



TRƯỜNG ĐH KHTN – TP HCM  
KHOA CÔNG NGHỆ THÔNG TIN

# KHAI THÁC ITEMSET PHỔ BIẾN SỬ DỤNG DIFFSET

DATA MINING

HCMUS - 2023



# Nội dung

---

1. Giới thiệu
2. Đặt vấn đề
3. Các phương pháp biểu diễn dữ liệu
4. Thuật toán Eclat
5. Sử dụng Diffset
6. Thuật toán dEclat, dCharm, dGenMax
7. So sánh Tidset và Diffset
8. Nhận xét

# 1. Giới thiệu

---

- ❖ Diffset là một cách biểu diễn dữ liệu được đưa ra bởi M. J. Zaki và K. Gouda.
- ❖ Phương pháp này theo vết sự thay đổi trong tidset của các ứng viên sau khi kết hợp với các ứng viên dùng phát sinh ra chúng.
- ❖ Mục tiêu của hướng tiếp cận là tiết kiệm bộ nhớ sử dụng để lưu các tidset và tăng hiệu suất thực thi thuật toán.

## 2. Đặt vấn đề

---

- ❖ Cho CSDL gồm các giao tác, tìm tất cả các itemset phổ biến – các itemset có số lần xuất hiện trong CSDL ít nhất thỏa giá trị do người dùng định nghĩa.
- ❖ Ví dụ: Cho CSDL D như sau.

| Transaction | Items         |
|-------------|---------------|
| 1           | A, C, T, W    |
| 2           | C, D, W       |
| 3           | A, C, T, W    |
| 4           | A, C, D, W    |
| 5           | A, C, D, T, W |
| 6           | C, D, T       |

## 2. Đặt vấn đề

Khai thác itemset phổ biến.

| Transaction | Items         |
|-------------|---------------|
| 1           | A, C, T, W    |
| 2           | C, D, W       |
| 3           | A, C, T, W    |
| 4           | A, C, D, W    |
| 5           | A, C, D, T, W |
| 6           | C, D, T       |

Minimum Support = 80%  
Minimum Support Count = 5

| Itemset phổ biến | Support | Support Count |
|------------------|---------|---------------|
| {C}              | 100%    | 6             |
| {W}              | 83%     | 5             |
| {CW}             | 83%     | 5             |

# 3. Các phương pháp biểu diễn

| Transaction | Items         |
|-------------|---------------|
| 1           | A, C, T, W    |
| 2           | C, D, W       |
| 3           | A, C, T, W    |
| 4           | A, C, D, W    |
| 5           | A, C, D, T, W |
| 6           | C, D, T       |

| Item | Support |
|------|---------|
| A    | 4       |
| C    | 6       |
| D    | 4       |
| T    | 4       |
| W    | 5       |

| Item | TIDSET      |
|------|-------------|
| A    | 1 3 4 5     |
| C    | 1 2 3 4 5 6 |
| D    | 2 4 5 6     |
| T    | 1 3 5 6     |
| W    | 1 2 3 4 5   |

a. Độ support

|   | A | C | D | T | W |
|---|---|---|---|---|---|
| 1 | 1 | 1 | 0 | 1 | 1 |
| 2 | 0 | 1 | 1 | 0 | 1 |
| 3 | 1 | 1 | 0 | 1 | 1 |
| 4 | 1 | 1 | 1 | 0 | 1 |
| 5 | 1 | 1 | 1 | 1 | 1 |
| 6 | 0 | 1 | 1 | 1 | 0 |

b. Tidset

c. Bitvector

# 3. Các phương pháp biểu diễn

---

## ***a. Lưu giá trị support***

- Mỗi itemset sẽ lưu kèm với một giá trị support. Phải đếm support cho từng itemset.

## ***b. Sử dụng Tidset (Mã giao tác)***

- Tổn bộ nhớ lưu các mã giao tác cho từng itemset.
- Duyệt CSDL một lần. Tính support bằng phép toán giao (intersection).

## ***c. Sử dụng Bitvector (vector nhị phân)***

- Tính toán support nhanh bằng phép toán AND.

# 4. Thuật toán Eclat

---

0. **Eclat**([ $P$ ]):

1. for all  $X_i \in [P]$  do

2.      $T_i = \emptyset$

3.     for all  $X_j \in [P]$ , with  $j > i$  do

4.          $R = X_i \cup X_j$ ;

5.          $t(R) = t(X_j) \cap t(X_i)$ ;

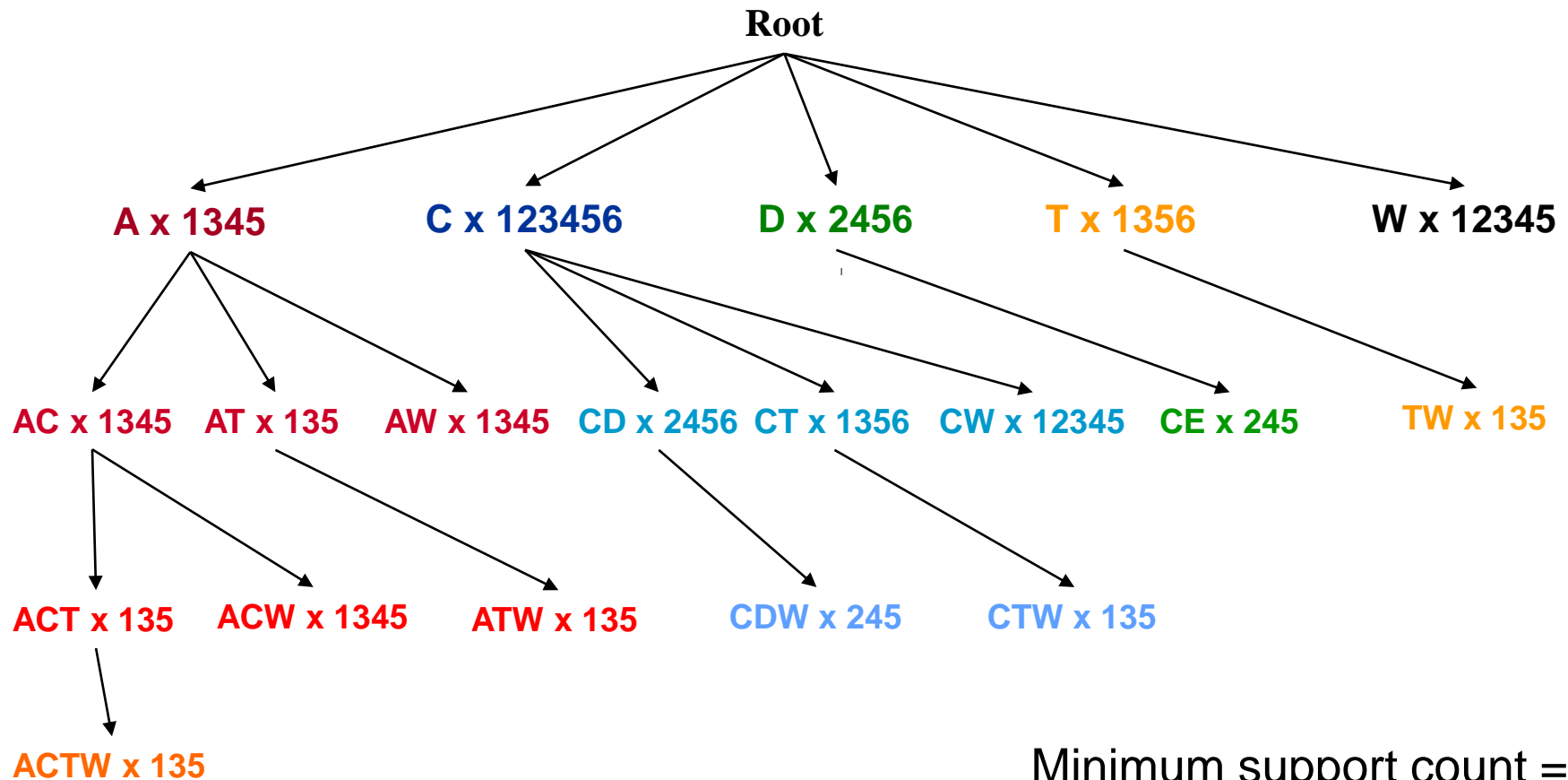
6.         if  $\sigma(R) \geq \textit{min\_sup}$  then

7.              $T_i = T_i \cup \{R\}$ ;  $F_{|R|} = F_{|R|} \cup \{R\}$ ;

8. for all  $T_i \neq \emptyset$  do **Eclat**( $T_i$ );

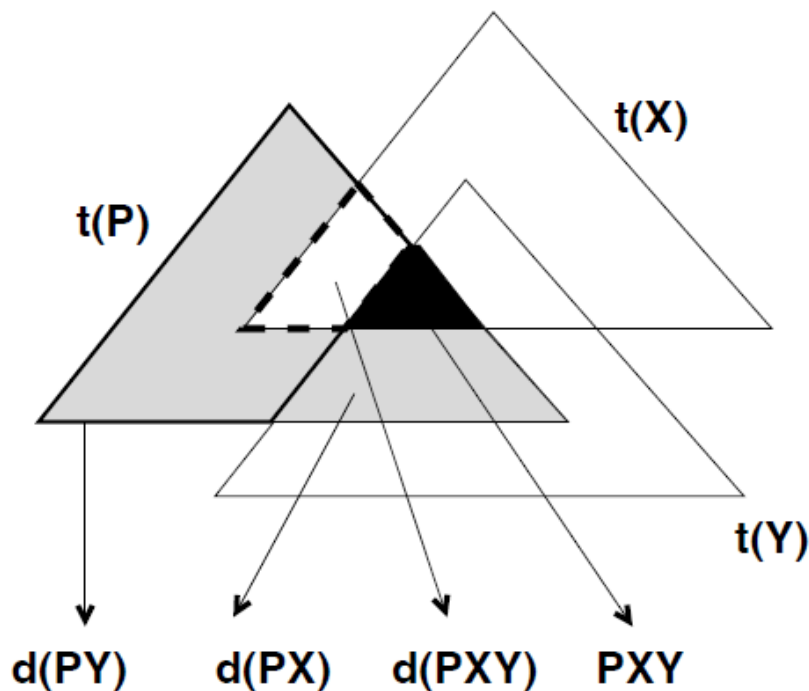


# 4. Thuật toán Eclat



# 5. Sử dụng Diffset (1)

❖ Lưu sự thay đổi Tiset của từng đối tượng cùng một lớp hoặc cùng tiền tố (prefix).



Ví dụ: Với  $P$  là tiền tố.  
Cho  $X, Y$  là các itemset.

- $t(P)$  là tidset của  $P$
- $d(PX)$  là diffset của  $PX$

$$d(PX) = t(P) - t(X)$$

$$d(PXY) = t(PX) - t(PY)$$

## 5. Sử dụng Diffset (2)

---

❖ Tính diffset của PXY:

$$d(PXY) = t(PX) - t(PY)$$

Tuy nhiên chúng ta chỉ lưu diffset của PX và PY là  $d(PX)$ ,  $d(PY)$ .

$$d(PXY) = t(PX) - t(PY) + \mathbf{t(P) - t(P)}$$

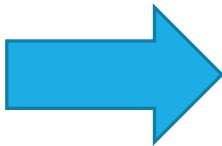
$$d(PXY) = (\mathbf{t(P) - t(PY)}) - (\mathbf{t(P) - t(PX)})$$

$$\Rightarrow \mathbf{d(PXY) = d(PY) - d(PX)}$$

❖ Tính support PXY:  $\sigma(PXY) = \sigma(PX) - |d(PXY)|$

## 5. Sử dụng Diffset (3)

| Item | TIDSET      |
|------|-------------|
| A    | 1 3 4 5     |
| C    | 1 2 3 4 5 6 |
| D    | 2 4 5 6     |
| T    | 1 3 5 6     |
| W    | 1 2 3 4 5   |



| Item | DIFFSET |
|------|---------|
| A    | 2 6     |
| C    |         |
| D    | 1 3     |
| T    | 2 4     |
| W    | 6       |

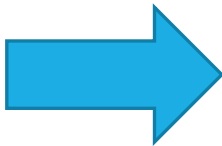
*Ví dụ:* Tính support của itemset AC

$$d(AC) = d(C) - d(A) = \{\emptyset\} - \{2, 6\} = \{\emptyset\}$$

$$\Rightarrow \sigma(AC) = \sigma(A) - |d(AC)| = 4 - 0 = 4$$

## 5. Sử dụng Diffset (4)

| Item | TIDSET      |
|------|-------------|
| A    | 1 3 4 5     |
| C    | 1 2 3 4 5 6 |
| D    | 2 4 5 6     |
| T    | 1 3 5 6     |
| W    | 1 2 3 4 5   |



| Item | DIFFSET |
|------|---------|
| A    | 2 6     |
| C    |         |
| D    | 1 3     |
| T    | 2 4     |
| W    | 6       |

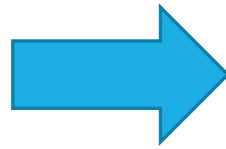
*Ví dụ:* Tính support của itemset AD

$$d(AD) = d(D) - d(A) = \{1, 3\} - \{2, 6\} = \{1, 3\}$$

$$\Rightarrow \sigma(AD) = \sigma(A) - |d(AD)| = 4 - 2 = 2$$

## 5. Sử dụng Diffset (5)

| Item | TIDSET      |
|------|-------------|
| A    | 1 3 4 5     |
| C    | 1 2 3 4 5 6 |
| D    | 2 4 5 6     |
| T    | 1 3 5 6     |
| W    | 1 2 3 4 5   |



| Item | DIFFSET |
|------|---------|
| A    | 2 6     |
| C    |         |
| D    | 1 3     |
| T    | 2 4     |
| W    | 6       |

Ví dụ: Tính support của itemset ACD kết hợp từ AC, AD

$$d(ACD) = d(AD) - d(AC) = \{1, 3\} - \{\emptyset\} = \{1, 3\}$$

$$\Rightarrow \sigma(ACD) = \sigma(AC) - |d(ACD)| = 4 - 2 = 2$$

# 5. Sử dụng Diffset (5)

$$d(A) = \{2, 6\} \Rightarrow \text{sup} = 4$$

$$d(C) = \{\} \Rightarrow \text{sup} = 6$$

$$d(D) = \{1, 3\} \Rightarrow \text{sup} = 4$$

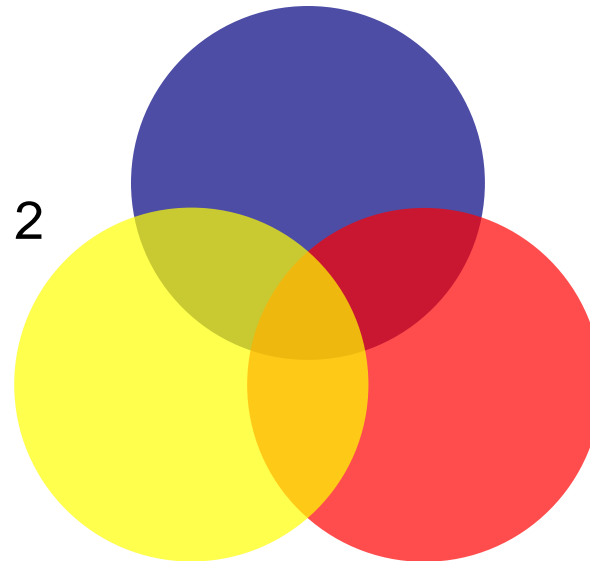
$$d(AC) = \{\} \Rightarrow \text{sup} = 4 - 0 = 4$$

$$d(AD) = \{1, 3\} \Rightarrow \text{sup} = 4 - 2 = 2$$

$$d(ACD) = \{1, 3\}$$

$$\Rightarrow \text{sup} = 4 - 2 = 2$$

$$\text{Tidset}(A) = \{1, 3, 4, 5\}$$



$$\text{Tidset}(C) = \{1, 2, 3, 4, 5, 6\}$$

$$\text{Tidset}(D) = \{2, 4, 5, 6\}$$

# 5. Sử dụng Diffset (5)

$$d(A) = \{2, 6\} \Rightarrow \text{sup} = 4$$

$$d(C) = \{\} \Rightarrow \text{sup} = 6$$

$$d(D) = \{1, 3\} \Rightarrow \text{sup} = 4$$

$$d(T) = \{2, 4\} \Rightarrow \text{sup} = 4$$

$$d(W) = \{6\} \Rightarrow \text{sup} = 5$$

$$*d(AC) = \{\} \Rightarrow \text{sup} = 4 - 0 = 4$$

$$d(AD) = \{1, 3\} \Rightarrow \text{sup} = 4 - 2 = 2$$

$$d(AT) = \{4\} \Rightarrow \text{sup} = 4 - 1 = 3$$

$$d(AW) = \{\} \Rightarrow \text{sup} = 4 - 0 = 4$$

$$*d(CD) = \{1, 3\} \Rightarrow \text{sup} = 6 - 2 = 4$$

$$d(CT) = \{2, 4\} \Rightarrow \text{sup} = 6 - 2 = 4$$

$$d(CW) = \{6\} \Rightarrow \text{sup} = 6 - 1 = 5$$

$$d(DT) = \{2, 4\} \Rightarrow \text{sup} = 4 - 2 = 2$$

$$d(DW) = \{6\} \Rightarrow \text{sup} = 4 - 1 = 3$$

| Item | DIFFSET |
|------|---------|
| A    | 2 6     |
| C    |         |
| D    | 1 3     |
| T    | 2 4     |
| W    | 6       |

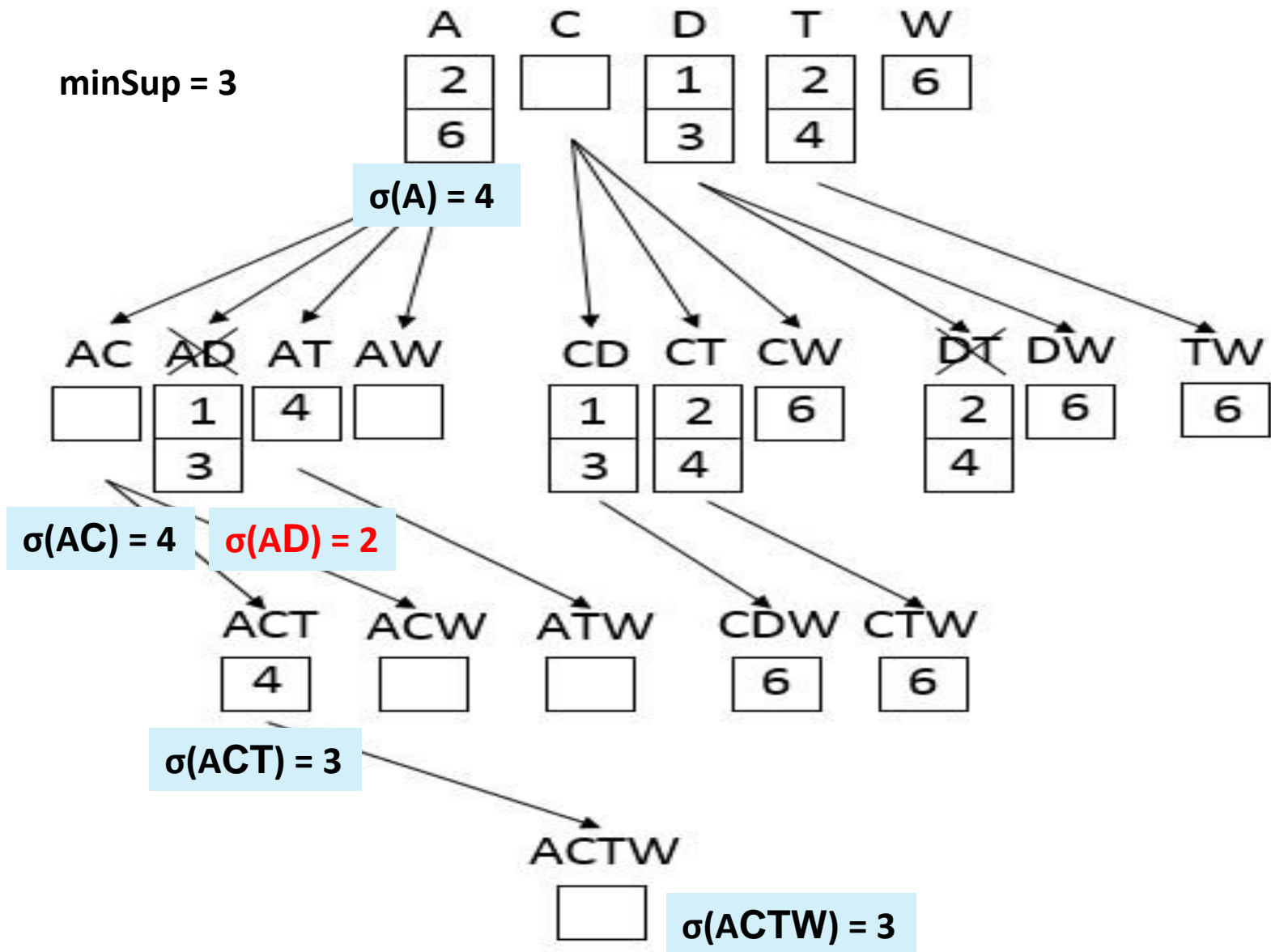
$$d(ACD) = \{1, 3\}$$

$$\Rightarrow \text{sup} = 4 - 2 = 2$$

....

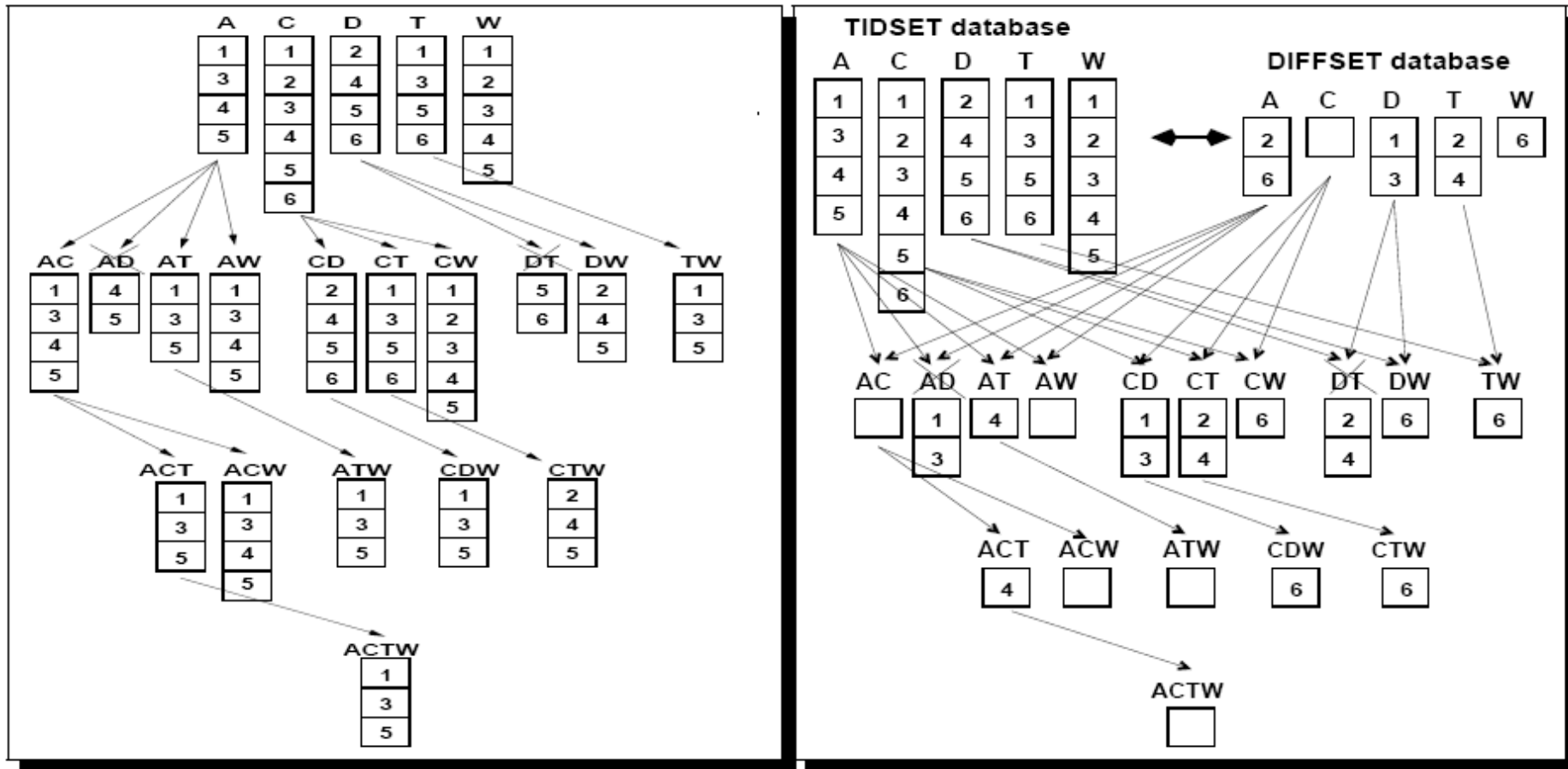


minSup = 3



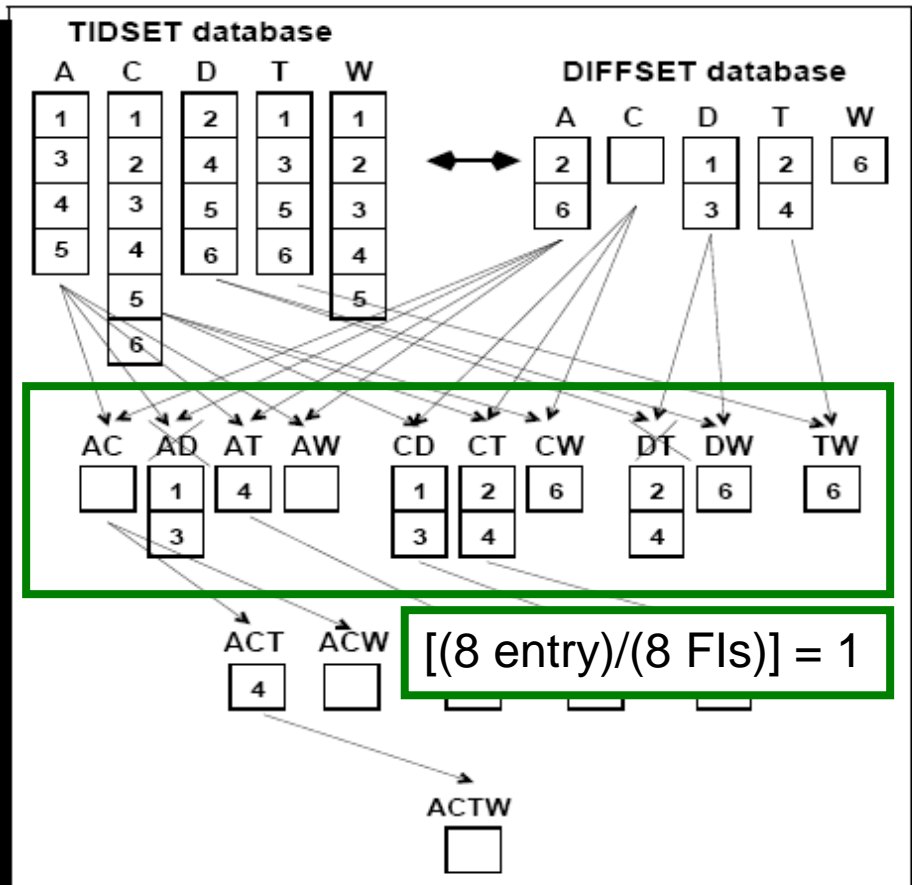
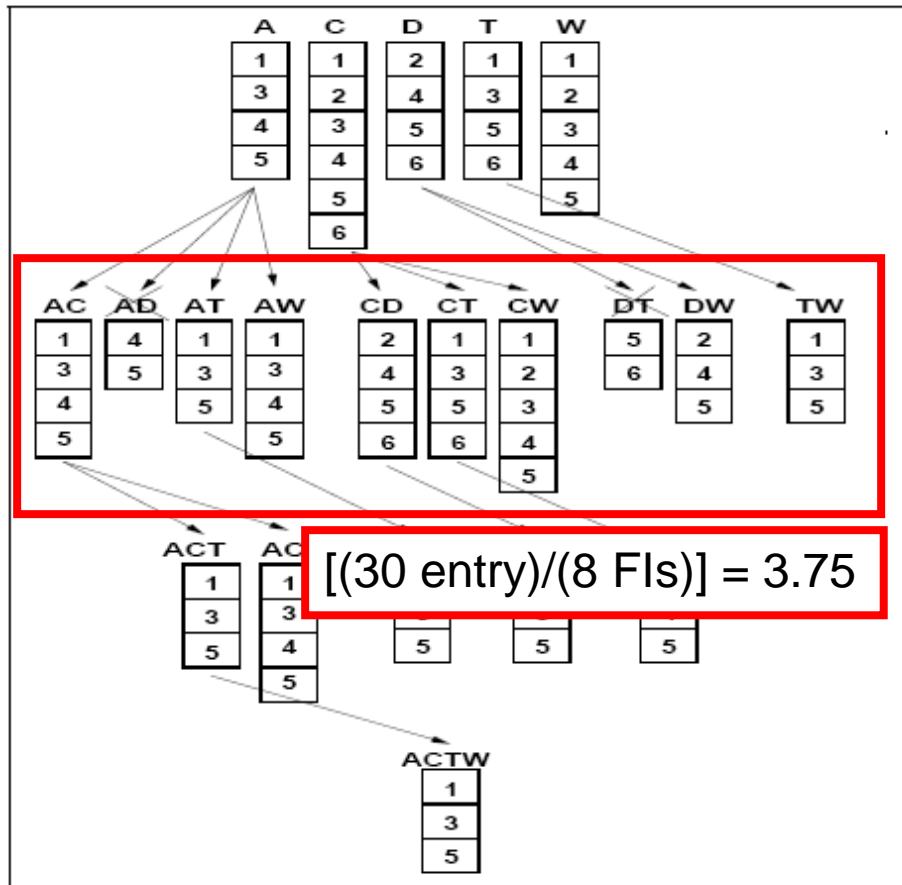
# 5. Sử dụng Diffset (6)

## (So sánh Tidsets và Diffsets)



# 5. Sử dụng Diffset (5)

(So sánh Tidsets và Diffsets)



# 5. Sử dụng Diffset (7)

---

Thông thường trong quá trình thực thi sẽ có một điểm giao để chuyển đổi giữa tidset và diffset.

- ❖ Sử dụng diffset đối với CSDL dày đặc.
- ❖ Bắt đầu với tidset đối với CSDL thưa thớt, và ở các giai đoạn sau có thể chuyển sang diffset.

**⇒ Khi nào nên chuyển đổi giữa tidset và diffset?**

# 5. Sử dụng Diffset (7)

---

## Giảm tỷ lệ (Reduction Ratio)

❖ Cho lớp  $P$

Gọi  $PX$  và  $PY$  là lớp thành viên với  $t(PX)$  và  $t(PY)$

Xét Itemset mới  $PXY$  trong lớp  $PX$

$PXY$  có thể được lưu trữ  $t(PXY)$  hoặc  $d(PXY)$

❖ Định nghĩa: giảm tỷ lệ  $r = t(PXY) / d(PXY)$

Đối với diffset sẽ có lợi nếu như  $r \geq 1$

hoặc  $t(PXY) / d(PXY) \geq 1$

## 5. Sử dụng Diffset (7)

$$r = t(PXY) / d(PXY)$$

- Thay  $d(PXY) \Rightarrow t(PXY) / (t(PX) - t(PY)) \geq 1$
- Khi  $t(PX) - t(PY) = t(PX) - t(PXY)$

Ta có  $t(PXY) = (t(PX) - t(PXY))$

- Chia cho  $t(PXY)$  được  $\frac{1}{\frac{t(PX)}{t(PXY)} - 1} \geq 1$
- Sau khi đơn giản được  $t(PX) / t(PXY) \leq 2$

**$\Rightarrow$  Điều đó có nghĩa là nếu độ hỗ trợ của  $PXY$  bằng ít nhất  $1/2$  của  $PX$  thì ta chuyển sang sử dụng diffset.**

## 6. Thuật toán dEclat

---

- ❖ Thuật toán áp dụng diffset vào phương pháp Eclat (state-of-the-art) trước đó sử dụng tidset.
- ❖ Thuật toán duyệt theo chiều sâu trước (DFS). Bắt đầu với các diffset của các items phổ biến.
- ❖ Vòng lặp đệ quy để tìm tất cả các itemset phổ biến ở cấp hiện tại. Tiến trình lặp lại cho đến khi tất cả các itemset phổ biến được khai thác.

# 6. Thuật toán dEclat

---

0. **dEclat**( $[P]$ ):

1. **for** all  $X_i \in [P]$  **do**

2.     **for** all  $X_j \in [P]$ , with  $j > i$  **do**

3.          $R = X_i \cup X_j$  ;

4.          $d(R) = d(X_j) - d(X_i)$ ;

5.         **if**  $\sigma(R) \geq \textit{min\_sup}$  **then**

6.              $T_i = T_i \cup \{R\}$ ; //  $T_i$  initially empty

7.     **if**  $T_i \neq \emptyset$  **then** **dEclat**( $T_i$ );



## 6. Thuật toán dCharm

---

- ❖ Thuật toán áp dụng diffset vào thuật toán Charm thay vì sử dụng tidset.
- ❖ Thuật toán duyệt theo chiều sâu trước (DFS). Bắt đầu với các diffset của các items phổ biến.
- ❖ Thuật toán sử dụng các bước tỉa nhánh dựa vào mối quan hệ của các tập con

# 6. Thuật toán dCharm

---

0. **dCharm**([ $P$ ]):

1. **for** all  $X_i \in [P]$  **do**

2.     **for** all  $X_j \in [P]$ , with  $j > i$  **do**

3.          $R = X_i \cup X_j$  ;

4.          $d(R) = d(X_j) - d(X_i)$ ;

5.         **if**  $\sigma(R) \geq \text{min\_sup}$  **then**

6.             **if**  $d(X_i) = d(X_j)$  **then**

7.                 Remove  $X_j$  from [ $P$ ];

8.                 Replace all  $X_i$  with  $R$ ;

9.             **else if**  $d(X_i) \supset d(X_j)$  **then**

10.                 Replace all  $X_i$  with  $R$ ;

11.             **else if**  $d(X_i) \subset d(X_j)$  **then**

12.                 Remove  $X_j$  from [ $P$ ]

13.                 Add  $R$  to  $NewN$ ;

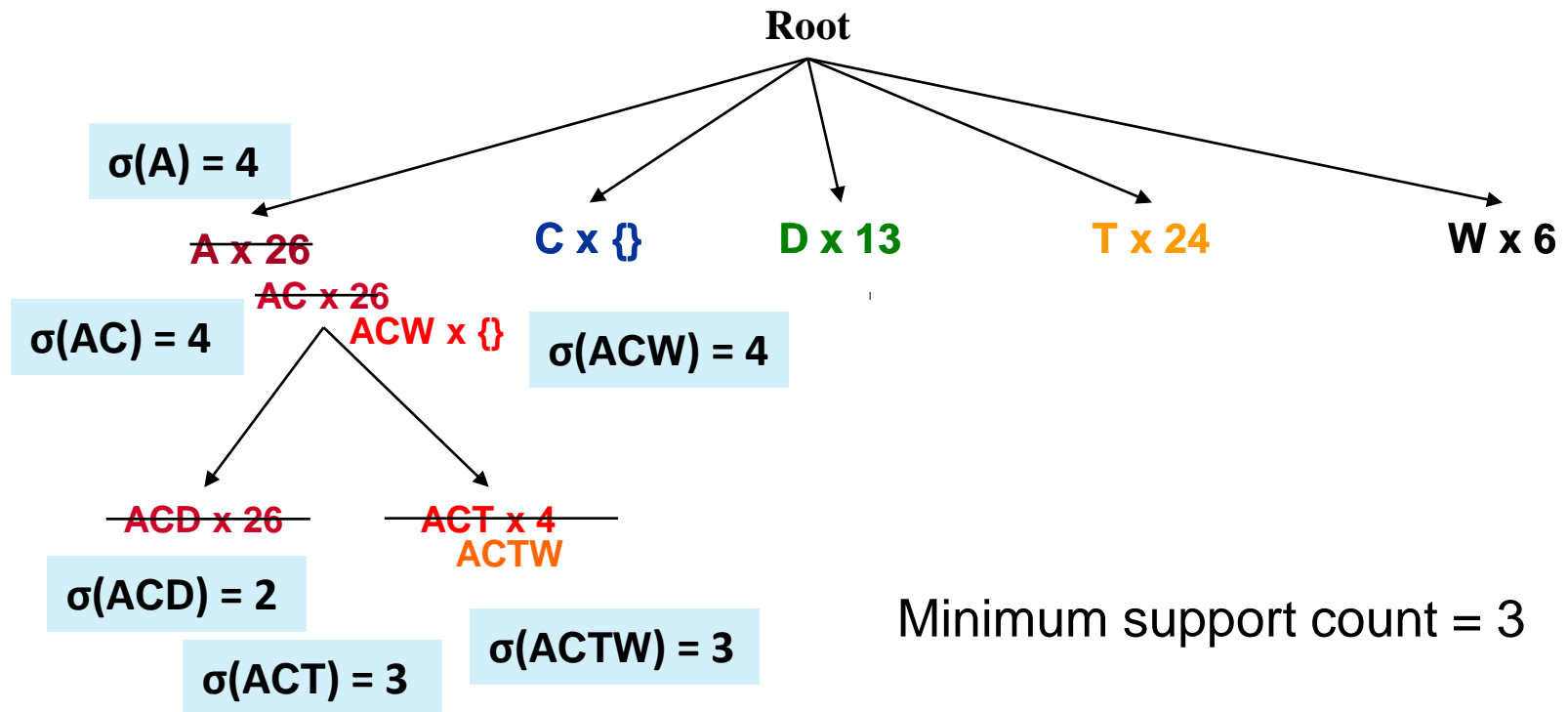
14.             **else if**  $d(X_i) \neq d(X_j)$  **then**

15.                 Add  $R$  to  $NewN$ ;

16.     Subsumption-Check( $C, R$ );

17. **if**  $NewN \neq \emptyset$  **then** **dCharm**( $NewN$ );

# 6. Thuật toán dCharm



# 6. Thuật toán dGenMax

- ❖ Thuật toán sử dụng tiến trình **backtrack** để tìm kiếm mẫu tối đại.
- ❖ Các cải tiến
  - Sắp item theo thứ tự **tăng dần** theo kích thước và độ *support* (i. đầu tiên khám phá item có kích thước nhỏ trước, ii. Bỏ một node càng sớm càng tốt trong cây tìm kiếm).
  - Kiểm tra các tập bao (**superset**) của itemset đang xét.
  - **CSDL theo chiều dọc** tối ưu việc kiểm tra phổ biến bằng cách sử dụng tidset, hoặc cải tiến hơn nữa là **diffsets**.
- ❖ Bộ nhớ
  - Lưu trữ nhiều nhất  $k = m + l$  tidsets (diffsets) trong bộ nhớ, với  $m$  là độ dài của tập kết hợp dài nhất và  $l$  là độ dài của itemset tối đại có chiều dài lớn nhất.

# 6. Thuật toán dGenMax

**GenMax** (Dataset  $T$ ):

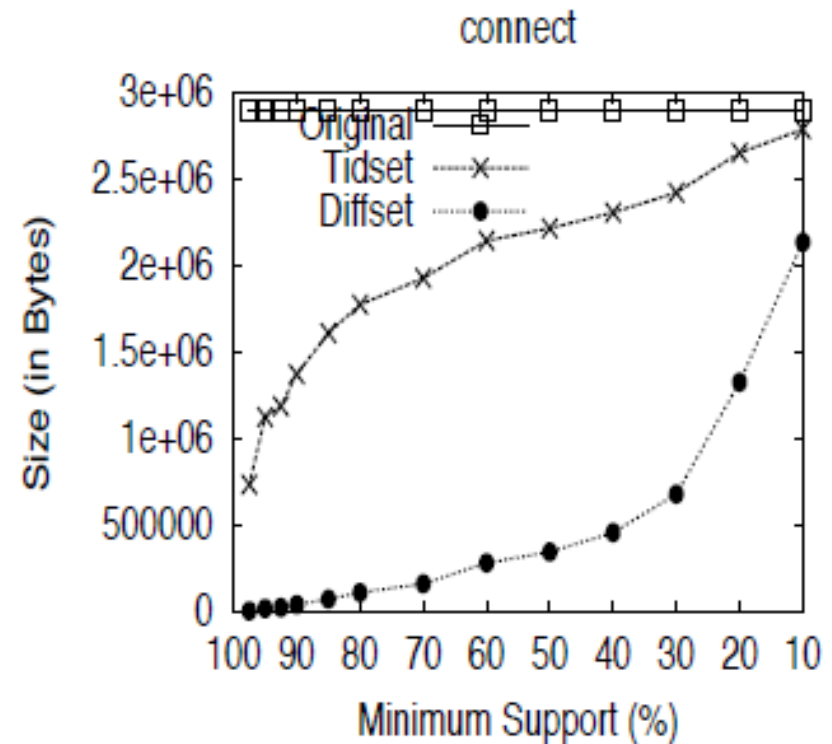
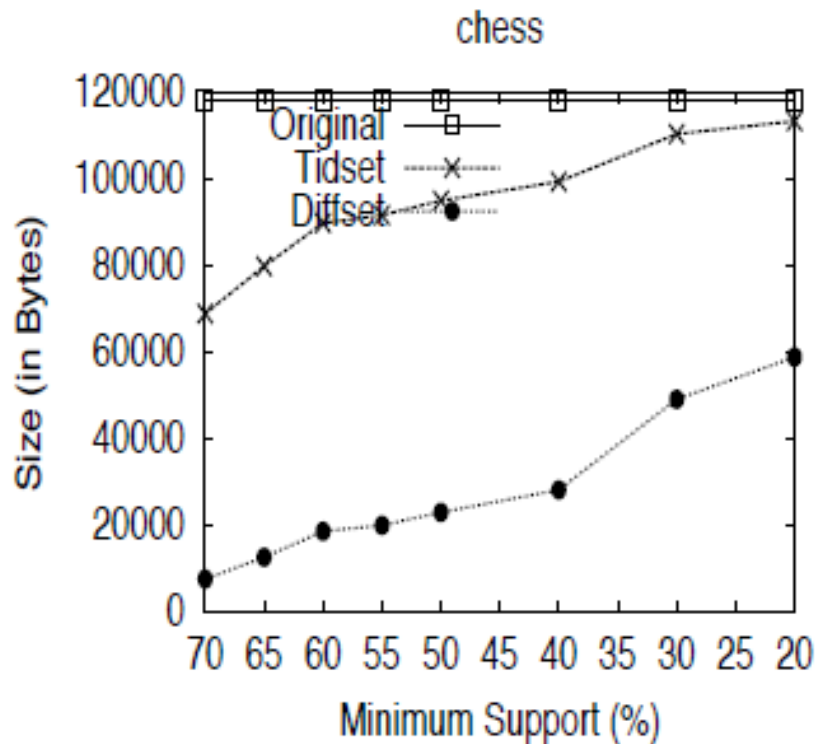
1. Calculate  $F_1$ , Calculate  $F_2$ .
2. For each item  $i \in F_1$  calculate  $c(i)$ , its combine-set.
3. Sort items in  $F_1$  in INCREASING cardinality of  $c(i)$  and then INCREASING  $\sigma(i)$ .
4. Sort each  $c(i)$  in order of  $F_1$ .
5.  $c(i) = c(i) - \{j : j < i \text{ in sorted order of } F_1\}$ .
6.  $M = \{\}$ ; // Maximal Frequent Itemsets.
7. **for each**  $i \in F_1$  **do**
8.    $Z = \{x \in M : i \in x\}$
9.   **for each**  $j \in c(i)$  **do**
10.      $H = \{x : x \text{ is } j \text{ or } x \text{ follows } j \text{ in } c(i)\}$
11.     **if**  $H$  has a super set in  $Z$  **then break**
12.      $I = \{i, j\}$
13.      $X = c(i) \cap c(j); d(X) = t(i) - t(j)$
14.      $Y = \{x \in Z : j \in x\}$
15.      $\text{Extend}(I, X, Y)$
16.      $Z = Z \cup Y$
17.    $M = M \cup Z$
18. **Return**  $M$

**Procedure**  $\text{Extend}(I, X, Y)$

//  $I$  is the itemset to be extended,  $X$  is the set of items  
// that can be added to  $I$ , i.e., the combine set, and  
//  $Y$  is the set of relevant maximal itemsets found so far  
// i.e., all maximal itemsets which contain  $I$ .

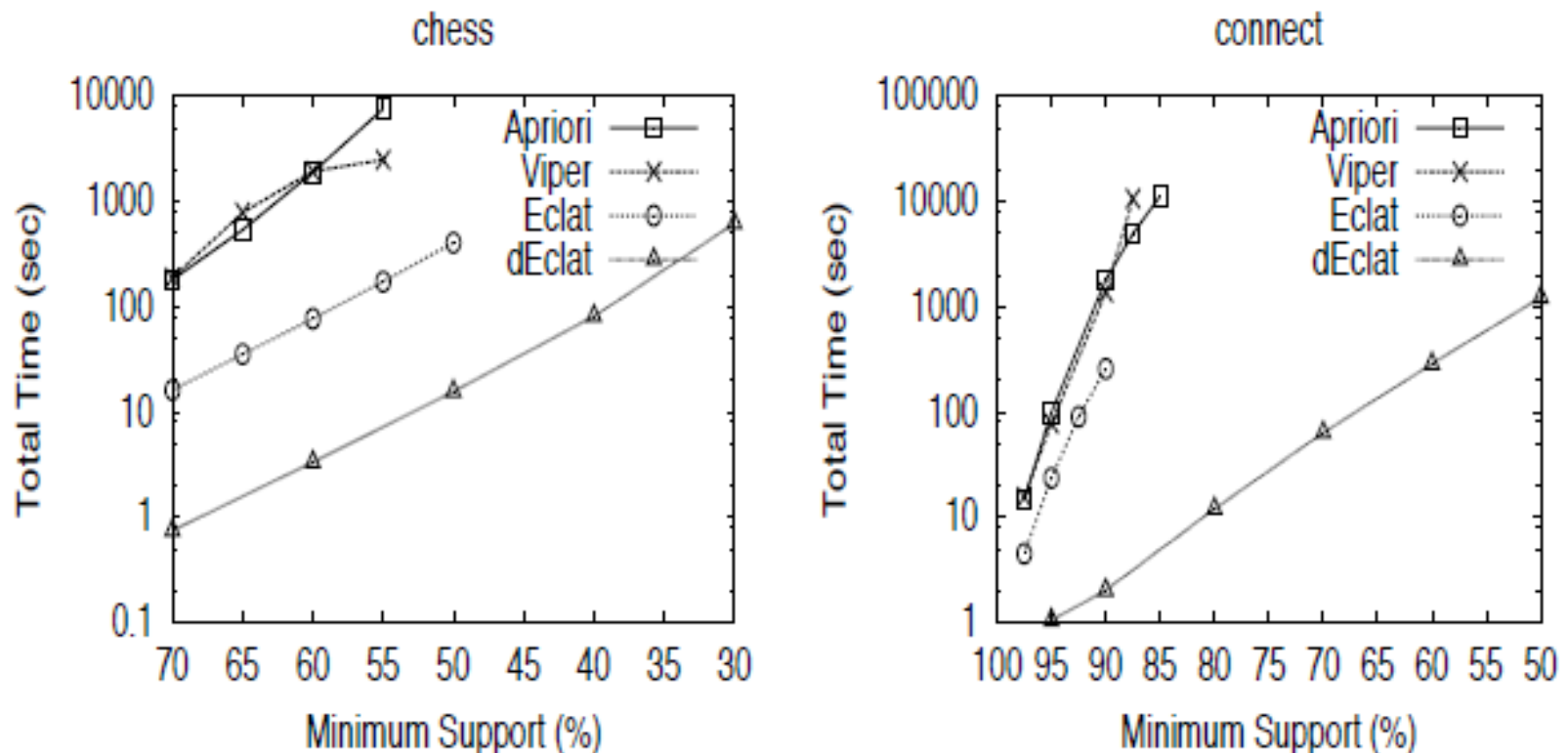
1.  $\text{extendflg} = 0$
2. **for each**  $j \in X$  **do**
3.   **if**  $|Y| > 0$  **then**
4.      $G = \{x : x \text{ is } j \text{ or } x \text{ follows } j \text{ in } X\}$
5.     **if**  $G$  has super set in  $Y$  **Then**
6.        $\text{extendflg} = 1$ ; **break**;
7.    $\text{NewI} = I \cup \{j\}; d(\text{NewI}) = d(j) - d(I)$
8.   **if** ( $\text{NewI}$  is frequent) **then**
9.      $\text{NewX} = X \cap c(j)$
10.      $\text{extendflg} = 1$
11.     **if** ( $\text{NewX} == \phi$ ) **then**
12.        $Y = Y \cup \{\text{NewI}\}$
13.     **else**
14.        $\text{NewY} = \{x \in Y : j \in x\}$
15.        $\text{Extend}(\text{NewI}, \text{NewX}, \text{NewY})$
16.      $Y = Y \cup \{\text{NewY}\}$
17. **if** ( $\text{extendflg} == 0$  and  $|Y| == 0$ ) **then**
18.    $Y = Y \cup \{I\}$

# 7. So sánh Tidset và Diffset



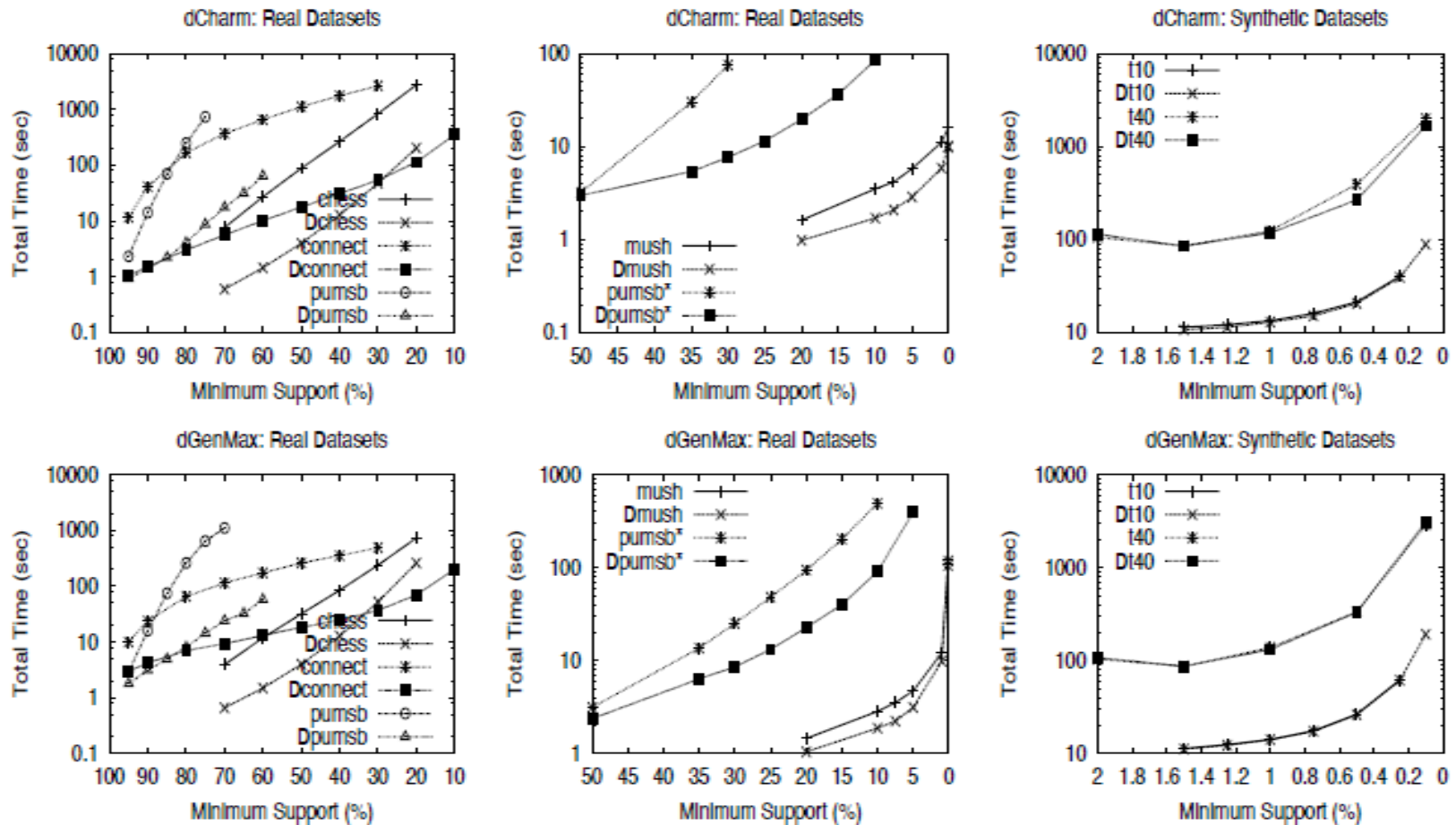
***So sánh bộ nhớ sử dụng  
trên dữ liệu dày đặc Chess và Connect***

# 7. So sánh Tidset và Diffset



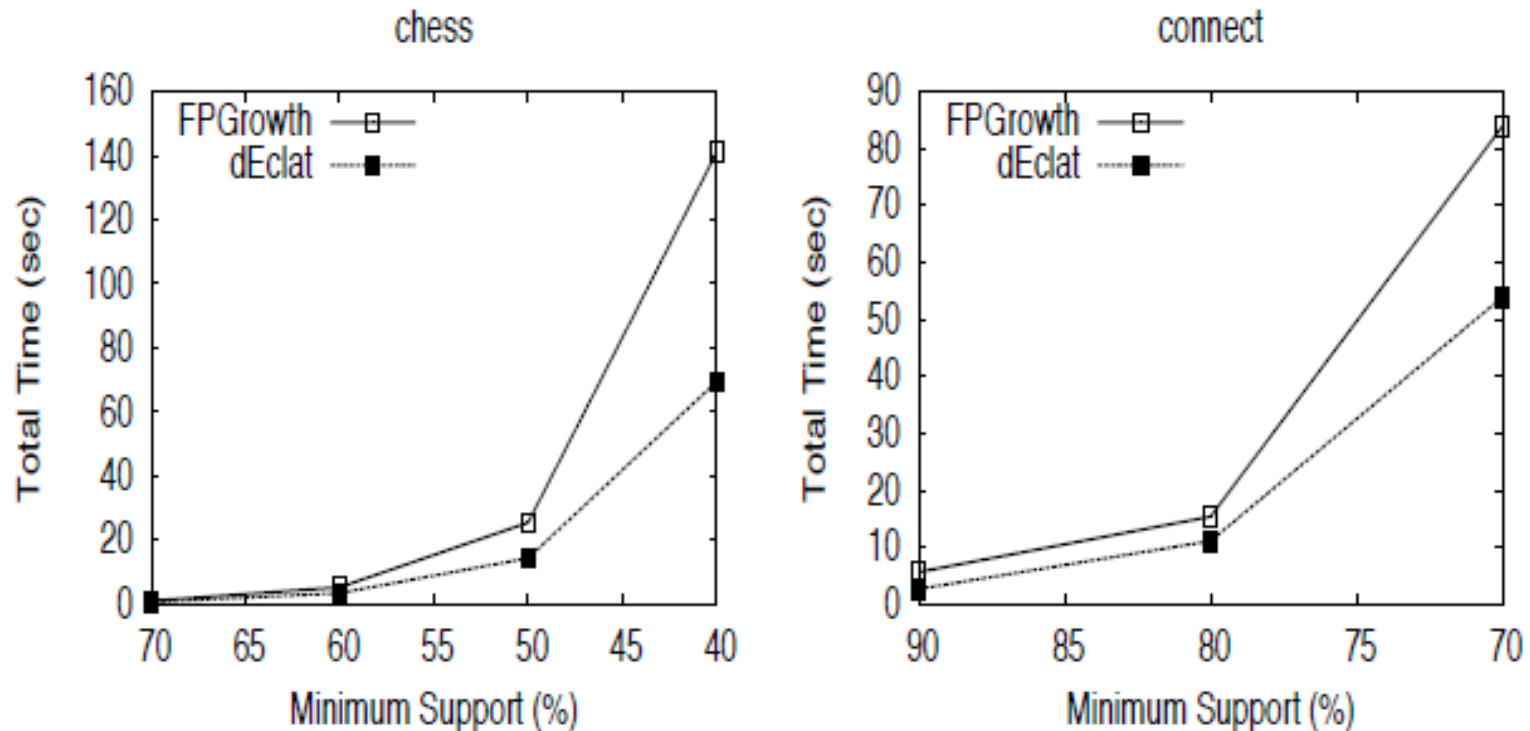
***So sánh thời gian thực thi  
trên dữ liệu dày đặc Chess và Connect***

# 7. So sánh Tidset và Diffset



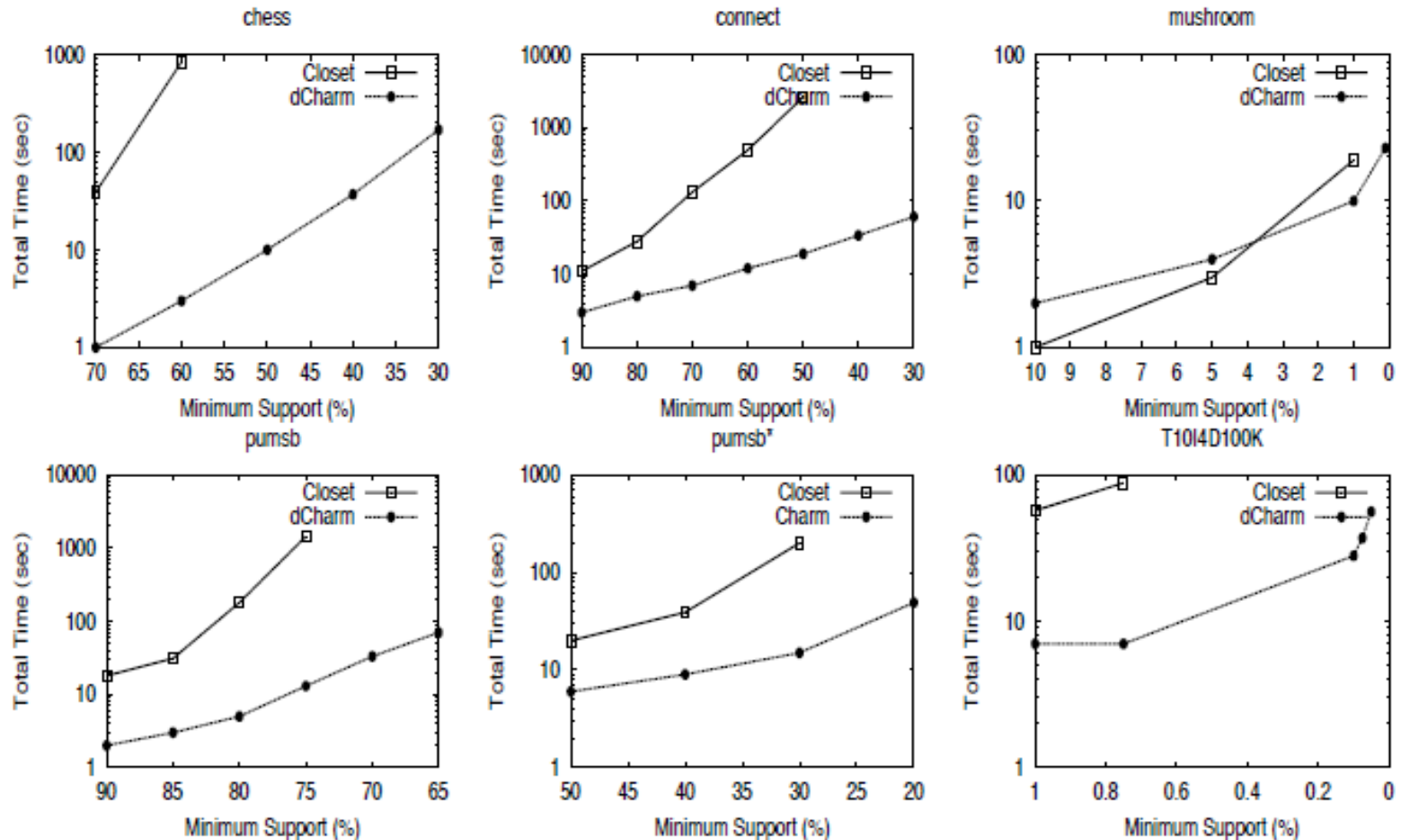


# 7. So sánh Tidset và Diffset

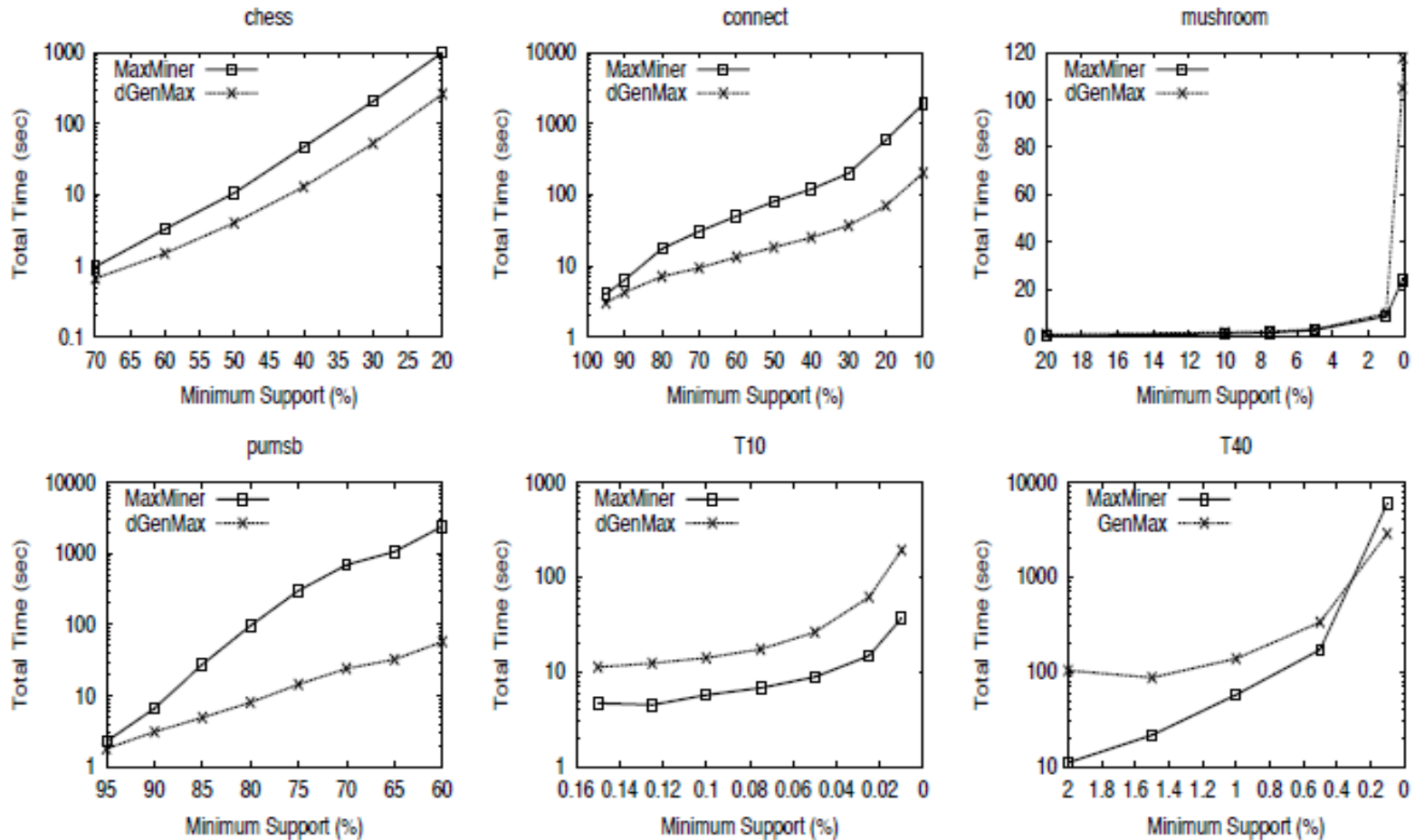


***So sánh thời gian thực thi  
giữ thuật toán FPGrowth và dEclat***

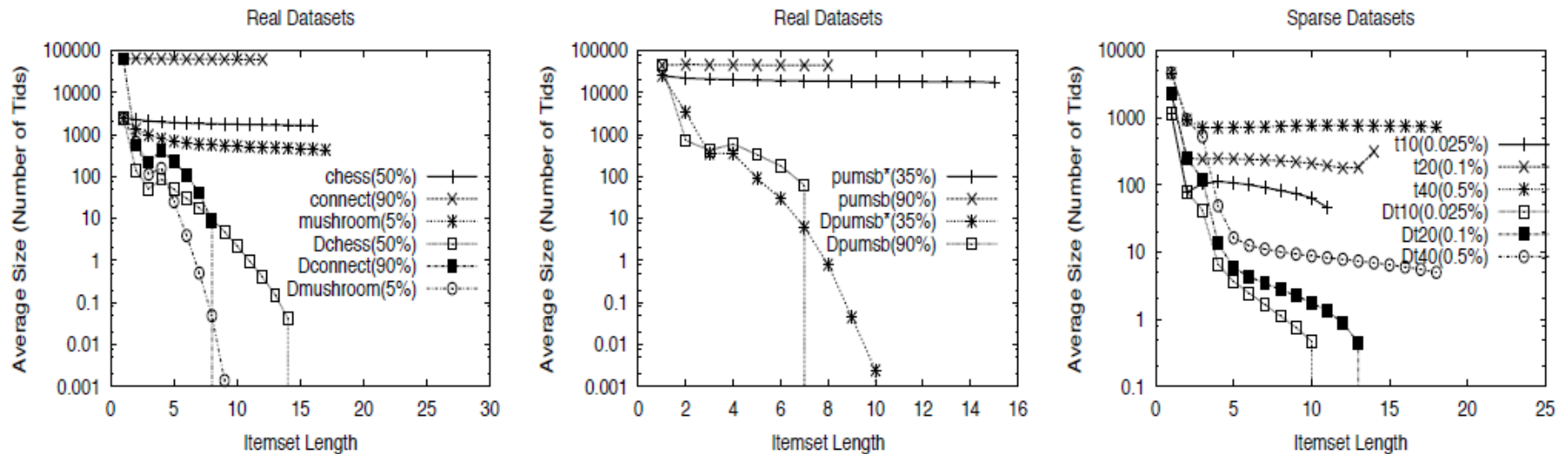
# 7. So sánh Tidset và Diffset



# 7. So sánh Tidset và Diffset



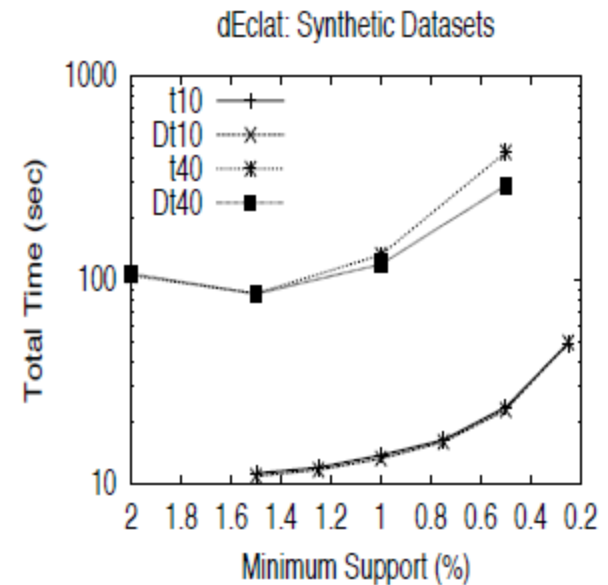
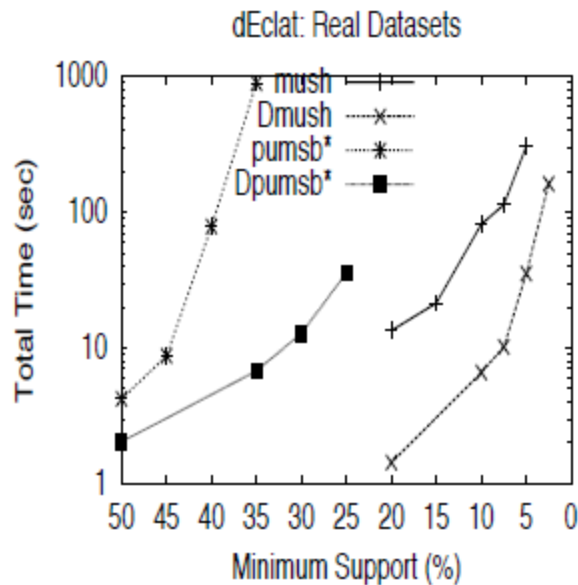
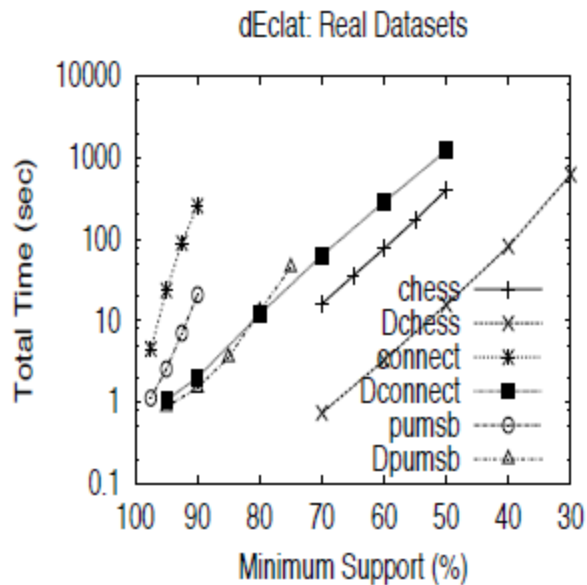
# 7. So sánh Tidset và Diffset



| Database    | <i>min_sup</i> | Max Length | Avg. Diffset Size | Avg. Tidset Size | Reduction Ratio |
|-------------|----------------|------------|-------------------|------------------|-----------------|
| chess       | 0.5%           | 16         | 26                | 1820             | 70              |
| connect     | 90%            | 12         | 143               | 62204            | 435             |
| mushroom    | 5%             | 17         | 60                | 622              | 10              |
| pumsb*      | 35%            | 15         | 301               | 18977            | 63              |
| pumsb       | 90%            | 8          | 330               | 45036            | 136             |
| T10I4D100K  | 0.025%         | 11         | 14                | 86               | 6               |
| T20I16D100K | 0.1%           | 14         | 31                | 230              | 11              |
| T40I10D100K | 0.5%           | 18         | 96                | 755              | 8               |

**Kích thước trung bình vòng lặp: Tidset vs Diffset**

# 7. So sánh Tidset và Diffset



**So sánh thời gian thực thi trên CSDL dày đặc (Real datasets) và thưa thớt (Synthetic datasets)**

## 8. Nhận xét

---

- ❖ Diffset giảm đáng kể kích thước bộ nhớ cần để lưu trữ trực tiếp kết quả.
- ❖ Diffset tăng hiệu quả thực thi khi đưa vào phương pháp khai thác dữ liệu theo chiều dọc.
- ❖ Diffset cung cấp tầm quan trọng về cải tiến hiệu suất so với các phương pháp tốt trước đó.

# Tài liệu tham khảo

---

- [1] M.J. Zaki and K. Gouda, ***Fast Vertical Mining Using Diffsets***, Proc. Ninth ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining, Aug. 2003.
- [2] M. J. Zaki, ***Scalable Algorithms for Association Mining***, IEEE Transactions on Knowledge and Data Engineering, 12(3), May/Jun 2000, pp. 372-390.
- [3] M. J. Zaki and C.-J. Hsiao, ***Efficient Algorithms for Mining Closed Itemsets and their Lattice Structure***. IEEE Transactions on Knowledge and Data Engineering, 17(4), Apr 2005, pp. 462-478.
- [4] M. J. Zaki and K. Gouda, ***GenMax: An Efficient Algorithm for Mining Maximal Frequent Itemsets***. Data Mining and Knowledge Discovery: An International Journal, 11(3), 2005, pp. 223-242.

Thanks for your listening !!  
Q & A