LÝ THUYẾT TÍNH TOÁN

BÀI 4: BIỂU THỨC CHÍNH QUY

Phạm Xuân Cường
Khoa Công nghệ thông tin
cuongpx@tlu.edu.vn

Nội dung bài giảng

Khái niệm

Khái niệm

- **Biểu thức chính quy:** Sử dụng các toán tử chính quy để biểu diễn một biểu thức mô tả ngôn ngữ
  Ví dụ: $\mathbf{(0 \cup 1)0^*}$
  $\rightarrow$ Tất cả các xâu bắt đầu bằng 1 ký tự 0 hoặc 1 và sau đó là một số nào đó các ký tự 0
- Vai trò của Biểu thức chính quy: Là một phương pháp mạnh để mô tả 1 mẫu văn bản nào đó
  $\rightarrow$ Trong một số ngôn ngữ lập trình đều ứng dụng kỹ thuật mô tả mẫu bằng biểu thức chính quy (**Regular Expression**)

Định nghĩa hình thức

Định nghĩa hình thức của biểu thức chính quy

Ta nói R là một biểu thức chính quy nếu R là:

1. a với a là ký hiệu nào đó trọng bộ chữ $\boldsymbol{\Sigma}$
2. ε
3. $\emptyset$
4. $(R_1 \cup R_2)$ trong đó R_1 và R_2 là các biểu thức chính quy
5. $(R_1 \circ R_2)$ trong đó R_1 và R_2 là các biểu thức chính quy
6. $(R_1{*})$ trong đó R_1 là một biểu thức chính quy

Độ ưu tiên của các toán tử chính quy

- Toán tử sao có độ ưu tiên cao nhất

  ab* = a(b*) $\neq$ (ab)*

- Toán tử ghép tiếp có độ ưu tiên cao hơn toán tử hợp

  a$\circ$b $\cup$ c = (a$\circ$b) $\cup$ c $\neq$ a(b $\cup$ c)

- Một số ký hiệu khác:
  - Hoặc (Union): ab|c = (ab)|c $\neq$ a(b|c)
  - Sao: a* = {a} = {a}*
  - 1 hoặc nhiều: a^+ = aa* = $\{a\}^+$
  - Tùy chọn: [a] = a|ε= (a∪ε) = a?

Ví dụ về độ ưu tiên toán tử chính quy

- aab$\cup$caab$\cup$caa = **????**
- aab$|$caab$|$caa = **????**
- d$\cup$ab* cd* = **????**
- d$|$ab* cd* = **????**

Ví dụ về độ ưu tiên toán tử chính quy

- aab∪caab∪caa = (aab)∪(caab)∪caa
- aab|caab|caa = (aab)|(caab)|(caa)
- d∪ab* cd* = d∪(a(b*)c(d*))
- d|ab* cd* = d|(a(b*)c(d*))

Ví dụ biểu thức chính quy

Giả thiết sử dụng bộ chữ $\boldsymbol{\Sigma} = \{0,1\}$

1. $0^*10^* = \{w|w$ chỉ có một ký hiệu $1\}$
2. $\Sigma^*1\Sigma^* = \{w|w$ có ít nhất một ký hiệu $1\}$
3. $\Sigma^*001\Sigma^* = \{w|w$ có chứa xâu con $001\}$
4. $1^*(01^+)^* = \{w|$sau mỗi ký hiệu 0 trong w sẽ có ít nhất 1 ký hiệu $1\}$
5. $(\Sigma\Sigma)^* = \{w|w$ là xâu có độ dài là một số chẵn$\}$
6. $01 \cup 10 = \{01,10\}$

Ví dụ biểu thức chính quy

7. $0\Sigma^*0 \cup 1\Sigma^*1 \cup 0 \cup 1$ = {w|w bắt đầu và kết thúc bởi cùng 1 ký hiệu}
8. $(0\cup\varepsilon)1^* = 01^*\cup 1^*$
9. $(0\cup\varepsilon)(1\cup\varepsilon) = \{\varepsilon, 0, 1, 01\}$
10. $1^*\emptyset = \emptyset \rightarrow$ Ghép tập trống với bất cứ tập nào cũng sinh ra tập trống
11. $\emptyset^* = \{\varepsilon\}$
12. $\emptyset | 01 = \{01\}$

Sự tương đương với Ôtômat hữu hạn

Ngôn ngữ của biểu thức chính quy

- Mỗi biểu thức chính quy R đều mô tả một ngôn ngữ $\rightarrow$ Ngôn ngữ gì?
  $L(a) = \{a\}$
  $L(R_1|R_2) = L(R_1) \cup L(R_2)$
  $L(R_1 \circ R_2) = L(R_1) \circ L(R_2)$
  $L(R_1*) = L(R_1)*$
  $L(\varepsilon) = \{\varepsilon\}$
  $L(\emptyset) = \{\}$

Ngôn ngữ của biểu thức chính quy

Định lý 1

Một ngôn ngữ là chính quy **nếu và chỉ nếu** có một biểu thức chính quy nào đó mô tả nó
$\Leftrightarrow$ Định lý này có 2 chiều. Ta phát biểu nó thành từng bổ đề sau

Bổ đề 1.1

Nếu một ngôn ngữ được mô tả bởi một biểu thức chính quy thì nó là chính quy

Bổ đề 1.2

Nếu một ngôn ngữ là chính quy, thì nó được mô tả bởi một biểu thức chính quy

Chứng minh Bổ đề 1.1

Từ Hệ quả 1.40 (Sách giáo trình): Nếu 1 NFA đoán nhận A thì A là chính quy $\rightarrow$ Chuyển đổi R thành một NFA N

1. $R = a \rightarrow L(R) = \{a\}$

2. $R = \varepsilon \rightarrow L(R) = \{\varepsilon\}$

3. $R = \emptyset \rightarrow L(R) = \emptyset$

4. $R = R_1 \cup R_2$
5. $R = R_1 \circ R_2$
6. $R = R_1$*

   Với 3 trường hợp cuối ta chứng minh tương tự với chứng minh tính đóng của 3 toán tử (Xem lại bài 3)

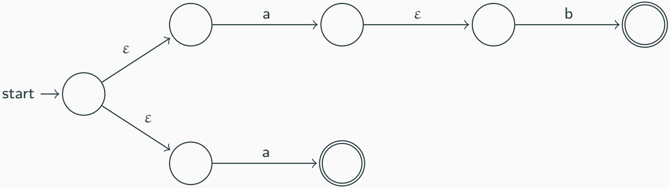
Chuyển đổi biểu thức chính quy sau thành NFA: **(ab∪a)***

a →

b →

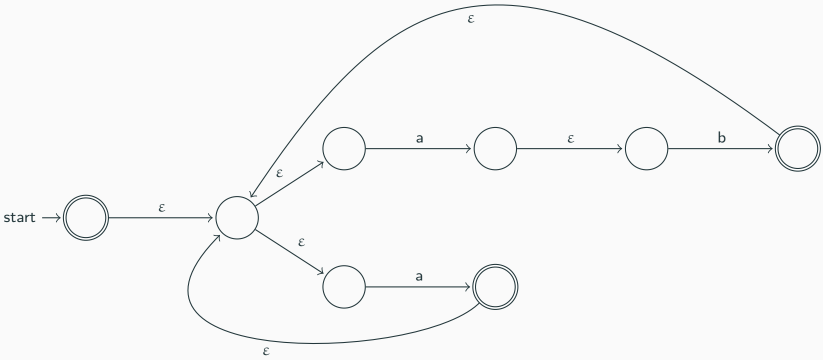
ab →

ab∪a →



Ví dụ: Chuyển đổi R → NFA

(ab∪a)*



Chứng minh Bổ đề 1.2

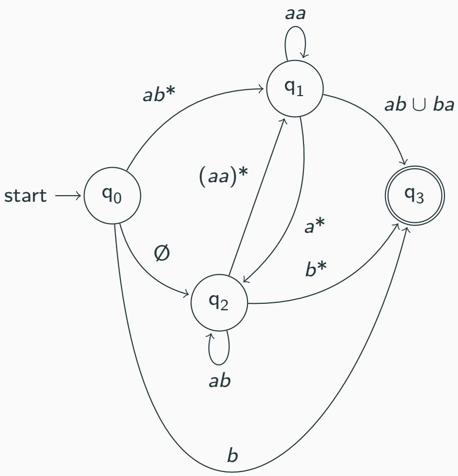
Ý TƯỞNG:

- Vì A là ngôn ngữ chính quy $\rightarrow$ Nó được đoán nhận bởi 1 DFA
- Chuyển đổi DFA thành biểu thức chính quy $\rightarrow$ Cần sử dụng **GNFA**. Vậy **GNFA** là gì?

Ôtômat hữu hạn không đơn định suy rộng (GNFA)

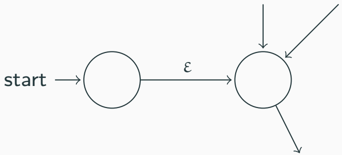
- GNFA = **G**eneralized **N**ondeterministic **F**inite **A**utomaton
  $\rightarrow$ Là Ôtômat hữu hạn không đơn định suy rộng
- GNFA giống NFA ngoại trừ:
  - Nhãn của các cạnh là các **biểu thức chính quy**
  - Chỉ có 1 trạng thái chấp thuận
  - Trạng thái chấp thuận không trùng với trạng thái bắt đầu
  - Không có cạnh nào nối tới trạng thái bắt đầu
  - Không có cạnh nào xuất phát từ trạng thái kết thúc
  - Loại trừ trạng thái bắt đầu và kết thúc, mọi mũi tên có thể đi từ 1 trạng thái đến các trạng thái còn lại hoặc là tới chính nó

Ví dụ GNFA

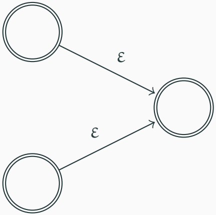


Chuyển đổi DFA $\rightarrow$ GNFA

- Thêm trạng thái bắt đầu mới

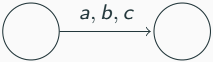


- Thêm trạng thái kết thúc mới

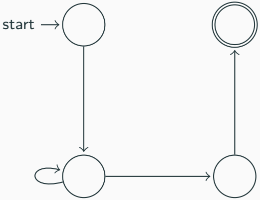


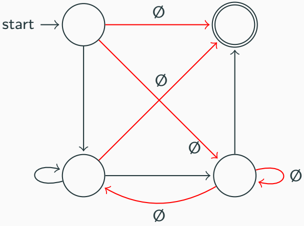
Chuyển đổi DFA → GNFA

- Cạnh có nhiều chuyển đổi → Hợp của các chuyển đổi



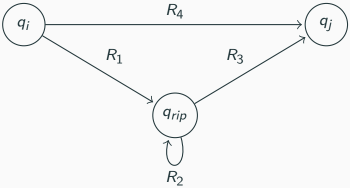
- Thêm các cạnh còn thiếu bằng các cạnh ∅ sao cho đầy đủ kết nối (**Fully connected**)

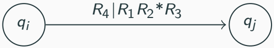




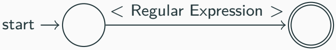
Chuyển đổi DFA → GNFA

- Chọn 1 trạng thái và tách nó ra khỏi máy. Chỉnh sửa phần còn lại sao cho ngôn ngữ tương tự vẫn được đoán nhận → Trạng thái bị tách ra được gọi là q_{rip}



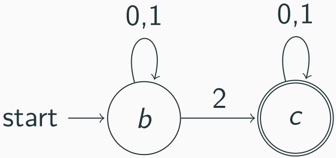


- Lặp lại bước trên cho đến khi máy chỉ còn 2 trạng thái bắt đầu và kết thúc



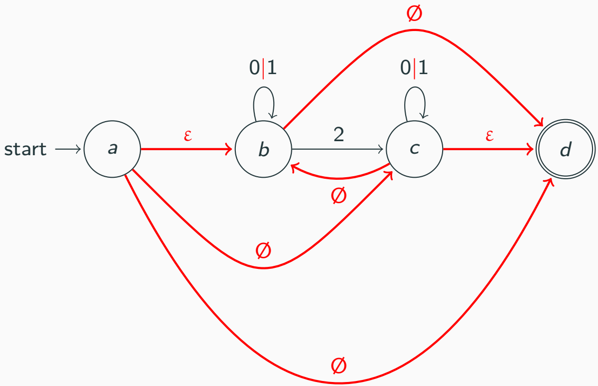
Ví dụ

- Cho bộ chữ $\boldsymbol{\Sigma} = \{0,1,2\}$
- Chuyển đổi DFA sau thành biểu thức chính quy

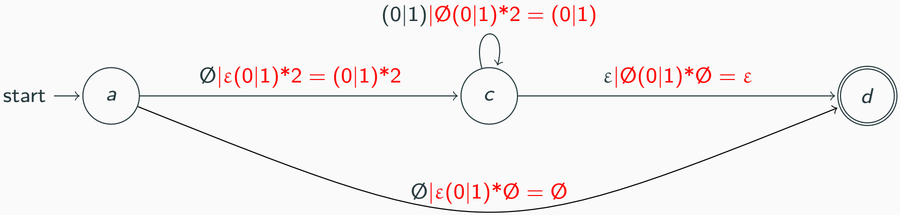


- Kết quả: (0|1)*2(0|1)* $\rightarrow$ **Làm như thế nào?**

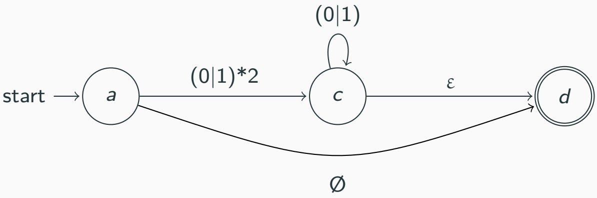
Chuyển từ DFA sang GNFA



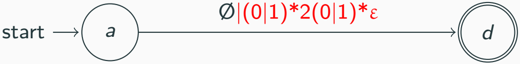
Loại bỏ nút b



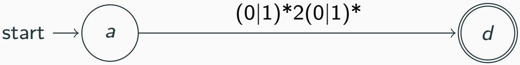
Thu gọn lại, ta được:



Loại bỏ nút c



Cuối cùng, ta được:



→ Mỗi biểu thức chính quy R đều mô tả một ngôn ngữ **chính quy**

Định nghĩa hình thức của GNFA

- Ôtômat hữu hạn không đơn định suy rộng (GNFA) $\equiv$ bộ 5 (hay 5 chiều)

  $$\mathbf{G = (Q, \Sigma, \delta, q_{start}, q_{accept})}$$

  Trong đó:

  - **Q:** Tập trạng thái (hữu hạn)
  - **Σ:** Bộ chữ, tập hữu hạn các ký tự
  - **δ:** Hàm dịch chuyển

    $$\boldsymbol{\delta}\textbf{: (Q-}\{\mathbf{q}_{accept}\}\textbf{) x (Q-}\{\mathbf{q}_{start}\}\textbf{)} \rightarrow \mathcal{R}$$

    $\mathcal{R}$ là tập tất cả các biểu thức chính quy trên bộ chữ $\mathbf{\Sigma}$

  - $q_{\mathbf{start}}$: Trạng thái bắt đầu
  - $q_{\mathbf{accept}}$: Trạng thái kết thúc

Questions?