

Efficient Frequent Pattern Mining Algorithm for Many Task Environments

Domain: Data Mining

Team:

Indu Sanka(21910104041)

Kiruppa Kalyanaraman(21910104055)

Guide:

Dr. V Vidhya

Professor

Department of Computer Science and Engineering

Abstract

- Mining frequent patterns - complex problem in data mining
- Research - FP using parallel and distributed techniques
- Mostly focused - single-task over multi-task environments
- Algorithms proposed can be used for different sized datasets
- Example: Inventory

Introduction

- FP Mining - searching for an FP more than a specified threshold
- FP Mining has 2 kinds of approach: Apriori-like Approach and FP Growth like approach
- Disadvantages of Apriori-like approach - cost of memory, scanning and scalability

Introduction (cont..)

- FP Growth uses - FP Trees in which transactions are compressed
- Advantage - better scalability and lesser execution time
- Large FP tree is a combination of 1 or more cause: data characteristics, user characteristics and mining parameters
- Large FP tree needs large DB, thus increase in computation time and memory

Current Scenario

- Multiprocessor architectures
- Tree projection technique
- 3rd method was distributing the DB across processors
- All have many disadvantages
- Thus proposed new method

Literature Survey

PAPER	PUBLISHED BY	APPROACH	LIMITATIONS
Frequent Pattern Mining on Message Passing Multiprocessor System	A. Javed, A. Khokhar(2004),Distributed and Parallel Database, vol. 16,pp. 321-334	<ul style="list-style-type: none">- Presents a scalable parallel algorithm for mining FP.- Technique based on FP-growth algorithm.- Frequent item list is partitioned over processors for minimum common overhead	<ul style="list-style-type: none">- Machines were found to be expensive.- Cannot preserve data privacy.

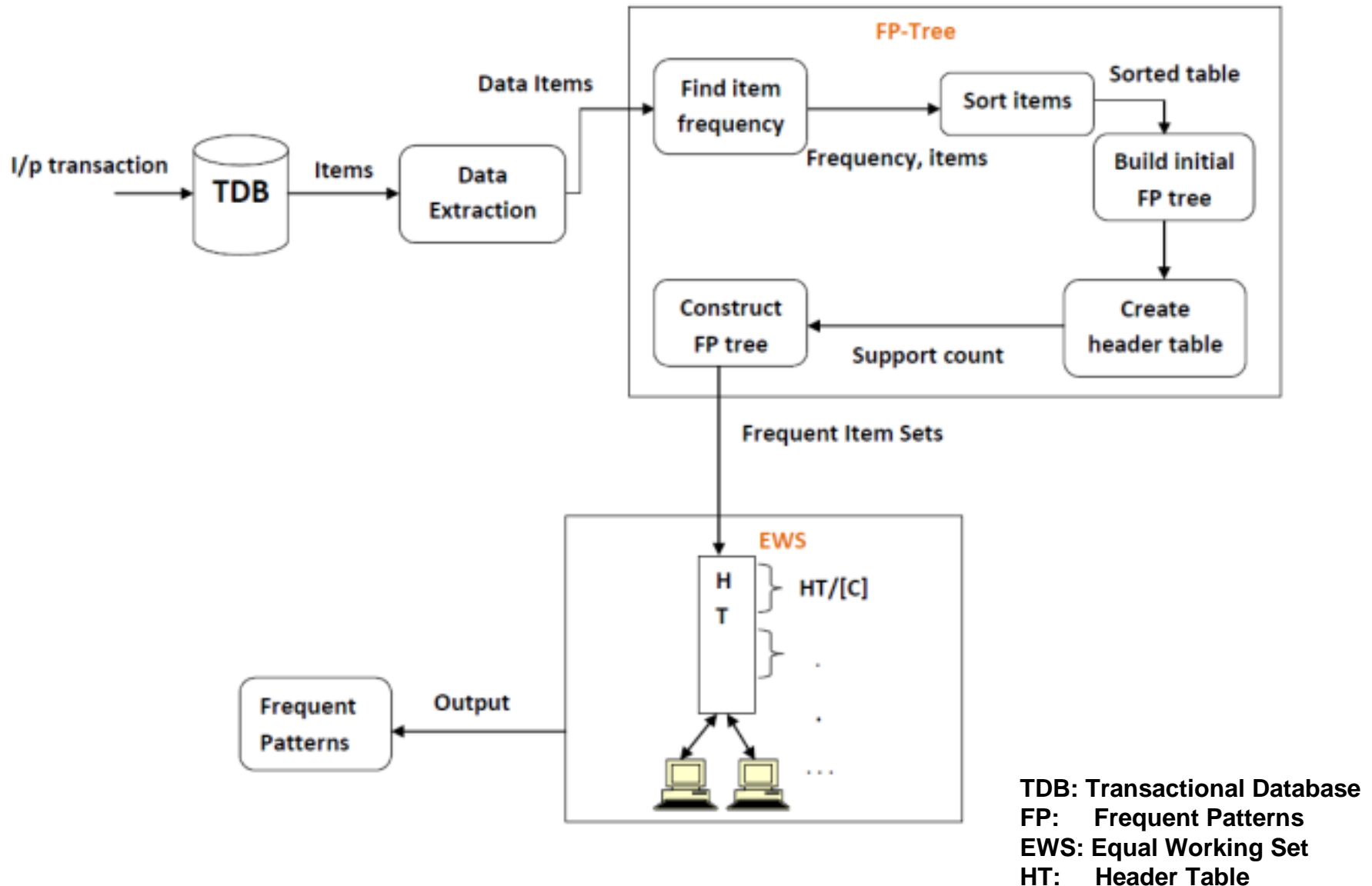
Literature Survey (cont..)

PAPER	PUBLISHED BY	APPROACH	LIMITATIONS
Scalable Parallel Data Mining for Association Rules	E.H.S. Han, G. Karypis, V. Kumar(2000), IEEE Transaction on Knowledge and Data Engineering, vol. 12, pp. 352-377	<ul style="list-style-type: none">-Computing association rules based on Apriori algorithm.-The algorithm partitions candidate set among processors to build hash tree.	<ul style="list-style-type: none">-Approach duplicates the database to nodes. Thus, risking leakage of data.

Literature Survey (cont..)

PAPER	PUBLISHED BY	APPROACH	LIMITATIONS
A novel parallel algorithm for frequent pattern mining with privacy preserved in cloud computing environment	K.W. Lin, D.J. Deng(2010), International Journal of Ad Hoc and Ubiquitous Computing, vol. 6,pp. 205-215	<ul style="list-style-type: none">-Proposed a data mining algorithm CARM to discover frequent pattern in cloud.-Data privacy is preserved.	<ul style="list-style-type: none">-Mining time for each conditional pattern base is different.-The required time cannot be known in advance.

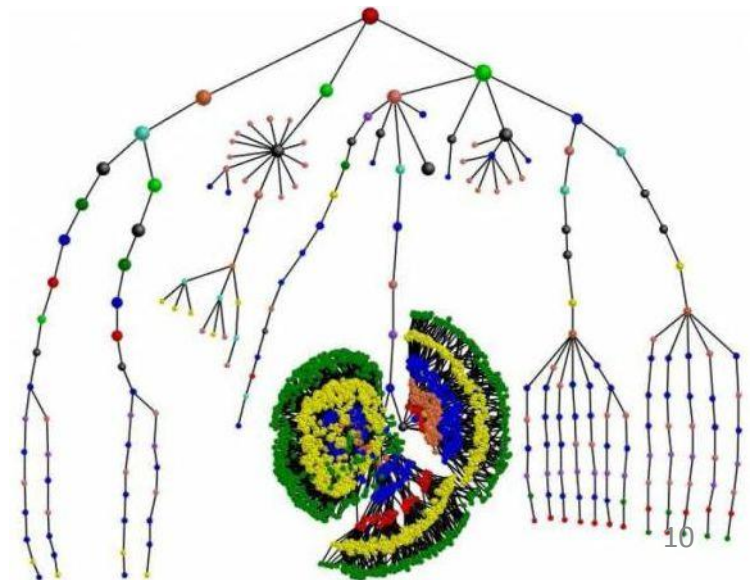
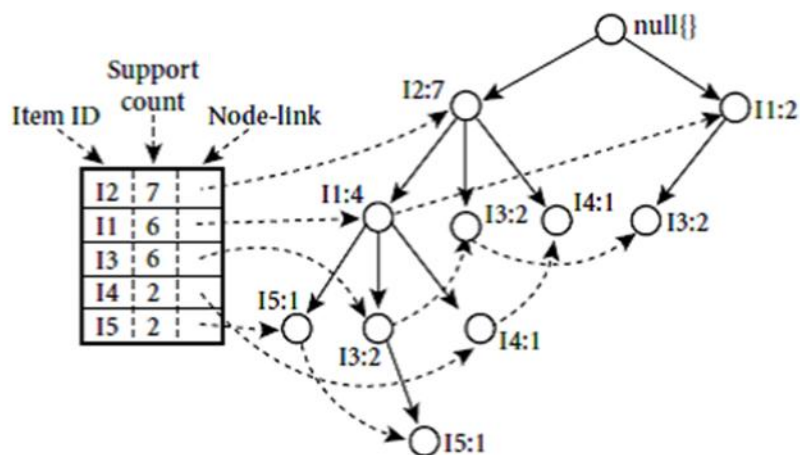
Block Diagram of Proposed System



Modules

- **Frequent Item Set Generation**

1. Uses support count to generate frequent item sets
2. Used as input for pattern mining
3. I/p transaction used – retail
4. FPMining Algorithm - frequent itemset generation
5. Constructs FP Trees



Frequent ItemSet Generation Example

TID	Items
1	E, A, D, B
2	D, A, C, E, B
3	C, A, B, E
4	B, A, D
5	D
6	D, B
7	A, D, E
8	B, C

Table 1 - Snapshot of the Database

Item	Frequency
A	5 3
B	6 1
C	3 5
D	6 2
E	4 4

Table2 -Frequency of Occurrence

TID	Items	Ordered Items
1	E, A, D, B	B,D,A,E
2	D, A, C, E, B	B,D,A,E,C
3	C, A, B, E	B,A,E,C
4	B, A, D	B,D,A
5	D	D
6	D, B	B,D
7	A, D, E	D,A,E
8	B, C	B,C

Table 3 - New version of the Table 1

Frequent ItemSet Generation

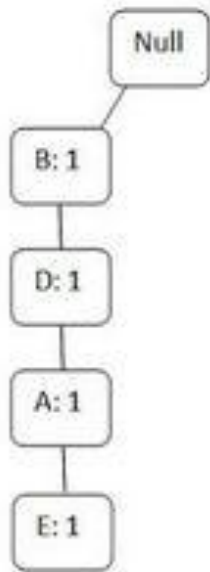


Figure 1- FP tree for Row 1

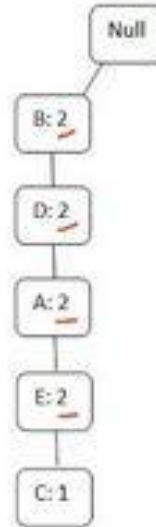


Figure 2- FP tree for Row 1,2

Row 5:

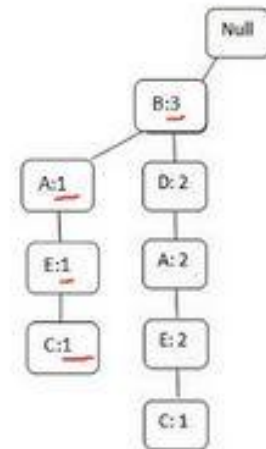


Figure 3 - After adding third row

Frequent ItemSet Generation

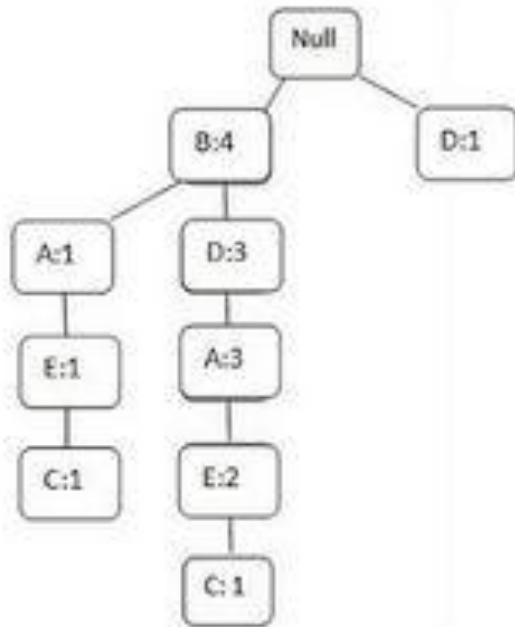


Figure 4- Connect D to null node

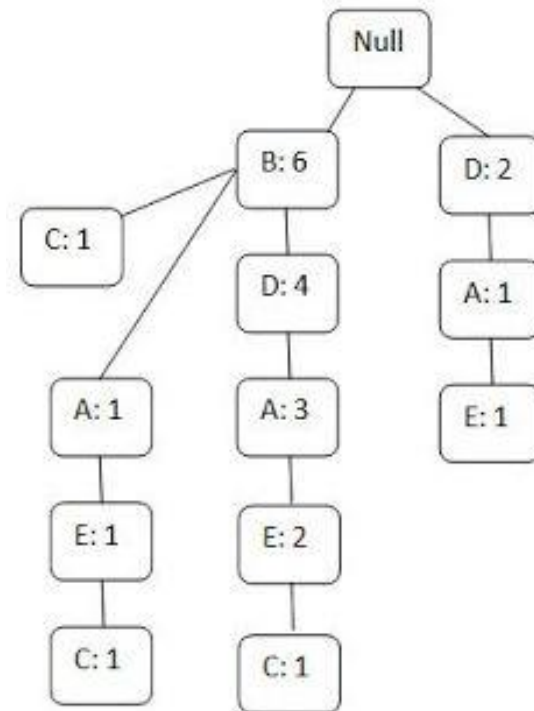


Figure 5 - Final FP tree

Modules cont..

- **EWS** (Equal Working Set)
 1. Partitions items in the HT
 2. Sends items to the node
 3. Generates Frequent Patterns from items
 4. Frequent Patterns collected from individual node
 5. Combined to form final Frequent Pattern

Modules cont..

EWS Algorithm:

Input: Transaction database DB, min. Supp threshold \$, computing nodes C

Output: Complete set of frequent patterns FP

```
HT=getHeaderTable(DB,$)
```

```
FPT=constructFPtree(DB,$)
```

```
L1=EqualPartitionHT(HT,|C|)
```

```
FOREACH L1i in L1
```

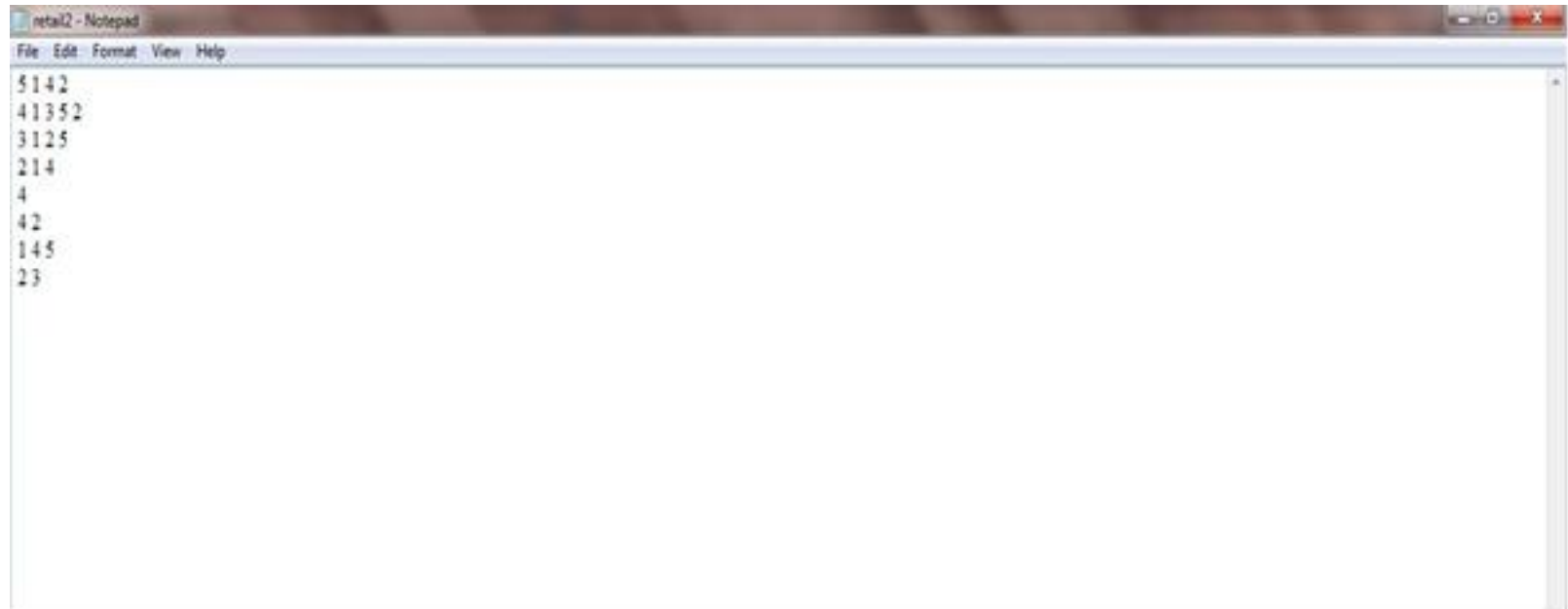
```
    n=getAvailableNode(C)
```

```
    FP=FPUFPM(L1i,FPT,n)
```

```
ENDFOR
```

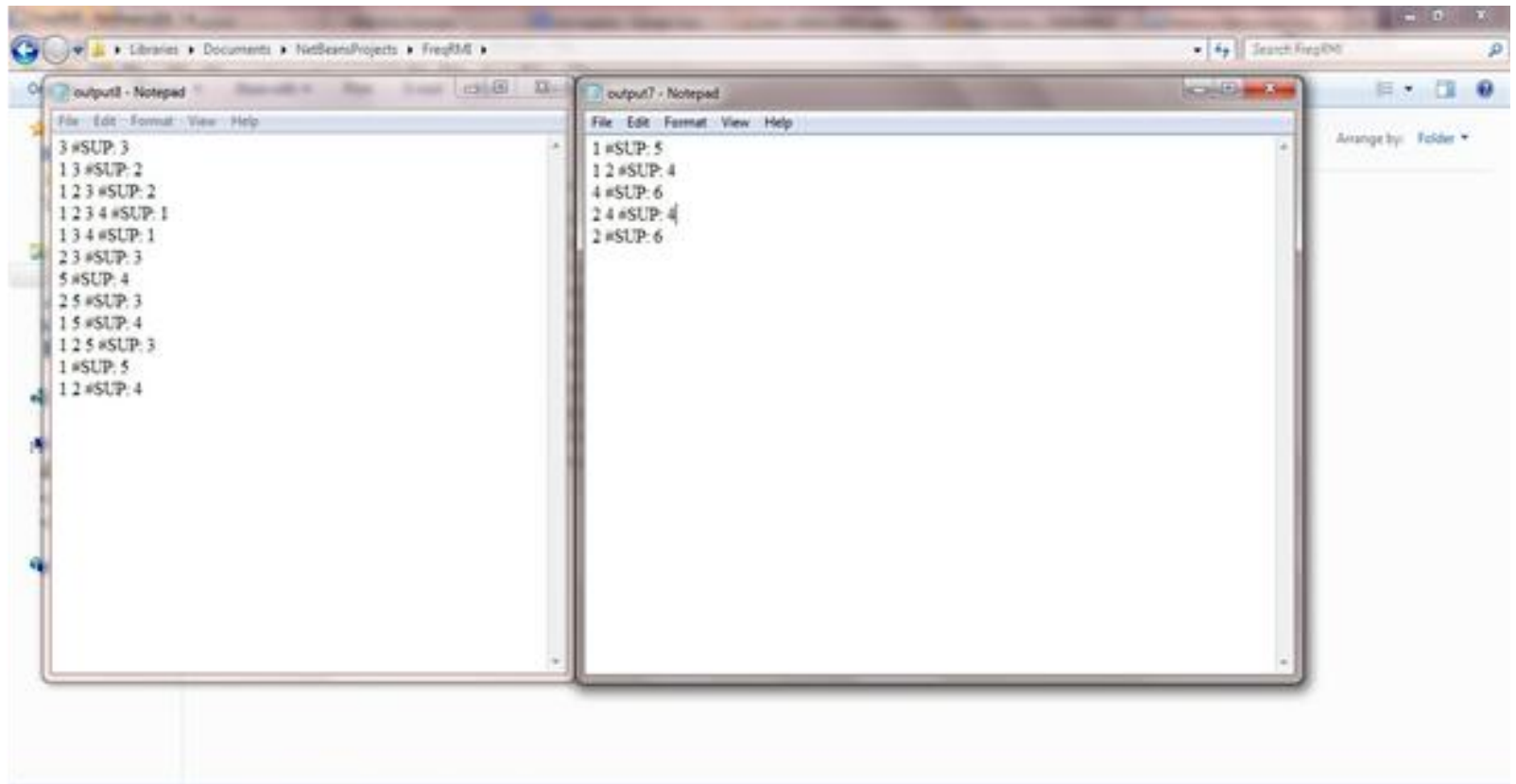
```
RETURN FP
```

Screenshots



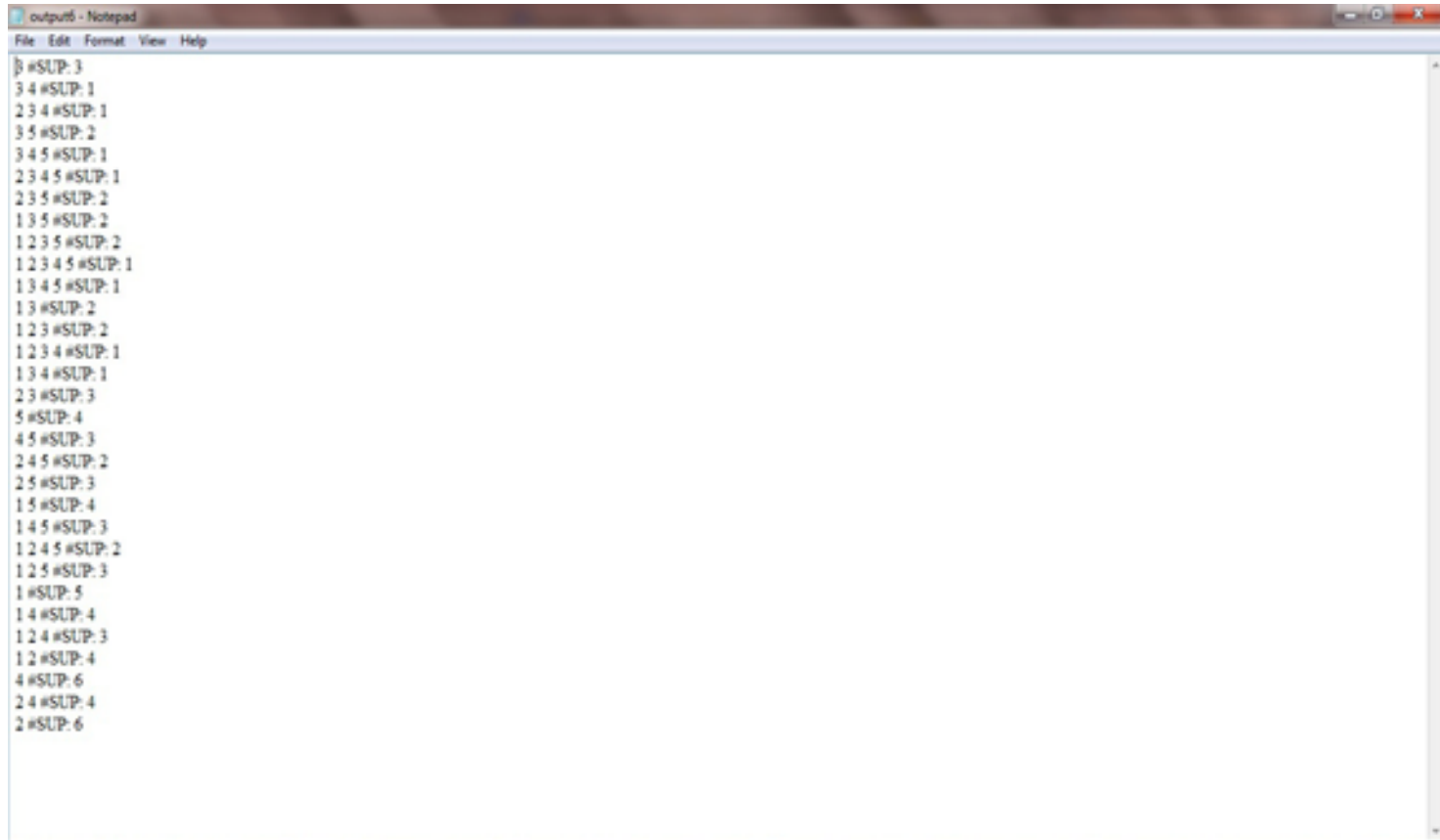
Sample retail shop transactions

Screenshots (cont..)



Frequent Patterns generated using EWS algorithm

Screenshots (cont..)



```
output6 - Notepad
File Edit Format View Help
[] #SUP: 3
3 4 #SUP: 1
2 3 4 #SUP: 1
3 5 #SUP: 2
3 4 5 #SUP: 1
2 3 4 5 #SUP: 1
2 3 5 #SUP: 2
1 3 5 #SUP: 2
1 2 3 5 #SUP: 2
1 2 3 4 5 #SUP: 1
1 3 4 5 #SUP: 1
1 3 #SUP: 2
1 2 3 #SUP: 2
1 2 3 4 #SUP: 1
1 3 4 #SUP: 1
2 3 #SUP: 3
5 #SUP: 4
4 5 #SUP: 3
2 4 5 #SUP: 2
2 5 #SUP: 3
1 5 #SUP: 4
1 4 5 #SUP: 3
1 2 4 5 #SUP: 2
1 2 5 #SUP: 3
1 #SUP: 5
1 4 #SUP: 4
1 2 4 #SUP: 3
1 2 #SUP: 4
4 #SUP: 6
2 4 #SUP: 4
2 #SUP: 6
```

Frequent patterns generated

Analysis

Algorithm	Data set size	Execution time (in millisec)
Apriori	8	95
FP-growth	8	78
EWS	8	16
Apriori	88162	1050
FP-growth	88162	920
EWS	88162	795

Conclusion

- Mining FP – important part of Data Mining
- Research still going on for optimization
- To balance workload – our proposed system
- Open website to upload datasets, find frequent patterns - future extension

References

- Kawuu W. Lin, Yu-Chin Lo(2013),”Efficient algorithms for frequent pattern mining in many-task computing environments”, Elsevier Knowledge-Based Systems, vol. 49,pp. 10-21.
- E.H.S. Han, G. Karypis, V. Kumar(2000),”Scalable parallel data mining for association rules”, IEEE Transaction on Knowledge and Data Engineering, vol. 12, pp. 352-377.
- A. Javed, A. Khokhar(2004),”Frequent pattern mining on message passing multiprocessor systems”, Distributed and Parallel Database, vol. 16,pp. 321-334.
- K.W. Lin, D.J. Deng(2010), ”A novel parallel algorithm for frequent pattern mining with privacy preserved in cloud computing environments”, International Journal of Ad Hoc and Ubiquitous Computing, vol. 6,pp. 205-215.
- R. Agrawal, R. Srikant, Fast algorithms for mining association rules, in: Proceedings of the 20th International Conference on Very Large Data Bases, 1994, pp. 487–499.