



TÀI LIỆU LÝ THUYẾT KTDL & UD

# KHAI THÁC MẪU PHỔ BIẾN và LUẬT KẾT HỢP (P2)

Giảng viên: ThS. Lê Ngọc Thành  
Email: [lnthanh@fit.hcmus.edu.vn](mailto:lnthanh@fit.hcmus.edu.vn)

# Nội dung

---

- Thuật toán FP-Growth
  - Xây dựng cây FP-Growth
  - Phát sinh mẫu phổ biến từ FP-Growth
- So sánh Fp-Growth và Apriori
- Độ đo tính lý thú của LKH

# Nhắc lại hạn chế Apriori

- Các hạn chế của thuật toán Apriori
  - Phải duyệt CSDL nhiều lần
  - Kiểm tra và tạo lượng lớn tập ứng viên
- Ví dụ:
  - Nếu có  $10^4$  mẫu phổ biến 1-hạng mục thì cần phát sinh nhiều hơn  $10^7$  ứng viên 2-hạng mục
  - Để tìm mẫu phổ biến  $i_1 i_2 \dots i_{100}$ :
    - Số lần duyệt CSDL: 100
    - Số lượng ứng viên ít nhất:  $2^{100}-1 = 10^{30}$

# Giới thiệu thuật toán FP-Growth

---

- Được đề xuất bởi J.Han, J.Pei và Y.Yin trong hội nghị SIGMOD năm 2000.
- Là thuật toán tìm kiếm theo chiều sâu (theo tư tưởng chia để trị)
- Khai thác mẫu phổ biến không sử dụng hàm tạo ứng viên.

# Ý tưởng FP-Growth

---



Nén CSDL thành cấu trúc cây  
FP (Frequent Pattern)

↓

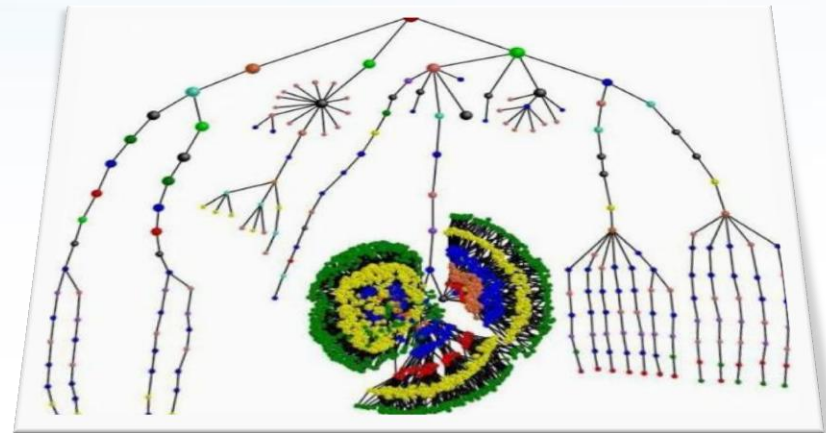
Chia dữ liệu đã nén thành  
các dữ liệu có điều kiện

↓

Thực hiện khai thác trên mỗi  
cơ sở dữ liệu riêng rẽ

# Thuật toán FP-Growth

- Input: dữ liệu giao tác D và minsup
- Output: tập mẫu phổ biến
- B1: Xây dựng cây FP
- B2: Khai thác cây FP



# B1. Xây dựng FP-tree

- **B1.1:** (Duyệt CSDL lần 1) Tìm tập phổ biến 1- hạng mục. Sắp xếp tập phổ biến giảm dần vào trong F-list .
- **B1.2:** Chọn lọc và sắp xếp mỗi giao tác trong CSDL lại theo thứ tự trong F-list.
- **B1.3:** Khởi tạo cây với gốc là “null”.
- **B1.4:** (Duyệt CSDL lần 2) Với mỗi giao tác ở B1.2, xuất phát từ gốc, lấy từng hạng mục theo thứ tự:
  - Nếu hạng mục chưa có ở vị trí con của node hiện tại thì thêm vào cây với đếm là 1.
  - Nếu hạng mục đã có ở vị trí con của node hiện tại thì tăng đếm lên 1.

# Ví dụ xây dựng cây FP (1/3)

| <i>TID</i> | <i>Items bought</i>      | <i>(ordered) frequent items</i> | <i>minsupp = 3</i> |
|------------|--------------------------|---------------------------------|--------------------|
| 100        | {f, a, c, d, g, i, m, p} | {f, c, a, m, p}                 |                    |
| 200        | {a, b, c, f, l, m, o}    | {f, c, a, b, m}                 |                    |
| 300        | {b, f, h, j, o, w}       | {f, b}                          |                    |
| 400        | {b, c, k, s, p}          | {c, b, p}                       |                    |
| 500        | {a, f, c, e, l, p, m, n} | {f, c, a, m, p}                 |                    |

B1.1: Tìm tập phổ biến 1-hạng mục. Sắp xếp tập phổ biến giảm dần vào trong F-list

**F-list=f-c-a-b-m-p**

B1.2: Chọn lọc và sắp xếp CSDL theo F-list.

**Header Table**

| <i>Item</i> | <i>Supp.count</i> |
|-------------|-------------------|
| <i>f</i>    | 4                 |
| <i>c</i>    | 4                 |
| <i>a</i>    | 3                 |
| <i>b</i>    | 3                 |
| <i>m</i>    | 3                 |
| <i>p</i>    | 3                 |



# Ví dụ xây dựng cây FP (2/3)

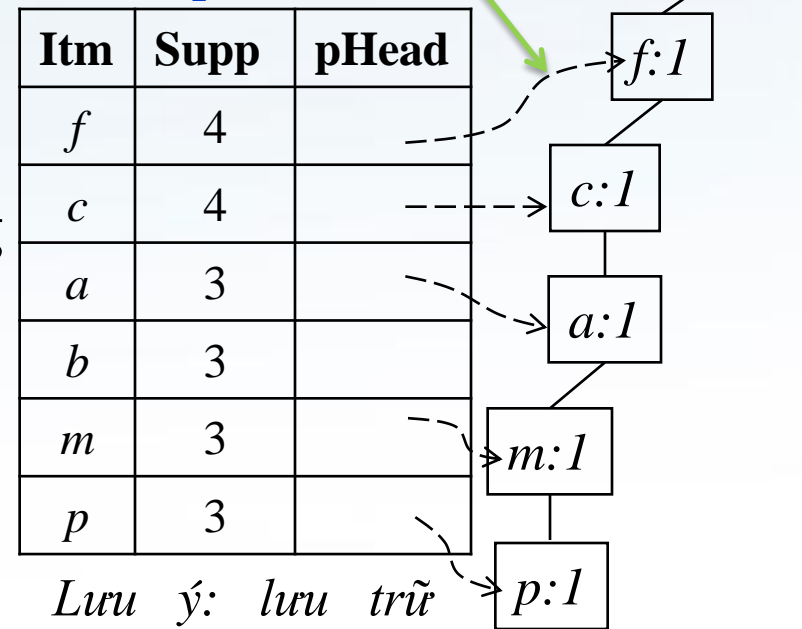
| <i>TID</i> | <i>Items bought</i>      | <i>(ordered) frequent items</i> |
|------------|--------------------------|---------------------------------|
| 100        | {f, a, c, d, g, i, m, p} | <b>{f, c, a, m, p}</b>          |
| 200        | {a, b, c, f, l, m, o}    | {f, c, a, b, m}                 |
| 300        | {b, f, h, j, o, w}       | {f, b}                          |
| 400        | {b, c, k, s, p}          | {c, b, p}                       |
| 500        | {a, f, c, e, l, p, m, n} | {f, c, a, m, p}                 |

*minsupp* = 3

B1.3: Khởi tạo cây với gốc “null”

B1.4: Với mỗi giao tác ở B1.2, xuất phát từ gốc, lấy từng hạng mục theo thứ tự:

- Nếu hạng mục chưa có ở vị trí con của node hiện tại thì thêm vào cây với đếm là 1.
- Nếu hạng mục đã có ở vị trí con của node hiện tại thì tăng đếm lên 1



Lưu ý: lưu trữ con trỏ pHead của DSLK để tiện truy cập về sau

# Ví dụ xây dựng cây FP (2/3)

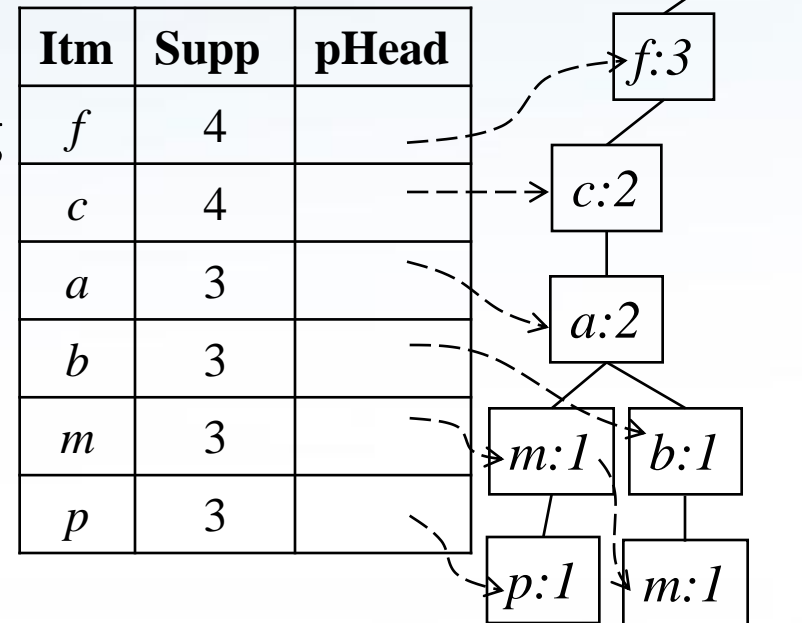
| <u>TID</u> | <u>Items bought</u>      | <u>(ordered) frequent items</u> |
|------------|--------------------------|---------------------------------|
| 100        | {f, a, c, d, g, i, m, p} | {f, c, a, m, p}                 |
| 200        | {a, b, c, f, l, m, o}    | {f, c, a, b, m}                 |
| 300        | {b, f, h, j, o, w}       | {f, b}                          |
| 400        | {b, c, k, s, p}          | {c, b, p}                       |
| 500        | {a, f, c, e, l, p, m, n} | {f, c, a, m, p}                 |

*minsupp* = 3

B1.3: Khởi tạo cây với gốc “null”

B1.4: Với mỗi giao tác ở B1.2, xuất phát từ gốc, lấy từng hạng mục theo thứ tự:

- Nếu hạng mục chưa có ở vị trí con của node hiện tại thì thêm vào cây với đếm là 1.
- Nếu hạng mục đã có ở vị trí con của node hiện tại thì tăng đếm lên 1



# Ví dụ xây dựng cây FP (2/3)

| <u>TID</u> | <u>Items bought</u>      | <u>(ordered) frequent items</u> |
|------------|--------------------------|---------------------------------|
| 100        | {f, a, c, d, g, i, m, p} | {f, c, a, m, p}                 |
| 200        | {a, b, c, f, l, m, o}    | {f, c, a, b, m}                 |
| 300        | {b, f, h, j, o, w}       | {f, b}                          |
| 400        | {b, c, k, s, p}          | {c, b, p}                       |
| 500        | {a, f, c, e, l, p, m, n} | {f, c, a, m, p}                 |

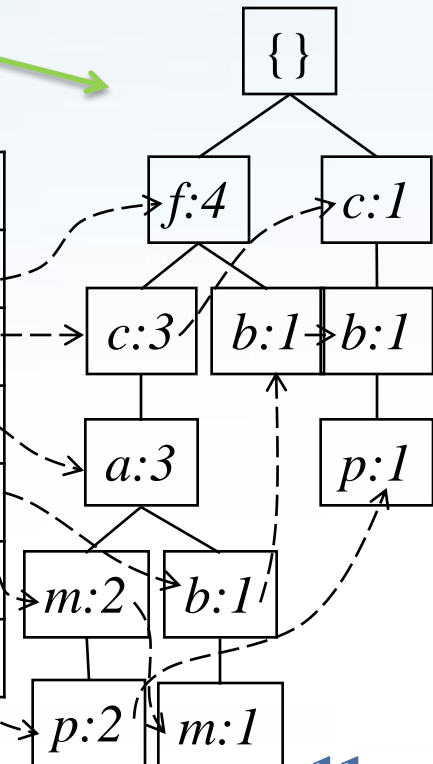
*minsupp* = 3

B1.3: Khởi tạo cây với gốc “null”

B1.4: Với mỗi giao tác ở B1.2, xuất phát từ gốc, lấy từng hạng mục theo thứ tự:

- Nếu hạng mục chưa có ở vị trí con của node hiện tại thì thêm vào cây với đếm là 1.
- Nếu hạng mục đã có ở vị trí con của node hiện tại thì tăng đếm lên 1

| Itm | Supp | pHead |
|-----|------|-------|
| f   | 4    |       |
| c   | 4    |       |
| a   | 3    |       |
| b   | 3    |       |
| m   | 3    |       |
| p   | 3    |       |



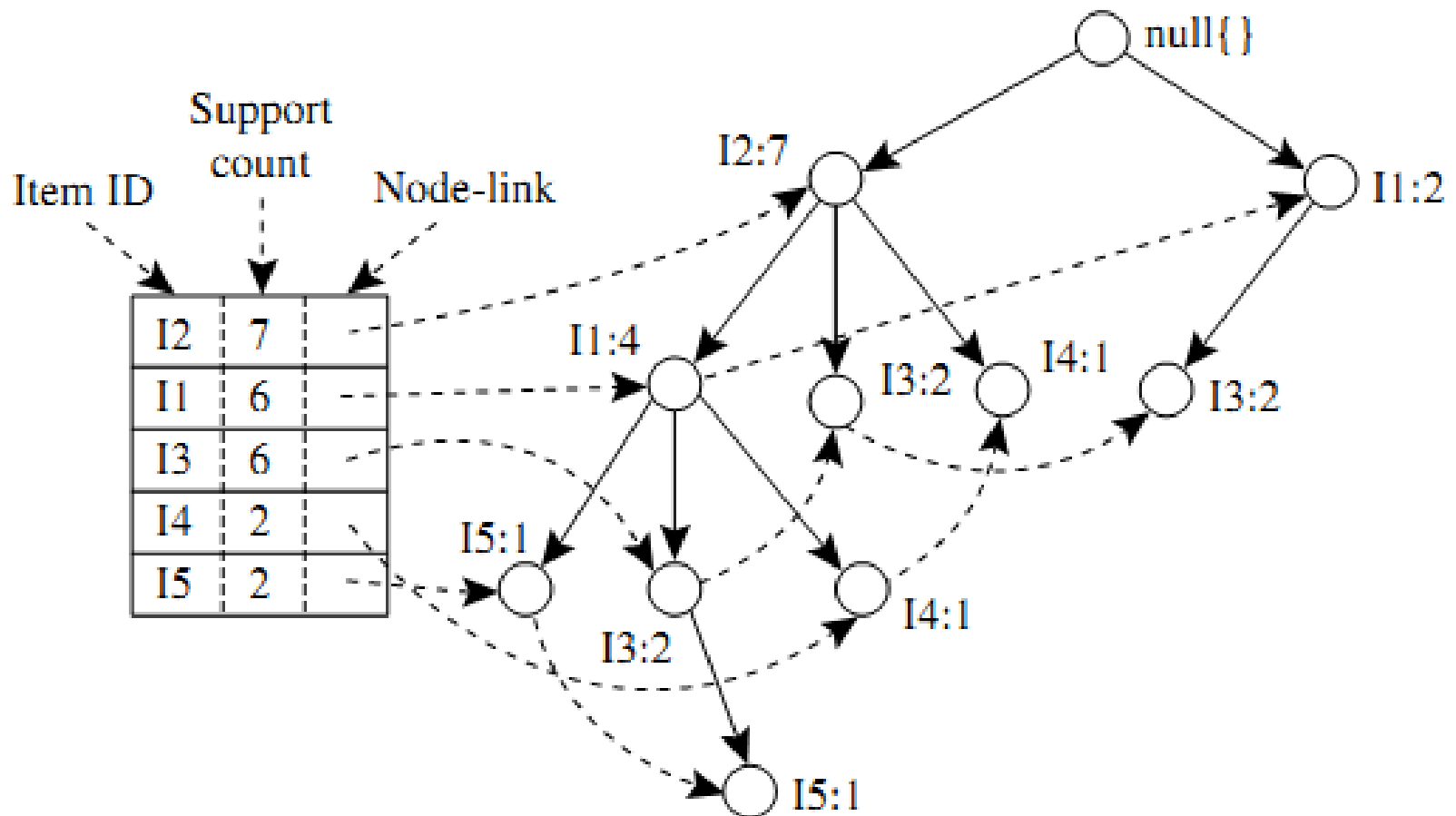
# Bài tập áp dụng 1

**Xây dựng cây FP cho dữ liệu giao tác sau với minsup = 2 hay 22%:**

| <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              |

# Bài tập áp dụng 1 – Đáp án

Cây FP được xây dựng:



## B2. Khai thác FP-tree

- FP-tree được khai thác bằng cách gọi hàm  $\text{FP\_growth}(\text{FP\_tree}, \text{null})$ :

procedure  $\text{FP\_growth}(\text{Tree}, \alpha)$

- (1) if  $\text{Tree}$  contains a single path  $P$  then
- (2)   for each combination (denoted as  $\beta$ ) of the nodes in the path  $P$
- (3)     generate pattern  $\beta \cup \alpha$  with  $\text{support\_count} = \text{minimum support count of nodes in } \beta$ ;
- (4) else for each  $a_i$  in the header of  $\text{Tree}$  {
- (5)   generate pattern  $\beta = a_i \cup \alpha$  with  $\text{support\_count} = a_i.\text{support\_count}$ ;
- (6)   construct  $\beta$ 's conditional pattern base and then  $\beta$ 's conditional FP-tree  $\text{Tree}_\beta$ ;
- (7)   if  $\text{Tree}_\beta \neq \emptyset$  then
- (8)     call  $\text{FP\_growth}(\text{Tree}_\beta, \beta)$ ; }

## B2. Khai thác FP-tree

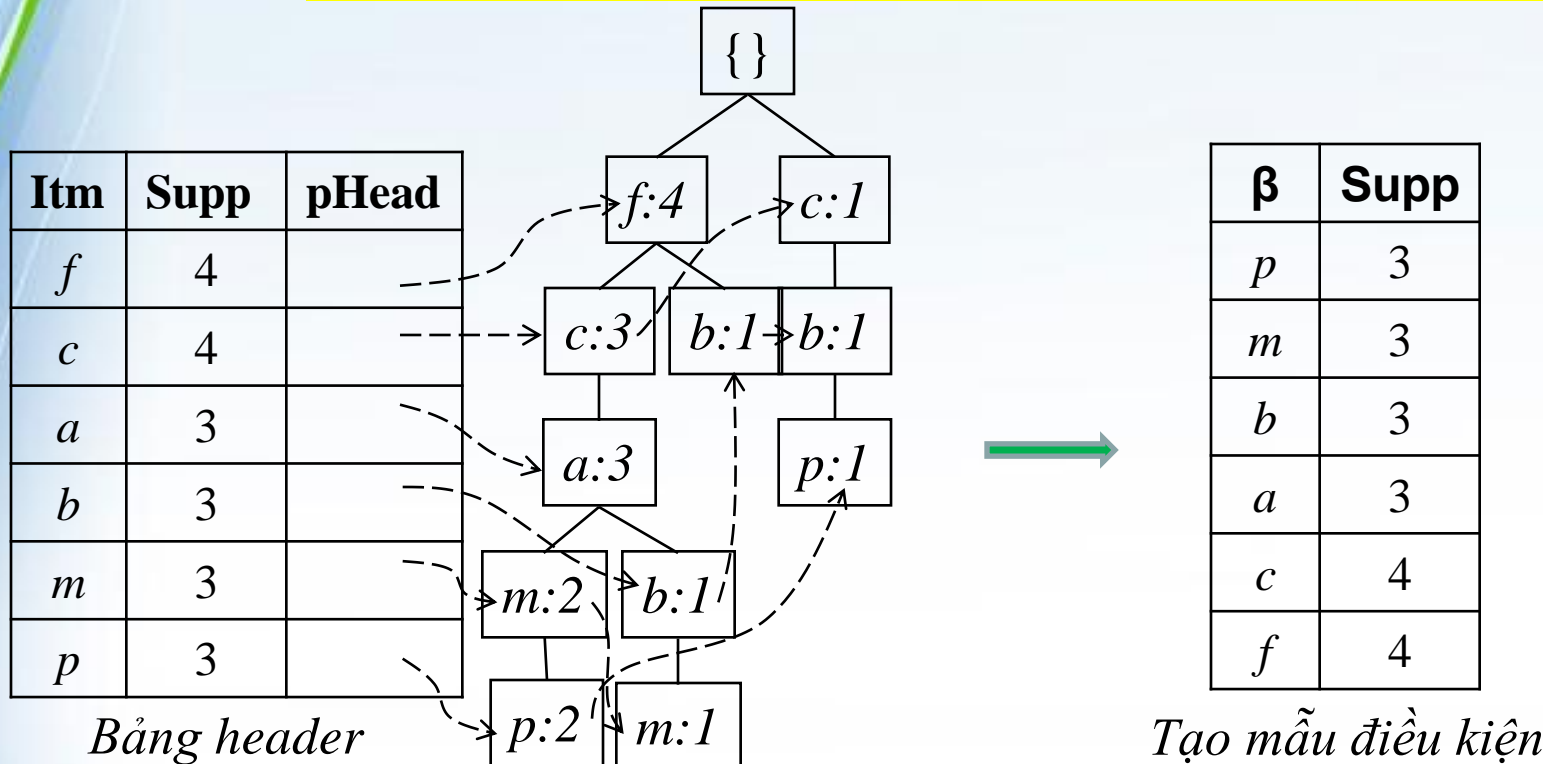
- FP-tree khai thác bằng cách duyệt từ dưới lên
  - Với mỗi mẫu phổ biến  $a_i$  trong bảng header (từ dưới lên)

```
procedure FP_growth
(1)  if Tree contains  $a_i$  then
(2)    for each condition  $\alpha$  in  $a_i$ 's header
(3)      generate pattern  $\beta = a_i \cup \alpha$  with  $support\_count = \alpha.support\_count$ ;
(4)    else for each  $a_i$  in the header of Tree {
(5)      generate pattern  $\beta = a_i \cup \alpha$  with  $support\_count = a_i.support\_count$ ;
(6)      construct  $\beta$ 's conditional pattern base and then  $\beta$ 's conditional FP-tree  $Tree_\beta$ ;
(7)      if  $Tree_\beta \neq \emptyset$  then
(8)        call FP_growth( $Tree_\beta, \beta$ ); }
```

# B2.1. Tạo mẫu điều kiện $\beta$

Với mỗi hàng mục phổ biến  $a_i$  trong bảng header (từ dưới lên)

1. Tạo mẫu điều kiện  $\beta$  bằng cách kết hợp hàng mục  $a_i$  với mẫu điều kiện trước đó và đếm trợ = đếm trợ của  $a_i$



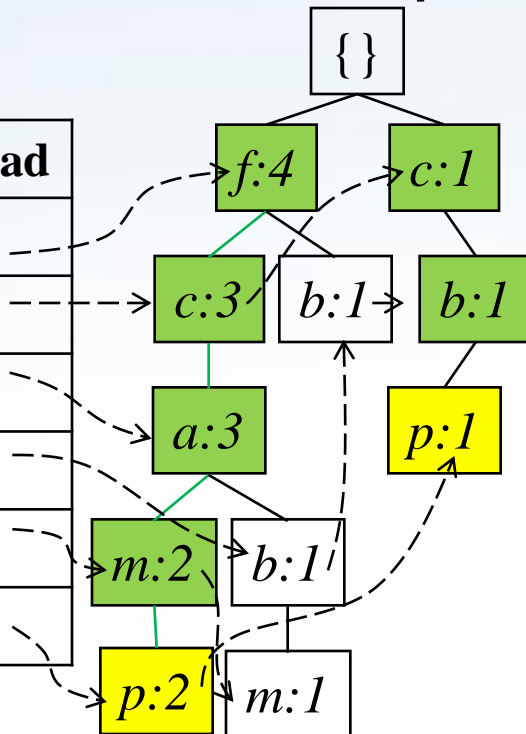


## B2.2. Xd cơ sở mẫu điều kiện $\beta$

- Dựa trên FP-tree và theo kết nối của mỗi mẫu.
- Gom tất cả đường dẫn tiền tố biến đổi (transformed prefix) của mẫu để tạo cơ sở mẫu điều kiện

| Itm      | Supp | pHead |
|----------|------|-------|
| <i>f</i> | 4    |       |
| <i>c</i> | 4    |       |
| <i>a</i> | 3    |       |
| <i>b</i> | 3    |       |
| <i>m</i> | 3    |       |
| <i>p</i> | 3    |       |

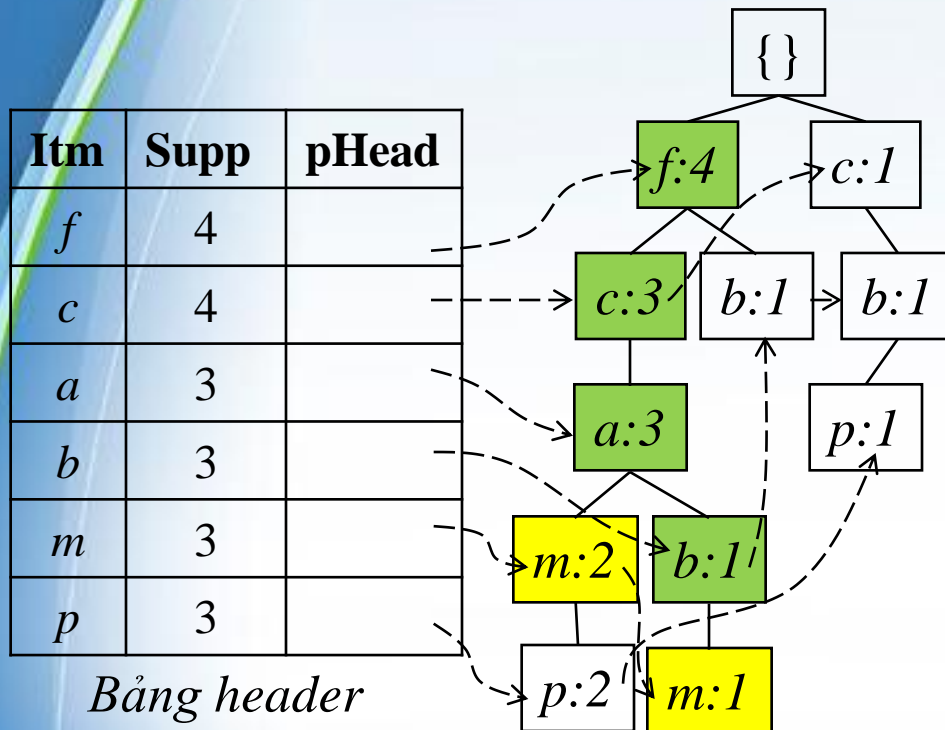
Bảng header



| $\beta$  | Supp | Cond.Pattern Base      |
|----------|------|------------------------|
| <i>p</i> | 3    | $\{fcam:2\}, \{cb:1\}$ |
| <i>m</i> | 3    |                        |
| <i>b</i> | 3    |                        |
| <i>a</i> | 3    |                        |
| <i>c</i> | 4    |                        |
| <i>f</i> | 4    |                        |

## B2.2. Xd cơ sở mẫu điều kiện $\beta$

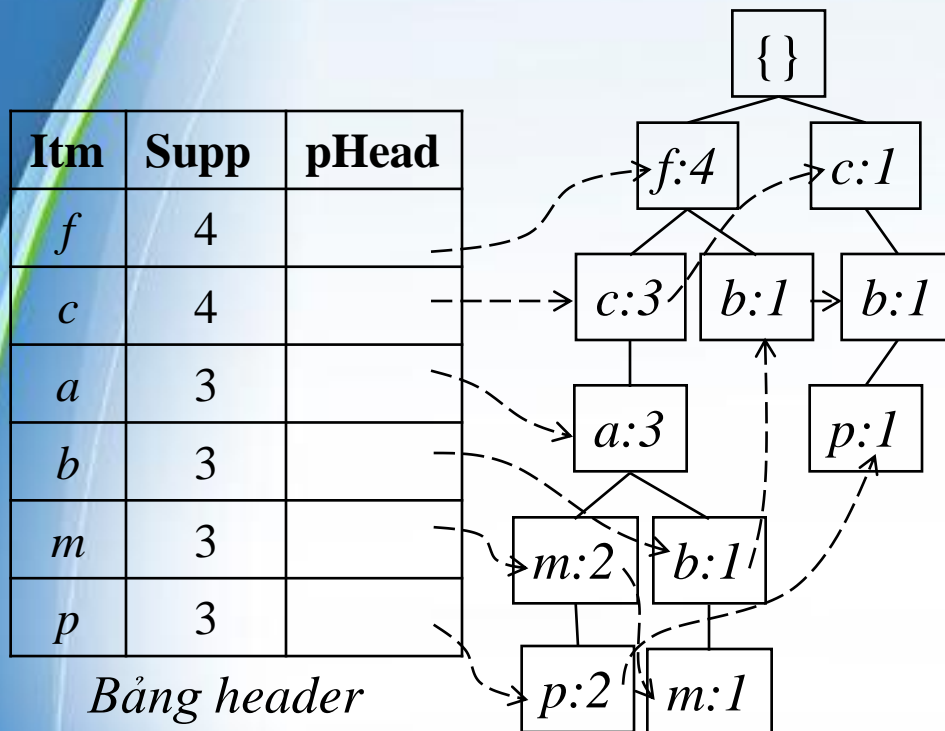
- Dựa trên FP-tree và theo kết nối của mỗi mẫu.
- Gom tất cả đường dẫn tiền tố biến đổi (transformed prefix) của mẫu để tạo cơ sở mẫu điều kiện



| $\beta$  | Supp | Cond.Pattern Base  |
|----------|------|--------------------|
| <i>p</i> | 3    | {fcam:2}, {cb:1}   |
| <i>m</i> | 3    | {fca:2}, {fcabm:1} |
| <i>b</i> | 3    |                    |
| <i>a</i> | 3    |                    |
| <i>c</i> | 4    |                    |
| <i>f</i> | 4    |                    |

## B2.2. Xd cơ sở mẫu điều kiện $\beta$

- Dựa trên FP-tree và theo kết nối của mỗi mẫu.
- Gom tất cả đường dẫn tiền tố biến đổi (transformed prefix) của mẫu để tạo cơ sở mẫu điều kiện



| $\beta$  | Supp | Cond.Pattern Base                                |
|----------|------|--|
| <i>p</i> | 3    | { <i>fcam:2</i> }, { <i>cb:1</i> }               |
| <i>m</i> | 3    | { <i>fca:2</i> }, { <i>fcabm:1</i> }             |
| <i>b</i> | 3    | { <i>fca:1</i> }, { <i>f:1</i> }, { <i>c:1</i> } |
| <i>a</i> | 3    | { <i>fc:3</i> }                                  |
| <i>c</i> | 4    | { <i>f:3</i> }                                   |
| <i>f</i> | 4    | { }  |

## B2.3. Xd FP-tree điều kiện $\beta$

- Xem cơ sở mẫu điều kiện như là những giao tác.
- Thực hiện xây dựng cây FP-tree cho từng cơ sở như bước B1 (chọn lọc hạng mục phổ biến, x/d) cây từ mỗi giao tác đã chọn lọc *minsupp = 3*

| $\beta$  | Supp | Cond.Pattern Base                                |
|----------|------|--|
| <i>p</i> | 3    | { <i>fcam:2</i> }, { <i>cb:1</i> }               |
| <i>m</i> | 3    | { <i>fca:2</i> }, { <i>fcabm:1</i> }             |
| <i>b</i> | 3    | { <i>fca:1</i> }, { <i>f:1</i> }, { <i>c:1</i> } |
| <i>a</i> | 3    | { <i>fc:3</i> }                                  |
| <i>c</i> | 4    | { <i>f:3</i> }                                   |
| <i>f</i> | 4    | { }  |

*Bảng header*

*Itm Supp pHead*

*c*

3

*f,a,m,b không thỏa minsup nên bị loại*

*p-conditional FP-tree*

{ }

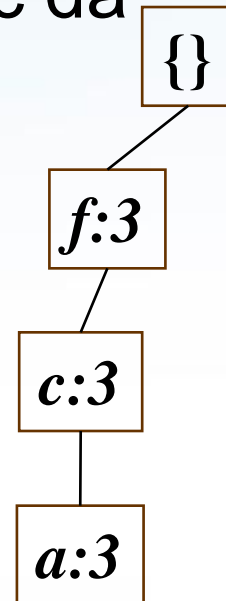
*c:3*

## B2.3. Xd FP-tree điều kiện $\beta$

- Xem cơ sở mẫu điều kiện như là những giao tác.
- Thực hiện xây dựng cây FP-tree cho từng cơ sở như bước B1 (chọn lọc hạng mục phổ biến, x/d) cây từ mỗi giao tác đã chọn lọc

*minsupp* = 3

| $\beta$  | Supp | Cond.Pattern Base                                | Cond.FP-Tree                             |
|----------|------|--|--|
| <i>p</i> | 3    | { <i>fcam</i> :2}, { <i>cb</i> :1}               | < <i>c</i> :3>                           |
| <i>m</i> | 3    | { <i>fca</i> :2}, { <i>fcabm</i> :1}             | < <i>f</i> :3, <i>c</i> :3, <i>a</i> :3> |
| <i>b</i> | 3    | { <i>fca</i> :1}, { <i>f</i> :1}, { <i>c</i> :1} |  |
| <i>a</i> | 3    | { <i>fc</i> :3}                                  |  |
| <i>c</i> | 4    | { <i>f</i> :3}                                   |  |
| <i>f</i> | 4    | { }  |  |



## B2.3. Xd FP-tree điều kiện $\beta$

- Xem cơ sở mẫu điều kiện như là những giao tác.
- Thực hiện xây dựng cây FP-tree cho từng cơ sở như bước B1 (chọn lọc hạng mục phổ biến, x/d) cây từ mỗi giao tác đã chọn lọc

*minsupp = 3*

| $\beta$  | Supp | Cond.Pattern Base                                | Cond.FP-Tree           |
|----------|------|--|------------------------|
| <i>p</i> | 3    | { <i>fcam:2</i> }, { <i>cb:1</i> }               | < <i>c:3</i> >         |
| <i>m</i> | 3    | { <i>fca:2</i> }, { <i>fcabm:1</i> }             | < <i>f:3,c:3,a:3</i> > |
| <i>b</i> | 3    | { <i>fca:1</i> }, { <i>f:1</i> }, { <i>c:1</i> } | < >                    |
| <i>a</i> | 3    | { <i>fc:3</i> }                                  | < <i>f:3,c:3</i> >     |
| <i>c</i> | 4    | { <i>f:3</i> }                                   | < <i>f:3</i> >         |
| <i>f</i> | 4    | { }  | < >                    |

## B2. Khai thác FP-tree

- FP-tree được khai thác bằng cách gọi hàm  $\text{FP\_growth}(\text{FP\_tree}, \text{null})$ :

procedure  $\text{FP\_growth}(\text{Tree}, \alpha)$

- (1) if  $\text{Tree}$  contains a single path  $P$  then
- (2)   for each combination (denoted as  $\beta$ ) of the nodes in the path  $P$
- (3)     generate pattern  $\beta \cup \alpha$  with  $\text{support\_count} = \text{minimum support count of nodes in } \beta$ ;
- (4) else for each
- (5)   generate p
- (6)   construct
- (7)   if  $\text{Tree}_\beta \neq \emptyset$  then
- (8)     call  $\text{FP\_growth}(\text{Tree}_\beta, \beta)$ ;

**Nếu cây FP chứa một nhánh đơn P**

- i. Với mỗi cách kết hợp các node trong nhánh  $P$  (kí hiệu  $\beta$ )  
Phát sinh mẫu phổ biến = mẫu  $\beta$  hợp với mẫu điều kiện  $\alpha$  trước đó và độ trợ = độ trợ của node nhỏ nhất trong  $\beta$

## B2.i. Phát sinh mẫu phổ biến

---

- Nếu cây FP chỉ có một nhánh thì phát sinh mẫu phổ biến bằng cách kết hợp các tổ hợp trên nhánh với mẫu điều kiện.
  - Độ trợ của mẫu là độ trợ nhỏ nhất trong tổ hợp
  - Lưu ý: với tổ hợp rỗng thì mẫu phổ biến chính là mẫu điều kiện và độ trợ bằng độ trợ mẫu điều kiện
- Nếu cây FP có nhiều nhánh thì quá trình đệ quy như các bước trước (x/d mẫu điều kiện, cơ sở mẫu điều kiện, cây FP điều kiện)



## B2.i. Phát sinh mẫu phổ biến

- Nếu cây FP chỉ có một nhánh thì phát sinh mẫu phổ biến bằng cách kết hợp các tổ hợp trên nhánh với mẫu điều kiện.
  - Độ trợ của mẫu là độ trợ nhỏ nhất trong tổ hợp
  - Lưu ý: với tổ hợp rỗng thì mẫu phổ biến chính là mẫu điều kiện và độ trợ bằng độ trợ mẫu điều kiện

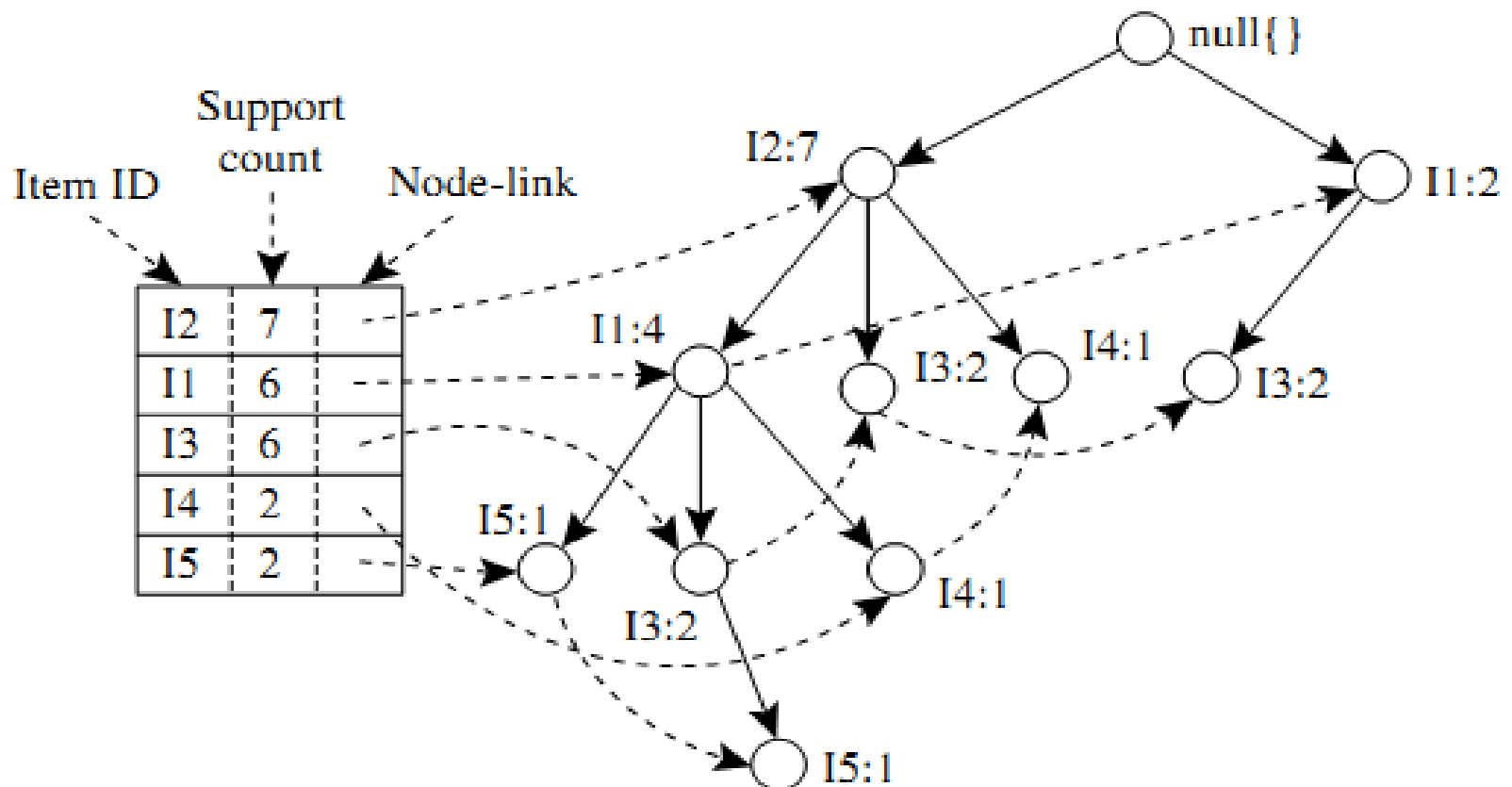
| $\alpha$ | Supp | Cond.Pattern Base             | Cond.FP-Tree                    | Frequent Patterns        |
|----------|------|-------------------------------|---------------------------------|--------------------------|
| $p$      | 3    | $\{fcam:2\}, \{cb:1\}$        | $\langle c:3 \rangle$           | $\{p:3\},$<br>$\{cp:3\}$ |
| $m$      | 3    | $\{fca:2\}, \{fcabm:1\}$      | $\langle f:3, c:3, a:3 \rangle$ |                          |
| $b$      | 3    | $\{fca:1\}, \{f:1\}, \{c:1\}$ | $\langle \rangle$               |                          |
| $a$      | 3    | $\{fc:3\}$                    | $\langle f:3, c:3 \rangle$      |                          |
| $c$      | 4    | $\{f:3\}$                     | $\langle f:3 \rangle$           |                          |
| $f$      | 4    | $\{ \}$                       | $\langle \rangle$               |                          |

## B2.i. Phát sinh mẫu phổ biến

| $\alpha$ | Supp | Cond.Pattern Base             | Cond.FP-Tree                    | Frequent Patterns   |
|----------|------|-------------------------------|---------------------------------|---|
| $p$      | 3    | $\{fcam:2\}, \{cb:1\}$        | $\langle c:3 \rangle$           | $\{p:3\},$<br>$\{cp:3\}$  |
| $m$      | 3    | $\{fca:2\}, \{fcabm:1\}$      | $\langle f:3, c:3, a:3 \rangle$ | $\{m:3\},$<br>$\{fm:3\}, \{cm:3\}, \{am:3\},$<br>$\{fcm:3\}, \{fam:3\}, \{cam:3\},$<br>$\{fcam:3\}$ |
| $b$      | 3    | $\{fca:1\}, \{f:1\}, \{c:1\}$ | $\langle \rangle$               | $\{b:3\}$   |
| $a$      | 3    | $\{fc:3\}$                    | $\langle f:3, c:3 \rangle$      | $\{a:3\},$<br>$\{fa:3\}, \{ca:3\},$<br>$\{fca:3\}$  |
| $c$      | 4    | $\{f:3\}$                     | $\langle f:3 \rangle$           | $\{c:4\},$<br>$\{fc:3\}$  |
| $f$      | 4    | $\{ \}$                       | $\langle \rangle$               | $\{f:4\}$   |

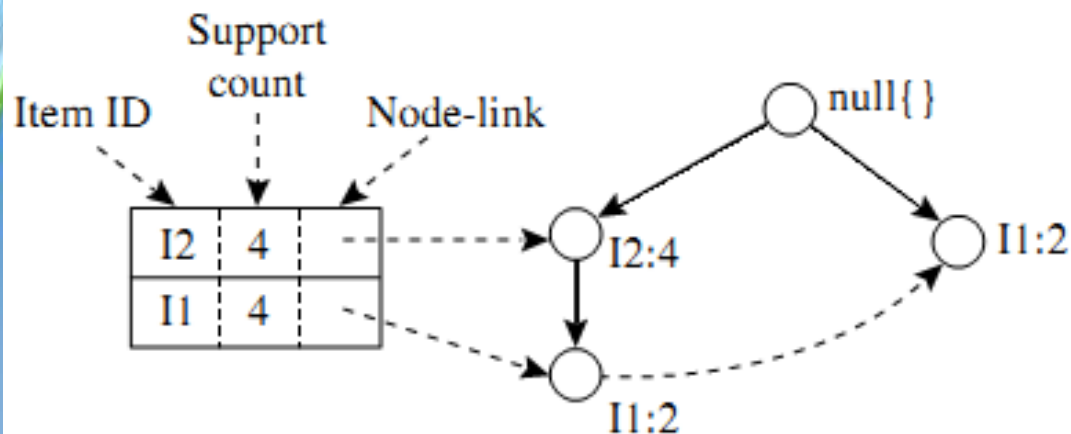
# Bài tập áp dụng 2

Tìm mẫu phổ biến từ cây FP đã xây dựng trong bài tập 1 (minsup = 2):



# Bài tập áp dụng 2 – Đáp án

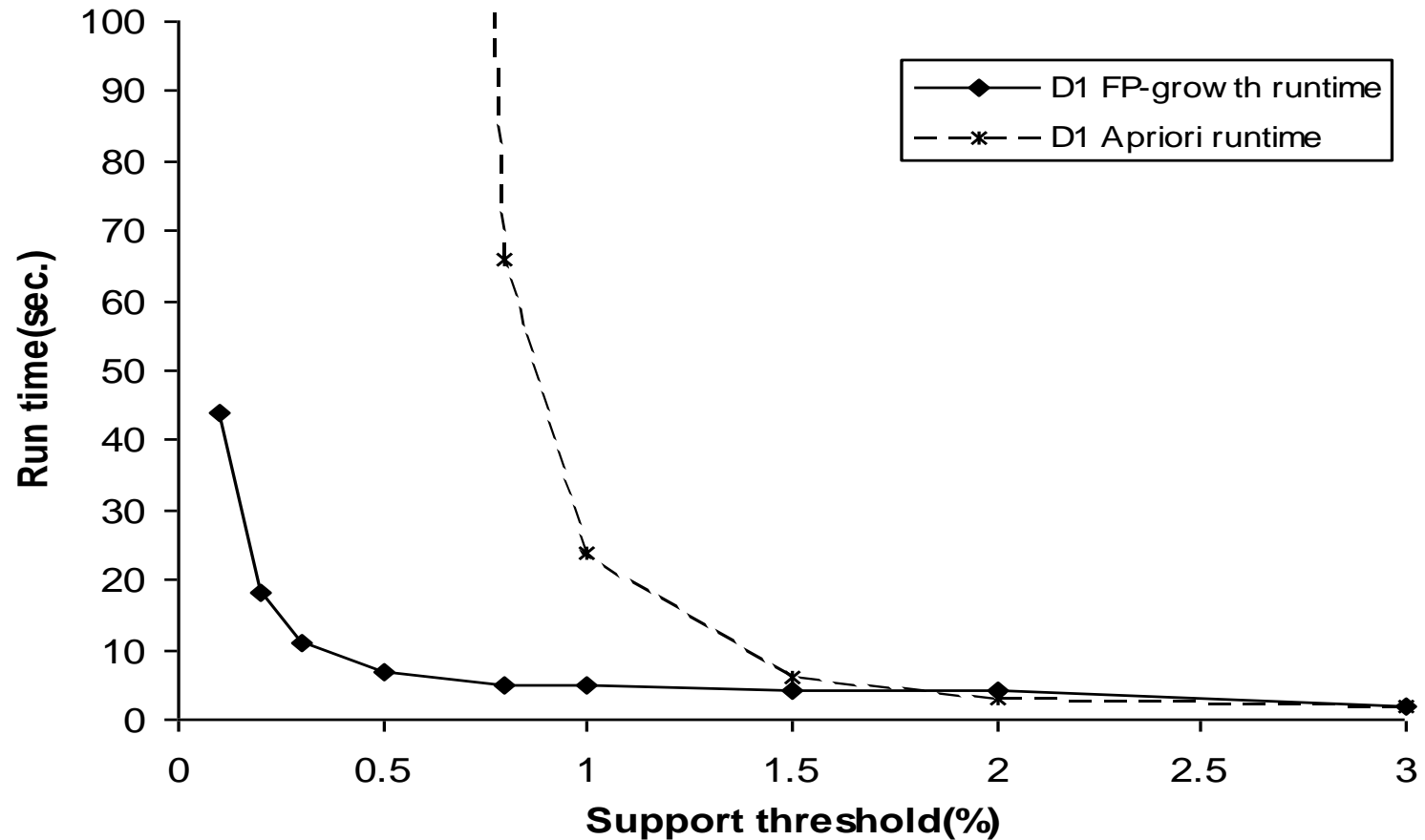
| Item | Conditional Pattern Base        | Conditional FP-tree                                   | Frequent Patterns Generated               |
|------|---------------------------------|---|---|
| I5   | {{I2, I1: 1}, {I2, I1, I3: 1}}  | $\langle I2: 2, I1: 2 \rangle$                        | {I2, I5: 2}, {I1, I5: 2}, {I2, I1, I5: 2} |
| I4   | {{I2, I1: 1}, {I2: 1}}          | $\langle I2: 2 \rangle$                               | {I2, I4: 2}                               |
| I3   | {{I2, I1: 2}, {I2: 2}, {I1: 2}} | $\langle I2: 4, I1: 2 \rangle, \langle I1: 2 \rangle$ | {I2, I3: 4}, {I1, I3: 4}, {I2, I1, I3: 2} |
| I1   | {{I2: 4}}                       | $\langle I2: 4 \rangle$                               | {I2, I1: 4}                               |



*Cây triển khai tiếp từ I3*

# So sánh Apriori và FP-Growth

Data set T25I20D10K



# Nội dung

---

- Thuật toán FP-Growth
  - Xây dựng cây FP-Growth
  - Phát sinh mẫu phổ biến từ FP-Growth
- So sánh Fp-Growth và Apriori
- Độ đo tính lý thú của LKH

# Vấn đề khi phát sinh luật

- Thuật toán khai thác LKH có xu hướng sinh ra quá nhiều luật khi minsup và minconf thấp.
- Trong đó có nhiều luật không hay hoặc bị thừa, thậm chí là sai.



# Ví dụ 1 về v/đ phát sinh luật

Thống kê số giao tác/hóa đơn của một cửa hàng bán nước:

|                       | <i>Coffee</i> | $\overline{Coffee}$ | $\Sigma_{\text{dòng}}$ |
|-----------------------|---------------|---------------------|------------------------|
| <i>Tea</i>            | 15            | 5                   | 20                     |
| $\overline{Tea}$      | 75            | 5                   | 80                     |
| $\Sigma_{\text{cột}}$ | 90            | 10                  | 100                    |

Luật kết hợp: **Tea**  $\rightarrow$  **Coffee** [sup = 15%; conf = 75%]  
 $\rightarrow$  Mặc dù độ tin cậy cao nhưng luật gây ra sự hiểu lầm vì thực tế  $P(\text{coffee}) = 90\%$  còn cao hơn conf của luật.

$$P(\text{coffee}|\overline{\text{tea}}) = \frac{75}{80} = 0.9375$$



# Ví dụ 2 về v/đ phát sinh luật

Thống kê về mối liên hệ giữa SV chơi bóng rổ và ăn sáng với ngũ cốc:

|                     | <i>Chơi bóng</i> | $\overline{Ko\ chơi\ bóng}$ | $\Sigma_{dòng}$ |
|---------------------|------------------|-----------------------------|-----------------|
| <i>Ăn</i>           | 4000             | 3500                        | 7500            |
| $\overline{Ko\ ăn}$ | 2000             | 500                         | 2500            |
| $\Sigma_{cột}$      | 6000             | 4000                        | 10,000          |

LKH: **Chơi bóng rổ  $\rightarrow$  Ăn ngũ cốc** [sup=40%; conf=66%]

- Ta thấy, phần trăm SV ăn ngũ cốc  $P(\text{Ăn ngũ cốc}) = 75\%$  lại còn cao hơn cả conf của luật.

- Thực tế, LKH: Chơi bóng rổ  $\rightarrow$  Không ăn ngũ cốc [20%; 34%] chính xác hơn.



# Độ đo tính lý thú của luật

- Luật thể nào là tốt/ly thú (interesting)?
  - tùy vào dữ liệu và mang tính chủ quan (ly thú với người này nhưng lại không ly thú với người khác)
  - Một cách tương đối, bằng pp thống kê hướng đến loại bỏ các luật không ly thú.
- Các độ đo ly thú là gì?
  - Độ đo tương quan giữa tiền đề và kết luận (correlation)
  - Độ đo khả năng dẫn xuất (implication)

# Các độ đo mối tương quan

| symbol    | measure             | range            | formula   |
|-----------|---------------------|------------------|---|
| $\phi$    | $\phi$ -coefficient | -1 ... 1         | $\frac{P(A,B) - P(A)P(B)}{\sqrt{P(A)P(B)(1-P(A))(1-P(B))}}$   |
| $Q$       | Yule's Q            | -1 ... 1         | $\frac{P(A,B)P(\bar{A},\bar{B}) - P(A,\bar{B})P(\bar{A},B)}{P(A,B)P(\bar{A},\bar{B}) + P(A,\bar{B})P(\bar{A},B)}$   |
| $Y$       | Yule's Y            | -1 ... 1         | $\frac{\sqrt{P(A,B)P(\bar{A},\bar{B})} - \sqrt{P(A,\bar{B})P(\bar{A},B)}}{\sqrt{P(A,B)P(\bar{A},\bar{B})} + \sqrt{P(A,\bar{B})P(\bar{A},B)}}$   |
| $k$       | Cohen's             | -1 ... 1         | $\frac{P(A,B) + P(\bar{A},\bar{B}) - P(A)P(B) - P(\bar{A})P(\bar{B})}{1 - P(A)P(B) - P(\bar{A})P(\bar{B})}$   |
| $PS$      | Piatetsky-Shapiro's | -0.25 ... 0.25   | $P(A, B) - P(A)P(B)$  |
| $F$       | Certainty factor    | -1 ... 1         | $\max\left(\frac{P(B A) - P(B)}{1 - P(B)}, \frac{P(A B) - P(A)}{1 - P(A)}\right)$   |
| $AV$      | added value         | -0.5 ... 1       | $\max(P(B A) - P(B), P(A B) - P(A))$  |
| $K$       | Klogsen's Q         | -0.33 ... 0.38   | $\frac{\sqrt{P(A, B) \max(P(B A) - P(B), P(A B) - P(A))}}{\sum_j \max_k P(A_j, B_k) + \sum_k \max_j P(A_j, B_k) - \max_k P(A_k)}$   |
| $g$       | Goodman-kruskal's   | 0 ... 1          | $\frac{2 - \max_j P(A_j) - \max_k P(B_k)}{\sum_i \sum_j P(A_i, B_j) \log \frac{P(A_i, B_j)}{P(A_i)P(B_j)}}$   |
| $M$       | Mutual Information  | 0 ... 1          | $\frac{\min(-\sum_i P(A_i) \log P(A_i) \log P(A_i), -\sum_i P(B_i) \log P(B_i) \log P(B_i))}{\max(P(A, B) \log(\frac{P(B A)}{P(B)}) + P(\bar{A}\bar{B}) \log(\frac{P(\bar{B} \bar{A})}{P(\bar{B})}))}$                                    |
| $J$       | J-Measure           | 0 ... 1          | $\frac{P(A, B) \log(\frac{P(A B)}{P(A)}) + P(\bar{A}\bar{B}) \log(\frac{P(\bar{A} \bar{B})}{P(\bar{A})})}{\max(P(A, B) \log(\frac{P(B A)}{P(B)}) + P(\bar{A}\bar{B}) \log(\frac{P(\bar{B} \bar{A})}{P(\bar{B})}))}$                       |
| $G$       | Gini index          | 0 ... 1          | $\max(P(A)[P(B A)^2 + P(\bar{B} \bar{A})^2] + P(\bar{A})[P(B \bar{A})^2 + P(\bar{B} \bar{A})^2] - P(B)^2 - P(\bar{B})^2, \\ P(B)[P(A B)^2 + P(\bar{A} B)^2] + P(\bar{B})[P(A \bar{B})^2 + P(\bar{A} \bar{B})^2] - P(A)^2 - P(\bar{A})^2)$ |
| $s$       | support             | 0 ... 1          | $P(A, B)$   |
| $c$       | confidence          | 0 ... 1          | $\max(P(B A), P(A B))$  |
| $L$       | Laplace             | 0 ... 1          | $\max\left(\frac{NP(A,B)+1}{NP(A)+2}, \frac{NP(A,B)+1}{NP(B)+2}\right)$   |
| $IS$      | Cosine              | 0 ... 1          | $\frac{P(A,B)}{\sqrt{P(A)P(B)}}$  |
| $\gamma$  | coherence(Jaccard)  | 0 ... 1          | $\frac{P(A,B)}{P(A)+P(B)-P(A,B)}$   |
| $\alpha$  | all_confidence      | 0 ... 1          | $\frac{\max(P(A), P(B))}{P(A, B)}$  |
| $o$       | odds ratio          | 0 ... $\infty$   | $\frac{P(A,B)P(\bar{A},\bar{B})}{P(\bar{A},B)P(A,\bar{B})}$   |
| $V$       | Conviction          | 0.5 ... $\infty$ | $\max\left(\frac{P(A)P(\bar{B})}{P(\bar{A}\bar{B})}, \frac{P(B)P(\bar{A})}{P(\bar{B}\bar{A})}\right)$   |
| $\lambda$ | lift                | 0 ... $\infty$   | $\frac{P(A,B)}{P(A)P(B)}$   |
| $S$       | Collective strength | 0 ... $\infty$   | $\frac{P(A,B) + P(\bar{A}\bar{B})}{P(A)P(B) + P(\bar{A})P(\bar{B})} \times \frac{1 - P(A)P(B) - P(\bar{A})P(\bar{B})}{1 - P(A,B) - P(\bar{A}\bar{B})}$  |
| $\chi^2$  | $\chi^2$            | 0 ... $\infty$   | $\sum_i \frac{(P(A_i) - E_i)^2}{E_i}$   |

# Thảo luận độ đo *lift* (1/2)

- Độ đo *lift* cho luật  $A \rightarrow B$ :

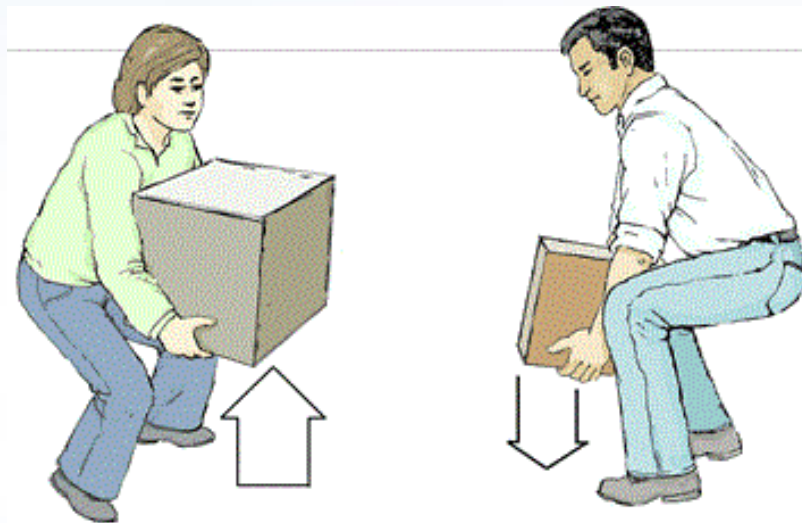
$$lift(A, B) = \frac{P(A \cup B)}{P(A)P(B)}$$

- Thể hiện hạng mục A có xảy ra độc lập với hạng mục B hay không?
  - Nếu độc lập:  $P(A, B) = P(A).P(B)$ , hay *lift* (A,B) = 1.
  - Tương quan thuận (negatively correlated): *lift* (A,B) < 1, nghĩa là cái này xảy ra dẫn đến cái kia sẽ không xảy ra.
  - Tương quan nghịch (positively correlated): *lift* (A,B) > 1, nghĩa là cái này xảy ra dẫn đến cái kia cũng xảy ra.

# Thảo luận độ đo *lift* (1/2)

- Nói cách khác:
  - A tăng cường hay “nâng” (*lift*) khả năng xảy ra của B với hệ số là giá trị của biểu thức *lift*(A,B)

$$lift(A, B) = \frac{P(A, B)}{P(A)P(B)} = \frac{P(B|A)}{P(B)} = \frac{conf(A \rightarrow B)}{sup(B)}$$



# Ví dụ độ đo *lift*

|                     | <i>Chơi bóng</i> | $\overline{Ko\ chơi\ bóng}$ | $\Sigma_{dòng}$ |
|---------------------|------------------|-----------------------------|-----------------|
| <i>Ăn</i>           | 4000             | 3500                        | 7500            |
| $\overline{Ko\ ăn}$ | 2000             | 500                         | 2500            |
| $\Sigma_{cột}$      | 6000             | 4000                        | 10,000          |

*lift* (*Chơi bóng*  $\rightarrow$  *Ăn ngũ cốc*) =

$$\frac{P(\text{Chơi bóng}, \text{Ăn})}{P(\text{Chơi bóng}) \times P(\text{Ăn})} = \frac{\frac{4000}{10000}}{\frac{6000}{10000} \times \frac{7500}{10000}} = 0.89 < 1$$

*lift* (*Chơi bóng*  $\rightarrow$   $\overline{\text{Ăn ngũ cốc}}$ ) = ?



# Bài tập áp dụng 3

- Phát sinh các luật dựa trên các mẫu phổ biến trong bài tập 2 thỏa  $\text{minsup} = 22\%$  và  $\text{minconf} = 70\%$
- Tính độ đo *lift* của từng luật trên.

| Item | Conditional Pattern Base        | Conditional FP-tree                                   | Frequent Patterns Generated               |
|------|---------------------------------|---|---|
| I5   | {{I2, I1: 1}, {I2, I1, I3: 1}}  | $\langle I2: 2, I1: 2 \rangle$                        | {I2, I5: 2}, {I1, I5: 2}, {I2, I1, I5: 2} |
| I4   | {{I2, I1: 1}, {I2: 1}}          | $\langle I2: 2 \rangle$                               | {I2, I4: 2}                               |
| I3   | {{I2, I1: 2}, {I2: 2}, {I1: 2}} | $\langle I2: 4, I1: 2 \rangle, \langle I1: 2 \rangle$ | {I2, I3: 4}, {I1, I3: 4}, {I2, I1, I3: 2} |
| I1   | {{I2: 4}}                       | $\langle I2: 4 \rangle$                               | {I2, I1: 4}                               |

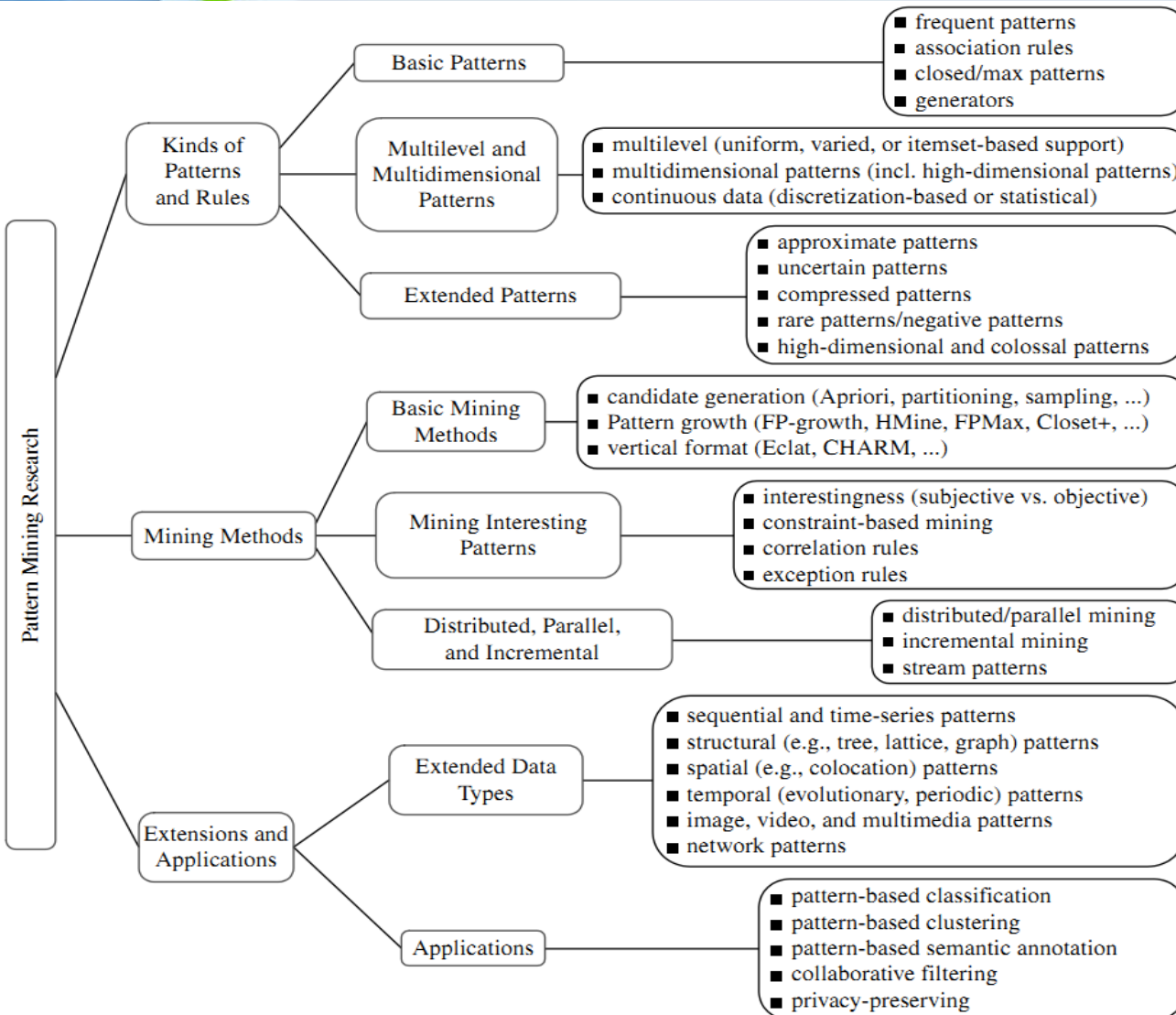
# So sánh các độ đo tương quan

- Đọc thêm trong phần 6.3.2 và 6.3.3 cuốn Data Mining: Concepts and Techniques (3<sup>rd</sup> Edition, J.Han) để nắm các độ đo tương quan khác và việc đánh giá các độ đo này.





# Road-map lĩnh vực khai thác mẫu



- Đọc thêm chương 7 cuốn Data Mining: Concepts and Techniques (3<sup>rd</sup> Edition, J.Han) để nắm thêm các hướng nghiên cứu liên quan đến khai thác mẫu.



# Bài tập tổng hợp 1

- Cho CSDL sau và minsup = 50%, minconf = 80%

| TID | Date    | Items_bought     |
|-----|---------|------------------|
| 100 | 15/1/03 | K, A, D, B, C, I |
| 200 | 15/1/03 | D, A, C, E, B    |
| 300 | 19/1/03 | C, A, B, E, D    |
| 400 | 25/1/03 | B, A, D, I       |

- a. Tìm tất cả các tập phổ biến, tập phổ biến tối đại, tập phổ biến đóng sử dụng thuật toán Apriori và FP-Growth
- b. Xây dựng LKH thỏa minsup và minconf
- c. Tính độ đo tương quan lift của các luật tìm được ở câu b

# Bài tập tổng hợp 2

| Transaction | Items                           |
|-------------|---------------------------------|
| $t_1$       | Blouse                          |
| $t_2$       | Shoes,Skirt,TShirt              |
| $t_3$       | Jeans,TShirt                    |
| $t_4$       | Jeans,Shoes,TShirt              |
| $t_5$       | Jeans,Shorts                    |
| $t_6$       | Shoes,TShirt                    |
| $t_7$       | Jeans,Skirt                     |
| $t_8$       | Jeans,Shoes,Shorts,TShirt       |
| $t_9$       | Jeans                           |
| $t_{10}$    | Jeans,Shoes,TShirt              |
| $t_{11}$    | TShirt                          |
| $t_{12}$    | Blouse,Jeans,Shoes,Skirt,TShirt |
| $t_{13}$    | Jeans,Shoes,Shorts,TShirt       |
| $t_{14}$    | Shoes,Skirt,TShirt              |
| $t_{15}$    | Jeans,TShirt                    |
| $t_{16}$    | Skirt,TShirt                    |
| $t_{17}$    | Blouse,Jeans,Skirt              |
| $t_{18}$    | Jeans,Shoes,Shorts,TShirt       |
| $t_{19}$    | Jeans                           |
| $t_{20}$    | Jeans,Shoes,Shorts,TShirt       |

- Cho CSDL bên và minsup=30%, minconf=50%.
- Yêu cầu tương tự bài tập tổng hợp 1

# Tóm tắt

---

- Các hạn chế của Apriori
- Thuật toán FP-Growth giảm thiểu việc duyệt dữ liệu nhiều lần bằng cách hình thành cây FP, dựa trên cây FP để chia dữ liệu thành các vùng nhỏ để tìm mẫu phổ biến.
- Các luật sinh ra có thể thừa hay gây nhầm lẫn, việc loại bỏ nó thông qua đánh giá các độ đo tương quan.

# Tài liệu tham khảo

---

1. J. Han, J. Pei, and Y. Yin. *Mining frequent patterns without candidate generation*. SIGMOD'00, 1-12, Dallas, TX, May 2000
2. Improvements on FP-growth,  
<http://www.cs.sfu.ca/CourseCentral/741/jpei/slides/FP-growth-Improvements.pdf>
3. J.Han, M.Kamber, Chương 6 – Mining Frequent Patterns, Associations, and Correlations và Chương 7 - Advanced Pattern Mining cuốn “*Data mining: Basic Concepts and Methods*”, 3<sup>rd</sup> edition

# Hỏi & Đáp

---

