

# **Khai phá mẫu phổ biến và luật kết hợp**

Phan Xuân Hiếu

Bài giảng của DSLab

Viện nghiên cứu cao cấp về Toán (VIASM)



Vietnam Institute for  
Advanced Study in Mathematics

# Câu chuyện “bỉm” và “bia”



# Mining association rules between sets of items in large databases

Full Text:  PDF  [Get this Article](#)

Authors: [Rakesh Agrawal](#) [IBM Almaden Research Center, 650 Harry Road, San Jose, CA](#)

[Tomasz Imieliński](#) [Computer Science Department, Rutgers University, New Brunswick, NJ](#)

[Arun Swami](#) [IBM Almaden Research Center, 650 Harry Road, San Jose, CA](#)

Published in:



- Proceeding

SIGMOD '93 Proceedings of the 1993 ACM SIGMOD international conference on Management of data

Pages 207-216



 1993 Article

## Bibliometrics

- Citation Count: 3,365
- Downloads (cumulative): 20,307
- Downloads (12 Months): 1,704
- Downloads (6 Weeks): 209

## Fast Algorithms for Mining Association Rules in Large Databases

Authors: [Rakesh Agrawal](#)

[Ramakrishnan Srikant](#)

Published in:

- Proceeding

VLDB '94 Proceedings of the 20th International Conference on Very Large Data Bases

Pages 487-499

September 12 - 15, 1994

Morgan Kaufmann Publishers Inc. San Francisco, CA, USA ©1994

[table of contents](#) ISBN:1-55860-153-8



 1994 Article

## Bibliometrics

- Citation Count: 4,164
- Downloads (cumulative): n/a
- Downloads (12 Months): n/a
- Downloads (6 Weeks): n/a



ute for  
ly in Mathematics

# Rakesh Agrawal

≡ Google Scholar



SIGN IN



Rakesh Agrawal

Technical Fellow, Microsoft Research  
Verified email at microsoft.com

Data Mining Web Search Education Privacy

FOLLOW

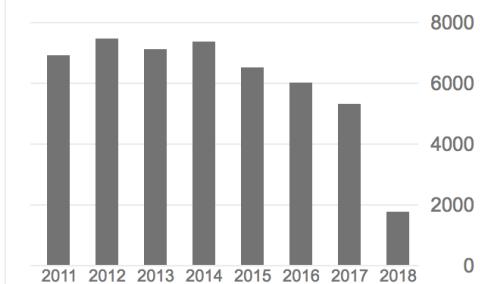
GET MY OWN PROFILE

Cited by

[VIEW ALL](#)

All Since 2013

Citations	114789	34215
h-index	100	57
i10-index	264	171



TITLE	CITED BY	YEAR
Fast algorithms for mining association rules R Agrawal, R Srikant Proc. 20th int. conf. very large data bases, VLDB 1215, 487-499	24041	1994
Mining association rules between sets of items in large databases R Agrawal, T Imielinski, A Swami Acm sigmod record 22 (2), 207-216	20661	1993
Mining sequential patterns R Agrawal, R Srikant Data Engineering, 1995. Proceedings of the Eleventh International Conference ...	7035	1995
Privacy-preserving data mining R Agrawal, R Srikant ACM Sigmod Record 29 (2), 439-450	3675	2000

# Nội dung

## I. Khái niệm và định nghĩa

- Tập mục, giao dịch, CSDL giao dịch
- Tập phổ biến (TPB) và luật kết hợp (LKH)

## 2. Các phương pháp khai phá TPB và LKH

- Phương pháp Apriori
- Phương pháp FP-Growth
- Các phương pháp khác

## 3. Đánh giá luật kết hợp

## 4. Các ứng dụng thực tiễn

# Nội dung

## I. Khái niệm và định nghĩa

- Tập mục, giao dịch, CSDL giao dịch
- Tập phổ biến (TPB) và luật kết hợp (LKH)

## 2. Các phương pháp khai phá TPB và LKH

- Phương pháp Apriori
- Phương pháp FP-Growth
- Các phương pháp khác

## 3. Đánh giá luật kết hợp

## 4. Các ứng dụng thực tiễn

# Khái niệm luật kết hợp (association rule)



- Luật kết hợp: *Mối 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%)

# 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:  $\mathbb{I} = \{sữa, trứng, đường, bánh mỳ, mật ong, mứt, bơ, thịt bò, giá, \dots\}$ .
  - ▶ 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:

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

## 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

## I. Khái niệm và định nghĩa

- Tập mục, giao dịch, CSDL giao dịch
- Tập phổ biến (TPB) và luật kết hợp (LKH)

## 2. Các phương pháp khai phá TPB và LKH

- Phương pháp Apriori
- Phương pháp FP-Growth
- Các phương pháp khác

## 3. Đánh giá luật kết hợp

## 4. Các ứng dụng thực tiễn

# Các phương pháp khai phá

1. Các bước trong khai phá luật kết hợp
2. Phương pháp brute-force
3. Phương pháp Apriori
4. Phương pháp FP-Growth
5. Các phương pháp khác



# 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  $\emptyset$ ).

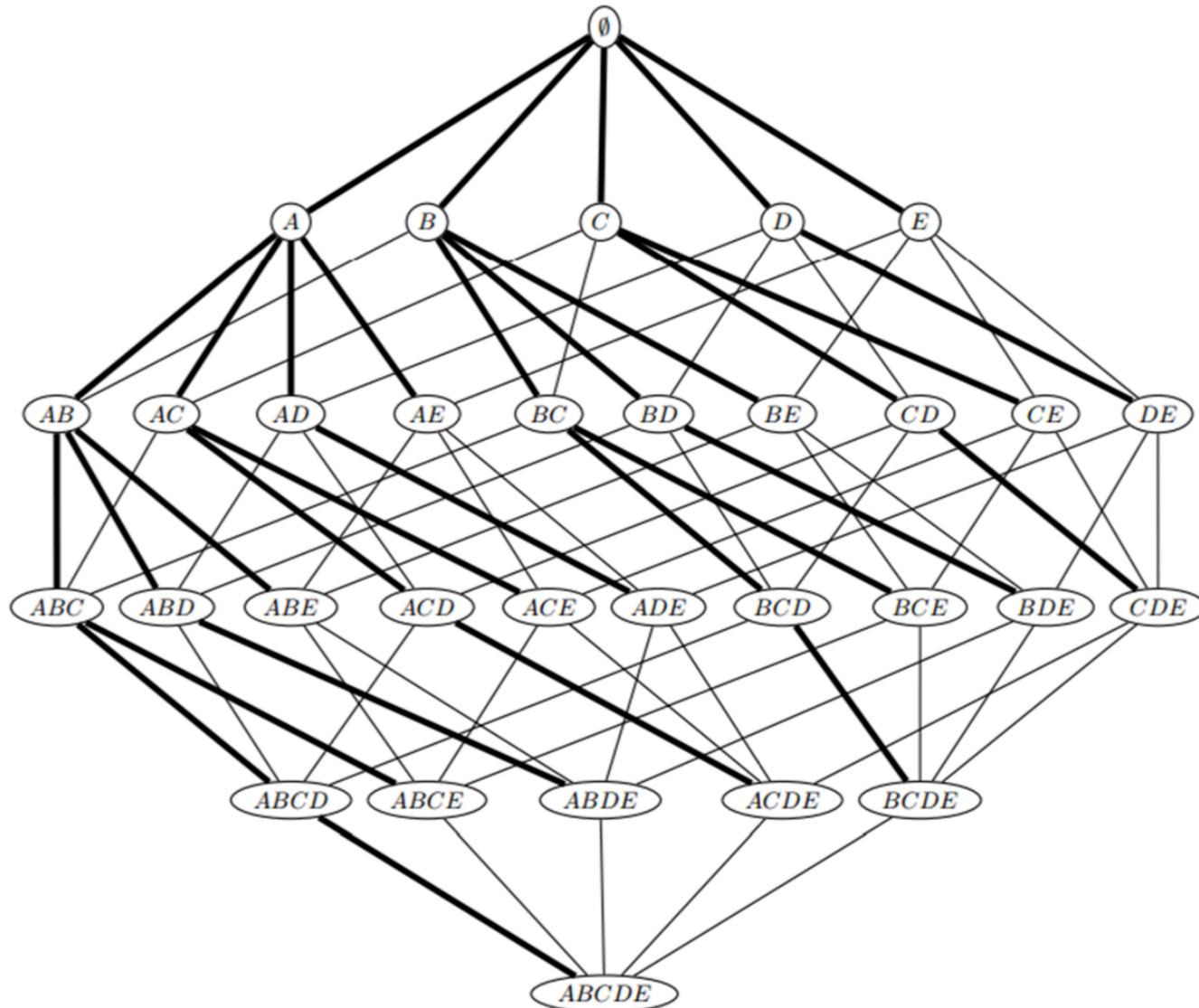
# Các phương pháp khai phá

1. Các bước trong khai phá luật kết hợp
2. Phương pháp brute-force
3. Phương pháp Apriori
4. Phương pháp FP-Growth
5. Các phương pháp khác

## 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 phổ 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 phổ 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 phô biến.

# Các phương pháp khai phá

1. Các bước trong khai phá luật kết hợp
2. Phương pháp brute-force
- 3. Phương pháp Apriori**
4. Phương pháp FP-Growth
5. Các phương pháp khác



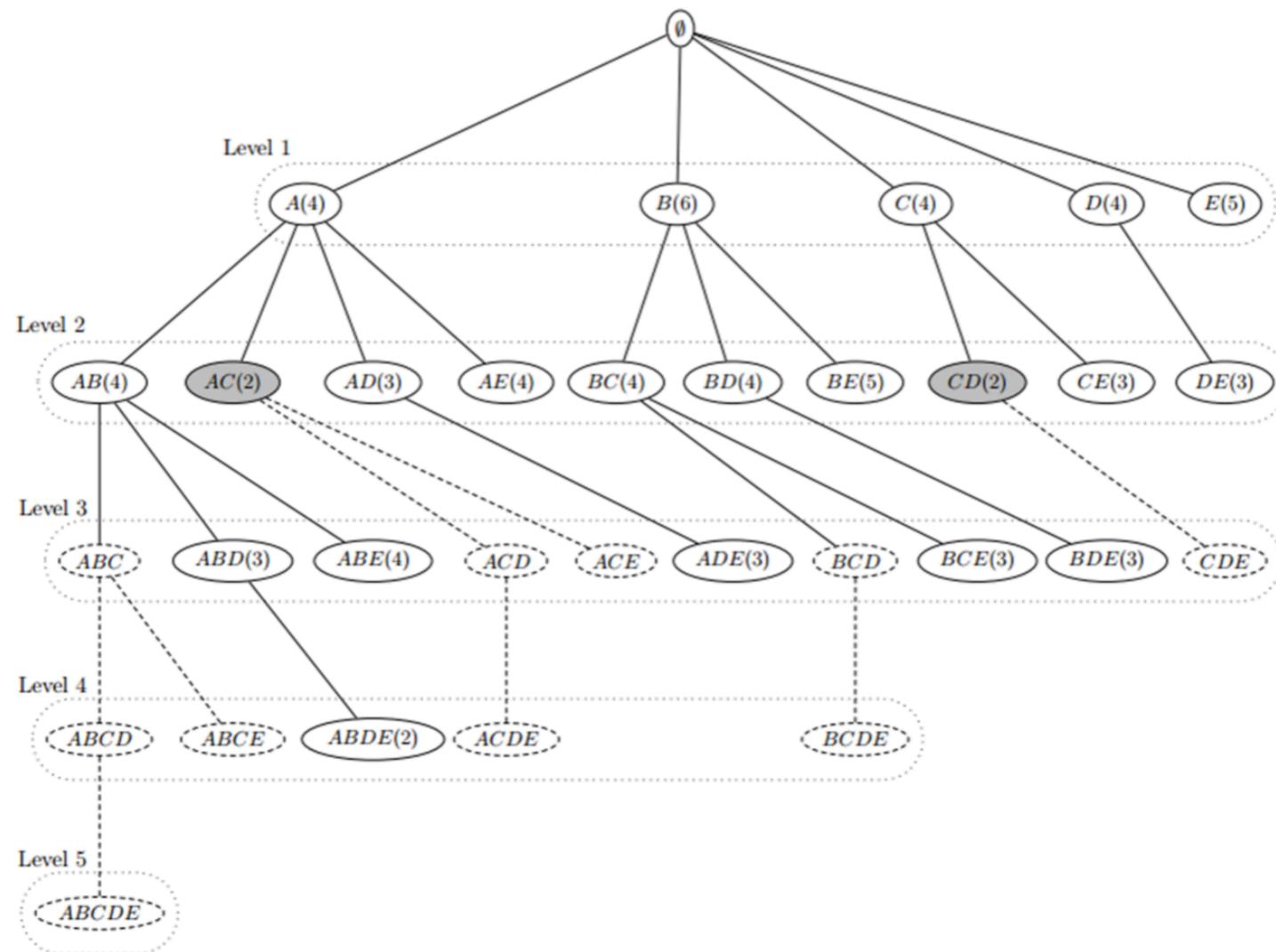
# Một số tính chất sử dụng trong phương pháp Apriori

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

## 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.

# CSDL giao dịch $\mathbb{D}$ minh họa cho thuật toán Apriori

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:

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

- 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  $F^{(1)}$

Tập các tập mục ứng viên  $C^{(2)}$

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

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

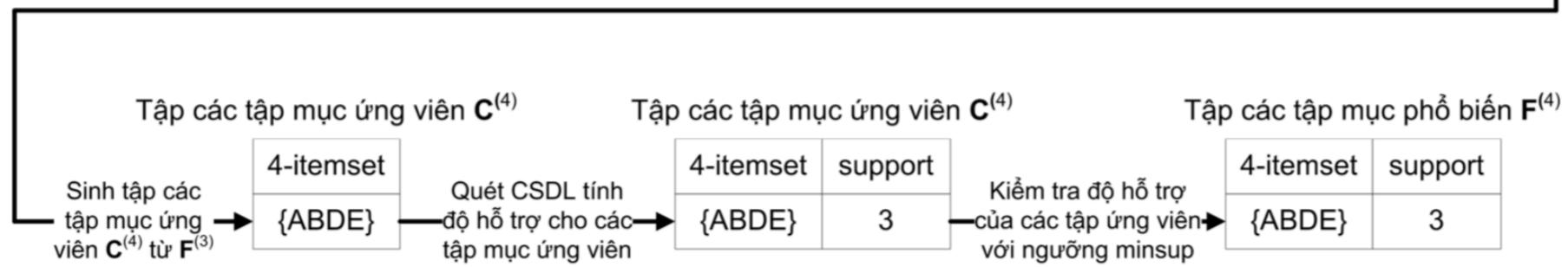
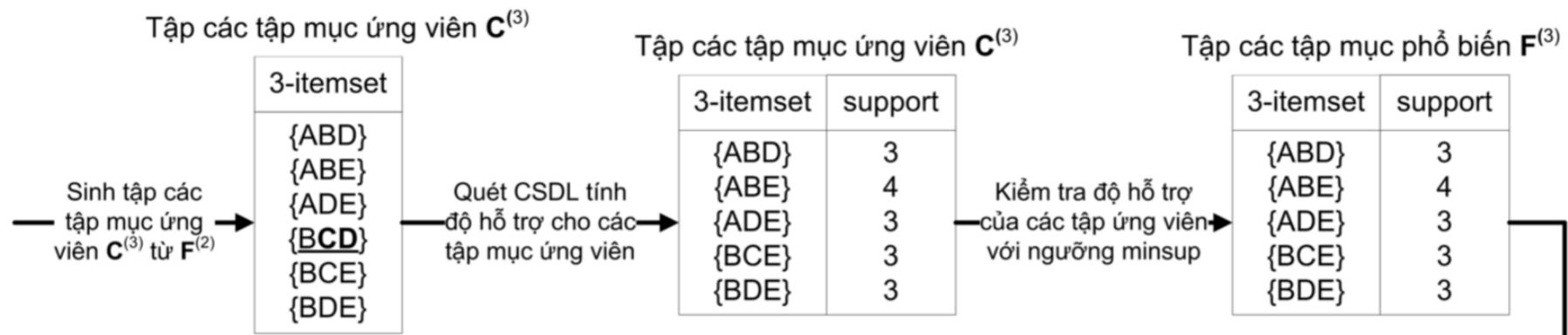
Tập các tập mục phổ biến  $F^{(2)}$

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 phổ 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 APRIORIGEN( $\mathbb{F}^{(k-1)}$ )
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 phổ biến  $ABDE$  có độ hỗ trợ bằng 3 với độ tin cậy tối thiểu  $\text{minconf} = 0.8$ :

- $\mathbb{X} = \{ABD(3), ABE(4), ADE(3), BDE(3), AB(4), AD(4), AE(4), BD(4), BE(5), DE(3), A(4), B(6), D(4), E(5)\}$ .
- $Y = ABD$ :  $\text{conf}(ABD \rightarrow E) = \frac{3}{3} = 1.0 \geq 0.8$  nên  $ABD \rightarrow E$  là luật mạnh.
- $Y = ABE$ :  $\text{conf}(ABE \rightarrow D) = \frac{3}{4} = 0.75 < 0.8$  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,  $\mathbb{X} = \{ADE(3), BDE(3), AD(4), BD(4), DE(3), D(4)\}$ .
- $Y = ADE$ :  $\text{conf}(ADE \rightarrow B) = \frac{3}{3} = 1.0 \geq 0.8$  nên  $ADE \rightarrow B$  là luật mạnh.
- $Y = BDE$ :  $\text{conf}(BDE \rightarrow A) = \frac{3}{3} = 1.0 \geq 0.8$  nên  $BDE \rightarrow A$  là luật mạnh.
- $Y = AD$ :  $\text{conf}(AD \rightarrow BE) = \frac{3}{4} = 0.75 < 0.8$  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,  $\mathbb{X} = \{BD(4), DE(3)\}$ .
- $Y = BD$ :  $\text{conf}(BD \rightarrow AE) = \frac{3}{4} = 0.75 < 0.8$  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,  $\mathbb{X} = \{DE(3)\}$ .
- $Y = DE$ :  $\text{conf}(DE \rightarrow AB) = \frac{3}{3} = 1.0 \geq 0.8$  nên  $DE \rightarrow AB$  là luật mạnh.

# 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.

# Các phương pháp khai phá

1. Các bước trong khai phá luật kết hợp
2. Phương pháp brute-force
3. Phương pháp Apriori
- 4. Phương pháp FP-Growth**
5. Các phương pháp khác



# 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 BUILDFP_TREE( $\mathbb{D} = \{T_1, T_2, \dots, T_m\}$ )
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 ( $cNode \in Children(pNode)$  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  $\mathbb{D}$ :

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

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

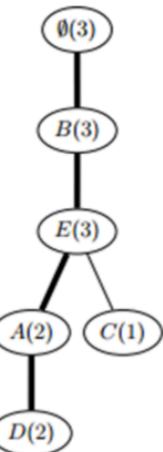
# Minh họa thuật toán sinh cây FP-Tree $\mathbb{T}$ từ CSDL $\mathbb{D}$



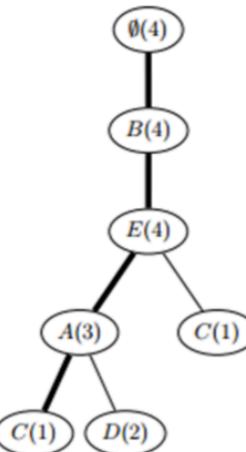
(a)  $\langle 1, BEAD \rangle$



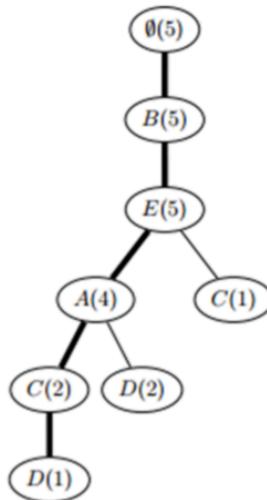
(b)  $\langle 2, BEC \rangle$



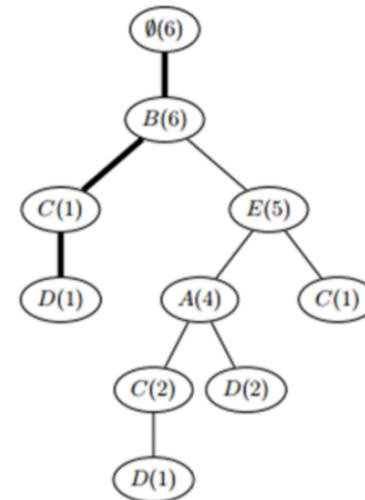
(c)  $\langle 3, BEAD \rangle$



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



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



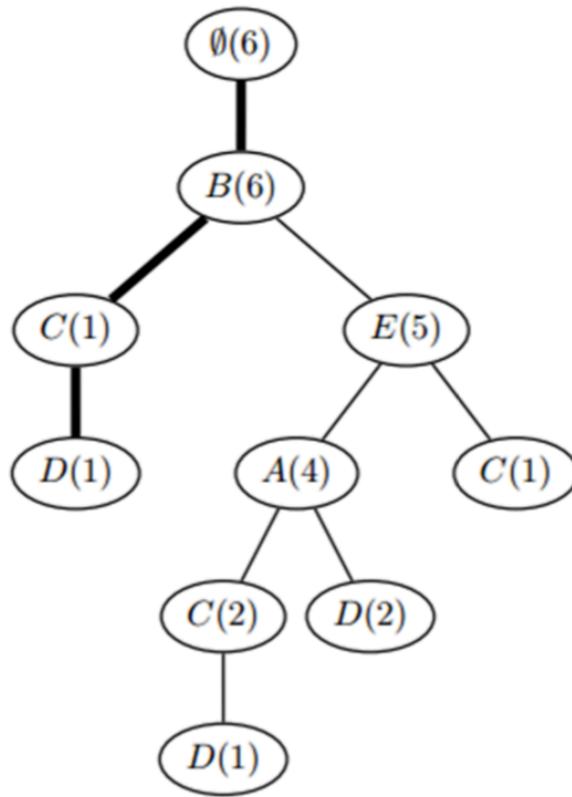
(f)  $\langle 6, BCD \rangle$

[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]

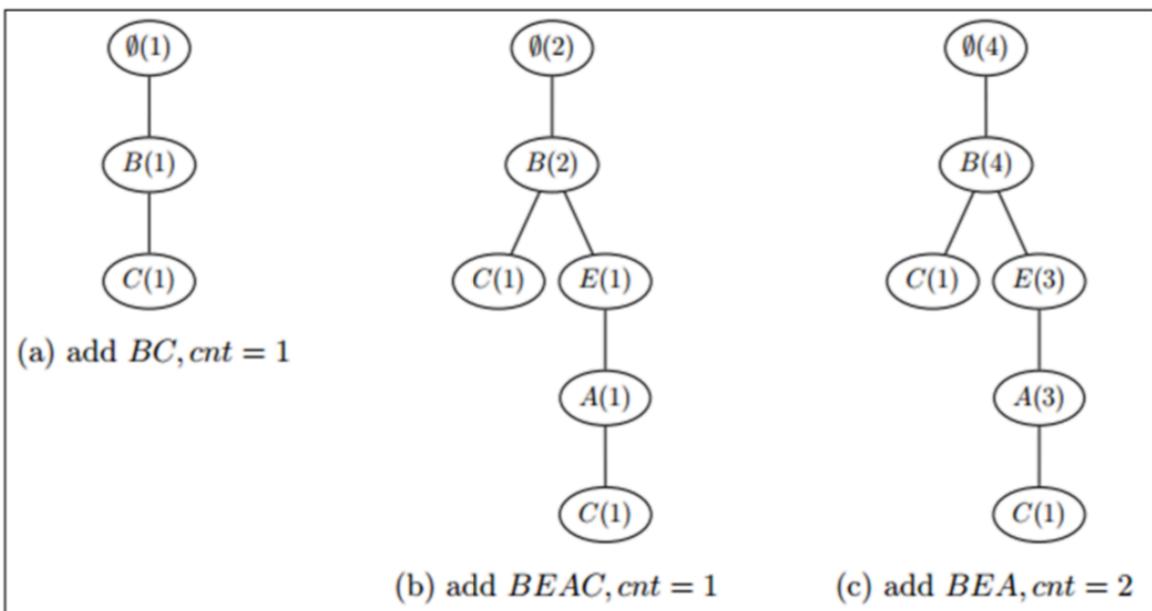
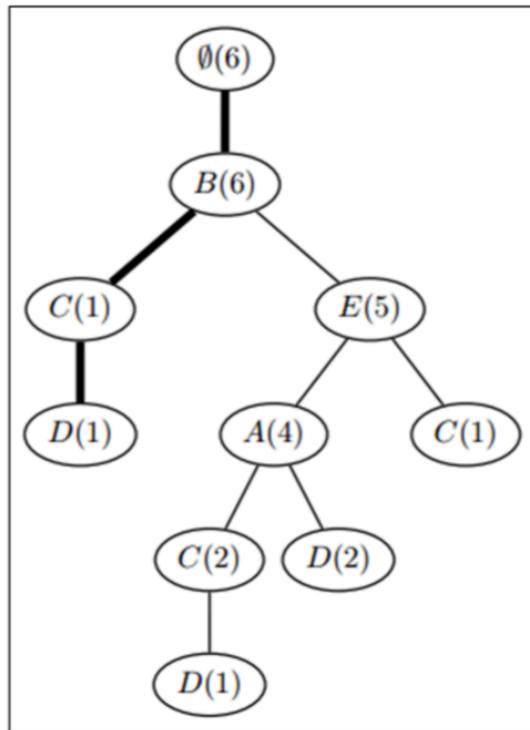
# Thuật toán đệ quy sinh các tập phổ biến từ cây FP-Tree $\mathbb{T}$

```
1: procedure FPGROWTH( $\mathbb{T}$ ,  $P$ ,  $\mathbb{F}$ ,  $minsup$ )
2:   Loại bỏ các mục không phổ 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:        FPGrowth( $\mathbb{T}_X$ ,  $X$ ,  $\mathbb{F}$ ,  $minsup$ );
21:      end if
22:    end for
23:  end if
24: end procedure
```

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

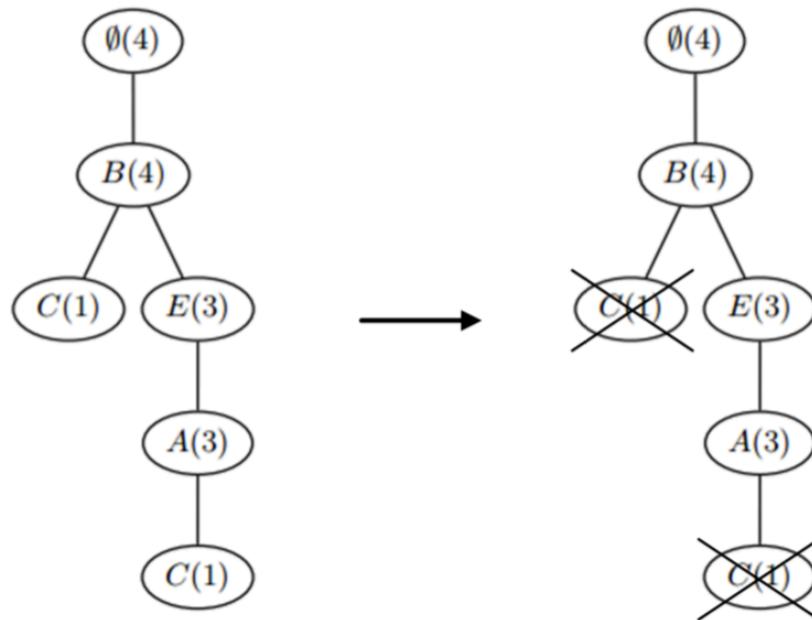
- Lời gọi hàm đầu tiên  $\text{FPGrowth}(\mathbb{T}, P \leftarrow \emptyset, \mathbb{F} \leftarrow \emptyset, \text{minsup})$ .
- Phép chiếu chọn (projection) cây FP–Tree  $\mathbb{T}$  theo một mục (item) nào đó.
- Cây FP–Tree  $\mathbb{T}$  có thể là một đường tuyến tính (*path*).
- Loại bỏ các mục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

# Cây FP-Tree chiểu 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  $\mathbb{T}_D$  chiếu 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 tuyến tính (*path*):  $B(4) - E(3) - A(3)$ :
    - ★ Liệt kê tất cả các tập con của đường tuyến 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).

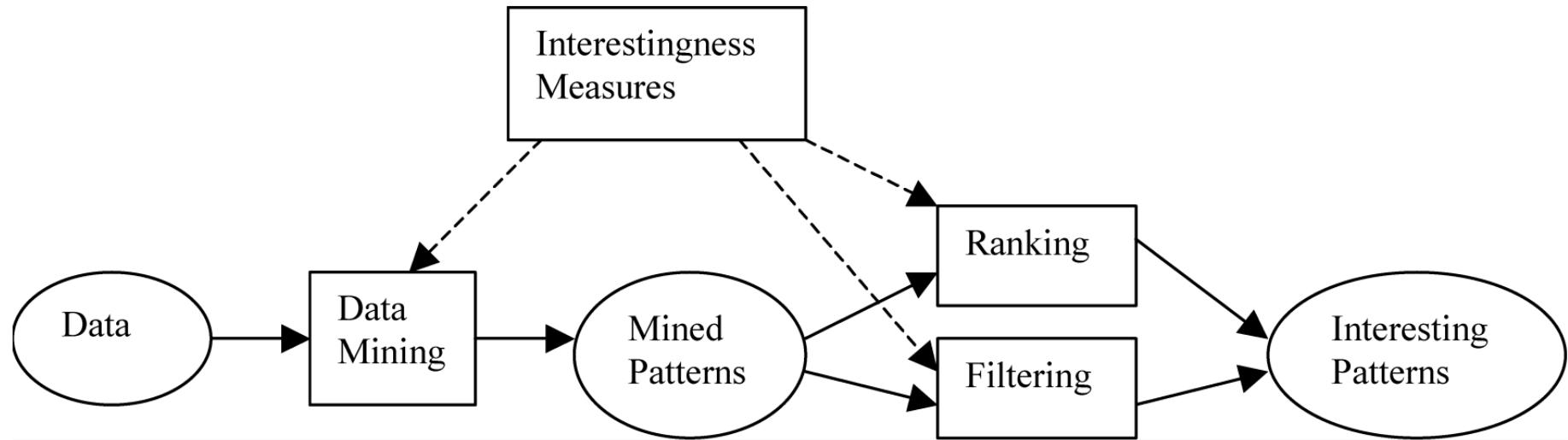
# Các phương pháp khai phá

1. Các bước trong khai phá luật kết hợp
2. Phương pháp brute-force
3. Phương pháp Apriori
4. Phương pháp FP-Growth
5. **Các phương pháp khác**

# Nội dung

- I. Khái niệm và định nghĩa
  - Tập mục, giao dịch, CSDL giao dịch
  - Tập phổ biến (TPB) và luật kết hợp (LKH)
2. Các phương pháp khai phá TPB và LKH
  - Phương pháp Apriori
  - Phương pháp FP-Growth
  - Các phương pháp khác
3. Đánh giá luật kết hợp
4. Các ứng dụng thực tiễn

# “Interestingness” và “Usefulness”



# Một số tiêu chí đánh giá

1. Conciseness
2. General/Coverage (Support)
3. Reliability (Confidence)
4. Peculiarity
5. Diversity
6. Novelty
7. Surprisingness/Unexpectedness
8. Utility
9. Actionability/Applicability

## 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:

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

- 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 mối quan hệ tương quan giữa  $A$  và  $B$ .
- Nếu  $lift(A \rightarrow B) > 1$ : luật  $A \rightarrow B$  có ý nghĩa (tương quan dương - positive correlation).
- Nếu  $lift(A \rightarrow B) < 1$ : luật  $A \rightarrow B$  và cả luật  $B \rightarrow A$  không có ý nghĩa (tương quan âm - negative correlation).

# Conviction

$$Conviction(A \rightarrow B) = \frac{P(A) P(\bar{B})}{P(A\bar{B})}.$$

- Giá trị trong khoảng  $[0, +\infty)$
- Conviction = I:A và B độc lập
- Conviction < I: luật ít ý nghĩa
- Conviction lớn: luật có nhiều ý nghĩa
- Nhận xét: điều kiện quá chặt (strict)

# Improve

$$Improve(A \rightarrow B) = [P(B | A) - P(B)].$$

	$B'$ occurring	$B'$ not occurring	Total		$D$ occurring	$D$ not occurring	Total
$A'$ occurring	8000	1000	9000	C occurring	3600	700	4300
$A'$ not occurring	500	500	1000	C not occurring	3700	2000	5700
Total	8500	1500	10000	Total	7300	2700	10000

$$Improve(A' \rightarrow B') = [P(B' | A') - P(B')] = 0.03,$$

$$Improve(C \rightarrow D) = [P(D | C) - P(D)] = 0.11.$$

- $P(B' | A') - P(B' | \text{not } A')$
- $P(D | C) - P(D | \text{not } C)$

# Nội dung

- I. Khái niệm và định nghĩa**
  - Tập mục, giao dịch, CSDL giao dịch
  - Tập phổ biến (TPB) và luật kết hợp (LKH)
- 2. Các phương pháp khai phá TPB và LKH**
  - Phương pháp Apriori
  - Phương pháp FP-Growth
  - Các phương pháp khác
- 3. Đánh giá luật kết hợp**
- 4. Các ứng dụng thực tiễn**



# Các ứng dụng thực tiễn

1. Phân lớp, phân loại (classification/decision rules)
2. Phân tích dữ liệu bán lẻ (market basket analysis)
3. Tư vấn trực tuyến (online recommendation)
4. Hiểu người dùng trực tuyến (user understanding)
5. Phân tích tìm ngoại lệ (outlier detection)
6. Ứng dụng trong các bài toán viễn thông  
(vd: churn prediction)
7. Phân tích dữ liệu di truyền.
8. Phân tích dữ cấu trúc mạng.



# Các ứng dụng thực tiễn

- I. Phân lớp, phân loại (classification/decision rules)
2. Phân tích dữ liệu bán lẻ (market basket analysis)
3. Tư vấn trực tuyến (online recommendation)
4. Hiểu người dùng trực tuyến (user understanding)
5. Phân tích tìm ngoại lệ (outlier detection)
6. Ứng dụng trong các bài toán viễn thông  
(vd: churn prediction)
7. Phân tích dữ liệu di truyền.
8. Phân tích dữ cấu trúc mạng.

# Các ứng dụng thực tiễn

- I. Phân lớp, phân loại (classification/decision rules)
2. **Phân tích dữ liệu bán lẻ (market basket analysis)**
3. Tư vấn trực tuyến (online recommendation)
4. Hiểu người dùng trực tuyến (user understanding)
5. Phân tích tìm ngoại lệ (outlier detection)
6. Ứng dụng trong các bài toán viễn thông  
(vd: churn prediction)
7. Phân tích dữ liệu di truyền.
8. Phân tích dữ cấu trúc mạng.





# Các ứng dụng thực tiễn

1. Phân lớp, phân loại (classification/decision rules)
2. Phân tích dữ liệu bán lẻ (market basket analysis)
3. **Tư vấn trực tuyến (online recommendation)**
4. Hiểu người dùng trực tuyến (user understanding)
5. Phân tích tìm ngoại lệ (outlier detection)
6. Ứng dụng trong các bài toán viễn thông  
(vd: churn prediction)
7. Phân tích dữ liệu di truyền.
8. Phân tích dữ cấu trúc mạng.





# Tư vấn, khuyến nghị trực tuyến

## I. Tư vấn, khuyến nghị

1. Sản phẩm phù hợp và khách hàng có thể quan tâm
2. Sản phẩm khách hàng có ý định mua
3. Cross-sell
4. Up-sell

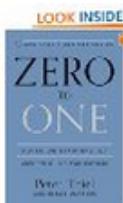
## 2. Phương pháp

1. Phân tích ý định mua sắm (purchase intent)
2. Lọc cộng tác (collaborative filtering)
3. Tư vấn dựa trên nội dung
4. Tư vấn dựa trên patterns mua sắm (frequent patterns, association rules)

## Frequently Bought Together



+



+



Price for all three: \$46.96

[Add all three to Cart](#)[Add all three to Wish List](#)[Show availability and shipping details](#)

- This item:** The Lean Startup: How Today's Entrepreneurs Use Continuous Innovation to Create Radically Successful ... by Eric Ries Hardcover \$14.64
- Zero to One: Notes on Startups, or How to Build the Future by Peter Thiel Hardcover \$16.20
- The Hard Thing About Hard Things: Building a Business When There Are No Easy Answers by Ben Horowitz Hardcover \$16.12

## Frequently Bought Together



+



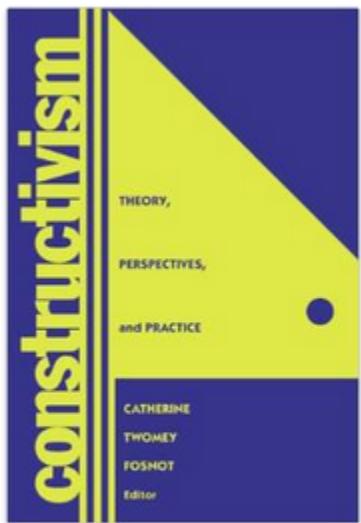
+



Price for all three: \$48.78

[Add all three to Cart](#)[Add all three to Wish List](#)[Show availability and shipping details](#)

- This item:** Golden Rose Matte Lipstick Set of 6 (SET1) \$24.99
- NYX Cosmetics Long Lasting Slim Lip Liner Pencils 6 Colors \$15.99
- Italia Eyeliners Set of 12 \$7.80



See this image

## Constructivism: Theory, Perspectives, and Practice Paperback – February, 1996

by Teachers College Press (Author), Catherine Twomey Fosnot (Editor)

Be the first to review this item

ISBN-13: 978-0807734889 ISBN-10: 0807734888



10 New from \$8.75 27 Used from \$0.01

Hardcover

Paperback

There is a newer edition of this item:



Constructivism: Theory, Perspectives And Practice

★★★★★ (1)

### FREE TWO-DAY SHIPPING FOR COLLEGE STUDENTS

[Learn more](#)



Part I, which covers the theoretical aspects of constructivism, includes chapters from Ernst von Glaserfeld, Catherine Twomey Fosnot, and Paul Cobb. In Part II, Candace Julyan, Eleanor Duckworth, Deborah Schifter, June S. Gould, Rheta DeVries, Betty Zan, and Maxine Greene provide perspectives from the field. Part III, which explores practices in the classroom, features work from Jill Bodner Lester, Susan Cowey, George Foman, Dewey Dykstra, Jr.

### Customers Who Viewed This Item Also Viewed



Constructivism: Theory,  
Perspectives And Practice  
Catherine Twomey...  
★★★★★ 1  
Paperback



Constructivism in Education  
Leslie P. Steffe  
Paperback  
\$96.85



Constructivism and  
Education  
Marie Larochele  
★★★★★ 1  
Paperback  
\$46.20



In Search of Understanding:  
The...  
Jacqueline Grennon...  
★★★★★ 14  
Paperback  
\$11.32

# Các mặt hàng thường được xem cùng nhau

Số lần được xem cùng nhau: 5



Aquaphor - Kem trị hâm, chàm hoặc khô da



Váng sữa MONTBLANC vani



MEKONGZON -  
Thuốc Lợi Sữa  
MOTHERLOVE  
More Milk Plus 60  
Viên



Thuốc bổ bà bầu  
Pregnacare  
Conception - 30  
viên



Thuốc Siro ho trái  
cây tiết xuất từ lá  
thường xuân  
PROSPAN.



THUỐC ELEVIT-  
VITAMIN CHO  
MẸ TRƯỚC KHI  
SINH



SET 2 TÃ VÀI  
(QUẦN) CHỐNG  
HẨM



Sữa bột Dumex  
Gold 1 - 800g (cho  
bé dưới 6 tháng)

Số lần được xem cùng nhau: 6



Đầm SN bông chân  
YD 186-Ke



Áo Sơ mi pha ren  
hồng Full size



Bộ đồ bay ngắn  
thắt eo



ÁO CHẤM BI TAY  
DÀI CÁCH ĐIỆU

# Các mặt hàng thường được mua cùng nhau

Số lần được mua cùng nhau: 6



Bộ quần áo bé gái  
Eko BN42



Bộ quần áo bé gái  
Eko BN44

Số lần được mua cùng nhau: 9



Áo Sơ mi pha ren  
hồng Full size



HAPPY F - Đầm  
xòe 3 màu

Số lần được mua cùng nhau: 9



SHOP THANH  
TAM- ÁO THUN  
NAM CÓ CỔ  
BURBERRY



Áo thun nam  
Abercrombie &  
Fitch

# Các ứng dụng thực tiễn

1. Phân lớp, phân loại (classification/decision rules)
2. Phân tích dữ liệu bán lẻ (market basket analysis)
3. Tư vấn trực tuyến (online recommendation)
4. **Hiểu người dùng trực tuyến (user understanding)**
5. Phân tích tìm ngoại lệ (outlier detection)
6. Ứng dụng trong các bài toán viễn thông  
(vd: churn prediction)
7. Phân tích dữ liệu di truyền.
8. Phân tích dữ cấu trúc mạng.



# Các ứng dụng thực tiễn

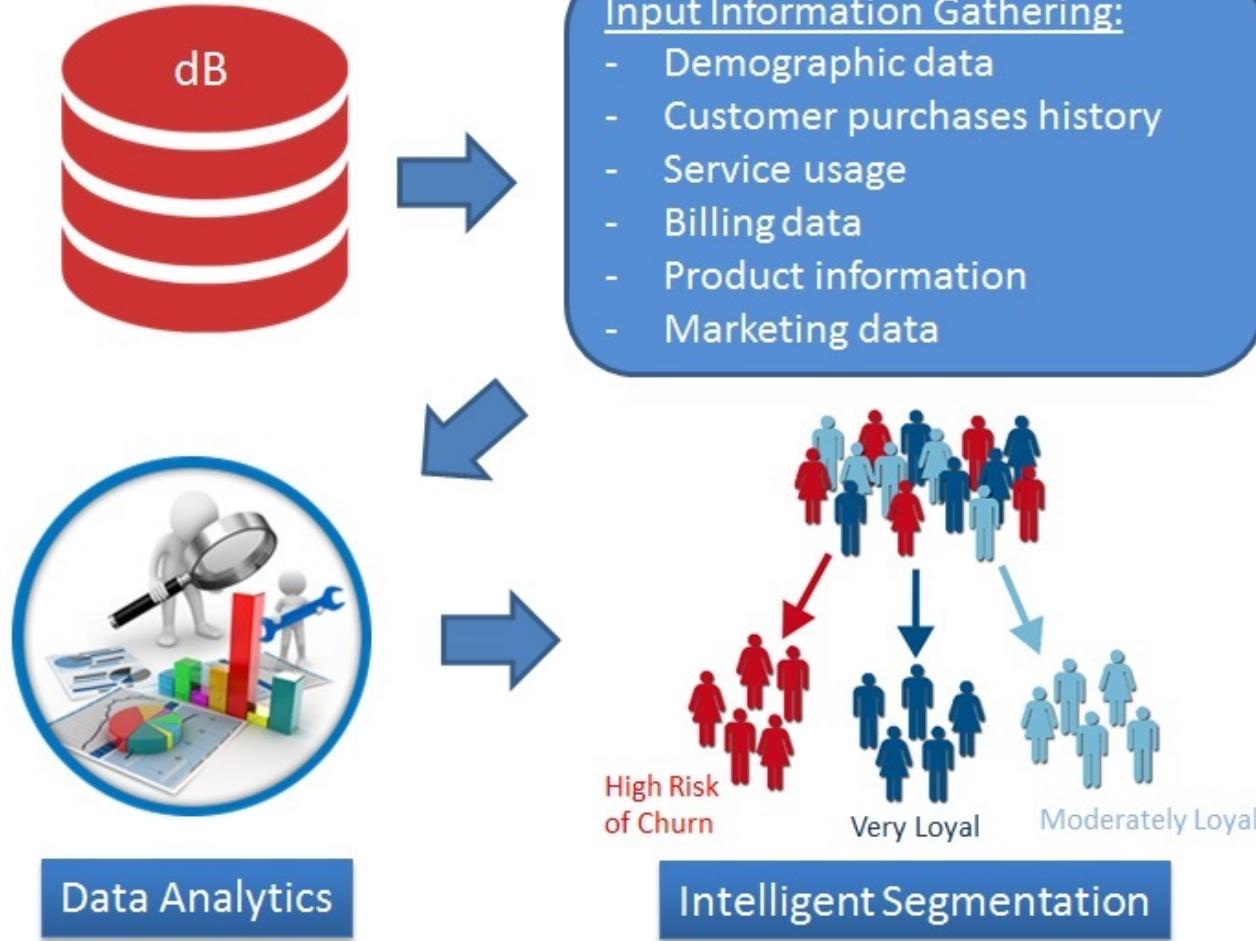
- I. Phân lớp, phân loại (classification/decision rules)
2. Phân tích dữ liệu bán lẻ (market basket analysis)
3. Tư vấn trực tuyến (online recommendation)
4. Hiểu người dùng trực tuyến (user understanding)
- 5. Phân tích tìm ngoại lệ (outlier detection)**
6. Ứng dụng trong các bài toán viễn thông  
(vd: churn prediction)
7. Phân tích dữ liệu di truyền.
8. Phân tích dữ cấu trúc mạng.



# Các ứng dụng thực tiễn

1. Phân lớp, phân loại (classification/decision rules)
2. Phân tích dữ liệu bán lẻ (market basket analysis)
3. Tư vấn trực tuyến (online recommendation)
4. Hiểu người dùng trực tuyến (user understanding)
5. Phân tích tìm ngoại lệ (outlier detection)
6. **Ứng dụng trong các bài toán viễn thông  
(vd: churn prediction)**
7. Phân tích dữ liệu di truyền.
8. Phân tích dữ cấu trúc mạng.

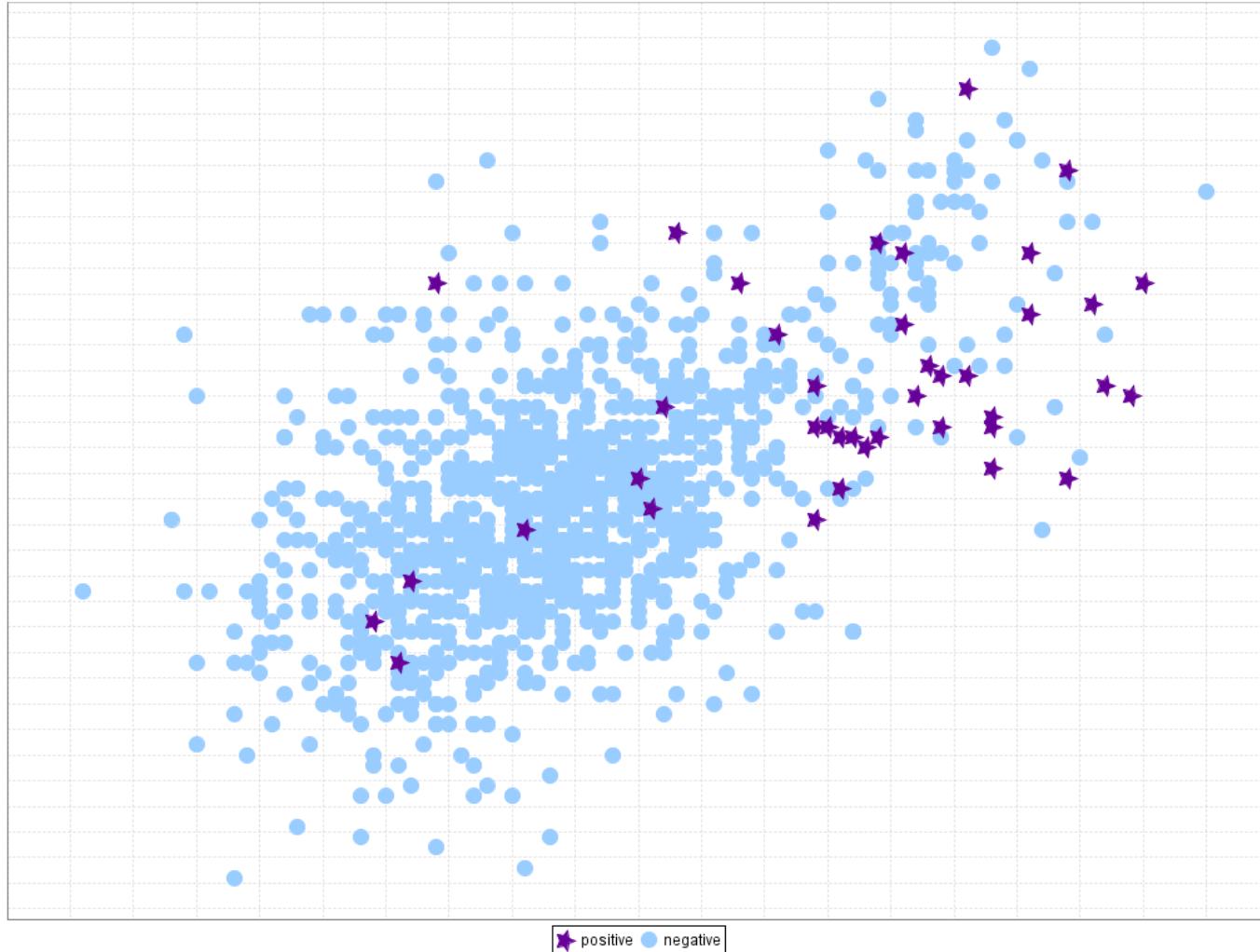
# Churn prediction



# Dữ liệu viễn thông

- I. Thông tin khách hàng (customer data)
  1. Thông tin tĩnh
  2. Thông tin động
2. Thông tin thuê bao (contract/plan data)
3. Thông tin sử dụng dịch vụ (call/service detail data)
  1. Hành vi sử dụng
  2. Thay đổi trong sử dụng dịch vụ
  3. Ngừng phát sinh cước
4. Lịch sử yêu cầu chăm sóc khách hàng (customer care history data)
- .v.v.

# Highly imbalanced data



# Nhận biết vấn đề và giải pháp

- I. Dữ liệu rất lớn
2. Quan tâm đặc biệt đến “True positive”
3. Khó khăn khi lấy mẫu (sampling)
4. Khó khăn khi xây dựng mô hình học (thống kê) cho dữ liệu mất cân bằng (nghiêm trọng)
5. Có thể ròi rạc hóa dữ liệu?
6. **Khai phá luật hiếm và tin cậy xấp xỉ**
  - I. Từ tập mẫu (sau khi sampling)
  2. Luật “xấp xỉ” cho lớp dương (positive)
  3. Whitebox: dễ hiểu, dễ đánh giá, và điều chỉnh

# Các ứng dụng thực tiễn

1. Phân lớp, phân loại (classification/decision rules)
2. Phân tích dữ liệu bán lẻ (market basket analysis)
3. Tư vấn trực tuyến (online recommendation)
4. Hiểu người dùng trực tuyến (user understanding)
5. Phân tích tìm ngoại lệ (outlier detection)
6. Ứng dụng trong các bài toán viễn thông  
(vd: churn prediction)
7. **Phân tích dữ liệu di truyền.**
8. **Phân tích dữ cấu trúc mạng.**



# Ứng dụng khai phá dữ liệu viễn thông

## I. Dữ liệu

1. Customer data
2. Call detail data
3. Log and content data
4. Network data

## 2. Các bài toán khai phá dữ liệu

1. Spam filtering
2. Churn prediction
3. Fraud detection (subscription vs. superimposition)
4. Customer profiling and segmentation (for marketing)
5. Network fault isolation and prediction
6. Service/content recommendation



# Tổng kết bài giảng

## I. Khái niệm và định nghĩa

- Tập mục, giao dịch, CSDL giao dịch
- Tập phổ biến (TPB) và luật kết hợp (LKH)

## 2. Các phương pháp khai phá TPB và LKH

- Phương pháp Apriori
- Phương pháp FP-Growth
- Các phương pháp khác

## 3. Đánh giá luật kết hợp

## 4. Các ứng dụng thực tiễn

# 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.