S4_A1_FPGrowth

FP-Growth

Given

Transactional Database

TransactionID	Items included
T1	ABDE
T2	BCE
T3	ABDE
T4	ABCE
T5	ABCDE
T6	BCD

minSupport = 3.

Task a: Find all frequent itemsets using FP-Growth

Overview FP-Growth

For a given transactional database

- Derive set of frequent 1-itemsets L
- Rearrange transactions in L-order
- Create FP-Tree from L-ordered transactions
- Create header table (items in decreasing frequency, node links)
- Mine FP-Tree
    - bottom to top through header table, for each (frequent) item:
        - create conditional pattern bases
        - create conditional FP-Tree
        - create frequent pattern combinations

1.1 Scan DB, derive frequent 1-itemsets

Item (1-itemset)	frequency	frequent?
A	4	yes
B	6	y
C	4	y
D	4	y
E	5	y

L = {B, E, A, C, D}

L-order refers to frequency beginning with the most frequent and decreasing. Additionally using lexicographical order on item(set)s of same frequency here.

1.2 Rearrange transactions following L-order

TransactionID	Items included	L-ordered
T1	ABDE	B E A D
T2	BCE	B E C
T3	ABDE	B E A D
T4	ABCE	B E A C
T5	ABCDE	B E A C D
T6	BCD	B C D

2 Create FP-Tree and Header Table

Header Table (L-ordered); see pictures of all (Conditional) FP-Trees below

Item (1-itemset)	frequency	node link
B	6	green
E	5	pink
A	4	blue
C	4	gray
D	4	yellow

3 Mine the FP-Tree

Item	Cond. Pattern Base	Conditional FP-Tree	Frequent Pattern Comb.
D	{{B, E, A : 2}, {B, E, A, C : 1}, {B, C : 1}}	< B : 4, E : 3, A : 3 >	{B, D : 4} {E, D : 3} {A, D : 3}, {B, E, D : 3}, {B, A, D : 3} {E, A, D : 3} {B, E, A, D : 3}
C	{{B, E, A : 2}, {B, E : 1} {B : 1} }	< B : 4, E : 3 >	{B, C : 4} {E, C : 3} {B, E, C : 3}
A	{B, E : 4}	< B : 4, E : 4 >	{B, A : 4} {E, A : 4} {B, E, A : 4}
E	{B : 5}	< B : 5 >	{B, E: 5}

Frequent Itemsets:

{B} {E} {A} {C} {D}

{B, E} {B, A} {B, C} {B, D} {E, A} {E, D} {E, C} {A, D}

{B, E, A} {B, E, C} {B, E, D} {E, A, D}

{B, E, A, D}

Task b

Closed frequent itemsets

- frequent itemset
- there is no proper superset, that has the same support count

{B} {B, E} {B, C} {B, D} {B, E, A} {B, E, C} {B, E, A, D}

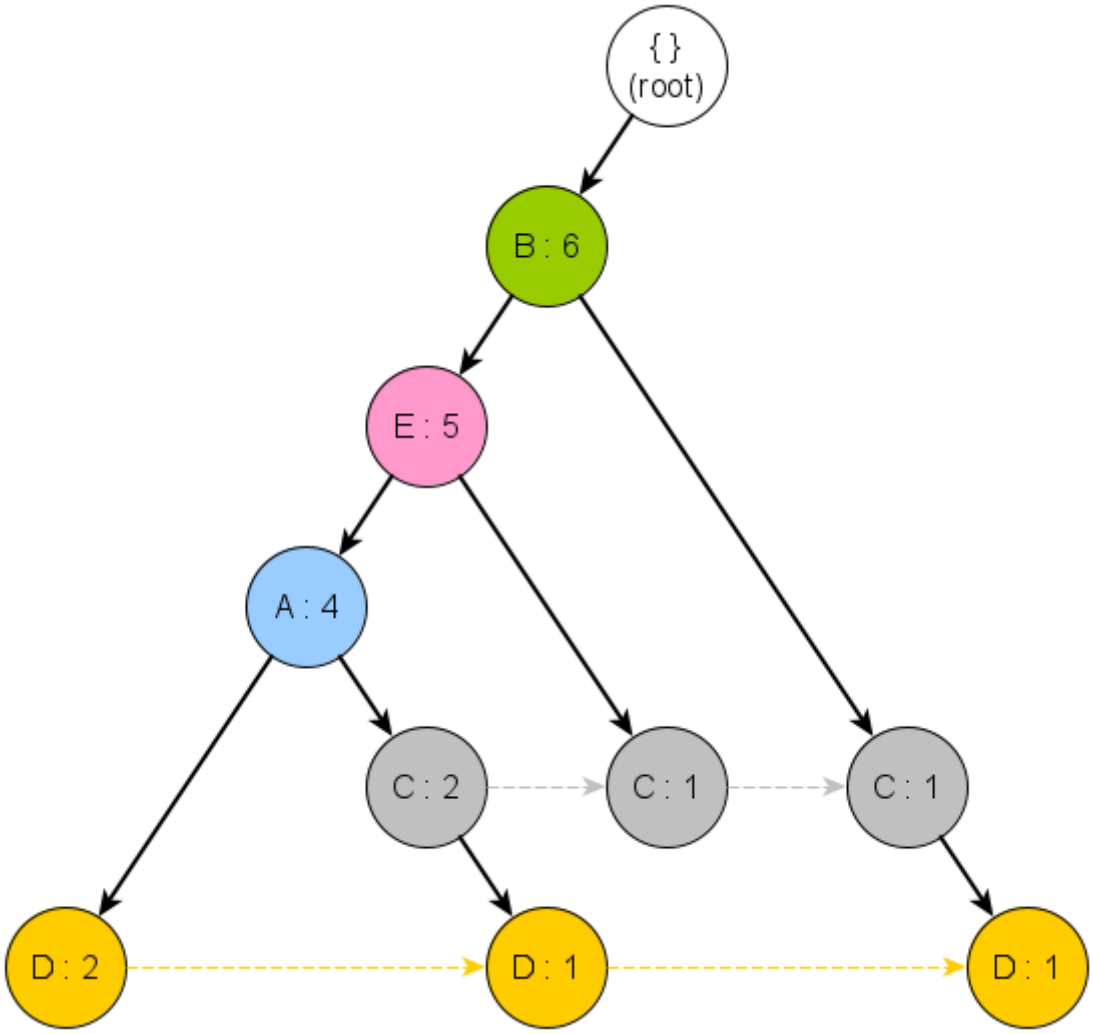
Maximal frequent itemsets

- frequent itemset
- there is no proper superset, that is frequent

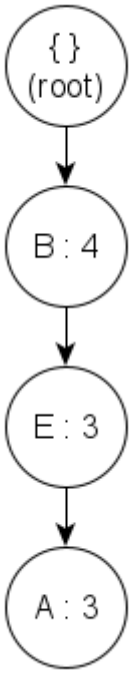
{B, E, C} {B, E, A, D}

Images

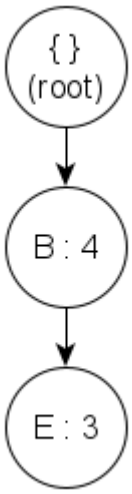
FP Tree



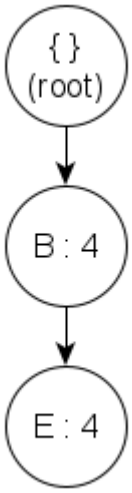
Conditional FP Tree (CFPT) for D



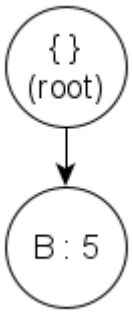
CFPT(C)



CFPT(A)



CFPT(E)



Additional Exercise

Found in Han, Kamber, Pei: Data Mining, 3rd edition, p. 257. (Example 6.5, referring to database on page 250 (table 6.1))

TID	Items
T1	i1, i2, i5
T2	i2, i4
T3	i2, i3
T4	i1, i2, i4
T5	i1, i3
T6	i2, i3
T7	i1, i3
T8	i1, i2, i3, i5
T9	i1, i2, i3

minSupport = 2

You could also try for minSupport = 3 and see the differences.