

# Khai phá mẫu phổ biến và luật kết hợp (Frequent Pattern and Association Rule Mining)

Phan Xuân Hiếu

Khoa Công nghệ Thông tin  
Trường ĐH Công nghệ (UET), ĐHQG Hà Nội (VNU)  
[hieupx@vnu.edu.vn](mailto:hieupx@vnu.edu.vn)

(last updated: 08–11–2015)

# Nội dung

## ① Các khái niệm và định nghĩa

- Tập mục, giao dịch, và cơ sở dữ liệu giao dịch
- Tập phổ biến và luật kết hợp
- Ứng dụng của tập phổ biến và luật kết hợp

## ② Khai phá tập phổ biến và luật kết hợp

- Các bước trong khai phá luật kết hợp
- Phương pháp brute-force
- Phương pháp Apriori
- Phương pháp FP-Growth
- Phương pháp Eclat

## ③ Các chủ đề khác

- Các tiêu chí đánh giá luật kết hợp (interestingness measures)
- Luật kết hợp tổng quát (generalized association rules)
- Luật kết hợp định lượng (quantitative association rules)
- Luật kết hợp cực đại (maximal association rules)
- Khai phá mẫu chuỗi (sequential pattern mining)

## ④ Tổng kết bài giảng

# Nội dung

## 1 Các khái niệm và định nghĩa

- Tập mục, giao dịch, và cơ sở dữ liệu giao dịch
- Tập phổ biến và luật kết hợp
- Ứng dụng của tập phổ biến và luật kết hợp

## 2 Khai phá tập phổ biến và luật kết hợp

- Các bước trong khai phá luật kết hợp
- Phương pháp brute-force
- Phương pháp Apriori
- Phương pháp FP-Growth
- Phương pháp Eclat

## 3 Các chủ đề khác

- Các tiêu chí đánh giá luật kết hợp (interestingness measures)
- Luật kết hợp tổng quát (generalized association rules)
- Luật kết hợp định lượng (quantitative association rules)
- Luật kết hợp cực đại (maximal association rules)
- Khai phá mẫu chuỗi (sequential pattern mining)

## 4 Tổng kết bài giảng

# Khái niệm luật kết hợp (association rule)



- Luật kết hợp: *Mỗi quan hệ kết hợp giữa các tập thuộc tính trong cơ sở dữ liệu.*
- Ví dụ:
  - ▶  $\{bánh mỳ, bơ, mứt dâu\} \rightarrow \{sữa tươi\}$  (phổ biến: 3%, tin cậy: 80%)
  - ▶  $\{tuổi > 45, gia đình có lịch sử tiểu đường, huyết áp cao\} \rightarrow \{mắc bệnh tiểu đường\}$  (phổ biến: 1.5%, tin cậy: 76%)

# Nội dung

## ① Các khái niệm và định nghĩa

- Tập mục, giao dịch, và cơ sở dữ liệu giao dịch
- Tập phổ biến và luật kết hợp
- Ứng dụng của tập phổ biến và luật kết hợp

## ② Khai phá tập phổ biến và luật kết hợp

- Các bước trong khai phá luật kết hợp
- Phương pháp brute-force
- Phương pháp Apriori
- Phương pháp FP-Growth
- Phương pháp Eclat

## ③ Các chủ đề khác

- Các tiêu chí đánh giá luật kết hợp (interestingness measures)
- Luật kết hợp tổng quát (generalized association rules)
- Luật kết hợp định lượng (quantitative association rules)
- Luật kết hợp cực đại (maximal association rules)
- Khai phá mẫu chuỗi (sequential pattern mining)

## ④ Tổng kết bài giảng

# Tập mục, giao dịch, và cơ sở dữ liệu giao dịch (Itemset, Transaction, and Transactional Database)

- Ký hiệu  $\mathbb{I} = \{x_1, x_2, \dots, x_n\}$  là tập  $n$  mục (item). Ví dụ:
  - ▶ Tập tất cả các mặt hàng thực phẩm trong siêu thị:  $\mathbb{I} = \{sữa, trứng, đường, bánh mỳ, mật ong, mút, bơ, thịt bò, giá, \dots\}$ .
  - ▶ Tập tất cả các bộ phim:  $\mathbb{I} = \{pearl harbor, fast and furious 7, fifty shades of grey, spectre, \dots\}$ .
- Một tập  $X \subseteq \mathbb{I}$  được gọi là một tập mục (itemset).
- Nếu  $X$  có  $k$  mục (tức  $|X| = k$ ) thì  $X$  được gọi là  $k$ -itemset.
- Ký hiệu  $\mathbb{D} = \{T_1, T_2, \dots, T_m\}$  là cơ sở dữ liệu gồm  $m$  giao dịch (transaction). Mỗi giao dịch  $T_i \in \mathbb{D}$  là một tập mục, tức  $T_i \subseteq \mathbb{I}$ .

# Ví dụ về cơ sở dữ liệu giao dịch

Tập tất cả các mục  $\mathbb{I}$ :

$$\mathbb{I} = \{A, B, C, D, E\}$$

Cơ sở dữ liệu giao dịch  $\mathbb{D}$ :

$\mathbb{D} = \{T_1, T_2, T_3, T_4, T_5, T_6\}$ , cụ thể:

- $T_1 = \{A, B, D, E\}$
- $T_2 = \{B, C, E\}$
- $T_3 = \{A, B, D, E\}$
- $T_4 = \{A, B, C, E\}$
- $T_5 = \{A, B, C, D, E\}$
- $T_6 = \{B, C, D\}$

# Nội dung

## ① Các khái niệm và định nghĩa

- Tập mục, giao dịch, và cơ sở dữ liệu giao dịch
- **Tập phổ biến và luật kết hợp**
- Ứng dụng của tập phổ biến và luật kết hợp

## ② Khai phá tập phổ biến và luật kết hợp

- Các bước trong khai phá luật kết hợp
- Phương pháp brute-force
- Phương pháp Apriori
- Phương pháp FP-Growth
- Phương pháp Eclat

## ③ Các chủ đề khác

- Các tiêu chí đánh giá luật kết hợp (interestingness measures)
- Luật kết hợp tổng quát (generalized association rules)
- Luật kết hợp định lượng (quantitative association rules)
- Luật kết hợp cực đại (maximal association rules)
- Khai phá mẫu chuỗi (sequential pattern mining)

## ④ Tổng kết bài giảng

## Tập/mẫu phổ biến (frequent itemset/pattern)

- Cho tập mục  $X$  ( $\subseteq \mathbb{I}$ ).
- Độ hỗ trợ (*support*) của  $X$ , ký hiệu là  $sup(X, \mathbb{D})$ , là số lượng giao dịch trong  $\mathbb{D}$  chứa  $X$ :

$$sup(X, \mathbb{D}) = |\{T | T \in \mathbb{D} \text{ và } X \subseteq T\}| \quad (1)$$

- Độ hỗ trợ tương đối (*relative support*) của  $X$ , ký hiệu là  $rsup(X, \mathbb{D})$ , là số phần trăm các giao dịch trong  $\mathbb{D}$  chứa  $X$ :

$$rsup(X, \mathbb{D}) = \frac{sup(X, \mathbb{D})}{|\mathbb{D}|} \quad (2)$$

- Tập mục  $X$  được gọi là tập (mục) phổ biến (frequent itemset) trong  $\mathbb{D}$  nếu  $sup(X, \mathbb{D}) \geq minsup$ , với  $minsup$  là một ngưỡng độ hỗ trợ tối thiểu (minimum support threshold) do người dùng định nghĩa.
- Ký hiệu  $\mathbb{F}$  là tập tất cả các tập phổ biến.
- Ký hiệu  $\mathbb{F}^{(k)}$  là tập tất cả các tập phổ biến có độ dài  $k$  (frequent  $k$ -itemsets).

## Các tập phổ biến (với $minsup = 3$ ) từ cơ sở dữ liệu $\mathbb{D}$

$$\mathbb{D} = \{T_1, T_2, T_3, T_4, T_5, T_6\}:$$

- $T_1 = \{A, B, D, E\}$
- $T_2 = \{B, C, E\}$
- $T_3 = \{A, B, D, E\}$
- $T_4 = \{A, B, C, E\}$
- $T_5 = \{A, B, C, D, E\}$
- $T_6 = \{B, C, D\}$

Tập tất cả các tập phổ biến  $\mathbb{F}$  và các  $\mathbb{F}^{(k)}$ :

- $\mathbb{F} = \{A, B, C, D, E, AB, AD, AE, BC, BD, BE, CE, DE, ABD, ABE, ADE, BCE, BDE, ABDE\}$
- $\mathbb{F}^{(1)} = \{A, B, C, D, E\}$
- $\mathbb{F}^{(2)} = \{AB, AD, AE, BC, BD, BE, CE, DE\}$
- $\mathbb{F}^{(3)} = \{ABD, ABE, ADE, BCE, BDE\}$
- $\mathbb{F}^{(4)} = \{ABDE\}$

## Luật kết hợp (association rule)

- Luật kết hợp có dạng:

$$X \rightarrow Y \quad (3)$$

với  $X$  và  $Y$  là hai tập mục ( $X, Y \subseteq \mathbb{I}$ ) và  $X \cap Y = \emptyset$ .

- Độ hỗ trợ (*support*) của luật  $X \rightarrow Y$  trong cơ sở dữ liệu  $\mathbb{D}$ , ký hiệu là  $sup(X \rightarrow Y, \mathbb{D})$ , là số giao dịch chứa cả  $X$  và  $Y$ :

$$sup(X \rightarrow Y, \mathbb{D}) = sup(X \cup Y, \mathbb{D}) \quad (4)$$

- Độ hỗ trợ tương đối (*relative support*) của luật  $X \rightarrow Y$  trong cơ sở dữ liệu  $\mathbb{D}$ , ký hiệu  $rsup(X \rightarrow Y, \mathbb{D})$ , là số phần trăm các giao dịch trong  $\mathbb{D}$  chứa cả  $X$  và  $Y$ :

$$rsup(X \rightarrow Y, \mathbb{D}) = \frac{sup(X \cup Y, \mathbb{D})}{|\mathbb{D}|} \quad (5)$$

- Luật  $X \rightarrow Y$  được gọi là phổ biến (frequent) nếu:

$$sup(X \rightarrow Y, \mathbb{D}) \geq minsup \quad (6)$$

## Luật kết hợp (association rule) - tiếp

- Độ tin cậy (*confidence*) của luật  $X \rightarrow Y$  trong  $\mathbb{D}$ , ký hiệu  $conf(X \rightarrow Y, \mathbb{D})$ , là tỉ lệ giữa số giao dịch chứa cả  $X$  và  $Y$  trên số giao dịch chỉ chứa  $X$ :

$$conf(X \rightarrow Y, \mathbb{D}) = \frac{sup(X \cup Y, \mathbb{D})}{sup(X, \mathbb{D})} \quad (7)$$

- Một cách diễn giải khác:  $conf(X \rightarrow Y, \mathbb{D})$  là xác suất có điều kiện mà một giao dịch trong  $\mathbb{D}$  chứa  $Y$  khi nó đã chứa  $X$ :  
 $conf(X \rightarrow Y, \mathbb{D}) = P(Y|X)$ . Tuy nhiên bản chất vẫn là mức độ tin cậy của luật.
- Luật  $X \rightarrow Y$  được gọi là mạnh (*strong*) nếu độ tin cậy của nó lớn hơn hoặc bằng một ngưỡng *minconf* nào đó do người dùng định nghĩa:

$$conf(X \rightarrow Y, \mathbb{D}) \geq minconf \quad (8)$$

- Ngoài độ tin cậy (độ mạnh) của luật kết hợp, còn các tiêu chí khác để đánh giá mức độ giá trị của luật (sẽ bàn luận sau).

# Ví dụ minh họa luật kết hợp

$$\mathbb{D} = \{T_1, T_2, T_3, T_4, T_5, T_6\}:$$

- $T_1 = \{A, B, D, E\}$
- $T_2 = \{B, C, E\}$
- $T_3 = \{A, B, D, E\}$
- $T_4 = \{A, B, C, E\}$
- $T_5 = \{A, B, C, D, E\}$
- $T_6 = \{B, C, D\}$

- Xét luật  $\{B, C\} \rightarrow \{E\}$  (ngắn gọn là  $BC \rightarrow E$ ):
  - ▶  $sup(BC \rightarrow E, \mathbb{D}) = sup(BCE, \mathbb{D}) = 3$
  - ▶  $conf(BC \rightarrow E, \mathbb{D}) = \frac{sup(BCE, \mathbb{D})}{sup(BC, \mathbb{D})} = \frac{3}{4} = 0.75$  (tức 75%)
- Xét luật  $\{A, D\} \rightarrow \{B, E\}$  (ngắn gọn là  $AD \rightarrow BE$ ):
  - ▶  $sup(AD \rightarrow BE, \mathbb{D}) = sup(ABDE, \mathbb{D}) = 3$
  - ▶  $conf(AD \rightarrow BE, \mathbb{D}) = \frac{sup(ABDE, \mathbb{D})}{sup(AD, \mathbb{D})} = \frac{3}{3} = 1.0$  (tức 100%)

# Nội dung

## ① Các khái niệm và định nghĩa

- Tập mục, giao dịch, và cơ sở dữ liệu giao dịch
- Tập phổ biến và luật kết hợp
- **Ứng dụng của tập phổ biến và luật kết hợp**

## ② Khai phá tập phổ biến và luật kết hợp

- Các bước trong khai phá luật kết hợp
- Phương pháp brute-force
- Phương pháp Apriori
- Phương pháp FP-Growth
- Phương pháp Eclat

## ③ Các chủ đề khác

- Các tiêu chí đánh giá luật kết hợp (interestingness measures)
- Luật kết hợp tổng quát (generalized association rules)
- Luật kết hợp định lượng (quantitative association rules)
- Luật kết hợp cực đại (maximal association rules)
- Khai phá mẫu chuỗi (sequential pattern mining)

## ④ Tổng kết bài giảng

# Ứng dụng của mẫu phổ biến và luật kết hợp

- Phân tích dữ liệu giao dịch bán lẻ (market basket analysis).
  - ▶ Tối ưu việc nhập các ngành hàng.
  - ▶ Sắp xếp vị trí các ngành hàng hợp lý (store layout).
  - ▶ Marketing và khuyến mại.
  - ▶ Gợi ý và khuyến nghị trực tuyến (online recommendation), ví dụ: “frequently bought together products” và “bought this also bought ...”
- Hiểu người dùng thông qua phân tích các mẫu phổ biến từ nhật ký duyệt web (web logs).
- Phân tích tìm ngoại lệ (outlier detection).
- Phân tích về tội phạm và an ninh.
- Khai phá các cấu trúc mạng xã hội (mẫu đồ thị phổ biến).
- Ứng dụng trong phân lớp, phân loại (decision rules).
- Ứng dụng trong khai phá dữ liệu text (text mining).
- Ứng dụng trong khai phá dữ liệu y/sinh học (biomedical data mining).
- Ứng dụng trong khai phá dữ liệu không/thời gian và dữ liệu dòng (stream data).
- ...

# Nội dung

## 1 Các khái niệm và định nghĩa

- Tập mục, giao dịch, và cơ sở dữ liệu giao dịch
- Tập phổ biến và luật kết hợp
- Ứng dụng của tập phổ biến và luật kết hợp

## 2 Khai phá tập phổ biến và luật kết hợp

- Các bước trong khai phá luật kết hợp
- Phương pháp brute-force
- Phương pháp Apriori
- Phương pháp FP-Growth
- Phương pháp Eclat

## 3 Các chủ đề khác

- Các tiêu chí đánh giá luật kết hợp (interestingness measures)
- Luật kết hợp tổng quát (generalized association rules)
- Luật kết hợp định lượng (quantitative association rules)
- Luật kết hợp cực đại (maximal association rules)
- Khai phá mẫu chuỗi (sequential pattern mining)

## 4 Tổng kết bài giảng

# Nội dung

## 1 Các khái niệm và định nghĩa

- Tập mục, giao dịch, và cơ sở dữ liệu giao dịch
- Tập phổ biến và luật kết hợp
- Ứng dụng của tập phổ biến và luật kết hợp

## 2 Khai phá tập phổ biến và luật kết hợp

### • Các bước trong khai phá luật kết hợp

- Phương pháp brute-force
- Phương pháp Apriori
- Phương pháp FP-Growth
- Phương pháp Eclat

## 3 Các chủ đề khác

- Các tiêu chí đánh giá luật kết hợp (interestingness measures)
- Luật kết hợp tổng quát (generalized association rules)
- Luật kết hợp định lượng (quantitative association rules)
- Luật kết hợp cực đại (maximal association rules)
- Khai phá mẫu chuỗi (sequential pattern mining)

## 4 Tổng kết bài giảng

# Các bước khai phá luật kết hợp

Hai bước khai phá luật kết hợp từ CSDL giao dịch  $\mathbb{D}$ :

- **Mining frequent itemsets/patterns:** Khai phá tất cả các tập phổ biến từ cơ sở dữ liệu  $\mathbb{D}$  với ngưỡng hỗ trợ tối thiểu  $minsup$ .
  - **Generating strong rules from mined frequent itemsets/patterns:** Sinh tất cả các luật mạnh từ các tập phổ biến được khai phá ở bước trước với ngưỡng tin cậy tối thiểu  $minconf$ .
- 
- Bước một có độ phức tạp tính toán cao hơn và thường chiếm phần lớn thời gian khai phá luật kết hợp.
  - Số lượng các tập mục (itemsets) là rất lớn. Ví dụ với  $\mathbb{I} = \{x_1, x_2, \dots, x_{100}\}$  chúng ta có  $2^{100} - 1 \approx 1.27 \times 10^{30}$  tập con (không tính tập  $\emptyset$ ).

# Các cách tiếp cận, phương pháp và thuật toán khai phá

# Nội dung

## 1 Các khái niệm và định nghĩa

- Tập mục, giao dịch, và cơ sở dữ liệu giao dịch
- Tập phổ biến và luật kết hợp
- Ứng dụng của tập phổ biến và luật kết hợp

## 2 Khai phá tập phổ biến và luật kết hợp

- Các bước trong khai phá luật kết hợp
- **Phương pháp brute-force**
- Phương pháp Apriori
- Phương pháp FP-Growth
- Phương pháp Eclat

## 3 Các chủ đề khác

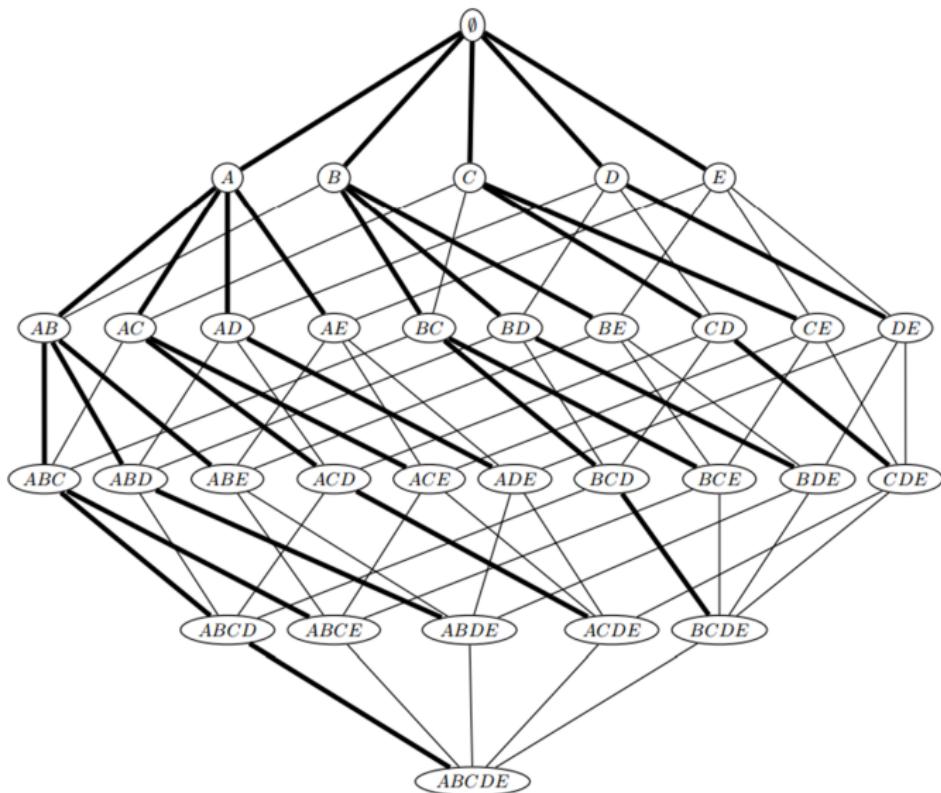
- Các tiêu chí đánh giá luật kết hợp (interestingness measures)
- Luật kết hợp tổng quát (generalized association rules)
- Luật kết hợp định lượng (quantitative association rules)
- Luật kết hợp cực đại (maximal association rules)
- Khai phá mẫu chuỗi (sequential pattern mining)

## 4 Tổng kết bài giảng

## Dàn các tập mục (itemset lattice)

- Cho tập các mục  $\mathbb{I} = \{x_1, x_2, \dots, x_n\}$ , có  $2^{|\mathbb{I}|} = 2^n$  tập mục (bao gồm cả tập rỗng).
- Các tập mục được kết nối với nhau thành một giàn các tập mục (itemset lattice):
  - Tập mục  $X$  và  $Y$  được kết nối với nhau trên giàn nếu và chỉ nếu  $X$  là tập con trực tiếp của  $Y$ , nghĩa là  $X \subseteq Y$  và  $|Y| = |X| + 1$ .
- Các tập mục trên giàn có thể được duyệt theo chiều rộng (breadth-first search – BFS) hoặc chiều sâu (depth-first search – DFS) trên cây tiền tố.
- Với tập các mục  $\mathbb{I} = \{A, B, C, D, E\}$ , chúng ta có giàn bao gồm  $2^5 = 32$  tập mục bao gồm tập rỗng ( $\emptyset$ ) và chính nó ( $ABCDE$ ) ở trang sau.

# Minh họa dàn các tập mục



[Nguồn: Data Mining and Analysis: Fundamental Concepts and Algorithms by Zaki and Jr]

# Tìm các tập phô biến bằng p.pháp vét cạn (brute-force)

```
1: procedure BRUTEFORCE( $\mathbb{D} = \{T_1, T_2, \dots, T_m\}$ ,  $\mathbb{I} = \{x_1, x_2, \dots, x_n\}$ ,  $minsup$ )
2:   Khởi tạo tập các tập phô biến:  $\mathbb{F} \leftarrow \emptyset$ ;
3:   for each  $X \subseteq \mathbb{I}$  do
4:      $sup(X, \mathbb{D}) \leftarrow \text{ComputeSupport}(X, \mathbb{D})$ ;
5:     if  $sup(X, \mathbb{D}) \geq minsup$  then
6:        $\mathbb{F} \leftarrow \mathbb{F} \cup \{X\}$ ;
7:     end if
8:   end for
9:   return  $\mathbb{F}$ ;
10: end procedure
```

```
1: procedure COMPUTESUPPORT( $X, \mathbb{D} = \{T_1, T_2, \dots, T_m\}$ )
2:   Khởi tạo:  $sup(X, \mathbb{D}) \leftarrow 0$ ;
3:   for each  $T \in \mathbb{D}$  do
4:     if  $X \subseteq T$  then
5:        $sup(X, \mathbb{D}) \leftarrow sup(X, \mathbb{D}) + 1$ ;
6:     end if
7:   end for
8:   return  $sup(X, \mathbb{D})$ ;
9: end procedure
```

# Kết quả khai phá các tập phổ biến

$\mathbb{D} = \{T_1, T_2, T_3, T_4, T_5, T_6\}$ :

- $T_1 = \{A, B, D, E\}$
- $T_2 = \{B, C, E\}$
- $T_3 = \{A, B, D, E\}$
- $T_4 = \{A, B, C, E\}$
- $T_5 = \{A, B, C, D, E\}$
- $T_6 = \{B, C, D\}$

Các tập phổ biến khai phá được từ  $\mathbb{D}$  với  $minsup = 3$ :

- $sup = 6: \{B\}$
- $sup = 5: \{E, BE\}$
- $sup = 4: \{A, C, D, AB, AE, BC, BD, ABE\}$
- $sup = 3: \{AD, CE, DE, ABD, ADE, BCE, BDE, ABDE\}$

## Hiệu quả của phương pháp vét cạn

- Thuật toán tính độ hỗ trợ (ComputeSupport) có độ phức tạp tính toán  $O(|\mathbb{I}| \cdot |\mathbb{D}|)$ .
- Vì có  $2^{|\mathbb{I}|}$  tập con của  $\mathbb{I}$  nên thuật toán BruteForce có độ phức tạp tính toán là  $O(|\mathbb{I}| \cdot |\mathbb{D}| \cdot 2^{|\mathbb{I}|})$ .
- Độ phức tạp vào ra (I/O complexity) là  $O(2^{|\mathbb{I}|})$  lần quét cơ sở dữ liệu giao dịch  $\mathbb{D}$ .
- Phải duyệt hết toàn bộ không gian các tập mục (tất cả các nút trên giàn).
- Thời gian tính toán và quét dữ liệu rất lớn.
- Kiểm tra rất nhiều tập mục không tiềm năng là tập phổ biến.

# Nội dung

## 1 Các khái niệm và định nghĩa

- Tập mục, giao dịch, và cơ sở dữ liệu giao dịch
- Tập phổ biến và luật kết hợp
- Ứng dụng của tập phổ biến và luật kết hợp

## 2 Khai phá tập phổ biến và luật kết hợp

- Các bước trong khai phá luật kết hợp
- Phương pháp brute-force
- **Phương pháp Apriori**
- Phương pháp FP-Growth
- Phương pháp Eclat

## 3 Các chủ đề khác

- Các tiêu chí đánh giá luật kết hợp (interestingness measures)
- Luật kết hợp tổng quát (generalized association rules)
- Luật kết hợp định lượng (quantitative association rules)
- Luật kết hợp cực đại (maximal association rules)
- Khai phá mẫu chuỗi (sequential pattern mining)

## 4 Tổng kết bài giảng

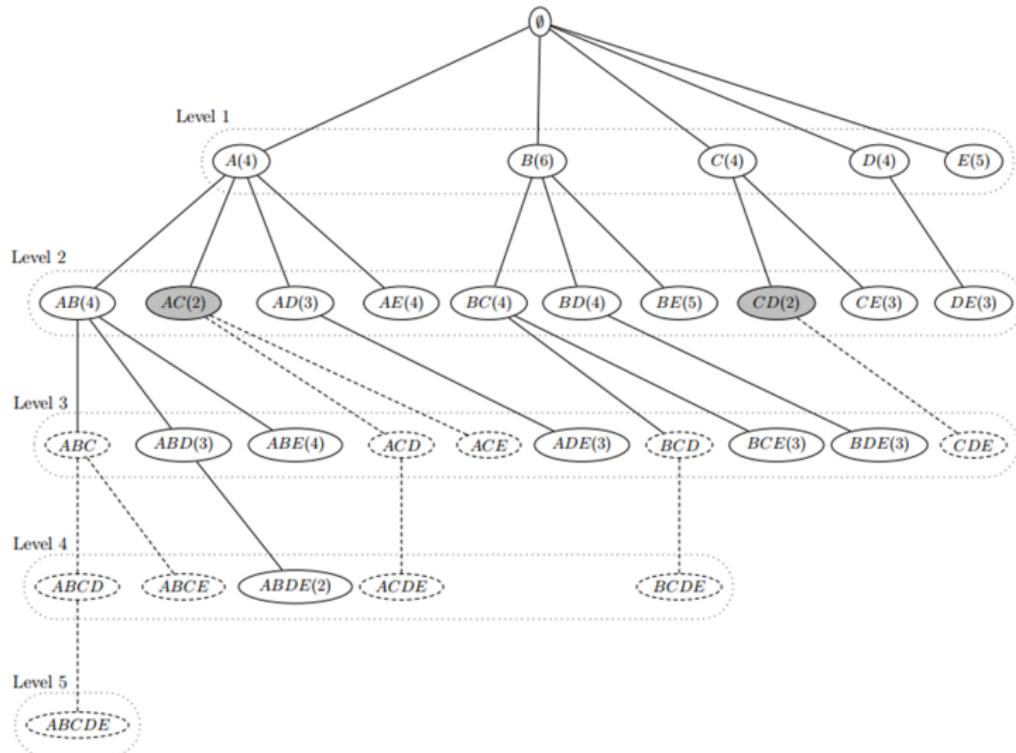
# Một số tính chất sử dụng trong phương pháp Apriori

- Cho hai tập mục  $X, Y \subseteq \mathbb{I}$  và cơ sở dữ liệu  $\mathbb{D}$ .
- Nếu  $X \subseteq Y$  thì  $\text{sup}(X, \mathbb{D}) \geq \text{sup}(Y, \mathbb{D})$ .

## Hai tính chất Apriori:

- Nếu  $Y$  là tập phổ biến (frequent) thì mọi tập con  $X$  ( $\subseteq Y$ ) của  $Y$  đều phổ biến.
- Nếu  $X$  là tập không phổ biến (infrequent) thì mọi tập cha  $Y$  ( $\supseteq X$ ) của  $X$  đều không phổ biến.
- Phương pháp Apriori dựa vào hai tính chất trên để cải tiến phương pháp vét cạn bằng cách cắt tỉa các nhánh không cần thiết trên giàn tập mục.
- Cụ thể, khi duyệt theo bề rộng (BFS) trên giàn tập mục, thuật toán Apriori cắt tỉa hết tất cả các tập cha của tập không phổ biến.

# Cắt tỉa trên giàn tập mục trong Apriori ( $minsup = 3$ )



[Nguồn: Data Mining and Analysis: Fundamental Concepts and Algorithms by Zaki and Jr]

## Cắt tỉa trên giàn tập mục trong Apriori ( $minsup = 3$ ) - tiếp

- Ở hình trước, các nút màu sậm là các tập mục không phổ biến.
- Tất cả các tập cha của chúng trên giàn (các nút vạch đứt) đều bị cắt tỉa, dẫn đến toàn bộ các nhánh vạch đứt được cắt tỉa.
- Ví dụ: tập  $AC$  có  $sup(AC, \mathbb{D}) = 2 < minsup$  nên các tập cha của  $AC$  có tiền tố là  $AC$  sẽ bị cắt tỉa, dẫn đến toàn bộ cây con dưới nút  $AC$  bị cắt tỉa.

Tập tất cả các mục  $\mathbb{I}$ :

$$\mathbb{I} = \{A, B, C, D, E\}$$

Cơ sở dữ liệu giao dịch  $\mathbb{D}$ :

$$\mathbb{D} = \{T_1, T_2, T_3, T_4, T_5, T_6\}, \text{ cụ thể:}$$

- $T_1 = \{A, B, D, E\}$
- $T_2 = \{B, C, E\}$
- $T_3 = \{A, B, D, E\}$
- $T_4 = \{A, B, C, E\}$
- $T_5 = \{A, B, C, D, E\}$
- $T_6 = \{B, C, D\}$

- Với  $minsup = 3$ .

# Minh họa thuật toán Apriori

Tập các tập mục ứng viên  $C^{(1)}$

1-itemset	support
{A}	4
{B}	6
{C}	4
{D}	4
{E}	5

Quét CSDL tính  
độ hỗ trợ cho các  
tập mục ứng viên

Tập các tập mục phổ biến  $F^{(1)}$

1-itemset	support
{A}	4
{B}	6
{C}	4
{D}	4
{E}	5

Kiểm tra độ hỗ trợ  
của các tập ứng viên  
với ngưỡng minsup

Tập các tập mục ứng viên  $C^{(2)}$

2-itemset
{AB}
{AC}
{AD}
{AE}
{BC}
{BD}
{BE}
{CD}
{CE}
{DE}

Sinh tập các  
tập mục ứng  
viên  $C^{(2)}$  từ  $F^{(1)}$

2-itemset	support
{AB}	4
{AC}	2
{AD}	3
{AE}	4
{BC}	4
{BD}	4
{BE}	5
{CD}	2
{CE}	3
{DE}	3

Quét CSDL tính  
độ hỗ trợ cho các  
tập mục ứng viên

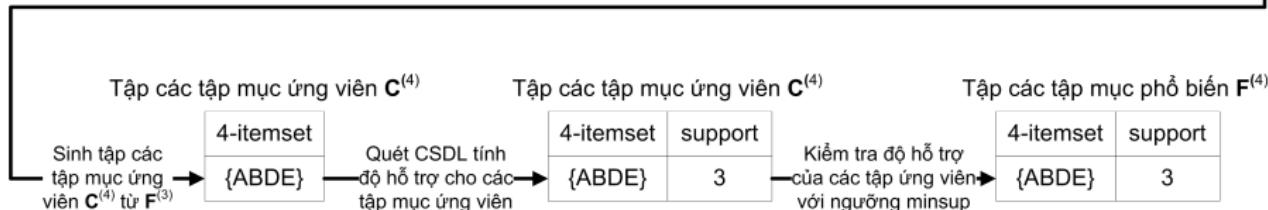
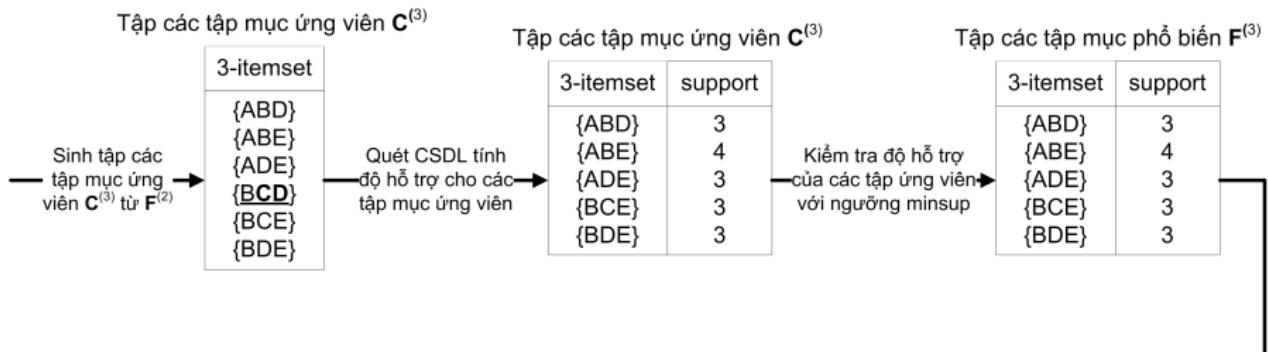
Tập các tập mục phổ biến  $F^{(2)}$

2-itemset	support
{AB}	4
{AD}	3
{AE}	4
{BC}	4
{BD}	4
{BE}	5
{CE}	3
{DE}	3

Kiểm tra độ hỗ trợ  
của các tập ứng viên  
với ngưỡng minsup

Sinh tập các  
tập mục ứng  
viên  $C^{(3)}$  từ  $F^{(2)}$

## Minh họa thuật toán Apriori (2)



# Thuật toán Apriori

```
1: procedure APRIORI( $\mathbb{D} = \{T_1, T_2, \dots, T_m\}$ ,  $\mathbb{I} = \{x_1, x_2, \dots, x_n\}$ ,  $minsup$ )
2:   Khởi tạo tập các tập phẩy biến:  $\mathbb{F} \leftarrow \emptyset$ ;
3:    $\mathbb{F}^{(1)} \leftarrow \text{FindFrequent1Itemsets}(\mathbb{D}, \mathbb{I}, minsup)$ ;
4:   for ( $k = 2$ ;  $\mathbb{F}^{(k-1)} \neq \emptyset$ ;  $k++$ ) do
5:      $\mathbb{C}^{(k)} \leftarrow \text{AprioriGen}(\mathbb{F}^{(k-1)})$ ;
6:     for (each transaction  $T \in \mathbb{D}$ ) do
7:        $\mathbb{C}_T \leftarrow \text{SubsetsOfT}(\mathbb{C}^{(k)}, T)$ ;
8:       for (each  $C \in \mathbb{C}_T$ ) do
9:          $C.count++$ ;
10:        end for
11:      end for
12:       $\mathbb{F}^{(k)} \leftarrow \{C \in \mathbb{C}^{(k)} | C.count \geq minsup\}$ ;
13:    end for
14:     $\mathbb{F} \leftarrow \mathbb{F}^{(1)} \cup \mathbb{F}^{(2)} \cup \dots \cup \mathbb{F}^{(k)}$ ;
15:    return  $\mathbb{F}$ ;
16: end procedure
```

## Thuật toán Apriori (2)

```
1: procedure APRIORI $\text{GEN}(\mathbb{F}^{(k-1)})$ 
2:   Khởi tạo tập các tập mục ứng viên:  $\mathbb{C}^{(k)} \leftarrow \emptyset$ ;
3:   for (each itemset  $F_1 \in \mathbb{F}^{(k-1)}$ ) do
4:     for (each itemset  $F_2 \in \mathbb{F}^{(k-1)}$ ) do
5:       if  $((F_1[1] = F_2[1]) \wedge \dots \wedge (F_1[k-2] = F_2[k-2]) \wedge (F_1[k-1] < F_2[k-1]))$  then
6:          $C \leftarrow F_1 \bowtie F_2$ ;
7:         if (HasInfrequentSubset( $C$ ,  $\mathbb{F}^{(k-1)}$ )) then
8:           remove  $C$ ;
9:         else
10:           $\mathbb{C}^{(k)} \leftarrow \mathbb{C}^{(k)} \cup \{C\}$ ;
11:        end if
12:      end if
13:    end for
14:  end for
15:  return  $\mathbb{C}^{(k)}$ ;
16: end procedure
```

## Thuật toán Apriori (3)

```
1: procedure HASINFREQUENTSUBSET( $C, \mathbb{F}^{(k-1)}$ )
2:   for (each  $(k - 1)$ -subset  $S$  of  $C$ ) do
3:     if ( $S \notin \mathbb{F}^{(k-1)}$ ) then
4:       return TRUE;
5:     end if
6:   end for
7:   return FALSE;
8: end procedure
```

# Sinh luật kết hợp phổ biến và mạnh từ các tập phổ biến

- **Input:** Tập tất cả các tập phổ biến  $\mathbb{F}$ .
- **Output:** Tập tất cả các luật phổ biến (frequent) và mạnh (strong):  $\mathbb{R}$ .

```
1: procedure GENFREQUENTSTRONGRULES( $\mathbb{F}$ ,  $minconf$ )
2:   Khởi tạo  $\mathbb{R} \leftarrow \emptyset$ ;
3:   for (với mỗi tập mục phổ biến  $F \in \mathbb{F}$  và  $|F| \geq 2$ ) do
4:      $\mathbb{X} \leftarrow \{X | X \subset F, X \neq \emptyset\}$ ;
5:     while ( $\mathbb{X} \neq \emptyset$ ) do
6:        $Y \leftarrow$  maximal element in  $\mathbb{X}$ ;
7:        $\mathbb{X} \leftarrow \mathbb{X} \setminus Y$ ;
8:       if ( $conf(Y \rightarrow F \setminus Y) \geq minconf$ ) then
9:          $\mathbb{R} \leftarrow \mathbb{R} \cup \{Y \rightarrow F \setminus Y\}$ ;
10:      else
11:         $\mathbb{X} \leftarrow \mathbb{X} \setminus \{Z | Z \subset Y\}$ 
12:      end if
13:    end while
14:  end for
15:  return  $\mathbb{R}$ ;
16: end procedure
```

## Minh họa thuật toán sinh luật

Sinh luật cho tập phỏng biến  $ABDE$  có độ hỗ trợ bằng 3 với độ tin cậy tối thiểu  $minconf = 0.8$ :

- $\mathbb{X} = \{ABD(3), ABE(4), ADE(3), BDE(3), AB(4), AD(4), AE(4), BD(4), BE(5), DE(3), A(4), B(6), D(4), E(5)\}$ .
- $Y = ABD$ :  $conf(ABD \rightarrow E) = \frac{3}{3} = 1.0 \geq 0.8$  nên  $ABD \rightarrow E$  là luật mạnh.
- $Y = ABE$ :  $conf(ABE \rightarrow D) = \frac{3}{4} = 0.75 < 0.8$  nên  $ABE \rightarrow D$  không tin cậy. Khi đó có thể loại bỏ khỏi  $\mathbb{X}$  tất cả các tập con của  $ABE$ . Do đó,  $\mathbb{X} = \{ADE(3), BDE(3), AD(4), BD(4), DE(3), D(4)\}$ .
- $Y = ADE$ :  $conf(ADE \rightarrow B) = \frac{3}{3} = 1.0 \geq 0.8$  nên  $ADE \rightarrow B$  là luật mạnh.
- $Y = BDE$ :  $conf(BDE \rightarrow A) = \frac{3}{3} = 1.0 \geq 0.8$  nên  $BDE \rightarrow A$  là luật mạnh.
- $Y = AD$ :  $conf(AD \rightarrow BE) = \frac{3}{4} = 0.75 < 0.8$  nên  $AD \rightarrow BE$  không tin cậy. Khi đó có thể loại bỏ khỏi  $\mathbb{X}$  tất cả các tập con của  $AD$ . Do đó,  $\mathbb{X} = \{BD(4), DE(3)\}$ .
- $Y = BD$ :  $conf(BD \rightarrow AE) = \frac{3}{4} = 0.75 < 0.8$  nên  $BD \rightarrow AE$  không tin cậy. Khi đó có thể loại bỏ khỏi  $\mathbb{X}$  tất cả các tập con của  $BD$ . Do đó,  $\mathbb{X} = \{DE(3)\}$ .
- $Y = DE$ :  $conf(DE \rightarrow AB) = \frac{3}{3} = 1.0 \geq 0.8$  nên  $DE \rightarrow AB$  là luật mạnh.

# Tối ưu cài đặt cho phương pháp Apriori

# Ưu và nhược điểm của phương pháp Apriori

- **Ưu điểm:**

- ▶ Nhờ các tính chất Apriori để cắt tỉa được nhiều nhánh trên giàn (lattice), giảm bớt đáng kể việc sinh các tập mục ứng viên và kiểm tra tính phổ biến của các tập ứng viên đó.

- **Nhược điểm:**

- ▶ Vẫn cần sinh ra một lượng lớn các tập ứng viên. Ví dụ, nếu có  $10^4$  tập mục phổ biến gồm một mục (1-itemsets), thuật toán Apriori cần sinh ra hơn  $10^7$  tập mục ứng viên có hai mục (2-itemsets).
  - ▶ Cần quét cơ sở dữ liệu nhiều lần để đếm độ hỗ trợ của các tập ứng viên trong quá trình thực hiện thuật toán.

# Nội dung

## 1 Các khái niệm và định nghĩa

- Tập mục, giao dịch, và cơ sở dữ liệu giao dịch
- Tập phổ biến và luật kết hợp
- Ứng dụng của tập phổ biến và luật kết hợp

## 2 Khai phá tập phổ biến và luật kết hợp

- Các bước trong khai phá luật kết hợp
- Phương pháp brute-force
- Phương pháp Apriori
- Phương pháp FP-Growth**
- Phương pháp Eclat

## 3 Các chủ đề khác

- Các tiêu chí đánh giá luật kết hợp (interestingness measures)
- Luật kết hợp tổng quát (generalized association rules)
- Luật kết hợp định lượng (quantitative association rules)
- Luật kết hợp cực đại (maximal association rules)
- Khai phá mẫu chuỗi (sequential pattern mining)

## 4 Tổng kết bài giảng

# Phương pháp FP–Growth

- Cấu trúc dữ liệu FP–Tree (Frequent Pattern Tree)
- Sinh cây FP–Tree từ cơ sở dữ liệu
- Sinh tập phổ biến từ FP-Tree
- Ưu và nhược điểm của phương pháp FP–Growth

# Cấu trúc dữ liệu FP-Tree

- Mỗi nốt trên cây được gắn nhãn là một mục (item).
- Các nốt con của một nốt đại diện cho các mục khác nhau.
- Mỗi nốt cũng lưu thông tin về độ hỗ trợ (support) của tập mục (itemset) bao gồm tất cả các mục trên đường đi từ nốt gốc đến nó.
- Có một bảng lưu tất cả các mục và con trỏ (node-link) để liên kết tất cả các vị trí xuất hiện của mỗi mục trong cây.

# Thuật toán sinh cây FP-Tree $\mathbb{T}$ từ CSDL giao dịch $\mathbb{D}$

```
1: procedure BUILDFPTREE( $\mathbb{D} = \{T_1, T_2, \dots, T_m\}$ )
2:   Khởi tạo cây FP-Tree  $\mathbb{T}$  chỉ chứa nốt gốc  $\emptyset$  và  $\emptyset.support \leftarrow 0$ ;
3:   for (với mỗi giao dịch  $T \in \mathbb{D}$ ) do
4:      $T' = \{x^1, \dots, x^h\} \leftarrow$  sắp xếp các mục phổ biến  $\in T$  giảm dần theo support;
5:      $pNode \leftarrow \emptyset$ ;
6:     for ( $i = 1; i \leq h; i++$ ) do
7:       if ( $cNode \in \text{Children}(pNode)$  and  $cNode.label = x^i$ ) then
8:          $cNode.support++$ ;
9:          $pNode \leftarrow cNode$ ;
10:      else
11:        Tạo nốt  $cNode$  là một nốt con mới của  $pNode$ ;
12:         $cNode.label \leftarrow x^i$ ;
13:         $cNode.support \leftarrow 1$ ;
14:         $pNode \leftarrow cNode$ ;
15:      end if
16:    end for
17:     $\emptyset.support++$ ;
18:  end for
19:  return cây FP-Tree  $\mathbb{T}$ ;
20: end procedure
```

# CSDL giao dịch $\mathbb{D}$ minh họa phương pháp FP–Growth

Tập tất cả các mục  $\mathbb{I}$ :

$$\mathbb{I} = \{A, B, C, D, E\}$$

Cơ sở dữ liệu giao dịch  $\mathbb{D}$ :

$$\mathbb{D} = \{T_1, T_2, T_3, T_4, T_5, T_6\}, \text{ cụ thể:}$$

- $T_1 = \{A, B, D, E\}$
- $T_2 = \{B, C, E\}$
- $T_3 = \{A, B, D, E\}$
- $T_4 = \{A, B, C, E\}$
- $T_5 = \{A, B, C, D, E\}$
- $T_6 = \{B, C, D\}$

- Với  $minsup = 3$ .

# Sắp xếp lại các mục (items) để xây dựng cây FP–Tree

Tập tất cả các mục  $\mathbb{I}$ :

$$\mathbb{I} = \{B(6), E(5), A(4), C(4), D(4)\}$$

Cơ sở dữ liệu giao dịch  $\mathbb{D}$ :

$$\mathbb{D} = \{T_1, T_2, T_3, T_4, T_5, T_6\}, \text{ cụ thể:}$$

- $T_1 = \{B, E, A, D\}$
- $T_2 = \{B, E, C\}$
- $T_3 = \{B, E, A, D\}$
- $T_4 = \{B, E, A, C\}$
- $T_5 = \{B, E, A, C, D\}$
- $T_6 = \{B, C, D\}$

# Minh họa thuật toán sinh cây FP–Tree $\mathbb{T}$ từ CSDL $\mathbb{D}$

$\emptyset(1)$

$B(1)$

$E(1)$

$A(1)$

$D(1)$

(a)  $\langle 1, BEAD \rangle$

$\emptyset(2)$

$B(2)$

$E(2)$

$A(1)$

$C(1)$

$D(1)$

(b)  $\langle 2, BEC \rangle$

$\emptyset(3)$

$B(3)$

$E(3)$

$A(2)$

$C(1)$

$D(2)$

(c)  $\langle 3, BEAD \rangle$

$\emptyset(4)$

$B(4)$

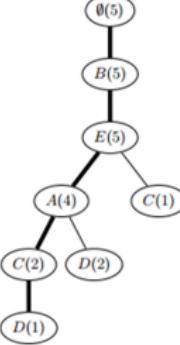
$E(4)$

$A(3)$

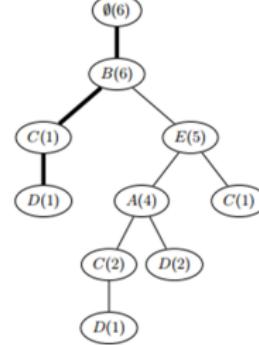
$C(1)$

$D(2)$

(d)  $\langle 4, BEAC' \rangle$



(e)  $\langle 5, BEACD \rangle$



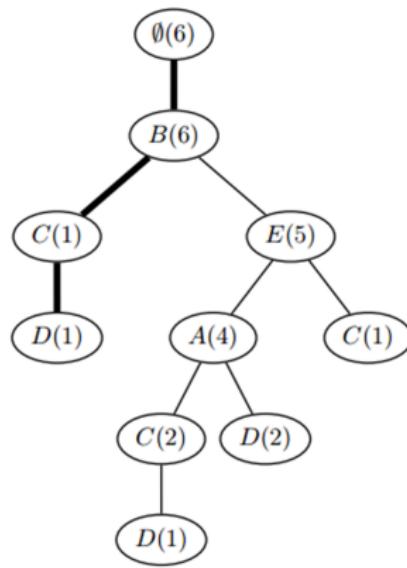
(f)  $\langle 6, BCD \rangle$

[Nguồn: Data Mining and Analysis: Fundamental Concepts and Algorithms by Zaki and Jr]

## Một vài đặc điểm của cây FP-Tree

- Chỉ cần quét toàn bộ cơ sở dữ liệu  $\mathbb{D}$  2 lần để xây dựng cây FP-Tree  $\mathbb{T}$ .
- Cây FP-Tree là một dạng biểu diễn cô đọng (compressed) của cơ sở dữ liệu giao dịch  $\mathbb{D}$ .
- Cây FP-Tree càng nhỏ gọn càng tốt.
- Các mục (items) càng phổ biến (có độ hỗ trợ cao) càng nằm phía gần gốc cây.
- Tất cả các tập phổ biến (frequent itemsets) có thể được khai phá trực tiếp từ cây FP-Tree  $\mathbb{T}$  thay vì từ CSDL  $\mathbb{D}$ .

# Cây FP-Tree $\mathbb{T}$ được xây dựng từ CSDL $\mathbb{D}$



[Nguồn: Data Mining and Analysis: Fundamental Concepts and Algorithms by Zaki and Jr]

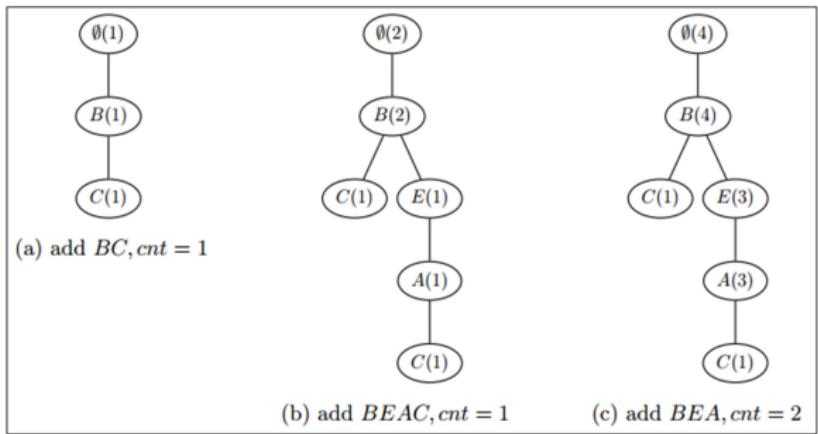
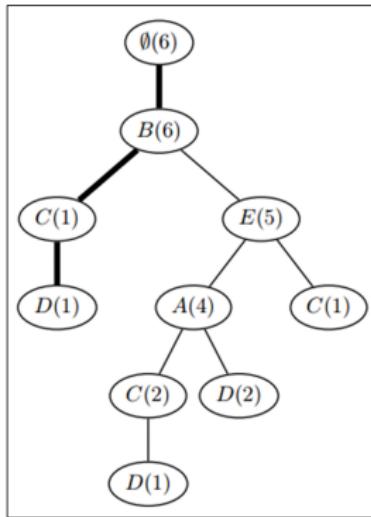
# Thuật toán đệ quy sinh các tập phô biến từ cây FP-Tree $\mathbb{T}$

```
1: procedure FPGROWTH( $\mathbb{T}$ ,  $P$ ,  $\mathbb{F}$ ,  $minsup$ )
2:   Loại bỏ các mục không phô biến (infrequent items) trong  $\mathbb{T}$ ;
3:   if (IsPath( $\mathbb{T}$ )) then
4:     for (với mỗi tập con  $Y \subseteq \mathbb{T}$ ) do
5:        $X \leftarrow P \cup Y$ ;
6:        $X.support \leftarrow \min_{x \in Y} \{cnt(x)\}$ ;
7:        $\mathbb{F} \leftarrow \mathbb{F} \cup \{X\}$ ;
8:     end for
9:   else
10:    for (mỗi mục  $y \in \mathbb{T}$  với thứ tự đã sắp xếp tăng dần theo  $sup(y)$ ) do
11:       $X \leftarrow P \cup \{y\}$ ;
12:       $X.support \leftarrow sup(y)$ ;     $\triangleright sup(y)$  là tổng  $cnt(y)$  tại mọi nốt có nhãn  $y$  trong  $\mathbb{T}$ 
13:       $\mathbb{F} \leftarrow \mathbb{F} \cup \{X\}$ ;
14:      Khởi tạo FP-Tree  $\mathbb{T}_X \leftarrow \emptyset$ ;
15:      for (với mỗi đường đi  $path$  từ gốc xuống nốt có nhãn  $y$  trong cây  $\mathbb{T}$ ) do
16:         $cnt(y) \leftarrow$  đếm tần suất của  $y$  trong  $path$ ;
17:        Chèn  $path$  (ngoại trừ nốt  $y$ ) vào cây FP-Tree  $\mathbb{T}_X$  với  $cnt(y)$ ;
18:      end for
19:      if ( $\mathbb{T}_X \neq \emptyset$ ) then
20:        FP_Growth( $\mathbb{T}_X$ ,  $X$ ,  $\mathbb{F}$ ,  $minsup$ );
21:      end if
22:    end for
23:  end if
24: end procedure
```

## Sinh tập phổ biến từ FP-Tree: một số khái niệm

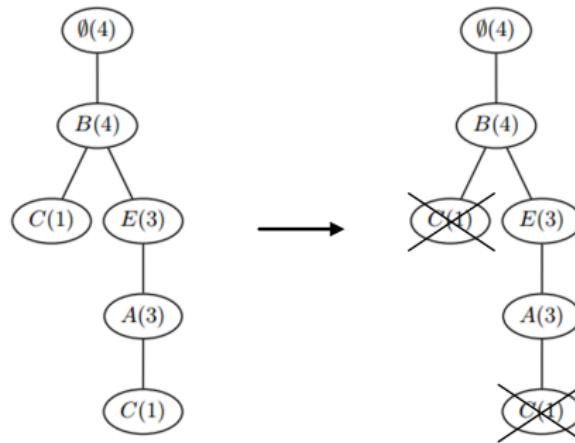
- Lời gọi hàm đầu tiên  $\text{FPGrowth}(\mathbb{T}, P \leftarrow \emptyset, \mathbb{F} \leftarrow \emptyset, \text{minsup})$ .
- Phép chiếu chọn (projection) cây FP–Tree  $\mathbb{T}$  theo một mục (item) nào đó.
- Cây FP–Tree  $\mathbb{T}$  có thể là một đường tuyến tính (*path*).
- Loại bỏ các mục không phổ biến (infrequent items) trong một cây FP–Tree.

# Cây FP-Tree chiề́u chọn (projected) theo mục (item) $D$



[Nguồn: Data Mining and Analysis: Fundamental Concepts and Algorithms by Zaki and Jr]

# Loại bỏ các mục không phổ biến (infrequent items) trong FP–Tree



- Bên trái: cây FP–Tree  $\mathbb{T}_D$  chiểu theo mục  $D$  từ cây FP–Tree  $\mathbb{T}$ .
- Bên phải: Cây FP–Tree  $\mathbb{T}_D$  sau khi đã loại bỏ mục  $C$  không phổ biến do  $cnt(C) = 1 + 1 = 2 < minsup = 3$ .

# Minh họa thuật toán FP–Growth

- Với lời gọi đầu tiên:  $\text{FPGrowth}(\mathbb{T}, P \leftarrow \emptyset, \mathbb{F} \leftarrow \emptyset, \text{minsup} = 3)$ .
  - ▶ Không xóa bỏ được mục không phổ biến nào (tất cả đều phổ biến).
  - ▶  $\mathbb{T}$  không phải dạng đường tuyến tính *path*.
  - ▶ Tiền tố (prefix)  $P = \emptyset$ .
  - ▶  $y$  sẽ lần lượt nhận  $D(4), C(4), A(4), E(5), B(6)$ .
  - ▶ Trước tiên  $y$  nhận  $D$ :
    - ★  $X \leftarrow P \cup \{y\} = \emptyset \cup \{D\} = \{D\}$ .
    - ★  $\mathbb{F} \leftarrow \mathbb{F} \cup \{X\} = \emptyset \cup \{\{D(4)\}\} = \{\{D(4)\}\}$ .
    - ★ Có 3 đường đi tuyến tính (*path*) từ gốc của  $\mathbb{T}$  đến nốt  $D$ :  $BCD$ ,  $\text{cnt}(D) = 1$ ;  $BEACD$ ,  $\text{cnt}(D) = 1$ ; và  $BEAD$ ,  $\text{cnt}(D) = 2$ .
    - ★ Tạo cây FP–Tree  $\mathbb{T}_{\{D\}}$  từ 3 paths nói trên.
    - ★ Gọi đệ quy hàm  $\text{FPGrowth}(\mathbb{T}_{\{D\}}, \{D\}, \{\{D(4)\}\}, \text{minsup} = 3)$ .
  - ▶  $y$  nhận  $C$ :
    - ★ ...
  - ▶  $y$  nhận  $A$ :
    - ★ ...
  - ▶  $y$  nhận  $E$ :
    - ★ ...
  - ▶  $y$  nhận  $B$ :
    - ★ ...

## Minh họa thuật toán FP–Growth (2)

- Với lời gọi  $\text{FPGrowth}(\mathbb{T}_{\{D\}}, P = \{D\}, \mathbb{F} = \{\{D(4)\}\}, \text{minsup} = 3)$ :
  - ▶ Loại bỏ tất cả nốt  $C$  ra khỏi  $\mathbb{T}_{\{D\}}$  vì  $\text{cnt}(C) = 1 + 1 = 2 < \text{minsup} = 3$ .
  - ▶ Cây FP–Tree  $\mathbb{T}_{\{D\}}$  bây giờ trở thành một đường truyền tính (*path*):  $B(4) - E(3) - A(3)$ :
    - ★ Liệt kê tất cả các tập con của đường truyền tính:  
 $B, E, A, BE, BA, EA, BEA$ .
    - ★ Ghép với tiền tố  $P = \{D\}$  tạo thành các tập phổ biến  $DB(4), DE(3), DA(3), DBE(3), DBA(3), DEA(3), DBEA(3)$ .
    - ★ Thêm các tập phổ biến vào trong  $\mathbb{F}$  ta được  $\mathbb{F} = \{D(4), DB(4), DE(3), DA(3), DBE(3), DBA(3), DEA(3), DBEA(3)\}$ .
    - ★ Lời gọi  $\text{FPGrowth}(\mathbb{T}_{\{D\}}, P = \{D\}, \mathbb{F} = \{\{D(4)\}\}, \text{minsup} = 3)$  kết thúc.

# Minh họa thuật toán FP–Growth (3)

- Khi y nhận các mục khác:

- ▶ y nhận C:

- ★  $\mathbb{F} = \{D(4), DB(4), DE(3), DA(3), DBE(3), DBA(3), DEA(3), DBEA(3)\} \cup \{C(4), CB(4), CE(3), CBE(3)\}.$

- ▶ y nhận A:

- ★  $\mathbb{F} = \{D(4), DB(4), DE(3), DA(3), DBE(3), DBA(3), DEA(3), DBEA(3)\} \cup \{C(4), CB(4), CE(3), CBE(3)\} \cup \{A(4), AE(4), AB(4), AEB(4)\}.$

- ▶ y nhận E:

- ★  $\mathbb{F} = \{D(4), DB(4), DE(3), DA(3), DBE(3), DBA(3), DEA(3), DBEA(3)\} \cup \{C(4), CB(4), CE(3), CBE(3)\} \cup \{A(4), AE(4), AB(4), AEB(4)\} \cup \{E(5), EB(5)\}.$

- ▶ y nhận B:

- ★  $\mathbb{F} = \{D(4), DB(4), DE(3), DA(3), DBE(3), DBA(3), DEA(3), DBEA(3)\} \cup \{C(4), CB(4), CE(3), CBE(3)\} \cup \{A(4), AE(4), AB(4), AEB(4)\} \cup \{E(5), EB(5)\} \cup \{B(6)\}.$

## Minh họa thuật toán FP–Growth (4)

- Vậy  $\mathbb{F}$  bao gồm các tập phổ biến với các mức hỗ trợ khác nhau:
  - ▶ Support = 6:  $B$
  - ▶ Support = 5:  $E, BE$
  - ▶ Support = 4:  $D, C, A, DB, CB, AE, AB, ABE$
  - ▶ Support = 3:  $DE, DA, CE, DBE, DBA, DAE, CBE, DBEA$

# Ưu và nhược điểm của phương pháp FP–Growth

- **Ưu điểm:**

- ▶ Nén được cơ sở dữ liệu trong một cấu trúc cây gọn nhẹ FP–Tree.
- ▶ Chỉ cần quét cơ sở dữ liệu 2 lần.
- ▶ Hiệu quả kể cả khi ngưỡng *minsup* bé.

- **Nhược điểm:**

- ▶ Thuật toán cài đặt phức tạp hơn so với Apriori.
- ▶ Khi cơ sở dữ liệu lớn: FP–Tree lớn và khó lưu vừa trong bộ nhớ.
- ▶ Sử dụng đệ quy (có thể khử đệ quy).

# Nội dung

## 1 Các khái niệm và định nghĩa

- Tập mục, giao dịch, và cơ sở dữ liệu giao dịch
- Tập phổ biến và luật kết hợp
- Ứng dụng của tập phổ biến và luật kết hợp

## 2 Khai phá tập phổ biến và luật kết hợp

- Các bước trong khai phá luật kết hợp
- Phương pháp brute-force
- Phương pháp Apriori
- Phương pháp FP-Growth
- **Phương pháp Eclat**

## 3 Các chủ đề khác

- Các tiêu chí đánh giá luật kết hợp (interestingness measures)
- Luật kết hợp tổng quát (generalized association rules)
- Luật kết hợp định lượng (quantitative association rules)
- Luật kết hợp cực đại (maximal association rules)
- Khai phá mẫu chuỗi (sequential pattern mining)

## 4 Tổng kết bài giảng

# Nội dung

## 1 Các khái niệm và định nghĩa

- Tập mục, giao dịch, và cơ sở dữ liệu giao dịch
- Tập phổ biến và luật kết hợp
- Ứng dụng của tập phổ biến và luật kết hợp

## 2 Khai phá tập phổ biến và luật kết hợp

- Các bước trong khai phá luật kết hợp
- Phương pháp brute-force
- Phương pháp Apriori
- Phương pháp FP-Growth
- Phương pháp Eclat

## 3 Các chủ đề khác

- Các tiêu chí đánh giá luật kết hợp (interestingness measures)
- Luật kết hợp tổng quát (generalized association rules)
- Luật kết hợp định lượng (quantitative association rules)
- Luật kết hợp cực đại (maximal association rules)
- Khai phá mẫu chuỗi (sequential pattern mining)

## 4 Tổng kết bài giảng

# Nội dung

## 1 Các khái niệm và định nghĩa

- Tập mục, giao dịch, và cơ sở dữ liệu giao dịch
- Tập phổ biến và luật kết hợp
- Ứng dụng của tập phổ biến và luật kết hợp

## 2 Khai phá tập phổ biến và luật kết hợp

- Các bước trong khai phá luật kết hợp
- Phương pháp brute-force
- Phương pháp Apriori
- Phương pháp FP-Growth
- Phương pháp Eclat

## 3 Các chủ đề khác

- Các tiêu chí đánh giá luật kết hợp (interestingness measures)
- Luật kết hợp tổng quát (generalized association rules)
- Luật kết hợp định lượng (quantitative association rules)
- Luật kết hợp cực đại (maximal association rules)
- Khai phá mẫu chuỗi (sequential pattern mining)

## 4 Tổng kết bài giảng

## Các tiêu chí đánh giá luật kết hợp (interestingness measures)

- Nhiều luật tuy phổ biến và tin cậy (mạnh) nhưng không có nhiều ý nghĩa hay hữu ích với người dùng.
- Khai phá luật kết hợp chỉ dựa trên độ hỗ trợ và độ tin cậy có thể cứng nhắc và chưa có tính sàng lọc cao.
- Một số tiêu chí/độ đo khác có thể bổ sung những khiếm khuyết của *support* và *confidence*.
- Phần này:
  - ▶ Giải thích tại sao khai phá luật dựa trên *support* và *confidence* là chưa đủ.
  - ▶ Giới thiệu một số tiêu chí/độ đo dựa trên phân tích tương quan (correlation analysis).
  - ▶ Dánh giá và so sánh giữa các độ đo/tiêu chí.

## Luật mạnh (strong) chưa chắc đã thú vị (interesting)

- Xét một cơ sở dữ liệu giao dịch về đồ điện tử *AllElectronics* trong đó có hai mặt hàng *games* và *videos*.
- Giả sử trong 10000 giao dịch được phân tích có:
  - ▶ 6000 giao dịch mua *games*,
  - ▶ 7500 giao dịch mua *videos*, và
  - ▶ 4000 giao dịch mua cả *games* lẫn *videos*.
- Giả sử độ hỗ trợ tối thiểu  $minsup = 30\%$  và độ tin cậy tối thiểu  $minconf = 60\%$ .
- Luật *games* → *videos* có  $sup = 40\%$  và  $conf = 66.67\%$  là luật phổ biến và mạnh (strong).
- Tuy nhiên, luật này không phản ánh đúng bản chất vì xác suất mua của *videos* trong CSDL là 75%, cao hơn cả độ tin cậy của luật này.

## Độ đo lift của luật

- Xét luật kết hợp  $A \rightarrow B$ .
- Độ đo  $lift(A \rightarrow B)$  được xác định như sau:

$$lift(A \rightarrow B) = \frac{con(A \rightarrow B)}{sup(B)} = \frac{sup(A \cup B)}{sup(A)sup(B)} \quad (9)$$

- Theo cách nhìn xác suất:

$$lift(A \rightarrow B) = \frac{P(A \cup B)}{P(A)P(B)} \quad (10)$$

- Nếu  $lift(A \rightarrow B) = 1$ :  $A$  và  $B$  độc lập, không nên có sự kết hợp giữa  $A$  và  $B$ .
- Nếu  $lift(A \rightarrow B) > 1$ : luật  $A \rightarrow B$  có ý nghĩa.
- Nếu  $lift(A \rightarrow B) < 1$ : luật  $A \rightarrow B$  và kề cả luật  $B \rightarrow A$  không có ý nghĩa (trong khuôn khổ tập dữ liệu  $\mathbb{D}$ ).

# So sánh giữa các phép đo/tiêu chí

# Nội dung

## 1 Các khái niệm và định nghĩa

- Tập mục, giao dịch, và cơ sở dữ liệu giao dịch
- Tập phổ biến và luật kết hợp
- Ứng dụng của tập phổ biến và luật kết hợp

## 2 Khai phá tập phổ biến và luật kết hợp

- Các bước trong khai phá luật kết hợp
- Phương pháp brute-force
- Phương pháp Apriori
- Phương pháp FP-Growth
- Phương pháp Eclat

## 3 Các chủ đề khác

- Các tiêu chí đánh giá luật kết hợp (interestingness measures)
- **Luật kết hợp tổng quát (generalized association rules)**
- Luật kết hợp định lượng (quantitative association rules)
- Luật kết hợp cực đại (maximal association rules)
- Khai phá mẫu chuỗi (sequential pattern mining)

## 4 Tổng kết bài giảng

# Nội dung

## 1 Các khái niệm và định nghĩa

- Tập mục, giao dịch, và cơ sở dữ liệu giao dịch
- Tập phổ biến và luật kết hợp
- Ứng dụng của tập phổ biến và luật kết hợp

## 2 Khai phá tập phổ biến và luật kết hợp

- Các bước trong khai phá luật kết hợp
- Phương pháp brute-force
- Phương pháp Apriori
- Phương pháp FP-Growth
- Phương pháp Eclat

## 3 Các chủ đề khác

- Các tiêu chí đánh giá luật kết hợp (interestingness measures)
- Luật kết hợp tổng quát (generalized association rules)
- **Luật kết hợp định lượng (quantitative association rules)**
- Luật kết hợp cực đại (maximal association rules)
- Khai phá mẫu chuỗi (sequential pattern mining)

## 4 Tổng kết bài giảng

# Nội dung

## 1 Các khái niệm và định nghĩa

- Tập mục, giao dịch, và cơ sở dữ liệu giao dịch
- Tập phổ biến và luật kết hợp
- Ứng dụng của tập phổ biến và luật kết hợp

## 2 Khai phá tập phổ biến và luật kết hợp

- Các bước trong khai phá luật kết hợp
- Phương pháp brute-force
- Phương pháp Apriori
- Phương pháp FP-Growth
- Phương pháp Eclat

## 3 Các chủ đề khác

- Các tiêu chí đánh giá luật kết hợp (interestingness measures)
- Luật kết hợp tổng quát (generalized association rules)
- Luật kết hợp định lượng (quantitative association rules)
- **Luật kết hợp cực đại (maximal association rules)**
- Khai phá mẫu chuỗi (sequential pattern mining)

## 4 Tổng kết bài giảng

# Nội dung

## 1 Các khái niệm và định nghĩa

- Tập mục, giao dịch, và cơ sở dữ liệu giao dịch
- Tập phổ biến và luật kết hợp
- Ứng dụng của tập phổ biến và luật kết hợp

## 2 Khai phá tập phổ biến và luật kết hợp

- Các bước trong khai phá luật kết hợp
- Phương pháp brute-force
- Phương pháp Apriori
- Phương pháp FP-Growth
- Phương pháp Eclat

## 3 Các chủ đề khác

- Các tiêu chí đánh giá luật kết hợp (interestingness measures)
- Luật kết hợp tổng quát (generalized association rules)
- Luật kết hợp định lượng (quantitative association rules)
- Luật kết hợp cực đại (maximal association rules)
- **Khai phá mẫu chuỗi (sequential pattern mining)**

## 4 Tổng kết bài giảng

# Nội dung

## 1 Các khái niệm và định nghĩa

- Tập mục, giao dịch, và cơ sở dữ liệu giao dịch
- Tập phổ biến và luật kết hợp
- Ứng dụng của tập phổ biến và luật kết hợp

## 2 Khai phá tập phổ biến và luật kết hợp

- Các bước trong khai phá luật kết hợp
- Phương pháp brute-force
- Phương pháp Apriori
- Phương pháp FP-Growth
- Phương pháp Eclat

## 3 Các chủ đề khác

- Các tiêu chí đánh giá luật kết hợp (interestingness measures)
- Luật kết hợp tổng quát (generalized association rules)
- Luật kết hợp định lượng (quantitative association rules)
- Luật kết hợp cực đại (maximal association rules)
- Khai phá mẫu chuỗi (sequential pattern mining)

## 4 Tổng kết bài giảng

# Tổng kết bài giảng

# Tài liệu tham khảo



J. Han, M. Kamber, and J. Pei.

*Data Mining: Concepts and Techniques (Chapter 6. Mining Frequent Patterns, Associations, and Correlations: Basic Concepts and Methods; Chapter 7. Advanced Pattern Mining).*

The 3rd Edition, Morgan Kaufmann, Elsevier, 2012.



A. Rajaraman, J. Leskovec, and J. D. Ullman.

*Mining of Massive Datasets (6. Frequent Itemsets).*

The 2nd Edition, Cambridge University Press, 2013.



M. J. Zaki and W. M. Jr.

*Data Mining and Analysis: Fundamental Concepts and Algorithms (II. Frequent Pattern Mining).*

Cambridge University Press, 2013.