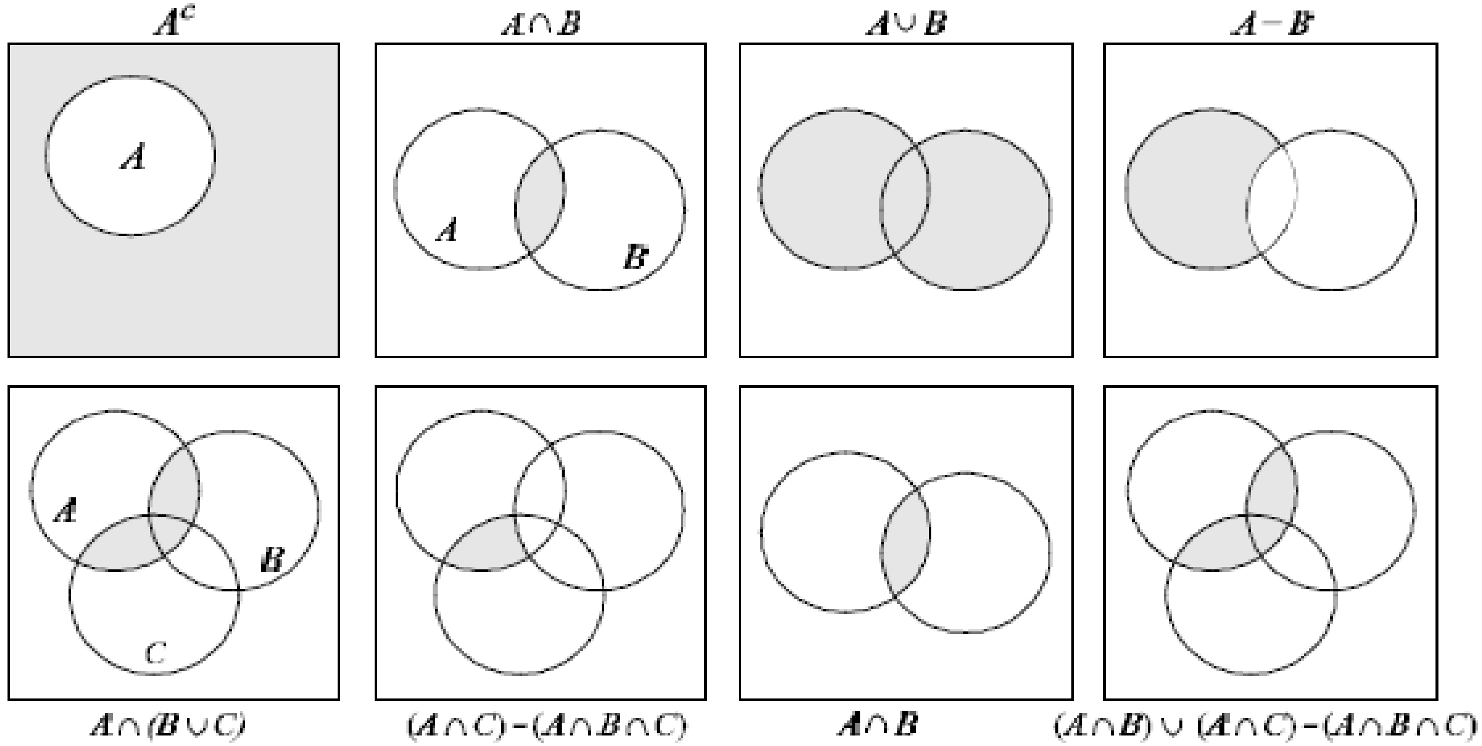
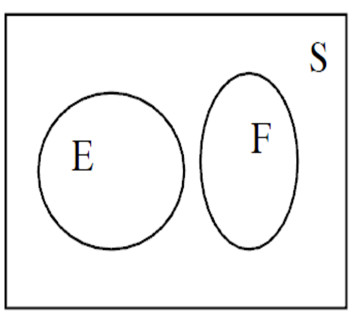
Markov Model– Hidden Markov Model

Các khái niệm

- Ngẫu nhiên: không biết chắc kết quả nào sẽ xảy ra, nhưng biết được các kết quả sẽ xảy ra. VD: tung đồng xu
- Không gian mẫu: tập hợp các kết quả có thể xảy ra trong thí nghiệm ngẫu nhiên, ký hiệu S
- Biến cố: tập hợp con của không gian mẫu, ký hiệu E
- Biến cố sơ đẳng: biến cố chỉ chứa một phần tử của S
- Xác suất: số lần xuất hiện của 1 biến cố E nào đó trên tổng số lần trong 1 thí nghiệm ngẫu nhiên

Các phép tính về biến cố





Xác suất

Xét N lần thử của một thí nghiệm ngẫu nhiên trong đó biến cố E xảy ra N_E lần, ta có N_E/ N gọi là tỷ lệ xuất hiện của biến cố E trong N lần thử.

Nếu N đủ lớn → tỷ lệ này gần như không đổi

→khái niệm tần xuất tương đối

Xác suất có điều kiện

- Xác suất của biến cố E được tính khi biến cố F đã xảy ra:

$$P(E/F) = \frac{P(E \cap F)}{P(F)}$$

Thuộc tính Markov

- Một dãy trạng thái ngẫu nhiên gọi là có thuộc tính Markov nếu như xác suất chuyển sang trạng thái tiếp theo chỉ phụ thuộc vào trạng thái hiện tại
- Dãy chuyển trạng quan sát được → Xích Markov/mô hình Markov.
- Dãy chuyển trạng không quan sát được → Mô hình Markov ẩn.

Định nghĩa mô hình Markov

Cho μ=(S, A)
Với S={s_1, s_2, s_3, …, s_N} - tập không gian trạng thái
X = (X_1, X_2, …, X_N) - dãy các biến ngẫu nhiên với các giá trị nằm trong tập S

$$P(X_{t+1} = s_k \mid X_1, X_2, \ldots, X_t) = P(X_{t+1} = s_k \mid X_t)$$

Xác xuất chuyển trạng thái độc lập với thời gian
Xác suất chuyển trạng thái được lưu ở ma trận A

$$a_{ij} = P(X_{t+1} = s_j \mid X_t = s_i)$$

➔Mô hình như trên gọi là mô hình Markov

VD Một mô hình Markov

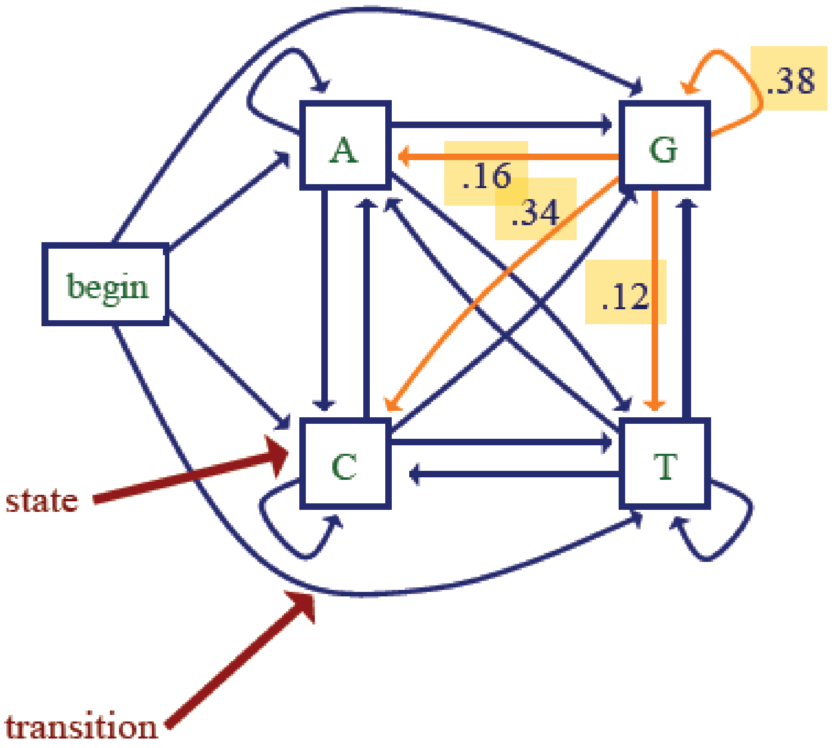
Có xác suất chuyển trạng thái:

$P(X_i = a \mid X_{i\text{-}1} = g) = 0.16$

$P(X_i = c \mid X_{i\text{-}1} = g) = 0.34$

$P(X_i = g \mid X_{i\text{-}1} = g) = 0.38$

$P(X_i = t \mid X_{i\text{-}1} = g) = 0.12$

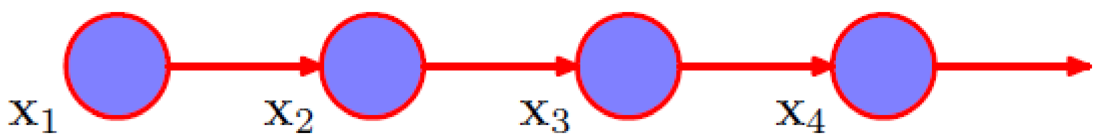


Mô hình Markov

- Cho X là một chuỗi N biến ngẫu nhiên $X_1 \ldots X_N$, ta có mô hình xác suất

$$\begin{aligned}\Pr(X) &= \Pr(X_N, X_{N-1}, \ldots, X_1) \\ &= \Pr(X_N | X_{N-1}, \ldots, X_1) \Pr(X_{N-1} | X_{N-2}, \ldots, X_1) \ldots \Pr(X_1)\end{aligned}$$

- Mô hình Markov bậc 1:

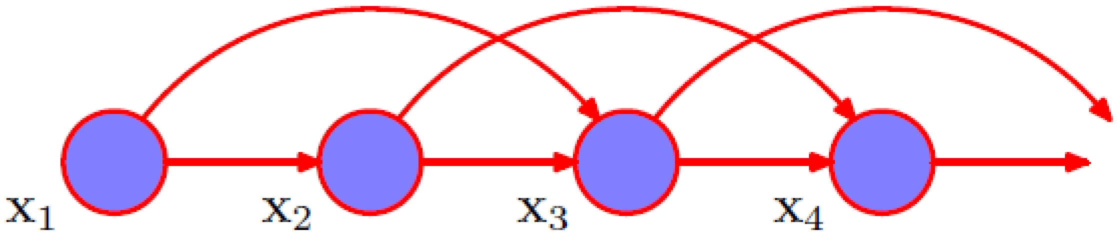


$$Pr(X) = Pr(X_N | X_{N-1}) Pr(X_{N-1} | X_{N-2}) \ldots Pr(X_2 | X_1) Pr(X_1)$$

$$= \Pr(X_1) \prod_{i=2}^{N} \Pr(X_i | X_{i-1})$$

Mô hình Markov

- Mô hình Markov bậc 2:



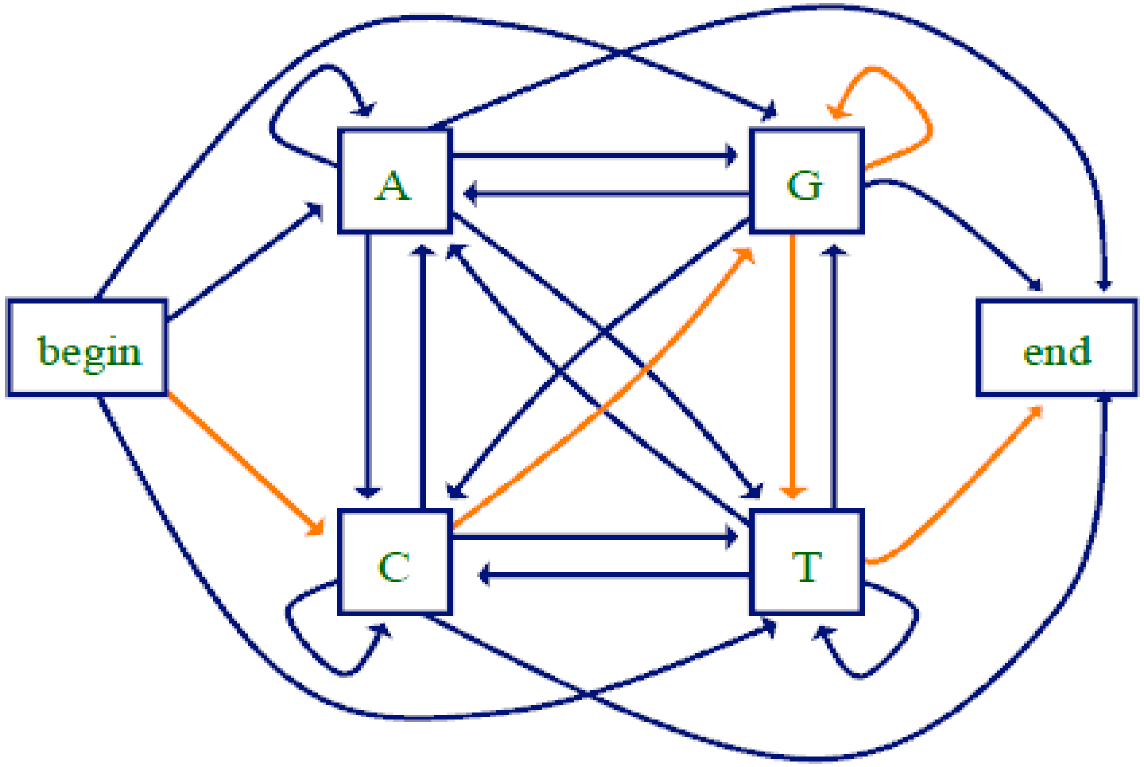
$\Pr(X) = \Pr(X_N \mid X_{N-1}, X_{N-2}) \Pr(X_{N-1} \mid X_{N-2}, X_{N-3}) \ldots \Pr(X_2 \mid X_1) \Pr(X_1)$

- Mô hình Markov n-bậc:

$\Pr(X_i \mid X_{i-1}, X_{i-2}, \ldots, X_1) = \Pr(X_i \mid X_{i-1}, \ldots, X_{i-n})$

VD

- <u>Ví dụ</u>



- $\Pr(cggt) = \Pr(c \mid begin)\Pr(g \mid c)\Pr(g \mid g)\Pr(t \mid g)\Pr(\text{end} \mid t)$

Xác suất chuyển trạng thái

- Tham số chuyển trạng thái, ký hiệu a_{ij}

  $a_{ij} = Pr(X_i = x_i \mid X_{i-1} = x_{i-1})$

- Xác suất của một chuỗi x

$$a_{Bx_1} \prod_{i=2}^{N} a_{x_{i-1}x_i} = Pr(x_1) \prod_{i=2}^{N} Pr(x_i \mid x_{i-1})$$

Với :

a_{Bx_1} là trạng thái bắt đầu chuỗi phân bố xác suất chiều dài N

VD

$P(X_1, X_2, \ldots, X_T) =$
$=P(X_1).P(X_2|X_1).P(X_3|X_1,X_2)\ldots P(X_T|X_1,X_2,\ldots,$
$X_{T-1}) = P(X_1).P(X_2|X_1).P(X_3|X_2)\ldots P(X_T|X_{T-1})$

$P(t,i,p) = P(X_1=t) * P(X_2=i|X_1=t) * P(X_3=p|X_2=i)$
$= 1.0 * 0.3 * 0.6$
$= 0.18$

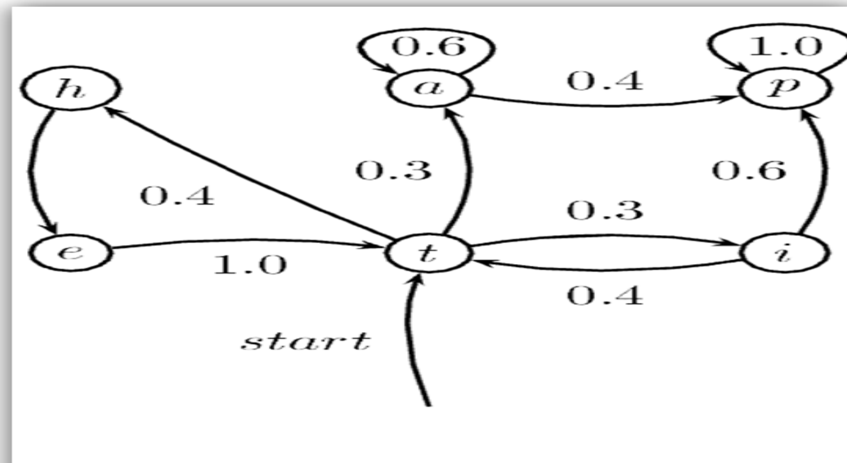


Figure 9.1 A Markov model.

Hạn chế của Markov

Một trong những thông số đặc trưng của mô hình Markov là các trạng thái ‘state’. Tùy thuộc vào việc xây dựng mô hình Markov với các đối tượng khác nhau thì sẽ có các ‘state’ khác nhau. Mô hình Markov có những hạn chế trong nhiều ứng dụng như các giá trị chuyển ‘state’ là các giá trị áp đặt sẵn, không thay đổi trong khi đối tượng, dữ liệu quan sát luôn biến đổi theo thời gian.

Để khắc phục tình trạng này, chúng ta sử dụng mô hình Hidden Markov (Hidden Markov Model – HMM).

HMM – Các thành phần

$$HMM = (S, V, B, A, \Pi)$$

N states

$$S = \{s_1, \ldots, s_N\}$$

M symbols

$$V = \{v_1, \ldots, v_M\}$$

Initial state probability:

$$\Pi = \{\pi_1, \ldots, \pi_N\} \quad \sum_{i=1}^{N} \pi_i = 1$$

$$\pi_i : \textit{prob of starting at state } s_i$$

State transition probability:

$$A = \{a_{ij}\} \quad 1 \le i, j \le N \quad \sum_{j=1}^{N} a_{ij} = 1$$

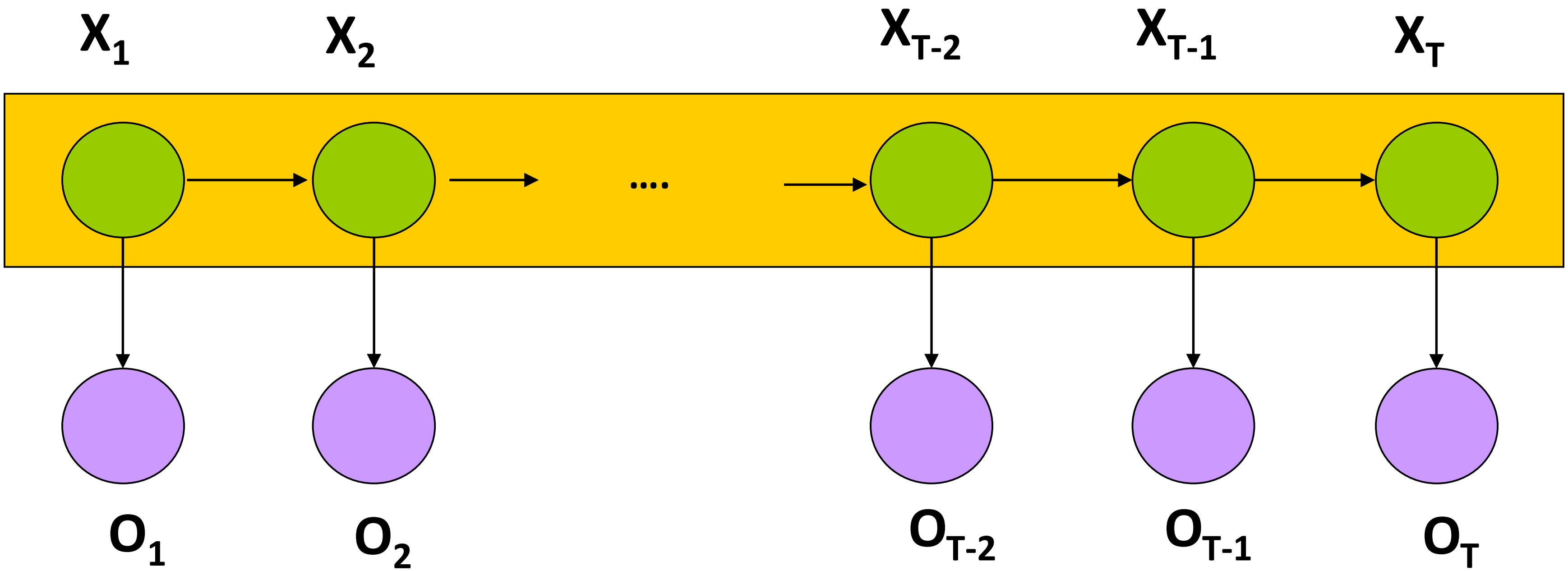
$$a_{ij} : \textit{prob of going } s_i \rightarrow s_j$$

Output probability:

$$B = \{b_i(v_k)\} \quad 1 \le i \le N, 1 \le k \le M \quad \sum_{k=1}^{M} b_i(v_k) = 1$$

$$b_i(v_k) : \textit{prob of "generating" } v_k \textit{ at } s_i$$

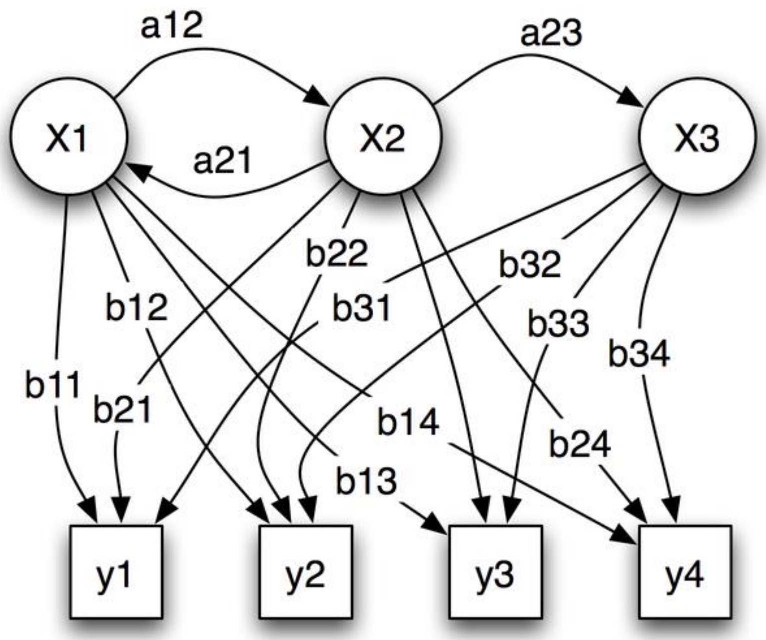
HMM – Cấu trúc

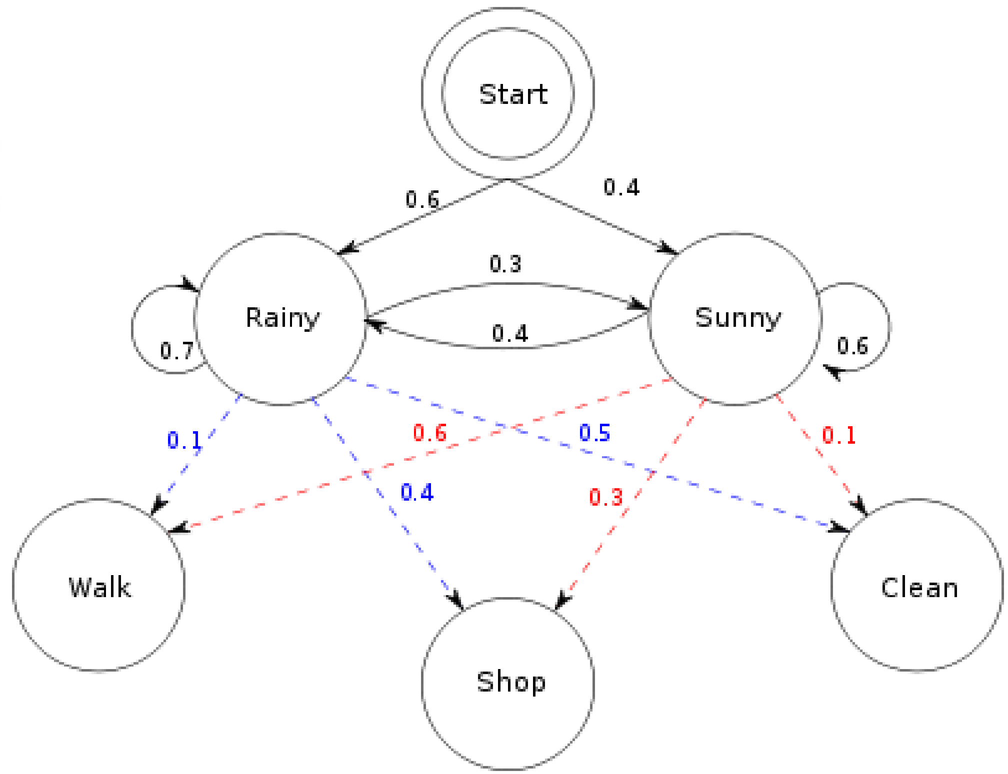


$X = (X_1, \ldots, X_T)$, $X_t \in \{s_1, \ldots, s_N\}$: Dãy chuyển trạng thái

$O = (o_1, \ldots, o_T)$, $o_t \in \{v_1, \ldots, v_M\}$: Dãy xuất (Observation)

HHM – trellis diagram





HMM - Các bài toán thường gặp

- Cho dãy quan sát $O=o_1o_2...o_T$ và các tham số của mô hình → Tính xác suất sinh dãy từ mô hình – $P(O|\mu)$
- Cho dãy quan sát $O=o_1o_2...o_T$ và các tham số của mô hình μ → Tìm dãy chuyển trạng thái $X=(X_1X_2...X_T)$ sao cho xác suất sinh ra O lớn nhất (optimal path)?
- Xác định các tham số của mô hình $\mu = (A, B, \Pi)$ (Training): Hiệu chỉnh μ để cực đại hoá xác suất sinh O – $P(O|\mu)$ (tìm mô hình "khớp" dãy quan sát nhất)

HMM - Tính xác suất sinh dãy từ mô hình – P(O| μ)

- Cho dãy quan sát $O=o_1o_2...o_T$ và các tham số của mô hình
- Các thuật giải:
  - Forward procedure.
  - Backward procedure.

Forward Procedure

$$\alpha(X_t) = P(o_1 o_2 \ldots o_t, X_t | \mu)$$

1. Khởi tạo

$$\alpha(X_1 = s_i) = \pi_i * b_i(o_1), \ 1 \le i \le N$$

2. Quy nạp

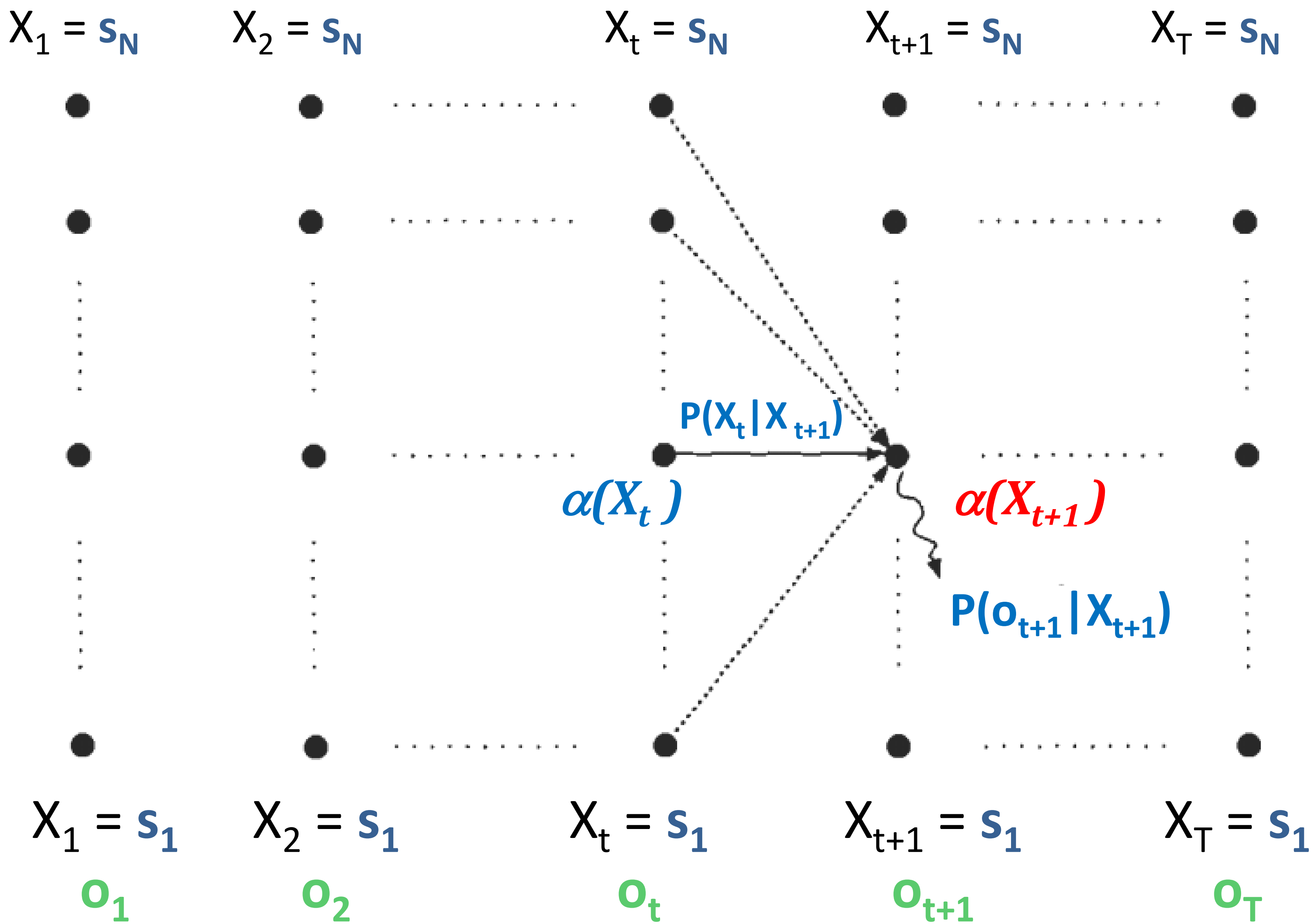
$$\alpha(X_{t+1} = s_j) = \sum_{i=1}^{N} \alpha(X_t = s_i) * a_{ij} * b_j(o_t),$$

$$1 \le t \le T, 1 \le i, j \le N$$

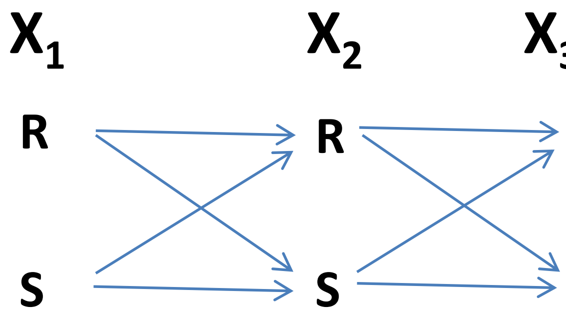
3. Tính tổng:

$$P(O|\mu) = \sum_{s=s_1}^{s_N} \alpha(X_T = s)$$

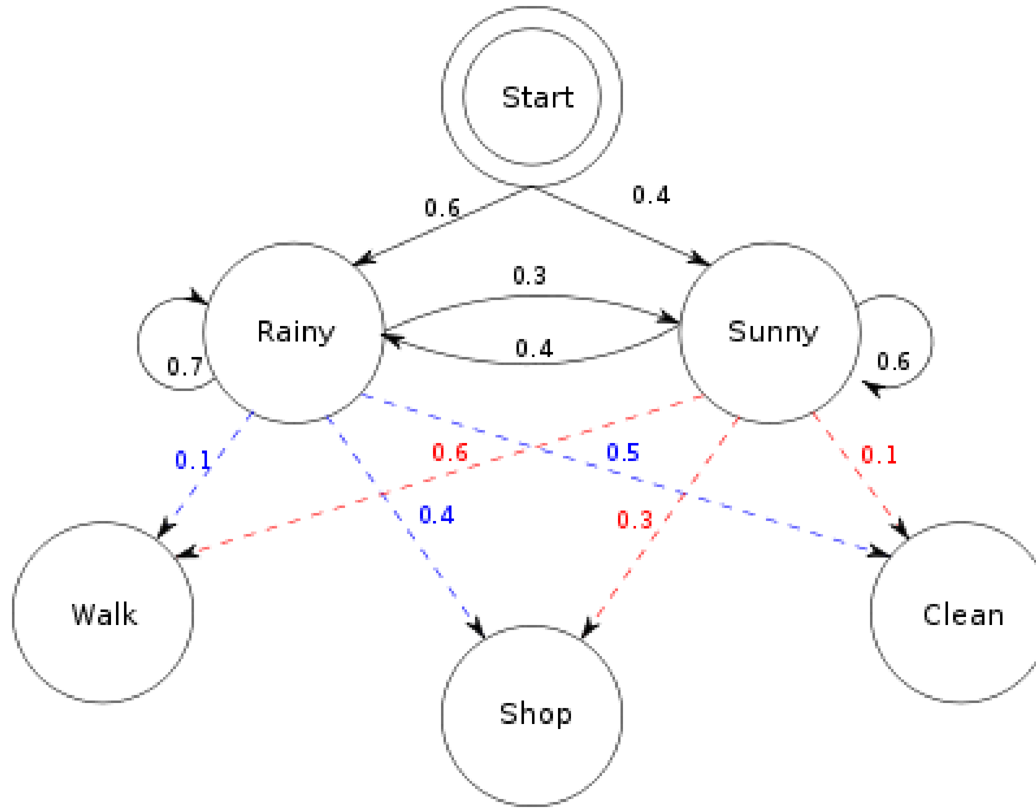
Forward Procedure - Tính $\alpha(X_t)$



Forward Procedure – ví dụ



O = **Walk** **Shop** **Clean**



- $\alpha(X_1 = R) = 0.6 * \mathbf{0.1} = \mathbf{0.06}$
- $\alpha(X_1 = S) = 0.4 * \mathbf{0.6} = \mathbf{0.24}$
- $\alpha(X_2 = R) = \mathbf{0.06} * 0.7 * \mathbf{0.4} + \mathbf{0.24} * 0.4 * \mathbf{0.4} = \mathbf{0.0384}$
- $\alpha(X_2 = S) = \mathbf{0.06} * 0.3 * \mathbf{0.3} + \mathbf{0.24} * 0.6 * \mathbf{0.3} = \mathbf{0.0432}$
- $\alpha(X_3 = R) = \mathbf{0.0384} * 0.7 * \mathbf{0.5} + \mathbf{0.0432} * 0.4 * \mathbf{0.5} = \mathbf{0.01344}$
- $\alpha(X_3 = S) = \mathbf{0.0384} * 0.3 * \mathbf{0.1} + \mathbf{0.0432} * 0.6 * \mathbf{0.1} = \mathbf{0.00259}$

$\longrightarrow$ P(O| μ) = 0.01344 + 0.0025 = 0.01594

Backward Procedure

$$\beta(X_t) = P(o_{t+1} \ldots o_T | X_t, \mu)$$

1. Khởi tạo

$$\beta(X_T = s_i) = 1, \qquad 1 \le i \le N$$

2. Quy nạp

$$\beta(X_t = s_i) = \sum_{j=1}^{N} \beta_j (X_{t+1} = s_j)\, a_{ij}\, b_j(o_{t+1}),$$
$$1 \le t \le T, 1 \le i, j \le N$$

3. Tính tổng:

$$P(O|\mu) = \sum_{i=1}^{N} \pi_i \beta_i(1)$$

HMM- Maximum likelihood

- Cho dãy quan sát $O = (o_1, \ldots, o_T)$.
- Tìm giá trị của các tham số mô hình $\mu = (A, B, \Pi)$ tốt nhất để sinh ra *O*.

$$\hat{\boldsymbol{\mu}} = \underset{\boldsymbol{\mu}}{\boldsymbol{argmax}}\, \boldsymbol{P}(\boldsymbol{O}_{\boldsymbol{training}}\,|\boldsymbol{\mu})$$

- Discrete observation HMM models
  - Baum-Welch reestimation (Forward–Backward algorithm)
  - Viterbi reestimation
  - Gradient – based optimizing techniques
  - Neural Network
- Continuous Observation HMM models

Thuật giải Baum-Welch

Khởi tạo các giá trị ban đầu cho các tham số của mô hình μ.

Tính P(O| μ)

- <u>Bước 1</u>:

  Ước lượng các tham số của mô hình mới $\hat{\mu}$

- <u>Bước 2</u>:

  Tính $P(O|\hat{\mu})$. Nếu $P(O|\hat{\mu}) - P(O|\mu) > \varepsilon$ thì cập nhật μ=$\hat{\mu}$ và quay lại bước 1, ngược lại thì dừng thuật giải.

Thuật giải Baum-Welch

Các công thức ước lượng:

$$\hat{\pi}_i = \gamma_1(s_i)$$

$$\hat{a}_{ij} = \frac{\sum_{t=1}^{T-1} \varepsilon_t(i,j)}{\sum_{t=1}^{T-1} \gamma_t(i)}$$

$$\hat{b}_i(vk) = \frac{\sum_{\{t:o_t=v_k,\mathbf{1}\le t\le T\}} \gamma_t(i)}{\sum_{t=1}^{T} \gamma_t(i)}$$

Thuật giải Baum-Welch

$$\varepsilon_t(i,j) \equiv \varepsilon_t(i,j|O,\mu) = \frac{\varepsilon_t(i,j,O|\mu)}{P(O|\mu)}$$

$$= \frac{\alpha(X_t = s_i)a_{ij}b_j(o_{t+1})\beta(X_{t+1} = s_j)}{P(O|\mu)}$$

$\varepsilon_t(i,j,O|\mu)$: xác suất để 1 dãy trạng thái X sinh ra chuỗi quan sát O với $X_t = s_i$ và $X_{t+1} = s_j$, với μ cho trước.

$$\gamma_t(i) \equiv \gamma_t(i|O,\mu) = \frac{\alpha(X_t = s_i)\beta(X_t = s_i)}{P(O|\mu)}$$

$\gamma_t(i,O|\mu)$: xác suất để 1 dãy trạng thái X sinh ra chuỗi quan sát O với $X_t = s_i$, với μ cho trước.

Chú ý

- Thuật giải có thể cho ra lời giải chỉ đạt cực trị địa phương → Cho thuật giải chạy 1 số lần nào đó mặc dù điều kiện $P(O|\hat{\mu}) - P(O|\mu) > \varepsilon$ đã hội tụ.
- Vì các tham số của mô hình là các giá trị xác xuất nên các giá trị khởi tạo ban đầu cho tham số của mô hình phải thỏa điều kiện dương và tổng bằng 1.
- Dữ liệu huấn luyện không đủ: Chuỗi huấn luyện phải đủ dài
- Scaling: Giá trị $\alpha(X_t)$ và $\beta(X_t)$ sẽ tiến về 0 rất nhanh

Thuật toán Viterbi

- Hệ thống có m phân lớp w_1, w_2,… w_M
- Một chuỗi các mẫu X: x_1, x_2, x_3,…, x_N sẽ được phân lớp vào một chuỗi lớp tương ứng Ω_i: w_{i1},w_{i2}, w_{i3},…., w_{iN} nếu :

$$p(X|\Omega_i)P(\Omega_i) = P(\omega_{i_1})p(\boldsymbol{x}_1|\omega_{i_1})\prod_{k=2}^{N} P(\omega_{i_k}|\omega_{i_{k-1}})p(\boldsymbol{x}_k|\omega_{i_k}) \qquad (9.6)$$

lớn nhất

- Tìm được Ω_i thỏa điều kiện phải tính (9.6) với M^N Ω_i => $O(NM^N)$

Thuật toán Viterbi

➢Ω_i và Ω_j chỉ khác nhau chỉ ở phân lớp cuối cùng : w_{iN} # w_{jN} và w_{ik} = w_{jk}, k = 1,2, . . . , N - 1 thì tính (9.6) chỉ khác nhau ở phân cuối.

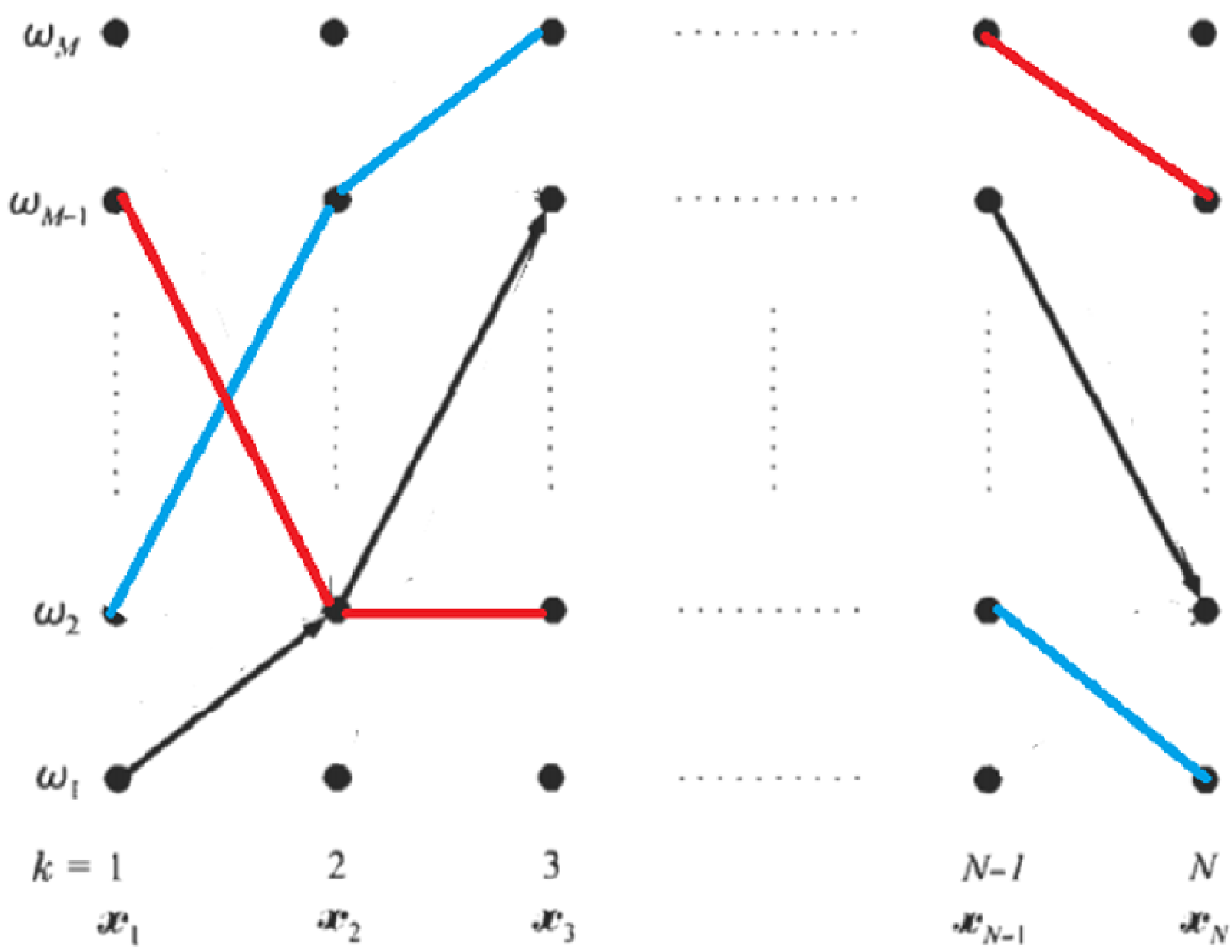
➢(9.6) có dạng đặc biệt:

$$p(X|\Omega_i)P(\Omega_i) = \underbrace{P(\omega_{i_1})p(\boldsymbol{x}_1|\omega_{i_1})}_{d(\omega_{i_1},\omega_{i_0})}\underbrace{P(\omega_{i_2}|\omega_{i_1})p(\boldsymbol{x}_2|\omega_{i_2})}_{d(\omega_{i_2},\omega_{i_1})}\underbrace{P(\omega_{i_3}|\omega_{i_2})p(\boldsymbol{x}_3|\omega_{i_3})}_{d(\omega_{i_3},\omega_{i_2})}\underbrace{P(\omega_{i_k}|\omega_{i_{k-1}})p(\boldsymbol{x}_k|\omega_{i_k})}_{d(\omega_{i_k},\omega_{i_{k-1}})}\dots\underbrace{P(\omega_{i_N}|\omega_{i_{N-1}})p(\boldsymbol{x}_N|\omega_{i_N})}_{d(\omega_{i_N},\omega_{i_{N-1}})}$$

$$= \prod_{K=1}^{N} d(\omega_{i_k},\omega_{i_{k-1}})$$

=> Xây dựng một cấu trúc mới để biểu diễn (9.6) để thuận tiện hơn trong việc tìm max

Thuật toán Viterbi



Thuật toán Viterbi

➢Chi phí di chuyển từ node $w_{i(k-1)}$ tới node w_{ik} trên đường biểu diễn cho chuỗi Ω_i

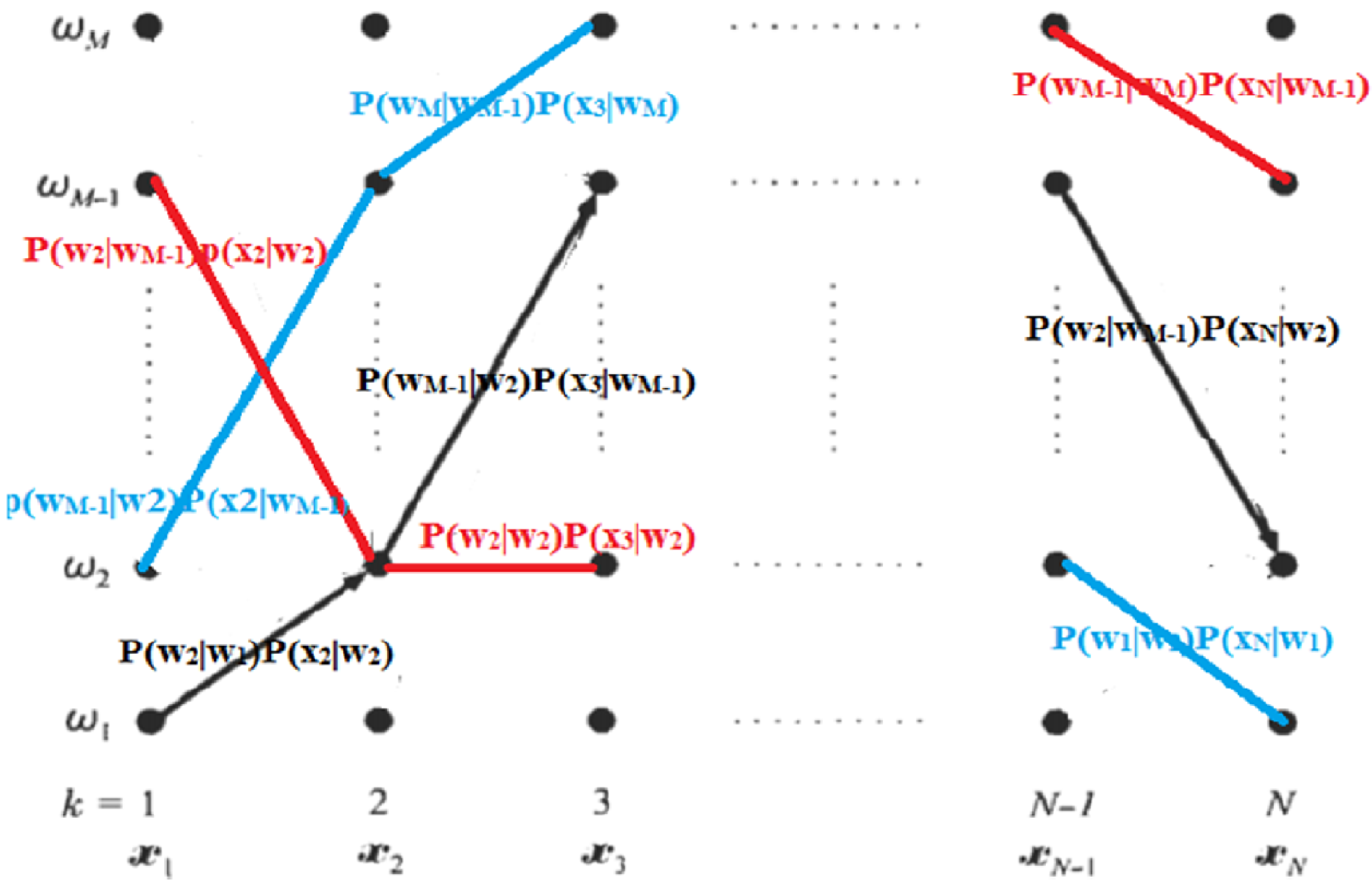
$$d(\omega_{i_k},\omega_{i_{k-1}}) = P(\omega_{i_k}|\omega_{i_{k-1}})\,p(x_k|\omega_{i_k})$$ với k =2…N

$$d(\omega_{i_1},\omega_{i_0}) = P(\omega_{i_1})\,p(x_1|\omega_{i_1})$$

▪$w_{i(k-1)}$ *: lớp* của mẫu x_{k-1} trong chuỗi Ω_i

▪w_{ik} : lớp của mẫu x_k trong chuỗi Ω_i

Thuật toán Viterbi



Thuật toán Viterbi

- Chi phí của một đường biểu diễn cho chuỗi Ω_i bằng tích các chi phí các di chuyển giữa các node liên tiếp (đường nối).

$$D = \prod_{k=1}^{N} d(\omega_{i_k}, \omega_{i_{k-1}})$$

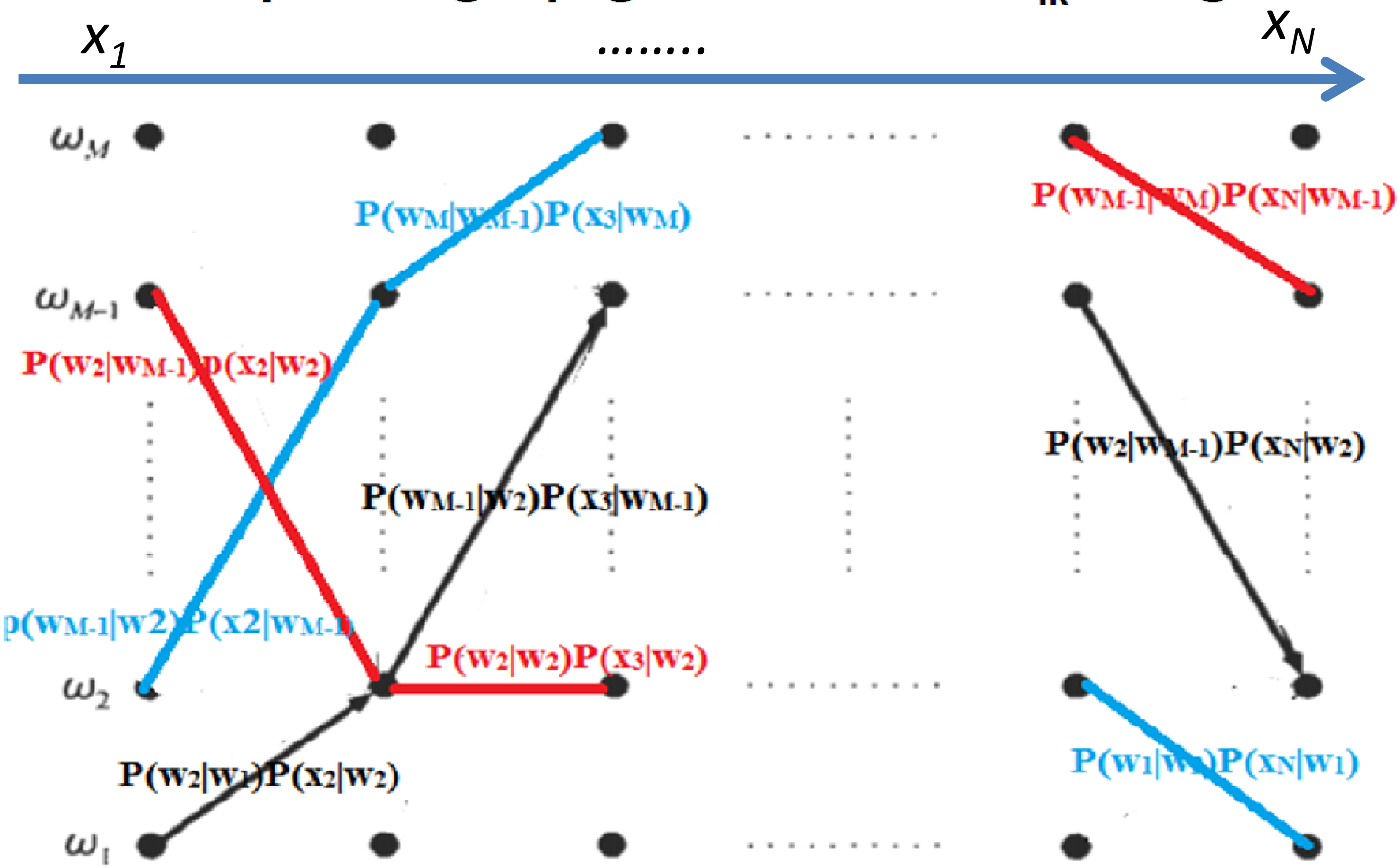
$$= \underbrace{P(\omega_{i_1})p(x_1|\omega_{i_1})}_{d(\omega_{i_1},\omega_{i_0})} \underbrace{P(\omega_{i_2}|\omega_{i_1})p(\boldsymbol{x}_2|\omega_{i_2})}_{d(\omega_{i_2},\omega_{i_1})} \underbrace{P(\omega_{i_3}|\omega_{i_2})p(\boldsymbol{x}_3|\omega_{i_3})}_{d(\omega_{i_3},\omega_{i_2})} \cdots \underbrace{P(\omega_{i_k}|\omega_{i_{k-1}})p(\boldsymbol{x}_k|\omega_{i_k})}_{d(\omega_{i_k},\omega_{i_{k-1}})} \cdots \underbrace{P(\omega_{i_N}|\omega_{i_{N-1}})p(\boldsymbol{x}_N|\omega_{i_N})}_{d(\omega_{i_N},\omega_{i_{N-1}})}$$

$$\equiv p(X|\Omega_i)P(\Omega_i)$$

Thuật toán Viterbi

- **Công thức (9.6) được chuyển thành bài toán tìm đường đi lớn nhất trên đồ thị của các nút $w_1, \ldots, w_k$ với**

$D_{max}(\omega_{i_k})$ **: chi phí tổng cộng lớn nhất tới w_{ik} trong chuỗi Ω_i.**



Thuật toán Viterbi

$$D_{max}(\omega_{i_k}) = \max_{i_{k-1}}[D_{max}(\omega_{i_{k-1}}) \text{ X } d(\omega_{i_k}, \omega_{i_{k-1}})].$$

$$D_{max}(\omega_{i_N}) = \max_{i_{N-1}}[D_{max}(\omega_{i_{N-1}}) \text{ X } d(\omega_{i_N}, \omega_{i_{N-1}})].$$

$$D_{max}(\omega_{i_{N-1}}) = \max_{i_{N-2}}[D_{max}(\omega_{i_{N-2}}) \text{ X } d(\omega_{i_{N-1}}, \omega_{i_{N-2}})].$$

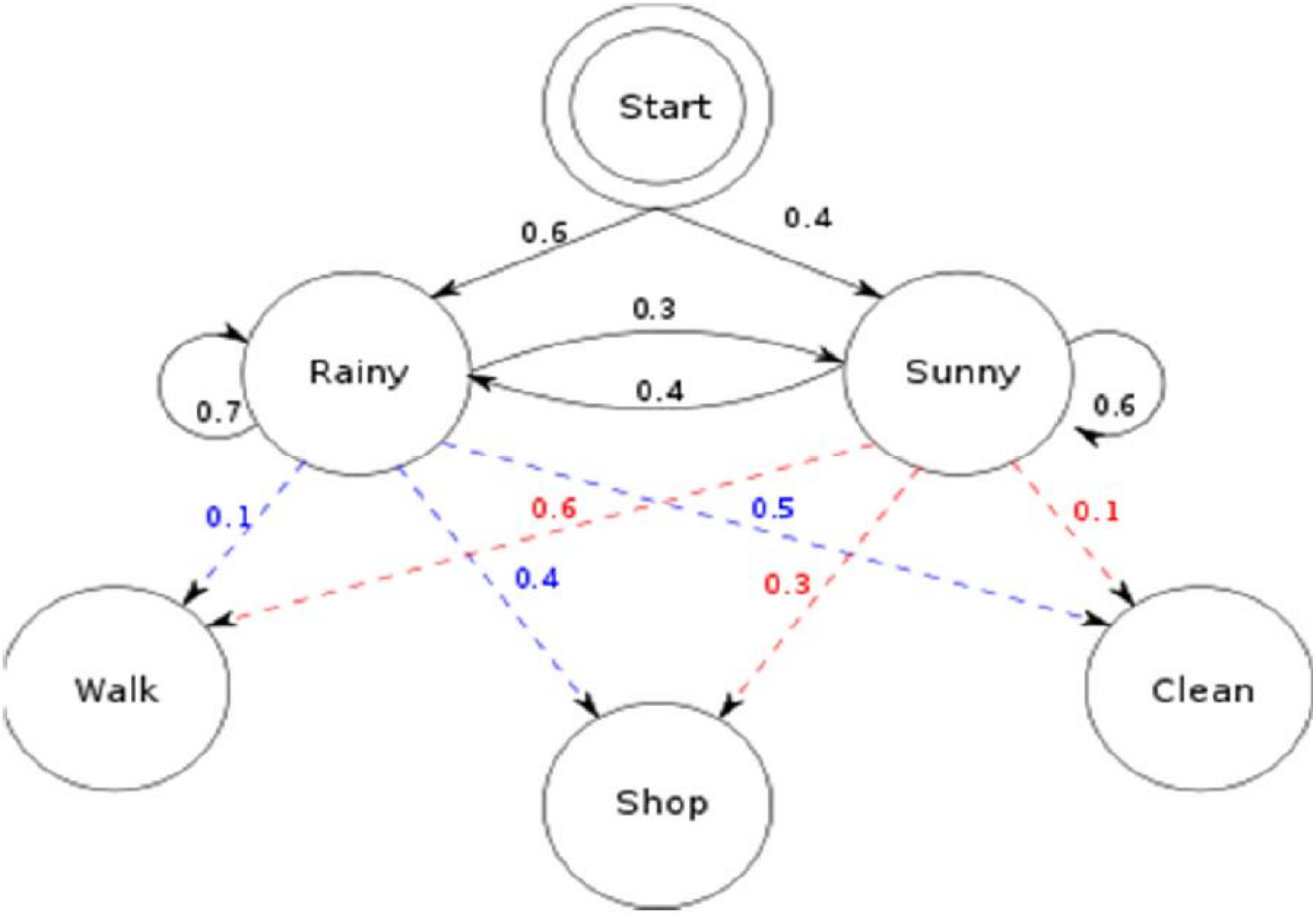
$$D_{max}(\omega_{i_2}) = \max_{i_1}[D_{max}(\omega_{i_1}) \text{ X } d(\omega_{i_2}, \omega_{i_1})].$$

$$D_{max}(\omega_{i_1}) = P(\omega_{i_1})P(x_1 | \omega_{i_1})$$

Ví dụ

- Alice và Bod sống ở xa nhau. Họ nói chuyện hằng ngày với nhau qua điện thoại về những hoạt động mà họ làm. Hằng ngày, Bob chỉ có 3 hoạt động: Walk, Shop, Clean. Những hoạt động mà họ làm thì phụ thuộc vào thời tiết của ngày ở nơi ở của họ. Dựa trên những hành động mà Bod nói hằng ngày , Alice sẽ đoán thời tiết ở nơi mà Alice sống.
- Họ nghĩ rằng thời tiết thì tuân theo chuỗi MaKov và có 2 trạng thái Rainy" và "Sunny".
- Trong 3 ngày liên tiếp thì Bob nói về Day1: Walk, Day2: Shop, Day3: Clean.

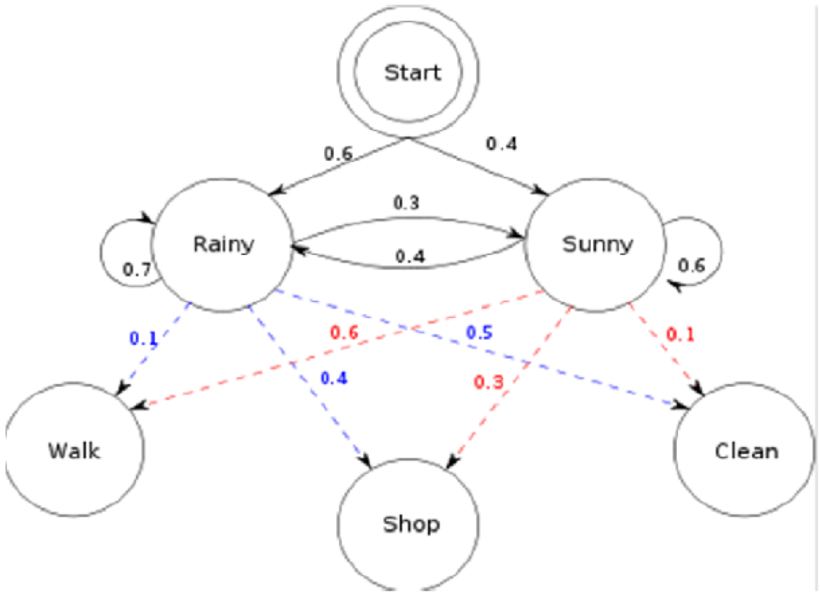
Ví dụ

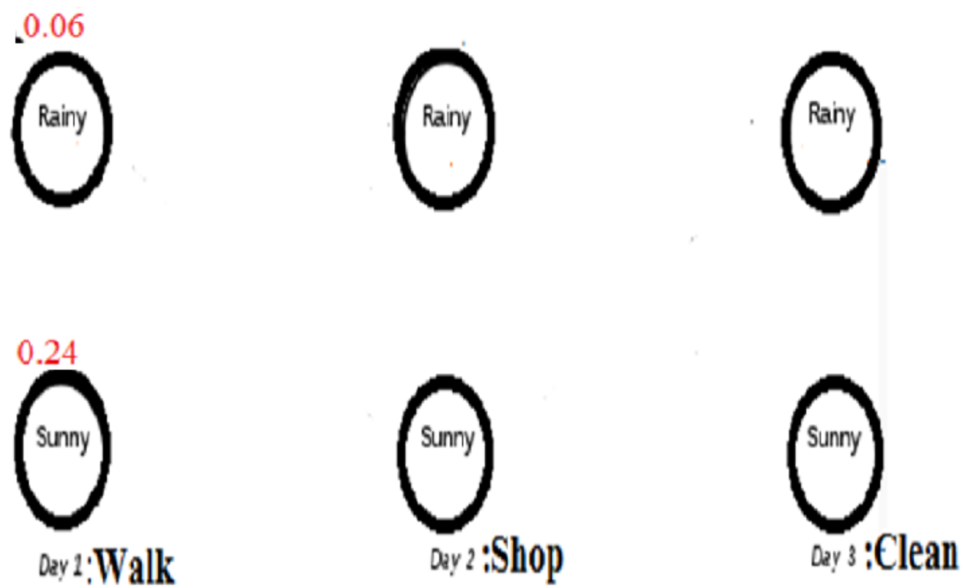


Start
0.6
0.4
0.3
Rainy
Sunny
0.4
0.7
0.6
0.1
0.6
0.5
0.1
0.4
0.3
Walk
Shop
Clean

Ví dụ

- tính $D_{max}(w_{i1}) = P(x_1|w_{i1}).P(w_{i1})$
  - $D_{max}(w_{i1}=Rainy) = P(Walk|Rainy)\ P(Rainy) = 0.1*0.6 = 0.06$
  - $D_{max}(w_{i1}=Sunny) = P(Walk|Sunny)\ P(Sunny) = 0.6*0.4 = 0.24$



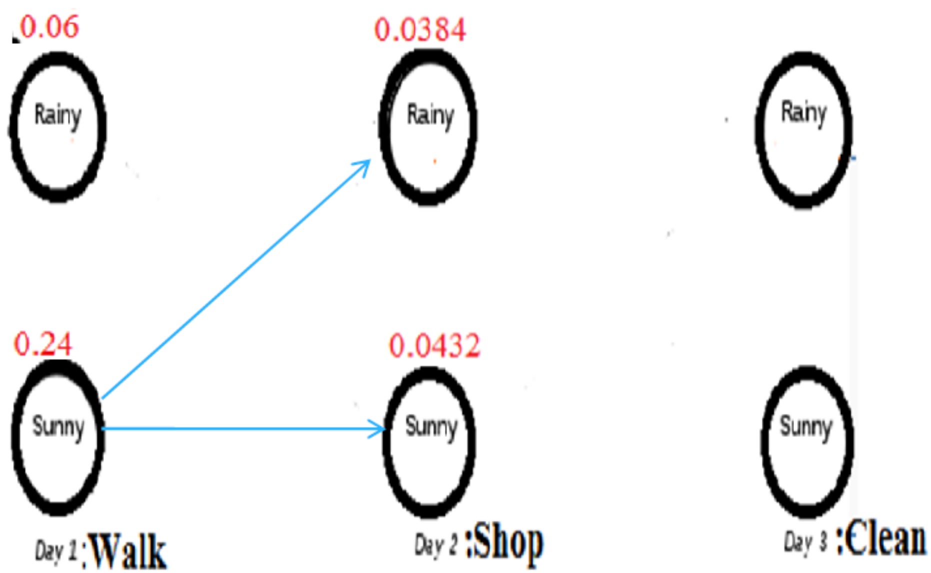


Ví dụ

- Tính $D_{max}(\omega_{i_2}) = \max_{i_1}[D_{max}(\omega_{i_1}) \times d(\omega_{i_2}, \omega_{i_1})]$.
  $= \max_{i_1}[D_{max}(\omega_{i_1}) \times p(\omega_{i_2}|\omega_{i_1})p(x_2|\omega_{i_2})]$.
- w_{i2}=Rainy
  - $D_{max}(w_{i1}$=Rainy) X p(Rainy|Rainy) X p(Shop|Rainy)=0.06*0.7*0.4=0.0168
  - $D_{max}(w_{i1}$=Sunny) X p(Rainy|Sunny) X p(Shop|Rainy)=0.24*0.4*0.4=0.0388

  => w_{i1}= Sunny.
- w_{i2}=Sunny
  - $D_{max}(w_{i1}$=Rainy) X p(Sunny|Rainy) X p(Shop|Sunny)=0.06*0.3*0.3=0.0054
  - $D_{max}(w_{i1}$=Sunny) X p(Sunny|Sunny) X p(Shop|Sunny)=0.24*0.6*0.3=0.0432

  => w_{i1}=Sunny



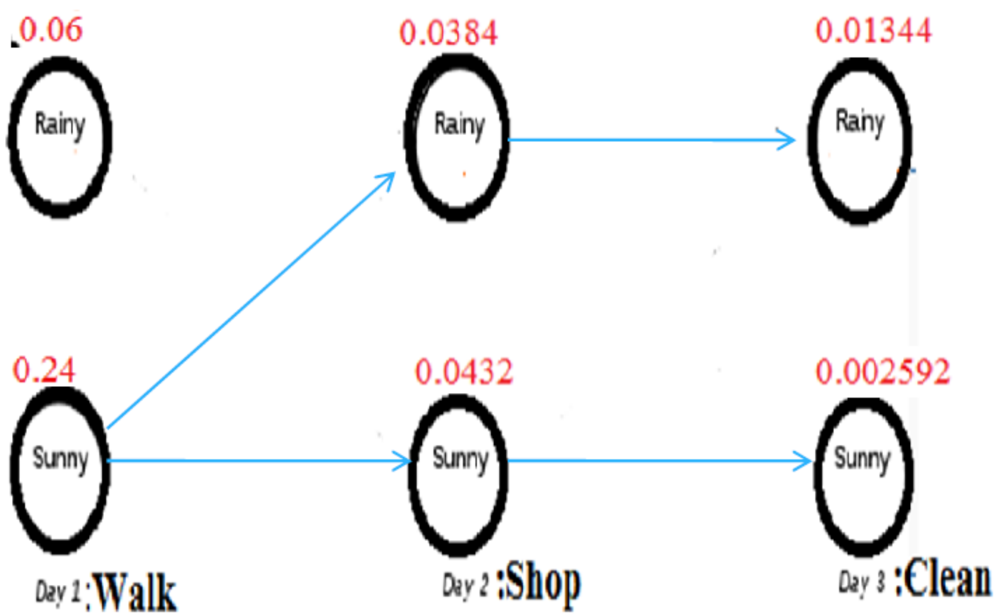
Ví dụ

Tính $$D_{max}(\omega_{i_3}) = \max_{i_2}[D_{max}(\omega_{i_2}) \times d(\omega_{i_3}, \omega_{i_2})].$$
$$= \max_{i_2}[D_{max}(\omega_{i_2}) \times p(\omega_{i_3}|\omega_{i_2})\, p(x_3|\omega_{i_3})].$$

- w_{i3}=Rainy
  - D_{max}(w_{i2}=Rainy) X p(Rainy|Rainy) X p(Clean|Rainy)=0.0384*0.7*0.5=0.01
  - D_{max}(w_{i2}=Sunny) X p(Rainy|Sunny) X P(Clean|Rainy)=0.0432*0.4*0.5=0.008

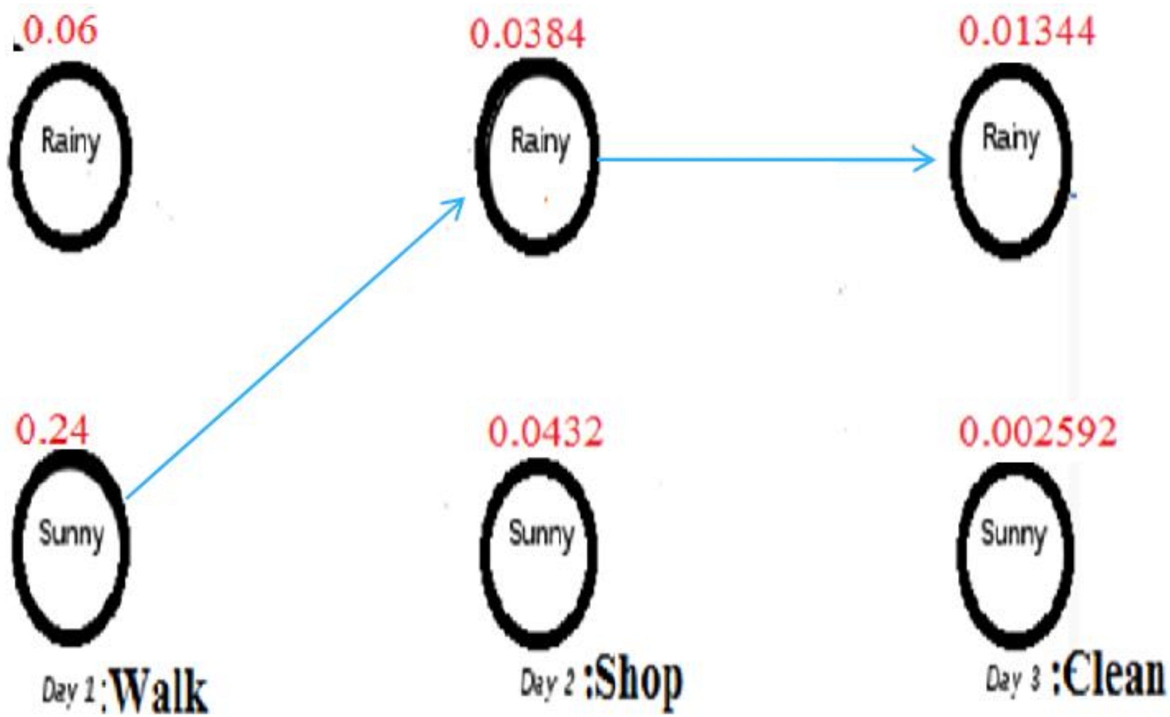
  => w_{i2}= Rainy.
- w_{i3}=Sunny
  - D_{max}(w_{i2}=Rainy) X p(Sunny|Rainy) X p(Clean|Sunny)=0.0384*0.3*0.1=0.001152
  - D_{max}(w_{i2}=Sunny) X p(Sunny|Sunny) X p(Clean|Sunny)=0.0432*0.6*0.1=0.002592

  => w_{i2}=Sunny



Ví dụ

- Tìm w_{i3} sao cho $D_{max}(w_{i3})$ lớn nhất =>w_{i3}=Rainy => w_{i2}=Rainy => w_{i1}=Sunny



Nhận xét

- Độ phức tạp: $O(T * Y^2)$

  Y: Số trạng thái (ẩn).

  T: Số sự kiện quan sát.
- Giảm chi phí tính toán

Tính thuật toán Viterbi từ HMM

- Tìm chuỗi trạng thái ẩn có xác xuất lớn nhất nhưng chi phí nhỏ nhất.

$$\delta_j(t+1) = \max_{1 \le i \le N} \delta_i(t) a_{ij} b_{ijo_t}, \quad \mathbf{1} \le j \le N$$