



CHƯƠNG 3

Khai phá tập phổ biến



Thuật toán Eclat

Các bước của thuật toán Eclat

Bước 1: Biểu diễn dữ liệu dưới dạng danh sách giao tác cho từng mục

Bước 2: Duyệt qua từng tập mục và xây dựng danh sách giao tác chứa tập mục đó.

Bước 3: Áp dụng phép giao giữa các danh sách giao tác để tìm số lần xuất hiện (support) của các tập mục lớn hơn.

Bước 4: Lọc các tập phổ biến theo ngưỡng tối thiểu (min_support).

Bước 5: Tiếp tục đệ quy với các tập phổ biến đã tìm được để mở rộng tập hợp con.



Thuật toán Eclat

Ví dụ:

Giả sử có tập giao tác (transaction database) như sau:

ID	Items
1	A, B, C
2	A, C
3	A, D
4	B, C
5	B, D



Thuật toán Eclat

Chuyển sang định dạng dọc:

Item	Transaction IDs
A	{1, 2, 3}
B	{1, 4, 5}
C	{1, 2, 4}
D	{3, 5}

Giả sử ngưỡng tối thiểu (**min_support**) là 2, chúng ta chỉ giữ lại các mục có **support** ≥ 2 :

- A (3), B (3), C (3), D (2) (tất cả đều thỏa mãn)



Thuật toán Eclat

Tìm tập phổ biến cấp 2

Itemset	Transaction IDs
AB	{1}
AC	{1, 2}
AD	{3}
BC	{1, 4}
BD	{5}
CD	\emptyset

Kết quả 2-itemsets phổ biến:

AC = {1, 2}, support = 2

BC = {1, 4}, support = 2



Thuật toán Eclat

Tìm tập phổ biến cấp 3

Tiếp tục giao các tập giao tác của 2-itemsets phổ biến:

Itemset	Transaction IDs
ABC	{1}

Không còn tập nào có **support** ≥ 2 , nên dừng lại.



Thuật toán Eclat

Kết quả:

Tập phổ biến 1 phần tử: A, B, C, D

Tập phổ biến 2 phần tử: AC, BC



Thuật toán Eclat

Nhận xét:

Thuật toán **Eclat** sử dụng giao tập giao tác để tính support thay vì quét dữ liệu nhiều lần như **Apriori**.

Chỉ cần lưu danh sách giao tác, giúp xử lý nhanh hơn khi dữ liệu nhỏ nhưng tốn bộ nhớ nếu tập giao tác quá lớn.



Thuật toán Eclat



Thuật toán FP-growth

Thuật toán FP-growth dựa trên nguyên tắc cơ bản sau:

- Nén tập dữ liệu vào cấu trúc cây (FP-Tree) nhờ đó giảm chi phí trong quá trình khai phá.
- Các hạng mục không phổ biến được loại bỏ sớm nhưng kết quả khai phá không ảnh hưởng.



Thuật toán FP-growth

- Quá trình khai phá dữ liệu được chia thành các công đoạn nhỏ hơn:
 1. Xây dựng cây FP-Tree.
 2. Khai phá các tập phổ biến dựa trên cây FP-Tree đã tạo.



Thuật toán FP-growth

Cây FP (Frequent Pattern tree).

- Là cấu trúc dữ liệu dạng cây được tổ chức như sau:
 - Nút gốc (root) được gán nhãn “null”.
 - Mỗi nút còn lại chứa các thông tin: item-name, count, node-link.
 - Bảng Header có số dòng bằng số hạng mục (item). Mỗi dòng chứa 3 thuộc tính: item-name, item-count, node-link.



Xây dựng cây FP - Tree

- **Input:**

Cơ sở dữ liệu giao dịch D.

Ngưỡng min-sup.

- **Output:**

Cây FP.




Xây dựng cây FP - Tree

Ví dụ: Xây dựng FP – tree cho CSDL giao dịch sau với $\text{minsupp} = 3$.

Mã giao dịch (TID)	Nội dung giao dịch
1	A, C, T, W
2	C, D, W
3	A, C, T, W
4	A, C, D, W
5	A, C, D, T, W
6	C, D, T

Xây dựng cây FP - Tree

Mã giao dịch (TID)	Nội dung giao dịch
1	A, C, T, W
2	C, D, W
3	A, C, T, W
4	A, C, D, W
5	A, C, D, T, W
6	C, D, T



Item	A	C	D	T	W
supp	4	6	4	4	5

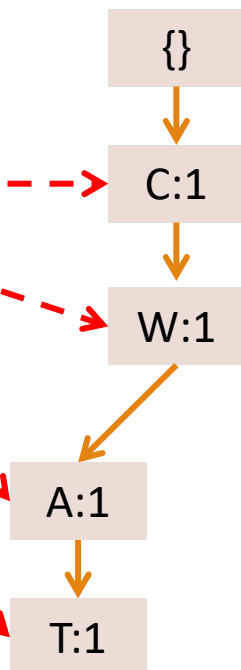
Sort 

Item	C	W	A	D	T
supp	6	5	4	4	4

Xây dựng cây FP - Tree

Bảng Header

item	Supp	link
C	6	- - - - ->
W	5	- - - - ->
A	4	- - - - ->
D	4	- - - - ->
T	4	- - - - ->

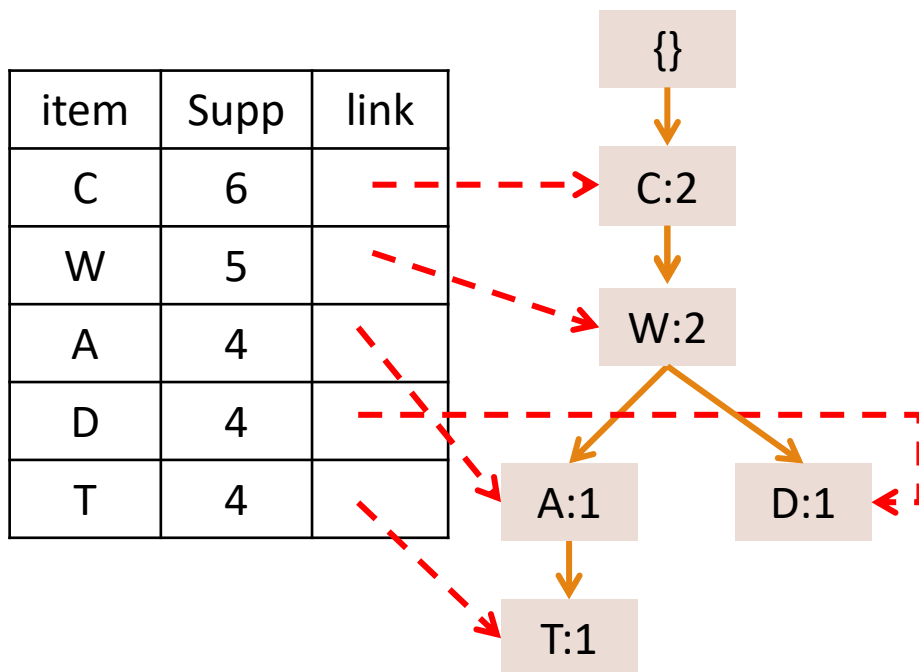


FP-Tree với giao dịch 1

Mã giao dịch (TID)	Nội dung giao dịch
1	A, C, T, W
2	C, D, W
3	A, C, T, W
4	A, C, D, W
5	A, C, D, T, W
6	C, D, T

C, W, A, T

Xây dựng cây FP - Tree

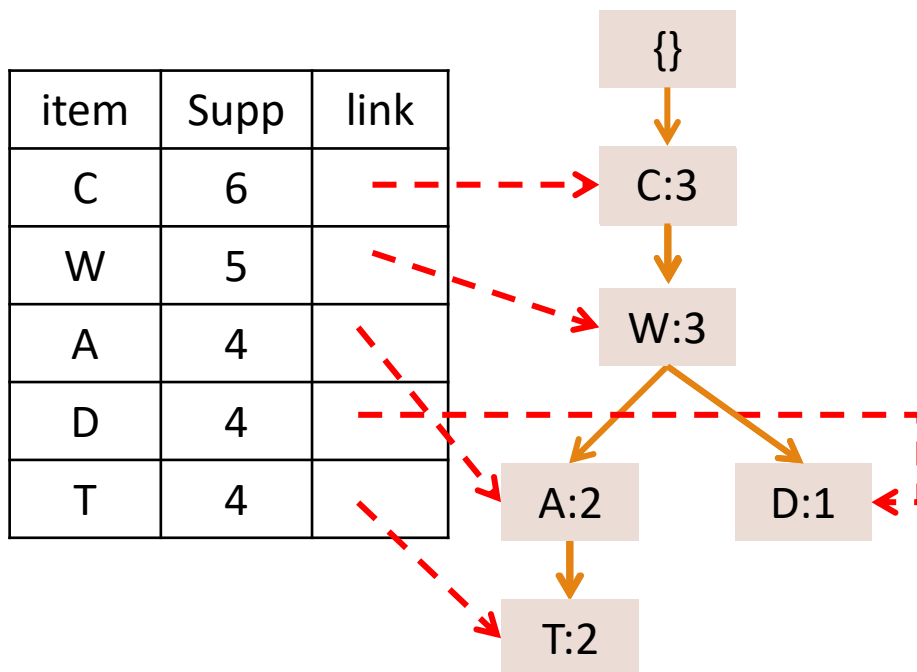


FP-tree với giao dịch 1 và 2

Mã giao dịch (TID)	Nội dung giao dịch
1	A, C, T, W
2	C, D, W
3	A, C, T, W
4	A, C, D, W
5	A, C, D, T, W
6	C, D, T

C, W, D

Xây dựng cây FP - Tree

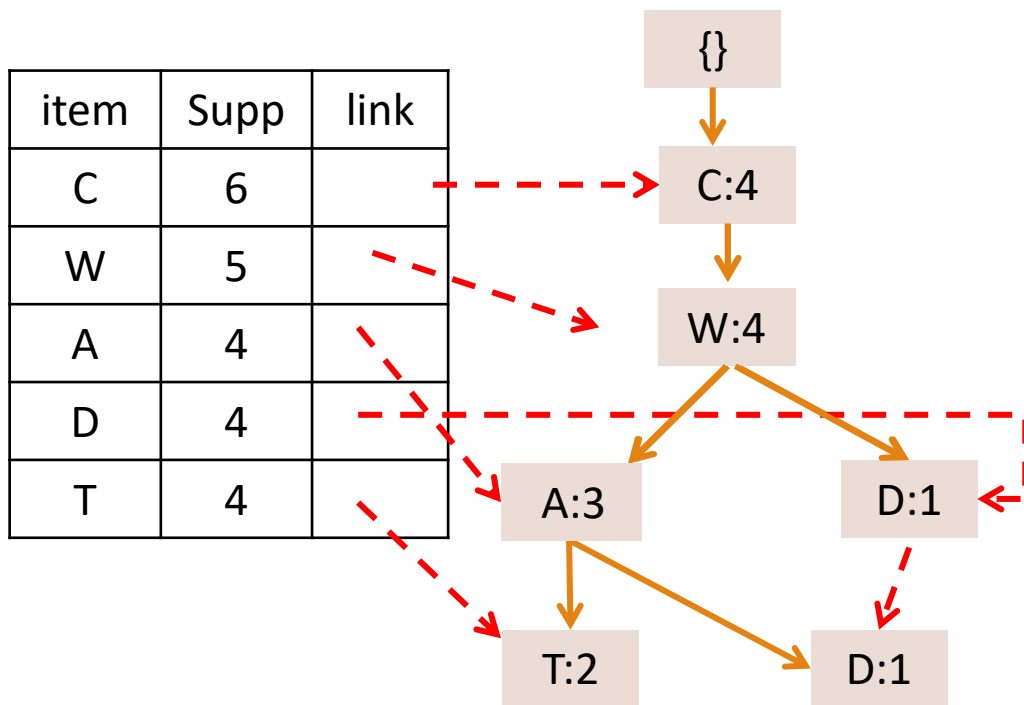


FP-tree với 3 giao dịch đầu

Mã giao dịch (TID)	Nội dung giao dịch
1	A, C, T, W
2	C, D, W
3	A, C, T, W
4	A, C, D, W
5	A, C, D, T, W
6	C, D, T

C, W, A, T

Xây dựng cây FP - Tree

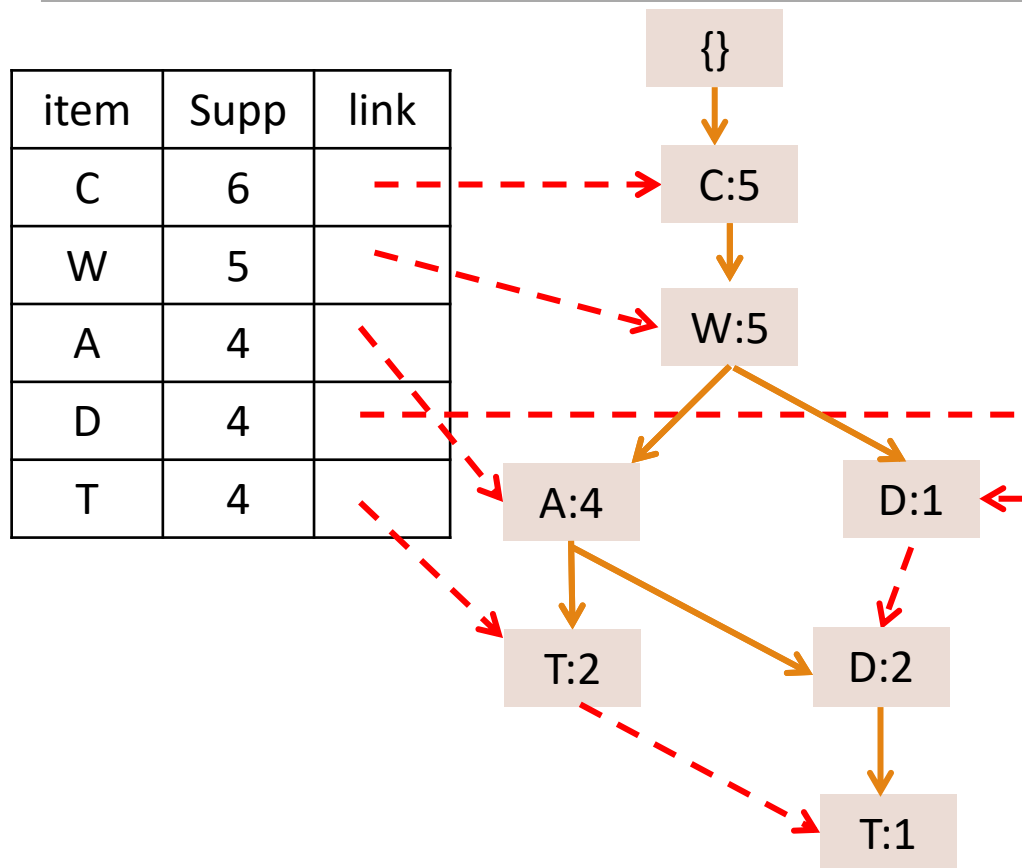


FP-tree với 4 giao dịch

Mã giao dịch (TID)	Nội dung giao dịch
1	A, C, T, W
2	C, D, W
3	A, C, T, W
4	A, C, D, W
5	A, C, D, T, W
6	C, D, T

C, W, A, D

Xây dựng cây FP - Tree



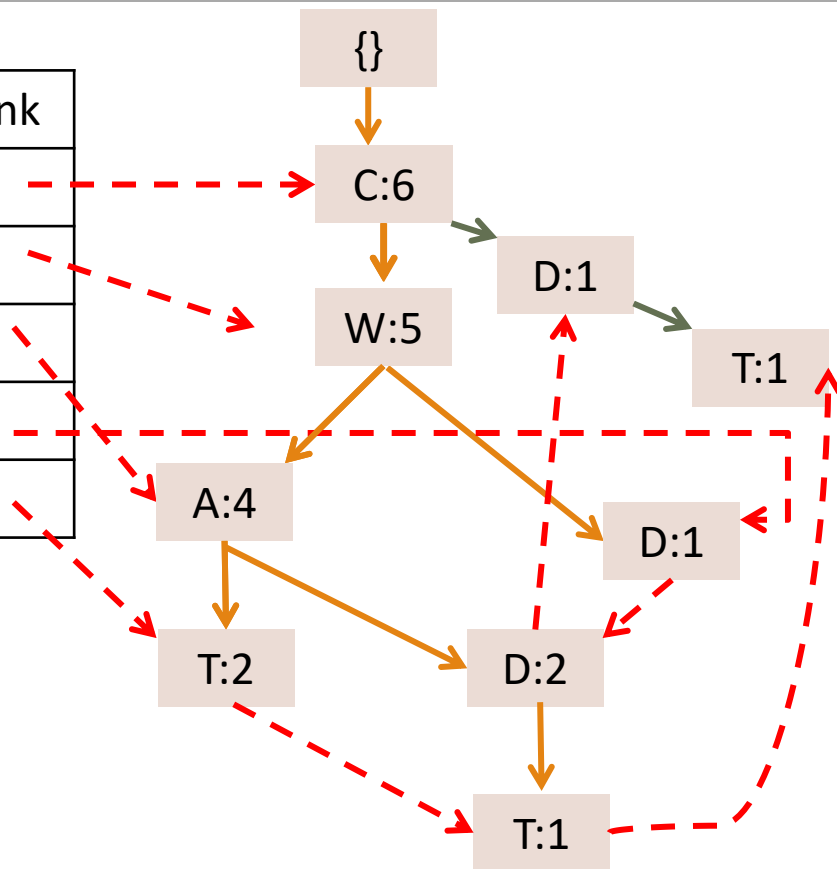
Mã giao dịch (TID)	Nội dung giao dịch
1	A, C, T, W
2	C, D, W
3	A, C, T, W
4	A, C, D, W
5	A, C, D, T, W
6	C, D, T

C, W, A, D, T

FP-tree với 5 giao dịch

Xây dựng cây FP - Tree

item	Supp	link
C	6	
W	5	
A	4	
D	4	
T	4	



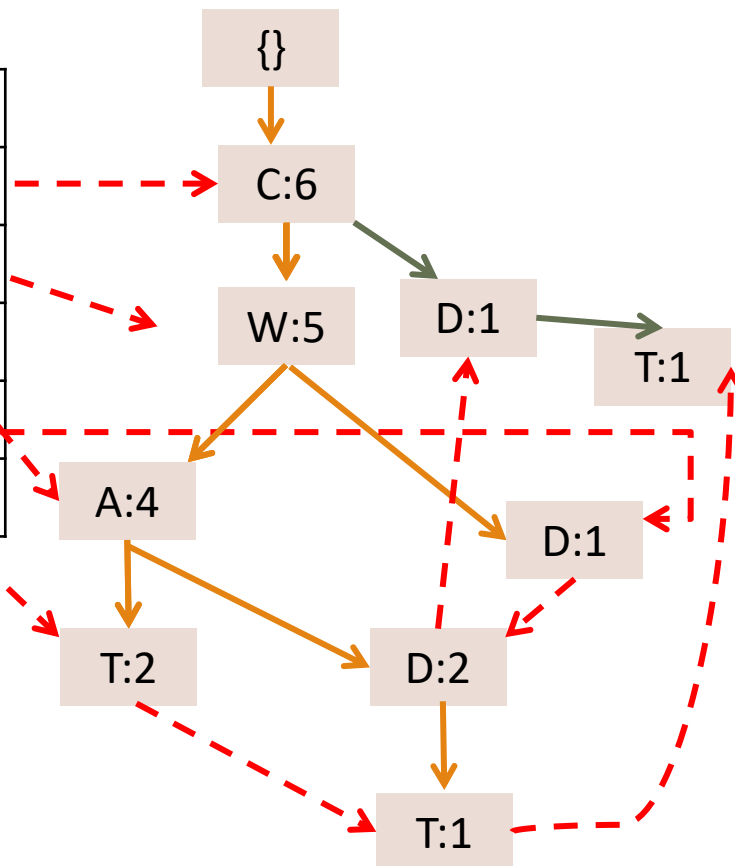
Mã giao dịch (TID)	Nội dung giao dịch
1	A, C, T, W
2	C, D, W
3	A, C, T, W
4	A, C, D, W
5	A, C, D, T, W
6	C, D, T

C, D, T

FP-tree với 6 giao dịch

Xây dựng cây FP - Tree

item	Supp	link
C	6	---
W	5	---
A	4	---
D	4	---
T	4	---



Mã giao dịch (TID)	Nội dung giao dịch
1	A, C, T, W
2	C, D, W
3	A, C, T, W
4	A, C, D, W
5	A, C, D, T, W
6	C, D, T

Hoàn thành FP – tree → tìm FIs dựa trên FP - tree



Xây dựng cây FP - Tree

Input: Cơ sở dữ liệu giao dịch T; Ngưỡng min-sup.

Output: Cây FP.

Procedure FP_TreeConstruction

{

Duyệt T được Fis và supp. Sắp xếp giảm dần theo supp ta được danh sách L.

Tạo nút gốc R và gán nhãn “null”.

Tạo bảng Header có |F| dòng và đặt tất cả các node-link chỉ đến null. //các items thỏa minsup

for each giao dịch $t \in T$

{

// Duyệt T lần 2

Chọn các item phổ biến của T đưa vào P; //Xét trên mỗi giao dịch chọn item thỏa minsup

Sắp các item trong P theo trật tự L;

Call Insert_Tree(P, R);

}

}



Xây dựng cây FP - Tree

Procedure Insert_Tree(P, R)

```
{      Đặt  $P=[p | P - p]$  , với p là phần tử đầu và  $P - p$  là phần còn lại của danh sách;  
      if R có một con N sao cho N.item-name = p then  
          N.count ++;  
  
      else  
      {      Tạo nút mới N;  
              N.count = 1;  
              N.item-name = p;  
              N. parent = R; //giữ node cha để tìm cơ sở mẫu điều kiện (tìm nhánh chứa node đó)  
              // Tạo node-link chỉ đến item, H là bảng Header  
              N.node-link = H[p].head; //Chèn vào đầu danh sách liên kết  
              H[p].head = N;  
      }  
      if ( $P - p$ ) != null then Call Insert_Tree( $P - p$ , N) ;  
}
```




Xây dựng cây FP - Tree

Để khai thác các mẫu phổ biến từ cây FP-Tree, ta sử dụng thủ tục FP-Growth:

Input: Cây FP-Tree của cơ sở giao dịch T, ngưỡng \min_sup , $\alpha = null$.

Output: Một tập đầy đủ các mẫu phổ biến F.

Procedure FP-Growth(Tree, α)

```
{
    F =  $\varnothing$ ;
    if Tree chỉ chứa một đường dẫn đơn P then
    {
        for each tổ hợp  $\beta$  của các nút trong P do
        {
            Phát sinh mẫu  $p = \beta \cup \alpha$ ;
             $supp(p) = \min\_sup$  các nút trong  $\beta$ ;
        }
    }
    else
    for each  $a_i$  in the header of Tree
    {
        Phát sinh mẫu  $\beta = a_i \cup \alpha$ ;
         $supp = a_i.supp$ ;
        Xây dựng cơ sở mẫu điều kiện của  $\beta$ ;
        Xây dựng FP-Tree điều kiện  $Tree_\beta$  của  $\beta$ ; //cây cục bộ
        if ( $Tree_\beta \neq \varnothing$ ) then Call FP_Growth( $Tree_\beta$ ,  $\beta$ );
    }
}
```



Khai thác cây FP - Tree

Cây FP được khai thác như sau:

- Xây dựng mẫu điều kiện (Conditional pattern base).
- Xây dựng cây FP điều kiện (cục bộ).



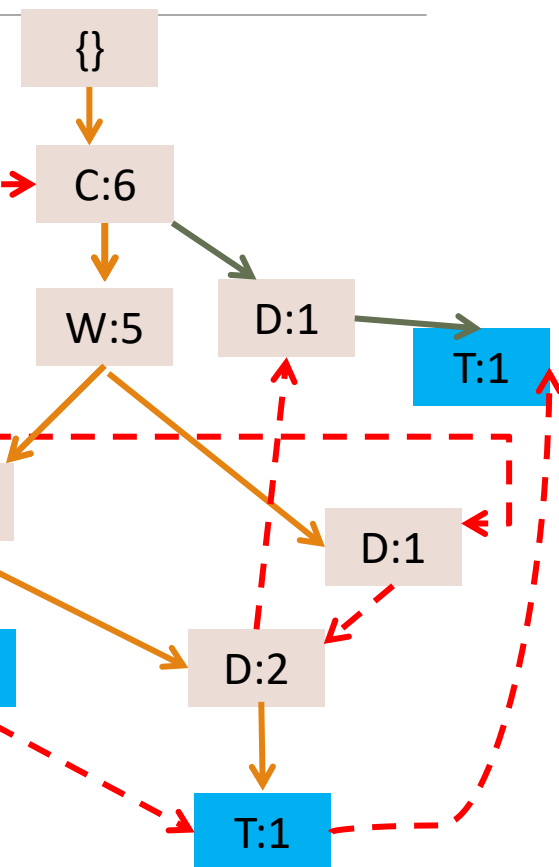
Khai thác cây FP - Tree

Xây dựng mẫu điều kiện

- Duyệt hạn mục p từ cuối bảng Header của cây FP.
- Duyệt cây FP tìm những nhánh có chứa hạng mục p .
- Gom tất cả đường dẫn đến p để tạo cơ sở mẫu điều kiện cho p .

Khai thác cây FP - Tree

item	Supp	link
C	6	
W	5	
A	4	
D	4	
T	4	

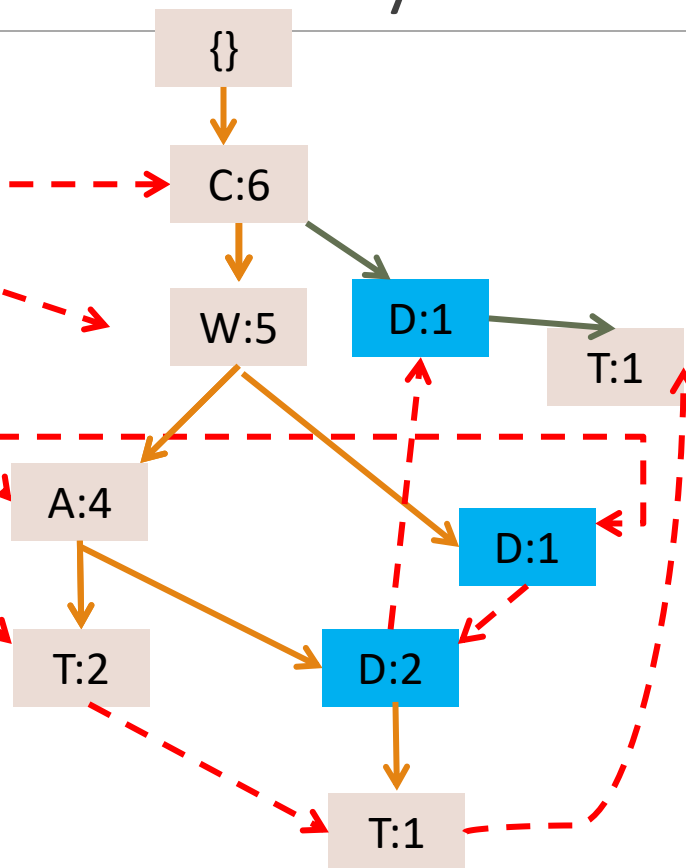


- Chiếu lên các nhánh chứa T trên cây FP:
 → cơ sở mẫu điều kiện của T là {CWA:2, CWAD:1, CD:1}



Khai thác cây FP - Tree

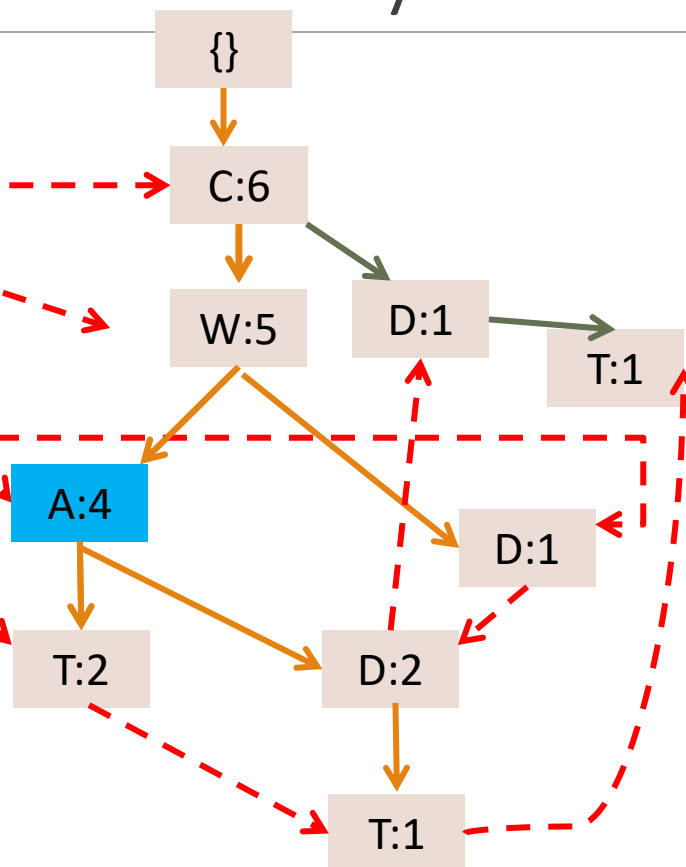
item	Supp	link
C	6	
W	5	
A	4	
D	4	
T	4	



- Chiều lên các nhánh chứa D trên cây FP → cơ sở mẫu điều kiện của D là {CWA:2, CW:1, C:1}

Khai thác cây FP - Tree

item	Supp	link
C	6	
W	5	
A	4	
D	4	
T	4	

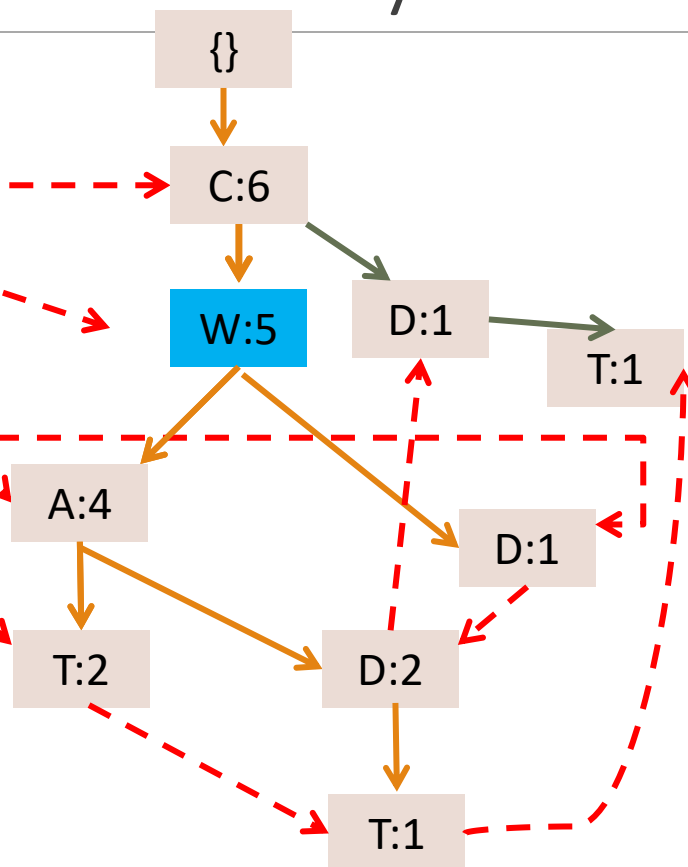


- Chiếu lên các nhánh chứa A trên cây FP \rightarrow cơ sở mẫu điều kiện của A là {CW:4}



Khai thác cây FP - Tree

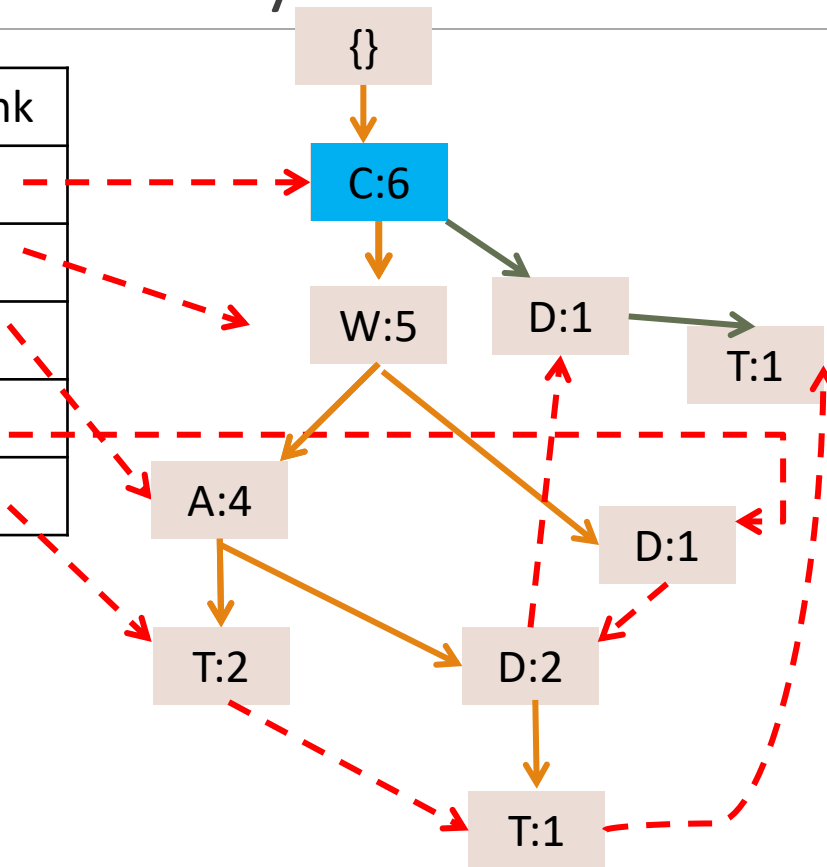
item	Supp	link
C	6	
W	5	
A	4	
D	4	
T	4	



- Chiếu lên các nhánh chứa W trên cây FP → cơ sở mẫu điều kiện của W là {C:5}

Khai thác cây FP - Tree

item	Supp	link
C	6	
W	5	
A	4	
D	4	
T	4	



- Chiếu lên các nhánh chứa C trên cây FP \rightarrow cơ sở mẫu điều kiện của C là $\{\emptyset\}$



Khai thác cây FP - Tree

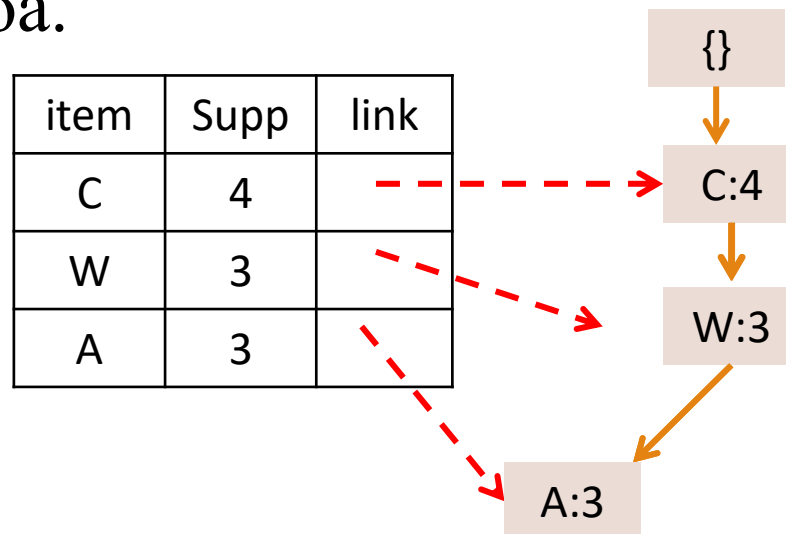
- Bảng cơ sở mẫu điều kiện cho mọi hạng mục

item	Cond Battern base
W	C:5
A	CW:4
D	CWA:2, CW:1,C:1
T	CWA:2, CWAD:1,CD:1



Khai thác cây FP - Tree

- Với mẫu điều kiện cho T là **CWA:2**, **CWAD:1**, **CD:1**
- Đếm số mẫu trong cơ sở mẫu: C:4, W:3, A:3 thỏa minsupp; D:2 không thỏa.



- Xây dựng cây FP điều kiện cho T: Chỉ có một đường đi đơn (C:4, W:3, A:3).



Khai thác cây FP - Tree

- Tập phổ biến với điều kiện cho $T: \{CT, AT, WT, CAT, AWT, CWT, ACWT\}$.
- Tương tự xây dựng tất cả cây FP điều kiện cho các hạng mục còn lại.



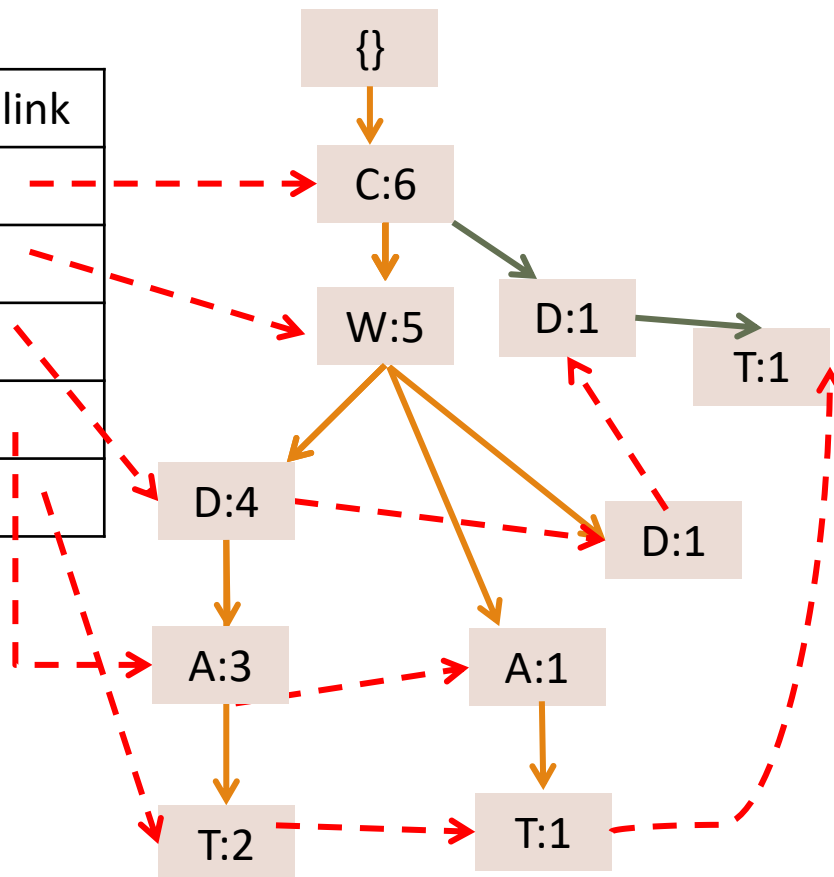
Khai thác cây FP - Tree

Xây dựng tập phổ biến

- Cây FP điều kiện chỉ có **đường đi đơn** → tập FIs xác định bằng cách liệt kê mọi **tổ hợp các nút** trên cây hợp với hạng mục đang xét.
- Cây FP điều kiện có **nhiều hơn một đường đi đơn** gọi **đệ quy thuật toán FP – Tree** từ bước xây dựng cơ sở mẫu cho cây này.

Cây FP điều kiện có nhiều hơn một đường đi đơn

item	Supp	link
C	6	
W	5	
D	5	
A	4	
T	4	



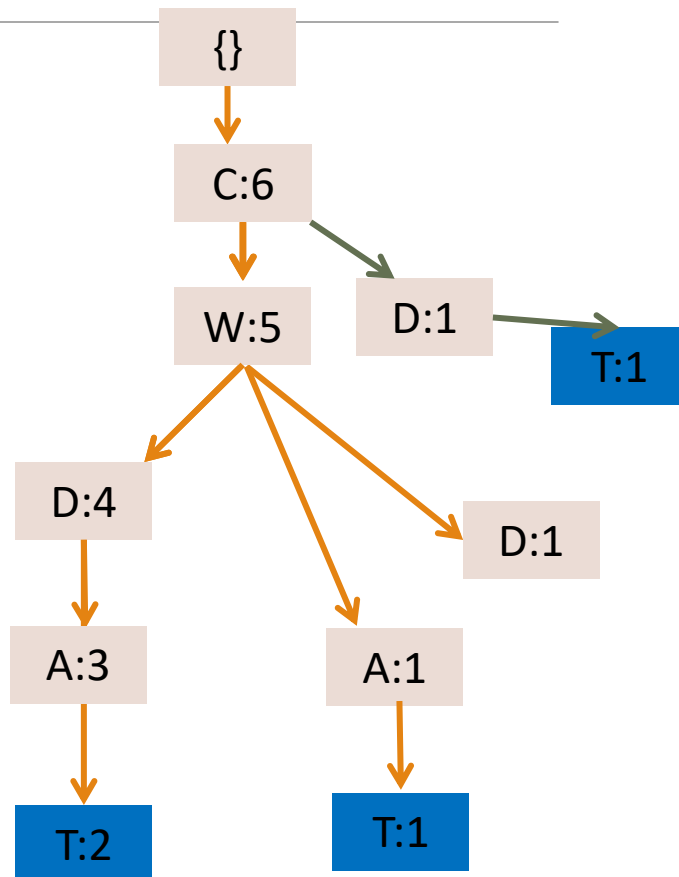
Mã giao dịch (TID)	Nội dung giao dịch
1	A, C, D , T, W
2	C, D, W
3	A, C, T, W
4	A, C, D, W
5	A, C, D, T, W
6	C, D, T

Hoàn thành FP – tree → tìm FIs dựa trên FP - tree



Khai thác cây FP - Tree

item	Supp	link
C	6	
W	5	
D	5	
A	4	
T	4	



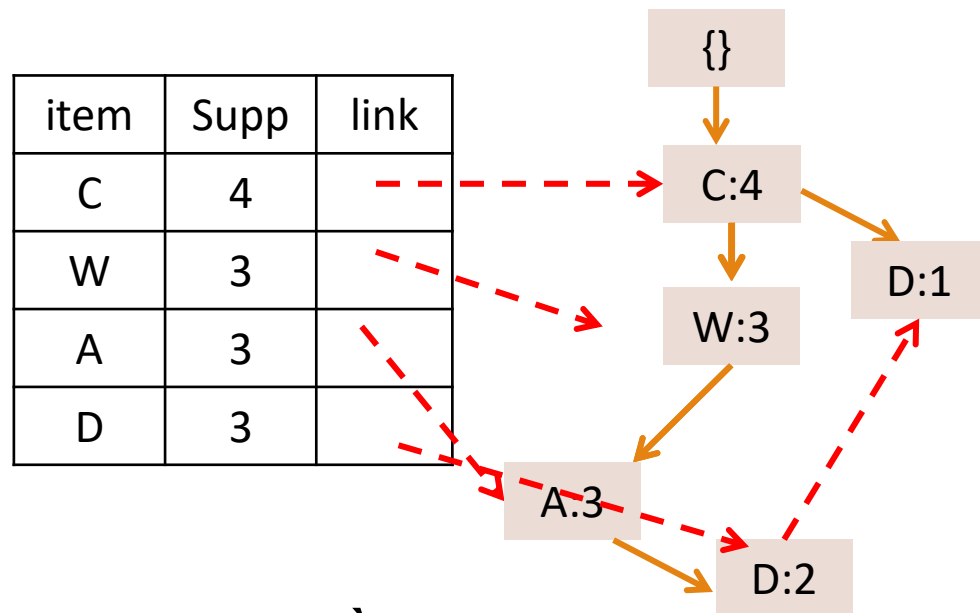
Chiếu lên các nhánh chứa T trên cây FP:

→ cơ sở mẫu điều kiện của T là {**CWAD**:2, **CWA**:1, **CD**:1}



Khai thác cây FP - Tree

- Với mẫu điều kiện cho T là **CWAD:2**, **CWA:1**, **CD:1**
- Đếm số mẫu trong cơ sở mẫu: C:4, W:3, A:3, D:3 thỏa minsupp



- Xây dựng cây FP điều kiện cho T: có nhiều hơn 1 đường đi nên lặp lại việc “xây dựng lại điều kiện mẫu cho D, A, W, C”



Bài tập chương 3

Bài 1. Cho cơ sở dữ liệu giao dịch như sau:

Sử dụng các giá trị ngưỡng minsupport = 30% (minsup = 2,4)

- a) Hãy liệt kê tất cả các tập phổ biến 1 phần tử
- b) Tìm độ phổ biến và độ tin cậy của luật $A \rightarrow B$
- c) Chạy từng bước thuật toán Apriori tìm tất cả tập phổ biến.
- d) Tìm tất cả các tập MFIs, FCIs trong cơ sở dữ liệu
- e) Xây dựng cây FP-Tree cho CSDL giao dịch trên

TID	Items
T01	A, B, C, D
T02	A, C, D, F
T03	C, D, E, G, A
T04	A, D, F, B
T05	B, C, G
T06	D, F, G
T07	A, B, G
T08	C, D, F, G



Bài tập chương 3

Bài 2. Cho cơ sở dữ liệu giao dịch như sau:

Sử dụng các ngưỡng support = 30%

TID	Items
T01	A1, B1, C2
T02	A2, C1, D1
T03	B2, C2, E2
T04	B1, C1, E1
T05	A3, C3, E2
T06	C1, D2, E2

- Hãy liệt kê tất cả các tập phổ biến 1 phần tử
- Chạy từng bước thuật toán Apriori tìm tất cả tập phổ biến.
- Tìm các tập phổ biến đóng (FCIs)
- Xây dựng cây FP-Tree cho CSDL giao dịch trên
- Chạy từng bước thuật toán Eclat tìm tất cả tập phổ biến.



Bài tập chương 3

Bài 3. Cho dữ liệu giao dịch sau và $\text{minsup} = 60\%$

TID	Items
01	a, c, d, f, g, i, m, p
02	a, b, c, f, l, m, o
03	b, f, h, j, o, w
04	b, c, k, s, p
05	a, c, e, f, l, m, n, p

Hãy liệt kê tất cả các tập phổ biến 1 phần tử

Tính độ phổ biến và độ tin cậy của luật $c \rightarrow a$.

Xây dựng cây FP-Tree cho CSDL giao dịch trên



Bài tập chương 3

Bài 4. Cho CSDL giao dịch sau:

<i>TID</i>	<i>List of item IDs</i>
T100	I1, I2, I5
T200	I2, I4
T300	I2, I3
T400	I1, I2, I4
T500	I1, I3
T600	I2, I3
T700	I1, I3
T800	I1, I2, I3, I5
T900	I1, I2, I3

Với $\text{misnupp} = 2$, chạy từng bước thuật toán FP growth tìm mọi tập FI, MFIs và FCIs.

Với $\text{misnupp} = 2$, Chạy từng bước thuật toán Eclat tìm tất cả tập phổ biến.