TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN
KHOA HỆ THỐNG THÔNG TIN

Tài liệu bài giảng:

KHAI THÁC DỮ LIỆU – IS252

Chương 4:

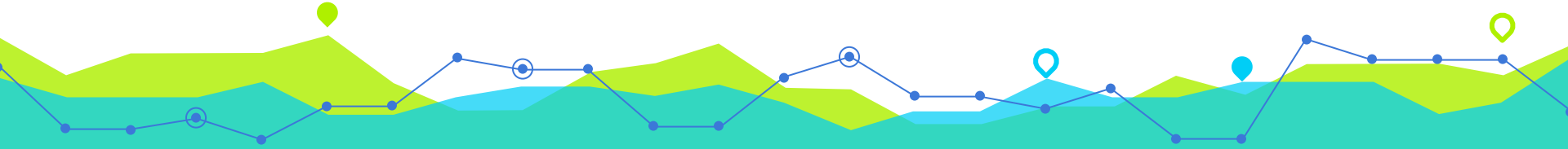
Dãy phổ biến (Episode)

ThS. Dương Phi Long – Email: longdp@uit.edu.vn

NỘI DUNG BÀI HỌC

01 Các khái niệm

02 Phương pháp WINEPI



1 Các khái niệm

1. Luật Episode
2. Dữ liệu cho bài toán
3. Dãy phổ biến (Episode)

1. Luật Episode

- Luật kết hợp trong bài toán dùng Episode mô tả các sự kiện xuất hiện cùng nhau trong dữ liệu.
- Các luật Episode mô tả quan hệ thời gian giữa các sự vật, hiện tượng.
- **VD:**

IF một tổ hợp các tín hiệu báo nguy xảy ra trong một khoảng thời gian **THEN** sẽ có một tổ hợp các tín hiệu báo nguy khác sẽ xảy ra trong một khoảng thời gian xác định khác

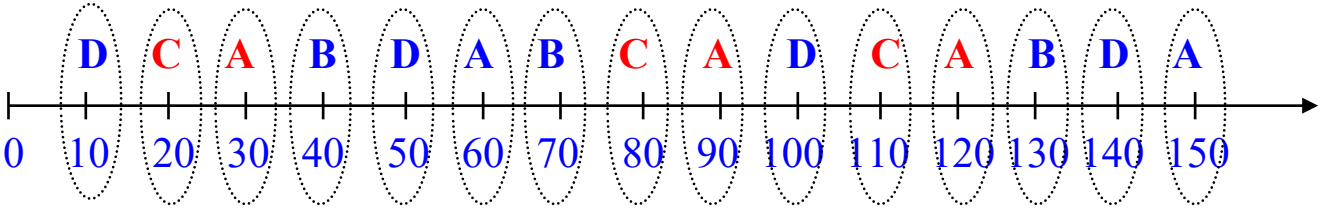
"Thâm đông, hồng tây, dựng may,
Ai ơi đợi đến ba ngày hãy đi"

2. Dữ liệu bài toán

- Tập **R** các **loại sự kiện, loại biến cố**. VD: R={A,B,C,D}
- Mỗi **sự kiện** là một cặp **(A, t)** với
  - **A**: loại sự kiện, $A \in R$
  - **t**: thời điểm xuất hiện của loại sự kiện, số nguyên
- **Chuỗi sự kiện S** trên R là bộ ba **(s, T_s, T_e)**
  - T_s : thời điểm bắt đầu chuỗi sự kiện, số nguyên
  - T_e: thời điểm kết thúc chuỗi sự kiện, số nguyên
  - $T_s < T_e$
  - $s = \langle (A_1, t_1), (A_2, t_2), \ldots, (A_n, t_n) \rangle$
  - $A_i \in R$ và $T_s \leq t_i < T_e$ với $i = 1, \ldots, n$

2. Dữ liệu bài toán

VD1: Cho 1 chuỗi tín hiệu báo động



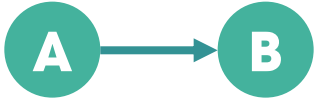
- *A, B, C, D:* các loại sự kiện báo động
- *10, 20, …, 150:* các thời điểm xảy ra
- $s = \langle (D, 10), (C, 20), \ldots, (A, 150) \rangle$
- Thời điểm bắt đầu: $T_s = 10$
- Thời điểm kết thúc: $T_e = 150$

3. Dãy phổ biến (Episode)

- **Episode**: cặp (V, ≤)
  - V: tập hợp các loại sự kiện. VD: loại tín hiệu báo động
  - ≤: thứ tự riêng phần trên V
- Episodes: Chứa các tín hiệu báo động có các tính chất nào đó và xảy ra theo một thứ tự riêng phần nào đó.

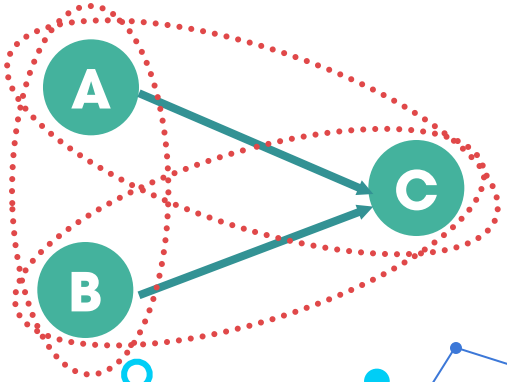
3. Dãy phổ biến (Episode)

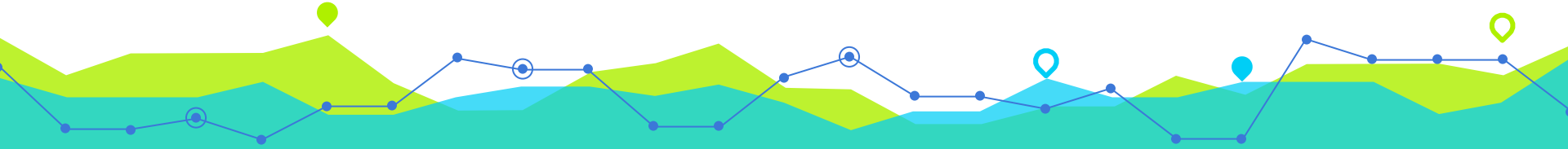
- **Phân loại**:
  - Episode tuần tự (có thứ tự)



  - Episode song song

  - Episode vừa song song vừa tuần tự





2 Phương pháp WINEPI

1. Cách tiếp cận & nguyên tắc
2. Tìm Episode phổ biến
3. Luật Episode và độ tin cậy

1. WINEPI: Cách tiếp cận & nguyên tắc

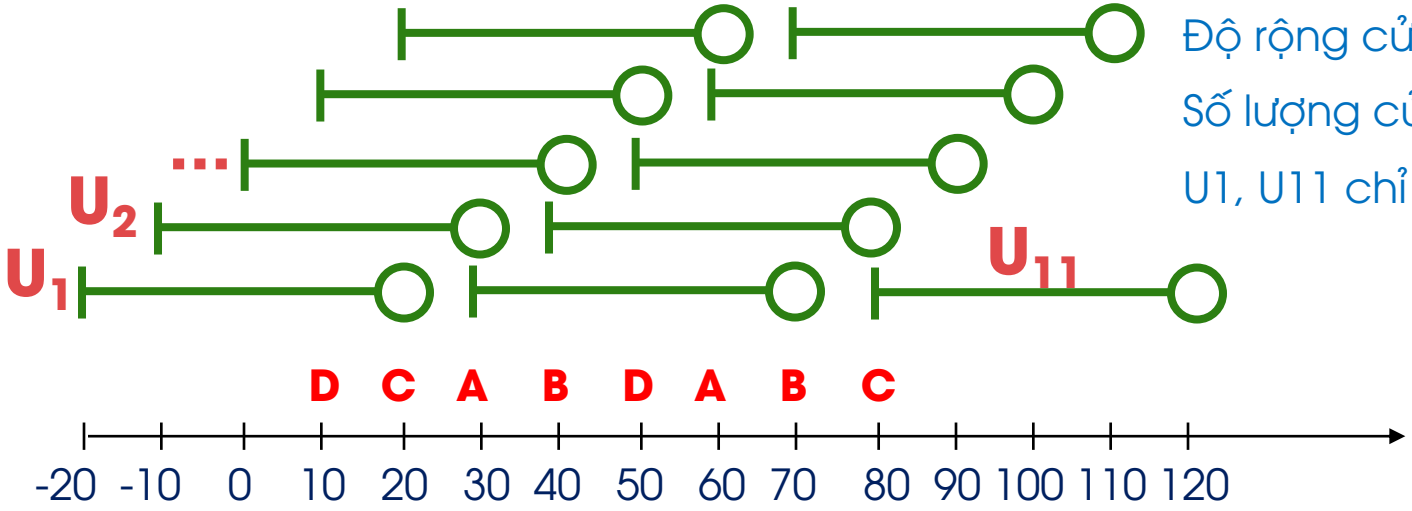
- **Cách tiếp cận:** Kỹ thuật sử dụng cửa sổ trượt
  - Cửa sổ được trượt qua chuỗi dữ liệu các sự kiện
  - Mỗi cửa sổ là một "khung ảnh" giống như một dòng của CSDL
  - Tập các "khung ảnh" tạo thành các dòng của CSDL

1. WINEPI: Cách tiếp cận & nguyên tắc

- **Nguyên tắc:**
  - Cửa sổ có độ rộng cố định
  - Cửa sổ đầu tiên (W1) chỉ chứa 1 sự kiện đầu tiên
  - Cửa sổ trượt sang phải lần lượt từng sự kiện
  - Cửa sổ cuối (Wcuối) chỉ chứa 1 sự kiện cuối cùng

1. WINEPI: Cách tiếp cận & nguyên tắc

VD2: Chuỗi tín hiệu báo động



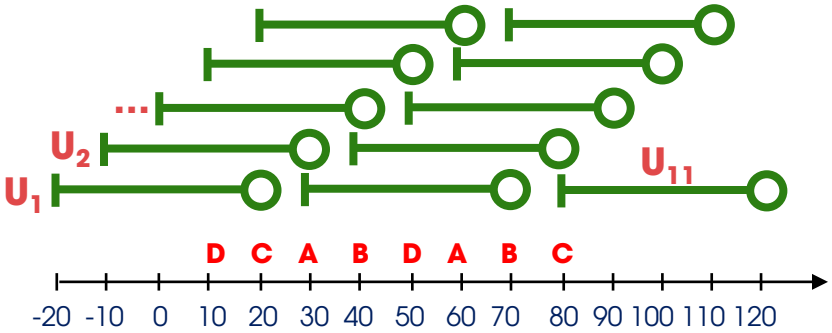
Độ rộng cửa sổ: 40 giây

Số lượng cửa sổ: 11 U_1, U_2,…, U_{11}

U1, U11 chỉ có 1 sự kiện

1. WINEPI: Cách tiếp cận & nguyên tắc

VD2: Chuỗi tín hiệu báo động



Cửa số U_i	Nội dung cửa sổ
U_1[-20,20)	_,_,_,D
U_2[-10,30)	_,_,D,C
U_3[0,40)	_,D,C,A
U_4[10,50)	D,C,A,B
U_5[20,60)	C,A,B,D
U_6[30,70)	A,B,D,A
U_7[40,80)	B,D,A,B
U_8[50,90)	D,A,B,C
U_9[60,100)	A,B,C,_
U_{10}[70,110)	B,C_,_
U_{11}[80,120)	C,_,_,_

2. WINEPI: Tìm Episode phổ biến

- **Tìm Episode phổ biến:**
  - Tìm các Episode theo độ rộng của cửa sổ trượt
  - Tính độ phổ biến của từng Episode
  - Episode phổ biến là Episode thỏa ngưỡng min_fr cho trước

2. WINEPI: Tìm Episode phổ biến

- **Độ phổ biến (tần suất) của Episode α :**

$$fr(\alpha, S, W) = \frac{|S_w \in W(S,W):\ \alpha \text{ xuất hiện trong } S_w|}{|W(S,W)|} \quad (1)$$

$$= \frac{\text{Số cửa sổ chứa episode } \alpha}{\text{Tổng số cửa sổ của chuỗi } S}$$

α: episode

S: chuỗi các sự kiện

W: bề rộng cửa sổ trượt

S_w: cửa sổ của chuỗi S

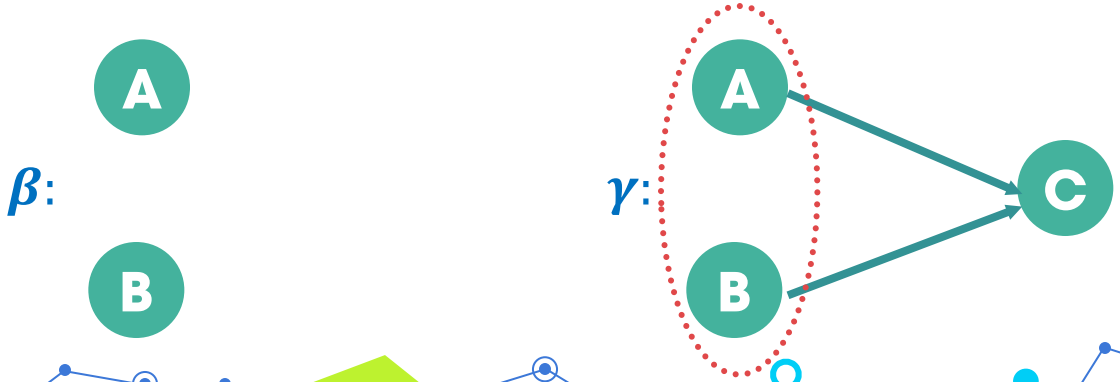
$W(S,W)$: tập các cửa sổ S_w của chuỗi S

2. WINEPI: Tìm Episode phổ biến

- Episode α là phổ biến nếu $fr(\alpha, S, W) \geq min\_fr$
- F(S, *W, min_fr*): tập hợp các episodes phổ biến trong chuỗi sự kiện S ứng với độ rộng *W* và ngưỡng *min_fr*
- Apriori: Nếu episode α là phổ biến trong chuỗi sự kiện S, thì tất cả các episodes con $\beta \prec \alpha$ là phổ biến.

3. WINEPI: Luật Episode và độ tin cậy

- Luật episode là biểu thức $\boldsymbol{\beta \rightarrow \gamma}$, với:
  - $\boldsymbol{\beta, \gamma}$: episode
  - $\boldsymbol{\beta}$ là episode con $\boldsymbol{\gamma}$ ($\boldsymbol{\beta \prec \gamma}$)
- $\beta \prec \gamma$: đồ thị biểu diễn β là con của đồ thị biểu diễn γ



3. WINEPI: Luật Episode và độ tin cậy

- **Độ tin cậy (conf):** Xác suất điều kiện của toàn bộ của γ xảy ra trong cửa sổ khi β xảy ra trước trong cửa sổ đó.

$$conf(\beta \rightarrow \gamma) = P(\gamma|\beta) = \frac{P(\beta \cup \gamma)}{P(\beta)} = \frac{fr(\square \cup \gamma, S, W)}{fr(\square, S, W)} \quad (2)$$

$$= \frac{\text{độ } phổ\ biến \text{ của } toàn\ bộ\ episode\ trong\ luật}{\text{độ } phổ\ biến \text{ của } episode\ vế\ trái\ trong\ luật}$$

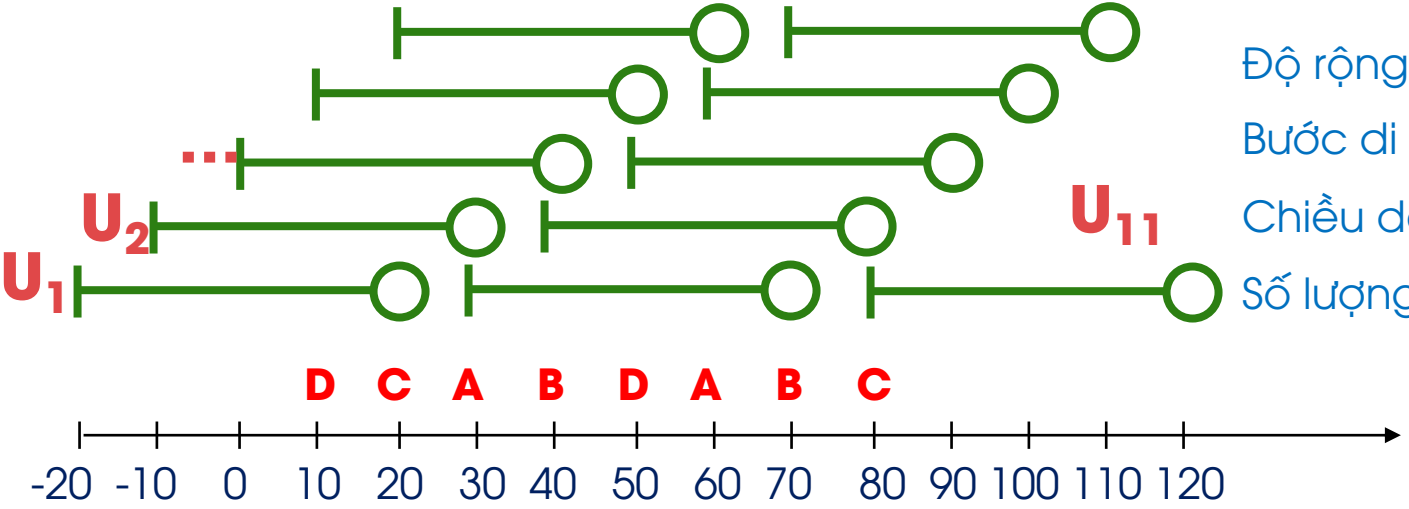
3. WINEPI: Luật Episode và độ tin cậy

- Các luật Episode giống luật kết hợp nhưng có thêm yếu tố thời gian
- Nếu sự kiện thỏa vế trái của luật xuất hiện theo thứ tự bên phải trong phạm vi W đơn vị thời gian, thì cũng xuất hiện trong phần kết luận (vế phải của luật), nó xuất hiện ở vị trí được mô tả bởi quan hệ thứ tự ≤, trong phạm vi W đơn vị thời gian.
- **Ký hiệu:** $\boldsymbol{\beta \rightarrow \gamma\ [W]\left(fr(\gamma, S, W), conf(\beta \rightarrow \gamma)\right)}$

WINEPI: Tìm Episode phổ biến song song

VD3: Chuỗi tín hiệu báo động. Giả sử W = 40, min_fr= 40%.

Tìm Episode phổ biến song song và một số luật Episode



Độ rộng cửa sổ: 40 giây

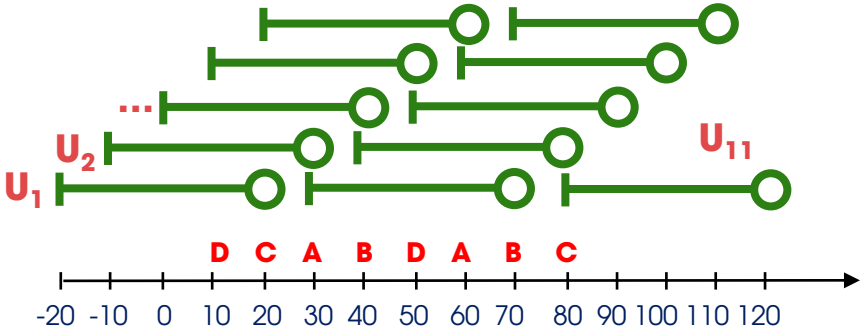
Bước di chuyển: 10 giây

Chiều dài chuỗi: 70 (10, …, 80)

Số lượng cửa sổ: 11 U_1, U_2,…, U_{11}

WINEPI: Tìm Episode phổ biến song song

VD3:



Cửa số U_i	Nội dung cửa sổ
U_1[-20,20)	_,_,_,D
U_2[-10,30)	_,_,D,C
U_3[0,40)	_,D,C,A
U_4[10,50)	D,C,A,B
U_5[20,60)	C,A,B,D
U_6[30,70)	A,B,D,A
U_7[40,80)	B,D,A,B
U_8[50,90)	D,A,B,C
U_9[60,100)	A,B,C,_
U_{10}[70,110)	B,C_,_
U_{11}[80,120)	C,_,_,_

WINEPI: Tìm Episode phổ biến song song

VD3:

- min_fr = 40% => Episode xảy ra **> 4** trong 11 cửa sổ (tần số)
- Các Episode song song có 1 phần tử: A (7), B (7) , C (8), D (8)
- Các Episode phổ biến song song có 1 phần tử: A, B, C, D

Cửa số U_i	Nội dung cửa sổ
U_1[-20,20)	_,_,_,D
U_2[-10,30)	_,_,D,C
U_3[0,40)	_,D,C,A
U_4[10,50)	D,C,A,B
U_5[20,60)	C,A,B,D
U_6[30,70)	A,B,D,A
U_7[40,80)	B,D,A,B
U_8[50,90)	D,A,B,C
U_9[60,100)	A,B,C,_
U_{10}[70,110)	B,C_,_
U_{11}[80,120)	C,_,_,_

WINEPI: Tìm Episode phổ biến song song

VD3:

- Từ các Episode phổ biến trên, tìm Episodes song song có 2 phần tử: AB (6), AC (5), AD (6), BC (5), BD (5), CD (5)
- Các Episode phổ biến song song có 2 phần tử: AB, AC, AD, BC, BD, CD

Cửa số U_i	Nội dung cửa sổ
U_1[-20,20)	_,_,_,D
U_2[-10,30)	_,_,D,C
U_3[0,40)	_,D,C,A
U_4[10,50)	D,C,A,B
U_5[20,60)	C,A,B,D
U_6[30,70)	A,B,D,A
U_7[40,80)	B,D,A,B
U_8[50,90)	D,A,B,C
U_9[60,100)	A,B,C,_
U_{10}[70,110)	B,C_,_
U_{11}[80,120)	C,_,_,_

WINEPI: Tìm Episode phổ biến song song

VD3:

- Từ các Episode phổ biến trên, tìm Episodes song song có 3 phần tử: ABC (4), ABD (5), ACD (4), BCD (3)
- Các Episode phổ biến song song có 3 phần tử: ABD
- Không có Episode có 4 phần tử

Cửa sổ U_i	Nội dung cửa sổ
U_1[-20,20)	_,_,_,D
U_2[-10,30)	_,_,D,C
U_3[0,40)	_,D,C,A
U_4[10,50)	D,C,A,B
U_5[20,60)	C,A,B,D
U_6[30,70)	A,B,D,A
U_7[40,80)	B,D,A,B
U_8[50,90)	D,A,B,C
U_9[60,100)	A,B,C,_
U_{10}[70,110)	B,C_,_
U_{11}[80,120)	C,_,_,_

WINEPI: Tìm Episode phổ biến song song

VD3: Các Episode phổ biến song song và độ phổ biến (Tần suất)

Episode PBSS	fr(Episode)
D	8/11 = 73%
C	8/11 = 73%
A	7/11 = 64%
B	7/11 = 64%
DC	5/11 = 45%
DA	6/11 = 55%
DB	5/11 = 45%
CA	5/11 = 45%
CB	5/11 = 45%
AB	6/11 = 55%
DAB	5/11 = 45%

Episode phổ biến tối đại: DAB

Một số luật Episode:

1. **Xét D → A:**

$$fr([DA]) = 55\%$$

$$conf(\mathrm{D} \rightarrow \mathrm{A}) = \frac{fr([DA])}{fr([D])} = \frac{55}{73} = 75\%$$

$$\boldsymbol{D \rightarrow A\ [40](55\%, 75\%)}$$

2. **Xét DA → B:**

$$fr([DAB]) = 45\%$$

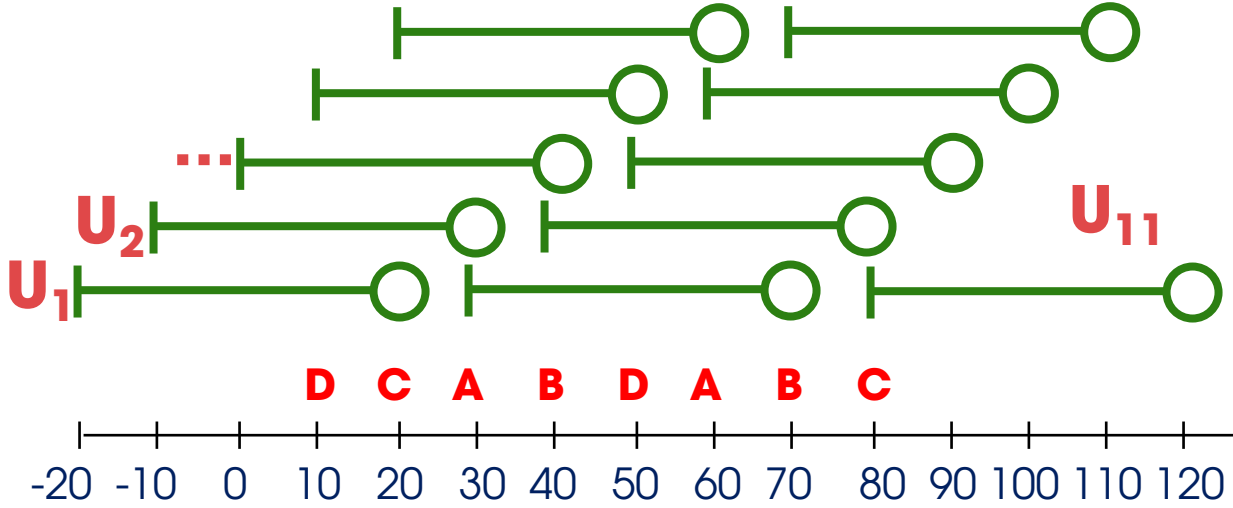
$$conf(\mathrm{DA} \rightarrow \mathrm{B}) = \frac{fr([DAB])}{fr([DA])} = \frac{45}{55} = 82\%$$

$$\boldsymbol{DA \rightarrow B\ [40](45\%, 82\%)}$$

WINEPI: Tìm Episode phổ biến tuần tự

VD4: Chuỗi tín hiệu báo động. Giả sử W = 40, min_fr= 40%.

Tìm Episode phổ biến tuần tự và một số luật Episode



Độ rộng cửa sổ: 40 giây

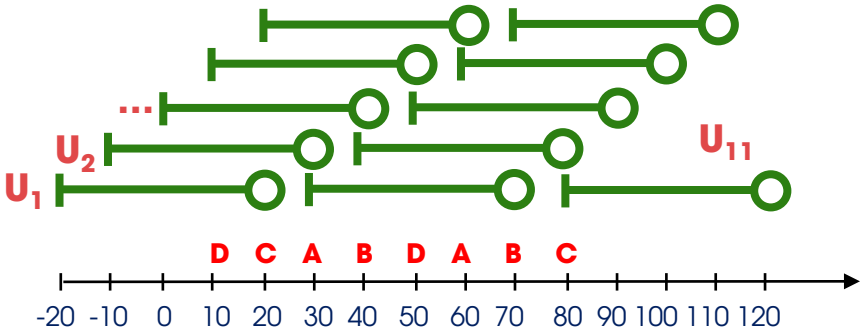
Bước di chuyển: 10 giây

Chiều dài chuỗi: 70 (10, …, 80)

Số lượng cửa sổ: 11 U_1, U_2,…, U_{11}

WINEPI: Tìm Episode phổ biến tuần tự

VD4:



Cửa số U_i	Nội dung cửa sổ
U_1[-20,20)	_,_,_,D
U_2[-10,30)	_,_,D,C
U_3[0,40)	_,D,C,A
U_4[10,50)	D,C,A,B
U_5[20,60)	C,A,B,D
U_6[30,70)	A,B,D,A
U_7[40,80)	B,D,A,B
U_8[50,90)	D,A,B,C
U_9[60,100)	A,B,C,_
U_{10}[70,110)	B,C_,_
U_{11}[80,120)	C,_,_,_

WINEPI: Tìm Episode phổ biến tuần tự

VD4:

- min_fr = 40% => Episode xảy ra **> 4** trong 11 cửa sổ (tần số)
- Các Episode tuần tự có 1 phần tử: A (7), B (7) , C (8), D (8)
- Các Episode phổ biến tuần tự có 1 phần tử: A, B, C, D

Cửa sổ U_i	Nội dung cửa sổ
U_1[-20,20)	_,_,_,D
U_2[-10,30)	_,_,D,C
U_3[0,40)	_,D,C,A
U_4[10,50)	D,C,A,B
U_5[20,60)	C,A,B,D
U_6[30,70)	A,B,D,A
U_7[40,80)	B,D,A,B
U_8[50,90)	D,A,B,C
U_9[60,100)	A,B,C,_
U_{10}[70,110)	B,C_,_
U_{11}[80,120)	C,_,_,_

WINEPI: Tìm Episode phổ biến tuần tự

VD4:

- Từ các Episode phổ biến trên, tìm Episodes tuần tự có 2 phần tử: AB (6), BA (2), AC (2), CA (3), AD (2), DA (5), BC (3), CB (2), BD (3), DB (3), CD (1), DC(4)
- Các Episode phổ biến tuần tự có 2 phần tử: AB, DA

Cửa số U_i	Nội dung cửa sổ
U_1[-20,20)	_,_,_,D
U_2[-10,30)	_,_,D,C
U_3[0,40)	_,D,C,A
U_4[10,50)	D,C,A,B
U_5[20,60)	C,A,B,D
U_6[30,70)	A,B,D,A
U_7[40,80)	B,D,A,B
U_8[50,90)	D,A,B,C
U_9[60,100)	A,B,C,_
U_{10}[70,110)	B,C_,_
U_{11}[80,120)	C,_,_,_

WINEPI: Tìm Episode phổ biến tuần tự

VD4:

- Từ các Episode phổ biến trên, tìm Episodes tuần tự có 3 phần tử: DAB (3)
- Các Episode phổ biến tuần tự có 3 phần tử: không có
- Không có Episode có 4 phần tử

Cửa số U_i	Nội dung cửa sổ
U_1[-20,20)	_,_,_,D
U_2[-10,30)	_,_,D,C
U_3[0,40)	_,D,C,A
U_4[10,50)	D,C,A,B
U_5[20,60)	C,A,B,D
U_6[30,70)	A,B,D,A
U_7[40,80)	B,D,A,B
U_8[50,90)	D,A,B,C
U_9[60,100)	A,B,C,_
U_{10}[70,110)	B,C_,_
U_{11}[80,120)	C,_,_,_

WINEPI: Tìm Episode phổ biến tuần tự

VD4: Các Episode tuần tự phổ biến và độ phổ biến (Tần suất)

Episode TTPB	fr(Episode)
D	8/11 = 73%
C	8/11 = 73%
A	7/11 = 64%
B	7/11 = 64%
AB	6/11 = 55%
DA	5/11 = 45%

Episode phổ biến tối đại: AB, DA

Một số luật Episode:

1. **Xét D → A:**

$$fr([DA]) = 45\%$$

$$conf(\mathrm{D} \rightarrow \mathrm{A}) = \frac{fr([DA])}{fr([D])} = \frac{45}{73} = 62\%$$

$$\boldsymbol{D \rightarrow \mathrm{A}\ [40](45\%, 62\%)}$$

2. **Xét A → B:**

$$fr([AB]) = 55\%$$

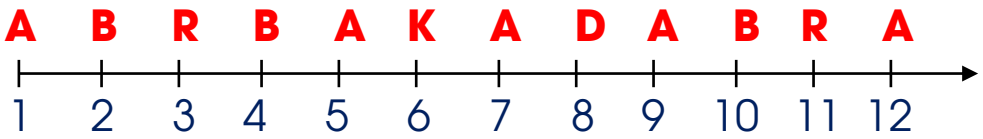
$$conf(\mathrm{A} \rightarrow \mathrm{B}) = \frac{fr([AB])}{fr([A])} = \frac{55}{64} = \mathbf{86\%}$$

$$\boldsymbol{\mathrm{A} \rightarrow \mathrm{B}\ [40](55\%, 86\%)}$$

WINEPI: Bài tập

BT07

Cho chuỗi sự kiện sau đây:



1. Có bao nhiêu cửa sổ có bề rộng là 5 được xử lý để tìm các episodes phổ biến theo tiếp cận WINEPI?
2. Giả sử ngưỡng min_fr là 0.4. Tìm các episode phổ biến tuần tự và song song trong chuỗi sự kiện trên?
3. Xác định các luật Episode và tính độ tin cậy từ các episode phổ biến tối đại tuần tự và song song tìm được thư câu 2.

Tổng kết chương

Các khái niệm

1.Luật Episode
2.Dữ liệu cho bài toán
3.Dãy phổ biến (Episode)

Phương pháp WINEPI

1.Cách tiếp cận & nguyên tắc
2.Tìm Episode phổ biến
3.Luật Episode và độ tin cậy

Tổng kết

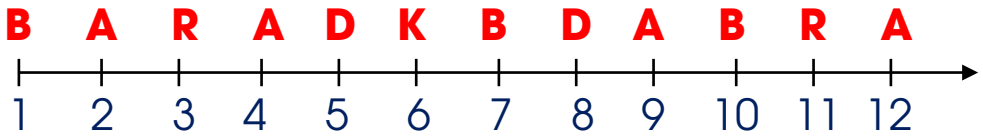
- Khai phá luật Episode:
  - Dựa trên kỹ thuật luật kết hợp
  - Dữ liệu hướng thời gian
- Hai cách tiếp cận:
  - WINEPI với cửa sổ trượt
  - MINEPI với việc tìm sự xuất hiện nhỏ nhất (*)

Các tiếp cận được dùng cho các mục tiêu khác nhau

(*): Tìm hiểu, seminar

Bài tập chương 4

4.1. Cho chuỗi sự kiện sau đây:



1. Có bao nhiêu cửa sổ có bề rộng là 5 được xử lý để tìm các episodes phổ biến theo tiếp cận WINEPI?
2. Giả sử ngưỡng min_fr là 0.4. Tìm các episode phổ biến tuần tự và song song trong chuỗi sự kiện trên?

THANKS!

Any questions?