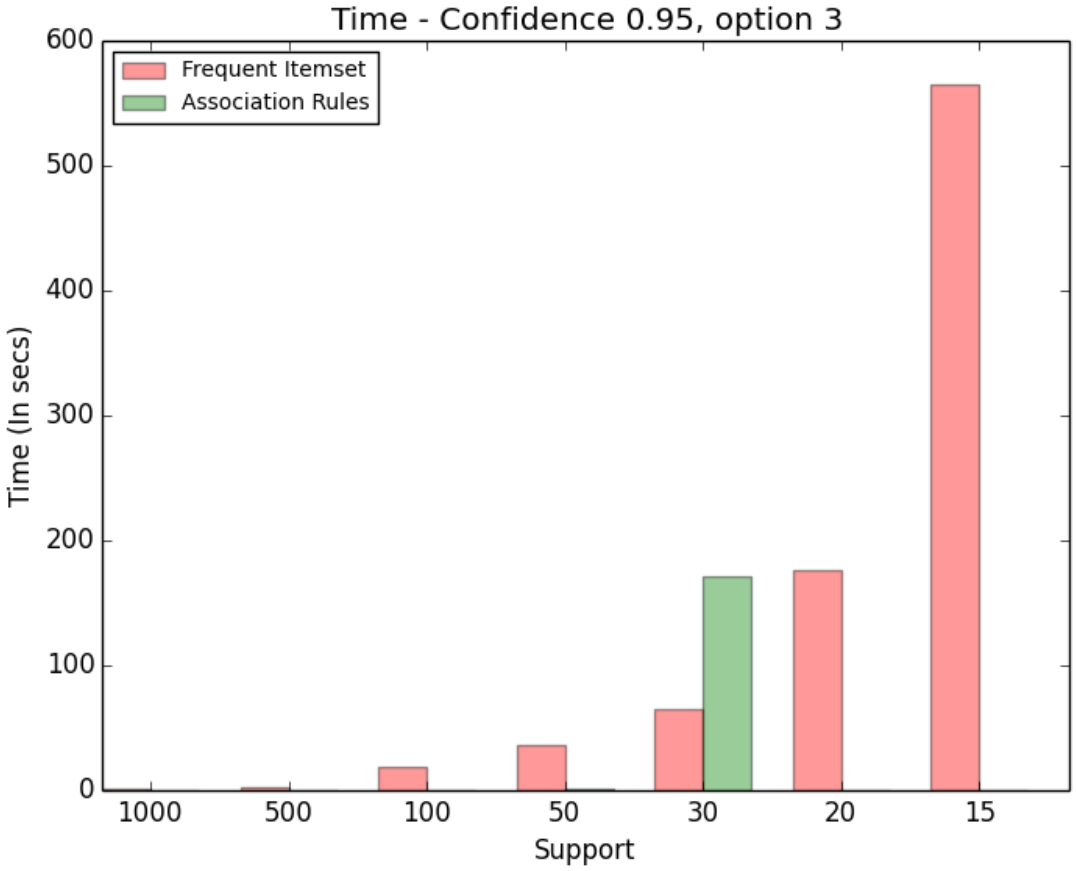

Frequency - Confidence 0.95, option 1
Frequent Itemset
Association Rules
Frequency (*1000)
80000
70000
60000
50000
40000
30000
20000
10000
0
1000
500
100
50
30
20
15
Support


Time - Confidence 0.95, option 1
Frequent Itemset
Association Rules
Time (In secs)
600
500
400
300
200
100
0
1000
500
100
50
30
20
15
Support


Frequency - Confidence 0.95, option 2
Frequent Itemset
Association Rules
Frequency (*1000)
80000
70000
60000
50000
40000
30000
20000
10000
0
1000
500
100
50
30
20
15
Support


Time - Confidence 0.95, option 2
Frequent Itemset
Association Rules
Time (ln secs)
200
150
100
50
0
1000
500
100
50
30
20
15
Support


Frequency - Confidence 0.95, option 3
Frequent Itemset
Association Rules
Frequency (*1000)
80000
70000
60000
50000
40000
30000
20000
10000
0
1000
500
100
50
30
20
15
Support


Time - Confidence 0.95, option 3
Frequent Itemset
Association Rules
Time (ln secs)
600
500
400
300
200
100
0
1000
500
100
50
30
20
15
Support