

TRƯỜNG ĐẠI HỌC BÁCH KHOA HÀ NỘI
VIỆN TOÁN ỨNG DỤNG VÀ TIN HỌC



**NGÔN NGỮ CHÍNH QUY, BIỂU THỨC
CHÍNH QUY VÀ SƠ LƯỢC VỀ LÝ THUYẾT
MÃ**

Giảng viên hướng dẫn: TS. Ngô Thị Hiền

Nhóm Sinh viên thực hiện:

Nguyễn Anh Tú

Phạm Anh Tuấn

Lớp: KSTN Toán Tin K60

HÀ NỘI - 12/2018

Mục lục

1	Giới thiệu	4
2	Kiến thức cơ sở	6
2.1	Xâu	6
2.2	Ngôn ngữ	8
2.3	Ngôn ngữ chính quy	10
2.3.1	Định nghĩa	10
2.4	Biểu thức chính quy	11
2.4.1	Định nghĩa	11
2.4.2	Các tính chất của biểu thức chính quy	11
2.4.3	Đồ thị biểu diễn biểu thức chính quy	12
2.5	Automat đơn định(Deterministic Finite Automata)	13
2.6	Automat đa định(Nondeterministic Finite Automata)	14
2.7	Văn phạm	15
2.7.1	Định nghĩa	15
2.7.2	Ngôn ngữ sinh bởi văn phạm	16
2.7.3	Văn phạm chính quy	16

2.8	Thuật Toán	16
3	Sơ lược về lý thuyết mã	17
3.1	Mã	17
3.2	Mã tối đại	17
3.3	Mã prefix	18
3.4	Tiêu chuẩn kiểm định mã	18
4	Ứng dụng: Single-Error Correcting Codes	19
5	Kết luận	20

Chương 1

Giới thiệu

Ngôn ngữ là phương tiện để giao tiếp, sự giao tiếp có thể hiểu là giao tiếp giữa con người với nhau, giao tiếp giữa người với máy, hay giao tiếp giữa máy với máy. Ngôn ngữ để con người có thể giao tiếp với nhau được gọi là ngôn ngữ tự nhiên, chẳng hạn như tiếng Anh, tiếng Việt... là các ngôn ngữ tự nhiên. Các quy tắc cú pháp của ngôn ngữ tự nhiên nói chung rất phức tạp nhưng các yêu cầu nghiêm ngặt về ngữ nghĩa thì lại thiếu chặt chẽ, chẳng hạn cùng một từ hay cùng một câu ta có thể hiểu chúng theo những nghĩa khác nhau tùy theo từng ngữ cảnh cụ thể. Con người muốn giao tiếp với máy tính tất nhiên cũng thông qua ngôn ngữ. Để có sự giao tiếp giữa người với máy hay giữa máy với nhau, cần phải có một ngôn ngữ với các quy tắc cú pháp chặt chẽ hơn so với các ngôn ngữ tự nhiên, nói cách khác, với một từ hay một câu thì ngữ nghĩa của chúng phải là duy nhất mà không phụ thuộc vào ngữ cảnh. Những ngôn ngữ như thế được gọi là ngôn ngữ hình thức. Con người muốn máy tính thực hiện công việc, phải

viết các yêu cầu đưa cho máy bằng ngôn ngữ máy hiểu được. Việc viết các yêu cầu như thế gọi là lập trình. Ngôn ngữ dùng để lập trình được gọi là ngôn ngữ lập trình. Các ngôn ngữ lập trình đều là các ngôn ngữ hình thức.

Trong bài báo cáo này, nhóm em sẽ trình bày về ngôn ngữ chính quy, biểu thức chính quy và sơ lược về lý thuyết mã.

Chương 2

Kiến thức cơ sở

2.1 Xâu

- Xâu là một dãy hữu hạn các kí tự
- Bảng chữ cái là một tập hợp các kí tự hữu hạn
- Độ dài xâu là số các kí tự trong xâu đó
- Xâu rỗng: ϵ , độ dài xâu rỗng bằng 0
- Các phép toán trên xâu:

1. Ghép nối xâu

Cho 2 xâu: $a = a_1a_2...a_n, b = b_1b_2...b_m$ trên bảng chữ cái A . Ghép nối 2 xâu trên ta được một xâu mới $c = ab = a_1a_2...a_nb_1b_2...b_m$

Nhận xét: Cho các xâu s, r, w trên bảng chữ cái A

- Xâu rỗng là phần tử đơn vị với phép nối xâu

$$s\epsilon = \epsilon s = s$$

- Phép ghép nối có tính chất kết hợp

$$(sr)w = s(rw)$$

- Kí hiệu w^n , với n là số tự nhiên

$$w^n = \begin{cases} \epsilon, n = 0 \\ w, n = 1 \\ w^{n-1}w, n > 1 \end{cases}$$

2. Xâu con:

s là xâu con của w nếu $\exists s_0, s_1$ sao cho $s_0ss_1 = w$

$s_0 = \epsilon$ thì s được gọi là prefix của w

$s_1 = \epsilon$ thì s được gọi là suffix của w

3. Phép đảo ngược xâu

w^R được gọi là xâu đảo ngược của w nếu:

$$w^R = \begin{cases} \epsilon, w = \epsilon \\ s_n s_{n-1} \dots s_0, w = s_0 \dots s_{n-1} s_n; s_i \in A, i = \overline{1, n} \end{cases}$$

Nhận xét: Cho xâu s, w . Phép đảo ngược xâu có các tính chất sau:

$$- (w^R)^R = w$$

$$- (sw)^R = w^R s^R$$

$$- |w^R| = |w|$$

2.2 Ngôn ngữ

- Ngôn ngữ là tập các xâu trên bảng chữ cái
- Các phép toán trên ngôn ngữ

Xét hai ngôn ngữ L_1, L_2 trên bảng chữ cái A

1. Phép hợp

Hợp của hai ngôn ngữ L_1, L_2 , kí hiệu $L_1 \cup L_2$ là một ngôn ngữ trên bảng chữ cái A :

$$L_1 \cup L_2 = \{w \in A^* | w \in L_1 \text{ hoặc } w \in L_2\}$$

Định nghĩa phép hợp có thể mở rộng cho hữu hạn các ngôn ngữ:

$$\bigcup_{i=1}^n L_i = \{w \in A^* | w \in L_i, i = \overline{1, n}\}$$

Nhận xét: Xét các ngôn ngữ L_1, L_2, L_3 . Phép hợp có các tính chất sau

- Tính chất giao hoán: $L_1 \cup L_2 = L_2 \cup L_1$
- Tính chất kết hợp: $(L_1 \cup L_2) \cup L_3 = L_1 \cup (L_2 \cup L_3)$
- $\forall L_1 : L_1 \cup \emptyset = \emptyset \cup L_1 = L_1$ và $L_1 \cup A^* = A^*$

2. Phép giao

Giao của hai ngôn ngữ L_1, L_2 , kí hiệu $L_1 \cap L_2$

$$L_1 \cap L_2 = \{w \in A^* | w \in L_1 \text{ và } w \in L_2\}$$

Định nghĩa phép giao có thể mở rộng cho hữu hạn các ngôn ngữ:

$$\bigcap_{i=1}^n = \{w \in A^* | w \in L_i, i = \overline{1, n}\}$$

Nhận xét: Xét các ngôn ngữ L_1, L_2, L_3 . Phép hợp có các tính chất sau

- Tính chất giao hoán: $L_1 \cap L_2 = L_2 \cap L_1$
- Tính chất kết hợp: $(L_1 \cap L_2) \cap L_3 = L_1 \cap (L_2 \cap L_3)$
- $\forall L_1 : L_1 \cap \emptyset = \emptyset \cup L_1 = \emptyset$ và $L_1 \cap A^* = L_1$
- Tính chất phân phối đối với phép hợp và phép giao:

$$(L_1 \cup L_2) \cap L_3 = (L_1 \cap L_3) \cup (L_2 \cap L_3)$$

$$(L_1 \cap L_2) \cup L_3 = (L_1 \cup L_3) \cap (L_2 \cup L_3)$$

3. Phép nhân ghép

Cho ngôn ngữ L_1 trên bảng chữ cái A_1 , L_2 trên bảng chữ cái A_2 . Phép nhân ghép của hai ngôn ngữ L_1, L_2 là một ngôn ngữ trên bảng chữ cái $A_1 \cup A_2$, kí hiệu $L_1 L_2$:

$$L_1 L_2 = \{ab | a \in L_1, b \in L_2\}$$

Nhận xét: Xét các ngôn ngữ L_1, L_2, L_3 . Phép nhân ghép có các tính chất sau

- Tính chất kết hợp: $(L_1 L_2) L_3 = L_1 (L_2 L_3)$
- $\forall L_1 : L_1 \emptyset = \emptyset L_1 = \emptyset$
- Tính chất phân phối đối với phép nhân ghép và phép hợp:

$$(L_1 \cup L_2) L_3 = (L_1 L_3) \cup (L_2 L_3)$$

$$L_1 (L_2 \cup L_3) = (L_1 L_2) \cup (L_1 L_3)$$

- Phép lấy phần bù Ngôn ngữ phần bù của ngôn ngữ L trên bảng chữ cái A , kí hiệu $C_A L$, là một ngôn ngữ trên bảng chữ cái A

$$C_A L = \{w \in A^* | w \notin L\}$$

Nhận xét: Phép lấy phần bù có các tính chất sau

- $C_A \{\epsilon\} = A^+, C_A A^+ = \{\epsilon\}$
- $C_A \emptyset = A^*, C_A A^* = \emptyset$
- $C_A (C_A L_1 \cup C_A L_2) = L_1 \cap L_2$

2.3 Ngôn ngữ chính quy

2.3.1 Định nghĩa

Ngôn ngữ chính quy được định nghĩa như sau:

1. \emptyset là một ngôn ngữ chính quy
2. $\forall a \in A, \{a\}$ là ngôn ngữ chính quy, A : bảng chữ cái
3. L_1, L_2 là các ngôn ngữ chính quy thì $L_1 \cup L_2, L_1 L_2, L^*$ là các ngôn ngữ chính quy
4. Không có bất kì ngôn ngữ chính quy nào khác ngoài 1,2,3.

2.4 Biểu thức chính quy

2.4.1 Định nghĩa

Biểu thức chính quy được định nghĩa như sau:

1. \emptyset là biểu thức chính quy
2. ϵ là biểu thức chính quy
3. a là ngôn ngữ chính quy với $a \in A$, A là bảng chữ cái hữu hạn
4. r, s là biểu thức chính quy

$$\left\{ \begin{array}{l} (r)(s) \rightarrow RS \\ r^* \rightarrow R^* \\ (r) + (s) \rightarrow R \cup S \end{array} \right. \text{ là các biểu thức chính quy}$$

2.4.2 Các tính chất của biểu thức chính quy

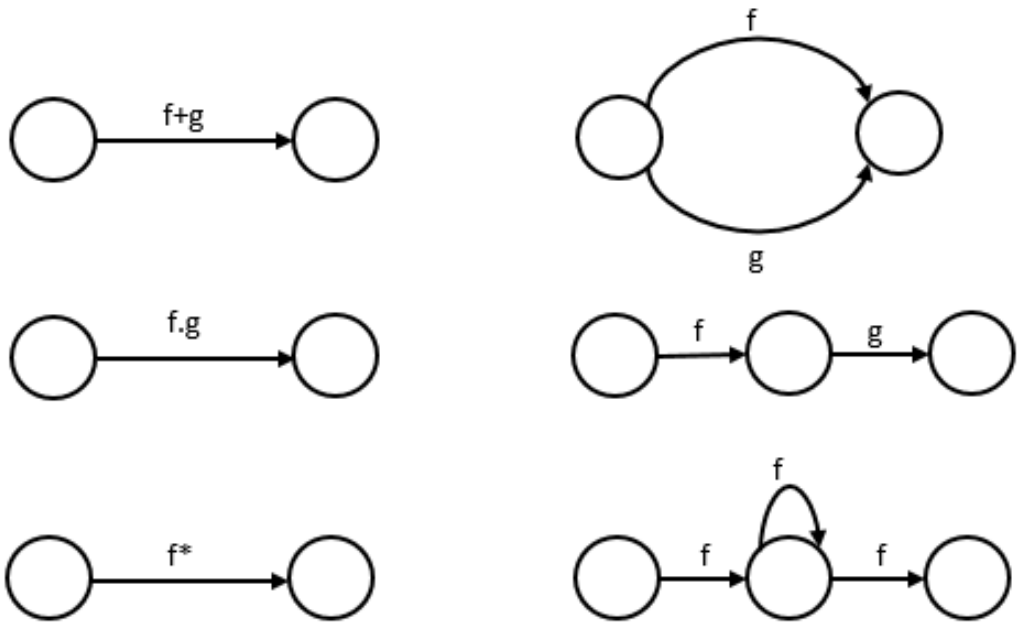
Cho r, s, w là các biểu thức chính quy

1. $r + s = s + r$
2. $(r + s) + w = r + (s + w)$
3. $r(s + w) = rs + rw$
4. $r\epsilon = \epsilon r = r$
5. $(\epsilon + r)^* = r^*$

6. $r + \emptyset = r$
7. $(r^*)^* = r^*$
8. $r + r = r$
9. $r(sw) = (rs)w$
10. $(r + s)w = rw + sw$
11. $\emptyset r = r\emptyset = \emptyset$
12. $\emptyset^* = \epsilon$
13. $r + r^* = r^*$
14. $(r^*s^*)^* = (r + s)^*$
15. $(rs)^*r = r(sr)^*$

2.4.3 Đồ thị biểu diễn biểu thức chính quy

- Mỗi đỉnh là một vòng tròn, đỉnh kết thúc được biểu diễn là một vòng tròn nhân đôi
- Đỉnh xuất phát thì mũi tên đi vào
- Các toán tử được biểu diễn như sau:



2.5 Automat đơn định(Deterministic Finite Automata)

1. Quy luật để chuyển trạng thái mới cho bởi hàm chuyển:

$$\delta : Q \times A \rightarrow Q$$

$$\delta(q, a) = p$$

$$q, p \in Q, a \in A$$

2. Định nghĩa

M là một DFA:

$$M = (A, Q, \delta, q_0, F)$$

Trong đó:

- A : bảng chữ cái hữu hạn
- Q : tập các trạng thái
- δ : hàm chuyển
- q_0 : trạng thái bắt đầu, $q_0 \in Q$
- F : tập trạng thái kết thúc $F \subseteq Q$

2.6 Automat đa định(Nondeterministic Finite Automata)

1. Quy luật để chuyển trạng thái mới cho bởi hàm chuyển:

$$\delta : Q \times (A \cup \epsilon) \rightarrow Q$$

$$\delta(q, a) = p$$

$$q, p \in Q, a \in A$$

2. Định nghĩa

M là một DFA:

$$M = (A, Q, \delta, q_0, F)$$

Trong đó:

- A : bảng chữ cái hữu hạn

- Q : tập các trạng thái
- δ : hàm chuyển
- q_0 : trạng thái bắt đầu, $q_0 \in Q$
- F : tập trạng thái kết thúc $F \subseteq Q$

2.7 Văn phạm

2.7.1 Định nghĩa

Văn phạm G là một bộ gồm 4 thành phần:

$$G = \{V_T, V_N, S, P\}$$

trong đó:

1. V_T là bảng chữ cái (hay bảng chữ cái kết thúc), mỗi kí tự của nó được gọi là một ký hiệu kết thúc.
2. V_N là bảng chữ cái $V_N \cap V_T = \emptyset$ (hay bảng chữ cái không kết thúc), mỗi phần tử của nó được gọi là một ký hiệu không kết thúc hay ký hiệu phụ
3. $S \in V_N$ được gọi là ký hiệu xuất phát
4. P là tập hợp các luật sinh có dạng $\alpha \rightarrow \beta$, α được gọi là vế trái, còn β được gọi là vế phải của một luật sinh, trong đó $\alpha, \beta \in (V_N \cup V_T)^*$ và α chứa ít nhất một ký tự không kết thúc.

2.7.2 Ngôn ngữ sinh bởi văn phạm

Định nghĩa. Cho văn phạm $G = \{V_T, V_N, S, P\}$ và $\eta, \omega \in (V_N \cup V_T)^*$. Ta nói ω suy dẫn trực tiếp từ η , ký hiệu $\eta \Rightarrow \omega$, nếu tồn tại luật sinh $\alpha \rightarrow \beta$, và $\gamma, \sigma \in (V_N \cup V_T)^*$ sao cho $\eta = \gamma\alpha\sigma$, $\omega = \gamma\beta\sigma$

Định nghĩa. Cho văn phạm $G = \{V_T, V_N, S, P\}$ và $\eta, \omega \in (V_N \cup V_T)^*$. Ta nói ω suy dẫn (hay suy dẫn gián tiếp) từ η , ký hiệu $\eta \Rightarrow^* \omega$, nếu $\eta = \omega$ hoặc tồn tại một dãy $D = \omega_0, \omega_1, \dots, \omega_k$ sao cho $\omega_0 = \eta$, $\omega_k = \omega$ và $\omega_i \Rightarrow \omega_{i+1}$

Định nghĩa. Cho văn phạm $G = \{V_T, V_N, S, P\}$. Xâu $\omega \in V_T^*$ được gọi là sinh bởi văn phạm G nếu tồn tại suy dẫn $S \Rightarrow^* \omega$. Ngôn ngữ sinh bởi văn phạm G , ký hiệu là $L(G)$, là tập hợp tất cả các xâu sinh bởi văn phạm G .

$$L(G) = \{\omega \in V_T^* | S \Rightarrow^* \omega\}$$

2.7.3 Văn phạm chính quy

Dựa vào đặc điểm của tập luật sinh mà người ta chia các văn phạm thành các nhóm khác nhau. Noam Chomsky đã phân loại văn phạm thành 4 nhóm:

1. Nhóm 0: Văn phạm không hạn chế
2. Nhóm 1: Văn phạm cảm ngữ cảnh
3. Nhóm 2: Văn phạm phi ngữ cảnh
4. Nhóm 3: Văn phạm chính quy

Trong khuôn khổ của bài báo cáo này, do ta chỉ quan tâm đến ngôn ngữ chính quy, nên bài báo cáo sẽ chỉ trình bày các kết quả liên quan đến văn phạm chính quy.

2.8 Thuật Toán

Chương 3

Sơ lược về lý thuyết mã

3.1 Mã

A là bảng chữ. X được gọi là mã nếu $\forall m, n$ và $x_1, x_2, \dots, x_n, y_1, y_2, \dots, y_m \in X$ thỏa mãn điều kiện:

$$x_1x_2\dots x_n = y_1y_2\dots y_m$$

Khi đó, $m = n$ và $x_i = y_i, i = \overline{1, n}$

3.2 Mã tối đại

Mã X được gọi là tối đại trên A nếu X không chứa thực sự trong một mã nào khác trên A .

Với mỗi lớp mã C trên A , một mã $X \in C$ là tối đại trong C (C không nhất thiết là mã tối đại) nếu nó không chứa thực sự trong một mã nào khác của C .

3.3 Mã prefix

Cho tập X là tập con của A^* . X được gọi là mã prefix nếu không có phần tử nào trên X nằm ở bên trái của một phần tử khác trong X

Nói cách khác, X là prefix nếu với mọi $x, x' \in X : x \leq x' \Rightarrow x = x'$

3.4 Tiêu chuẩn kiểm định mã

Bài toán:

- Input: 1 ngôn ngữ chính quy
- Output: Có phải là mã không?

Cho $X \subseteq A^+ : U_1 = X^{-1}X - \{\epsilon\}, \dots; U_{n+1} = X^{-1}U_n \cup U_n^{-1}X$ với $n \geq 1$

- Mệnh đề: Nếu X là một ngôn ngữ chính quy thì tập tất cả các U_n là hữu hạn với $n \geq 1$
- Điều kiện dừng của thuật toán:

Nếu $\epsilon \in U_i$ thì X không phải là mã

Nếu $U_i = U_n$ hoặc $U_n = \emptyset$ thì X là mã

Chương 4

Ứng dụng: Single-Error Correcting Codes

Chương 5

Kết luận

Tài liệu tham khảo