Θεωρία Υπολογισμού

Αυτόματα, Κανονικές Εκφράσεις και Κανονικές Γλώσσες

2003-04

Αυτόματα, Κανονικές Εκφράσεις και Κανονικές Γλώσσες

- Ιδιότητες των γλωσσών που γίνονται δεκτές από αυτόματα.
- Κανονικές εκφράσεις και κανονικές γλώσσες.
- Πεπερασμένα αυτόματα και κανονικές γλώσσες.
- Μη κανονικές γλώσσες.

Πράξεις σε Γλώσσες

- **Ένωση, τομή** και **διαφορά** όπως για οποιαδήποτε σύνολα.
- **Συμπλήρωμα**
  Αν L είναι μια γλώσσα με αλφάβητο Σ, τότε το συμπλήρωμα $\bar{L}$ της γλώσσας L ορίζεται ως εξής:
  $$\bar{L} = \Sigma^* \setminus L$$

Πράξεις σε Γλώσσες

- **Παράθεση**
  Αν L_1 και L_2 είναι γλώσσες του Σ, η παράθεση τους L_1L_2 ορίζεται ως εξής:
  $$L_1L_2 = \{w : \ w = xy \text{ για κάποιο } x \in L_1 \text{ και κάποιο } y \in L_2\}$$
- **Κλειστότητα** (closure ή Kleene closure ή Kleene star)
  Αν L είναι μια γλώσσα, ορίζουμε
  $$L^0 = \{e\} \text{ και } L^i = LL^{i-1}.$$
  Τότε η κλειστότητα L^* της γλώσσας L ορίζεται ως εξής:
  $$L^* = \cup_{i=0}^{\infty} L^i$$

Πράξεις σε Γλώσσες

- **Θετική κλειστότητα**
  Αν L είναι μια γλώσσα, η θετική κλειστότητα L^+ της γλώσσας L ορίζεται ως εξής:
  $$L^+ = \cup_{i=1}^{\infty} L^i$$
- **Ερώτηση:** Σε ποιές περιπτώσεις περιέχει η L^+ το e;
- **Ιδιότητα:** $L^+ = LL^*$

Γλώσσες Δεκτές από Πεπερασμένα Αυτόματα

- **Θεώρημα:** Η κλάση των γλωσσών που γίνονται δεκτές από πεπερασμένα αυτόματα είναι κλειστή ως προς τις ακόλουθες πράξεις: **ένωση**, **τομή**, **συμπλήρωμα**, **παράθεση** και **κλειστότητα**.
- **Απόδειξη:**
  - Εξετάζουμε κάθε μια πράξη.

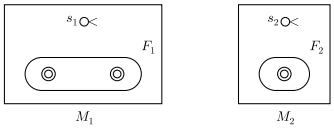
Ένωση Αυτομάτων

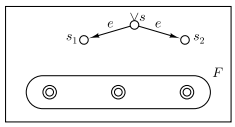
- Αν L_1 και L_2 είναι γλώσσες δεκτές από τα μη ντετερμινιστικά αυτόματα $M_1 = (K_1, \Sigma, \Delta_1, s_1, F_1)$ και $M_2 = (K_2, \Sigma, \Delta_2, s_2, F_2)$, τότε το αυτόματο $M = (K, \Sigma, \Delta, s, F)$ όπου
  - $K = K_1 \cup K_2 \cup \{s\}$
  - s είναι μια νέα αρχική κατάσταση
  - $F = F_1 \cup F_2$
  - $\Delta = \Delta_1 \cup \Delta_2 \cup \{(s, e, s_1), (s, e, s_2)\}$

  δέχεται τη γλώσσα $L_1 \cup L_2$.

Ένωση Αυτομάτων

- Η περίπτωση της ένωσης με γραφικό τρόπο:





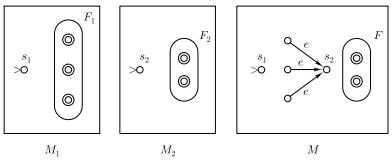
Παράθεση Αυτομάτων

- Αν L_1 και L_2 είναι γλώσσες δεκτές από τα μη ντετερμινιστικά αυτόματα $M_1 = (K_1, \Sigma, \Delta_1, s_1, F_1)$ και $M_2 = (K_2, \Sigma, \Delta_2, s_2, F_2)$, τότε το αυτόματο $M = (K, \Sigma, \Delta, s, F)$ όπου
  - $K = K_1 \cup K_2 \cup \{s\}$
  - $s = s_1$
  - $F = F_2$
  - $\Delta = \Delta_1 \cup \Delta_2$
  - $(F_1 \times \{e\} \times \{s_2\})$

  δέχεται τη γλώσσα $L_1 L_2$.

Παράθεση Αυτομάτων

- Η περίπτωση της παράθεσης με γραφικό τρόπο:



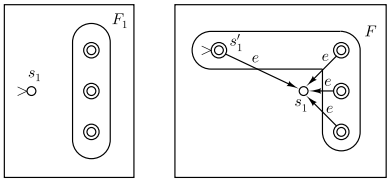
Κλειστότητα Αυτομάτων

- Αν L_1 είναι μια γλώσσα δεκτή από το μη ντετερμινιστικό αυτόματο $M_1 = (K_1, \Sigma, \Delta_1, s_1, F_1)$ τότε το αυτόματο $M = (K, \Sigma, \Delta, s'_1, F)$ όπου
  - $K = K_1 \cup \{s'_1\}$
  - s'_1 είναι μια καινούρια αρχική κατάσταση
  - $F = F_1 \cup \{s'_1\}$
  - $\Delta = \Delta_1 \cup (F_1 \times \{e\} \times \{s_1\}) \cup \{(s'_1, e, s_1)\}$

  δέχεται τη γλώσσα L_1^*.

Κλειστότητα Αυτομάτων

- Η περίπτωση της κλειστότητας με γραφικό τρόπο:



Συμπλήρωμα Αυτομάτων

- Αν L είναι μια γλώσσα που γίνεται δεκτή από το μη ντετερμινιστικό πεπερασμένο αυτόματο $M = (K, \Sigma, \Delta, s, F)$, τότε η γλώσσα $\Sigma^* \setminus L$ γίνεται δεκτή από το αυτόματο

$$\bar{M} = (K, \Sigma, \Delta, s, K \setminus F).$$

Τομή Αυτομάτων

- Αν L_1 και L_2 είναι γλώσσες, τότε

$$L_1 \cap L_2 = \Sigma^* \setminus ((\Sigma^* \setminus L_1) \cup (\Sigma^* \setminus L_2)).$$

Μπορούμε λοιπόν να χρησιμοποιήσουμε τις τεχνικές από τις προηγούμενες πράξεις.

Άσκηση

- Να βρείτε αλγόριθμους για τα ακόλουθα προβλήματα:
  - Αν M είναι ένα αυτόματο και w μια συμβολοσειρά, ισχύει $w \in L(M)$;
  - Αν M είναι ένα αυτόματο, ισχύει $L(M) = \emptyset$;
  - Αν M είναι ένα αυτόματο, ισχύει $L(M) = \Sigma^*$;
  - Αν M_1 και M_2 είναι πεπερασμένα αυτόματα, ισχύει $L(M_1) \subseteq L(M_2)$;
  - Αν M_1 και M_2 είναι πεπερασμένα αυτόματα, ισχύει $L(M_1) = L(M_2)$;

Κανονικές Εκφράσεις και Κανονικές Γλώσσες

- Αν Σ είναι ένα αλφάβητο, οι **κανονικές εκφράσεις** του Σ, και οι γλώσσες που παριστάνουν (**κανονικές γλώσσες**), ορίζονται ως εξής:
  - Το $\emptyset$ είναι μια κανονική έκφραση και παριστάνει την κενή γλώσσα.
  - Το **e** είναι μια κανονική έκφραση και παριστάνει τη γλώσσα $\{e\}$.
  - Για κάθε $a \in \Sigma$, **a** είναι μια κανονική έκφραση και παριστάνει τη γλώσσα $\{a\}$.
  - Αν r και s είναι κανονικές εκφράσεις που παριστάνουν τις γλώσσες R και S, τότε οι εκφράσεις $r \cup s$, rs και r^* είναι κανονικές εκφράσεις που παριστάνουν τις γλώσσες $R \cup S$, RS, και R^* αντίστοιχα.

Κανονικές Εκφράσεις

- Οι κανονικές εκφράσεις είναι ένας κομψός τρόπος για να παριστάνουμε κανονικές γλώσσες (πεπερασμένες ή άπειρες).
- Ποιές γλώσσες παριστάνουν οι παρακάτω κανονικές εκφράσεις του αλφάβητου $\{0,1\}$;
  - $\mathbf{00}$
  - $(\mathbf{0} \cup \mathbf{1})^*$
  - $(\mathbf{0} \cup \mathbf{1})^*\mathbf{00}(\mathbf{0} \cup \mathbf{1})^*$
  - $(\mathbf{1} \cup \mathbf{10})^*$
  - $(\mathbf{0} \cup \mathbf{1})^*\mathbf{011}$

Κανονικές Εκφράσεις

- Αν το αλφάβητο είναι

$$\Sigma = \{a, b, \ldots, z, A, B, \ldots, Z, 0, 1, \ldots, 9\}$$

ποιά γλώσσα παριστάνει η παρακάτω κανονική έκφραση;

$$(letter)(letter \cup digit)^*$$

όπου

$$letter \;=\; a \cup b \cup \cdots \cup z \cup A \cup B \cup \cdots \cup Z$$

και

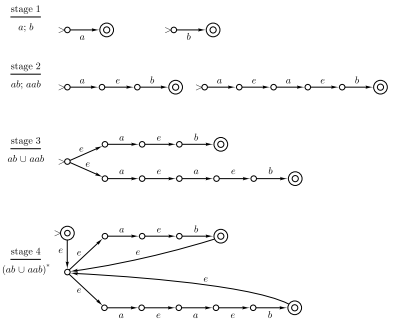
$$digit \;=\; 0 \cup 1 \cup \cdots \cup 9.$$

Πεπερασμένα Αυτόματα και Κανονικές Εκφράσεις

- **Θεώρημα:** Μια γλώσσα είναι κανονική αν και μόνο αν είναι δεκτή από ένα πεπερασμένο αυτόματο.
- **Απόδειξη:** (**Μόνο αν**)
  Είναι εύκολο να βρούμε αυτόματα για τις κανονικές γλώσσες $\emptyset$, $\{e\}$ και $\{\sigma\}$ $(\sigma \in \Sigma)$.
  Για όλες τις υπόλοιπες γλώσσες μπορούμε να κατασκευάσουμε αυτόματα αρχίζοντας με αυτόματα για τις παραπάνω.

Παράδειγμα

- Σταδιακή κατασκευή αυτομάτου για την κανονική έκφραση $(ab \cup aab)^*$:



Πεπερασμένα Αυτόματα και Κανονικές Εκφράσεις

- **Απόδειξη:** (**Αν**)
  Έστω

$$M = (\{q_1, \ldots, q_n\}, \Sigma, \delta, q_1, F)$$

ένα ντετερμινιστικό πεπερασμένο αυτόματο.
Ονομάζουμε $R(i,j,k)$ το σύνολο όλων των συμβολοσειρών του Σ^* που οδηγούν το M από τη κατάσταση q_i στην q_j χωρίς να περάσει από κατάσταση με αριθμό k ή μεγαλύτερο.

Πεπερασμένα Αυτόματα και Κανονικές Εκφράσεις

- Το σύνολο $R(i,j,k)$ είναι κανονικό για οποιαδήποτε $i,j = 1,\ldots,n$ και $k = 1,\ldots,n+1$ διότι:

$$R(i,j,1) = \begin{cases} \{\sigma \in \Sigma :\ \delta(q_i,\sigma) = q_j\} & \text{αν } i \neq j \\ \{e\} \cup \{\sigma \in \Sigma :\ \delta(q_i,\sigma) = q_j\} & \text{αν } i = j \end{cases}$$

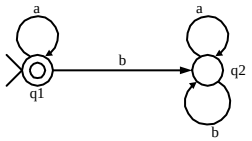
$$R(i,j,k+1) = R(i,j,k) \cup R(i,k,k)R(k,k,k)^*R(k,j,k)$$

Με λίγη σκέψη βλέπουμε ότι

$$L(M) = \bigcup \{R(1,j,n+1) :\ q_j \in F\}$$

Παράδειγμα

- Σταδιακή κατασκευή κανονικής έκφρασης για το αυτόματο:



Παράδειγμα

- Αφού $F = \{q_1\}$, πρέπει να βρούμε τη γλώσσα

$$L(M) = R(1,1,3).$$

Θα χρησιμοποιήσουμε την ισότητα

$$R(1,1,3) = R(1,1,2) \cup R(1,2,2)R(2,2,2)^*R(2,1,2)$$

και θα υπολογίσουμε τα σύνολα του δεύτερου μέλους.

Παράδειγμα

- Δηλαδή

$$R(1,1,2) = R(1,1,1) \cup R(1,1,1)R(1,1,1)^*R(1,1,1) =$$
$$= \{e,a\} \cup \{e,a\}\{e,a\}^*\{e,a\} = a^*$$
$$R(1,2,2) = R(1,2,1) \cup R(1,1,1)R(1,1,1)^*R(1,2,1) =$$
$$= \{b\} \cup \{e,a\}\{e,a\}^*\{b\} = a^*b$$
$$R(2,2,2) = R(2,2,1) \cup R(2,1,1)R(1,1,1)^*R(1,2,1) =$$
$$= \{a,b,e\} \cup \emptyset\, R(1,1,1)^*\, R(1,2,1) = \{a,b,e\}$$
$$R(2,1,2) = R(2,1,1) \cup R(2,1,1)R(1,1,1)^*R(1,1,1) =$$
$$= \emptyset \cup \emptyset\, R(1,1,1)^*\, R(1,1,1) = \emptyset$$

Αρα

$$L(M) = R(1,1,3) = a^*.$$

Κανονικές Γλώσσες

- **Θεώρημα:** Η κλάση των κανονικών γλωσσών είναι κλειστή ως πρός τις ακόλουθες πράξεις: ένωση, τομή, συμπλήρωμα, παράθεση, και κλειστότητα.
- **Απόδειξη:** Προφανής, αφού η κλάση των κανονικών γλωσσών είναι ίδια με την κλάση των γλωσσών που γίνονται δεκτές από ένα πεπερασμένο αυτόματο.

Άσκηση

- Αν $S = \{0,1,\ldots,9\}$, τότε αποδείξτε ότι η παρακάτω γλώσσα είναι κανονική:

$$L = \{w \in \Sigma^* : \ w \ \text{είναι δεκαδική παράσταση ενός φυσικού}$$
$$\text{αριθμού που διαιρείται με το 2 ή το 3}\}$$

Λύση

- Η κανονική γλώσσα που περιγράφει τις δεκαδικές παραστάσεις των φυσικών αριθμών είναι $L_1 = \{0\} \cup \{1,\ldots,9\}\Sigma^*$.
- Η κανονική γλώσσα που περιγράφει τις δεκαδικές παραστάσεις των φυσικών αριθμών που διαιρούνται με το 2 είναι

$$L_2 = L_1 \cap \Sigma^*\{0,2,4,6,8\}.$$

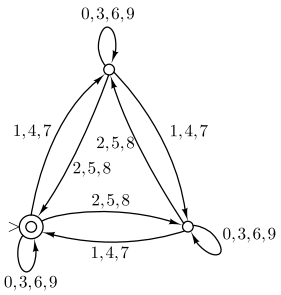
- Η κανονική γλώσσα που περιγράφει τις δεκαδικές παραστάσεις των φυσικών αριθμών που διαιρούνται με το 3 είναι

$$L_3 = L_1 \cap L(M)$$

όπου M είναι το αυτόματο της επόμενης σελίδας.

- Τελικά $L = L_2 \cup L_3$. Αρα η L είναι κανονική.

Λύση



Άσκηση

- Αν $S = \{a,b\}$, τότε αποδείξτε ότι η παρακάτω γλώσσα είναι κανονική:

$$L = \{w \in \Sigma^* : \ w \ \text{είναι συμβολοσειρά περιττού μήκους που}$$
$$\text{περιέχει άρτιο αριθμό από } a\}$$

Λύση

- Η κανονική γλώσσα που περιγράφει τις συμβολοσειρές περιττού μήκους είναι

$$L_1 = \Sigma(\Sigma\Sigma)^*$$

- Η κανονική γλώσσα που περιγράφει τις συμβολοσειρές με άρτιο αριθμό από a είναι

$$L_2 = b^*(ab^*ab^*)^*$$

- Τελικά $L = L_1 \cap L_2$. Αρα η L είναι κανονική.

Μη Κανονικές Γλώσσες

- Μια πεπερασμένη γλώσσα είναι πάντα κανονική. Γιατί;
- Πως μπορούμε να αποδείξουμε ότι μια **άπειρη** γλώσσα δεν είναι κανονική;
- **Παρατηρήσεις:**
  - Οι άπειρες κανονικές γλώσσες έχουν κάποια επαναληπτική δομή που προέρχεται από την πράξη *.
  - Οι κανονικές γλώσσες χρειάζονται πεπερασμένη "μνήμη" για να αναγνωριστούν. Η μνήμη αυτή πρέπει να εξαρτάται από τη γλώσσα και όχι από κάθε συμβολοσειρά της!

Μη Κανονικές Γλώσσες

- **Ερώτηση:**
  Αν $\Sigma = \{a, b\}$, είναι οι γλώσσες
  - $L = \{a^n b^n : n \geq 0\}$
  - $L = \{a^n : n \text{ είναι πρώτος αριθμός}\}$

  κανονικές;

Μη Κανονικές Γλώσσες

- Το παρακάτω **θεώρημα άντλησης** μπορεί να χρησιμοποιηθεί για να δείξουμε ότι μια γλώσσα δεν είναι κανονική.
- **Θεώρημα:** Έστω L μια άπειρη κανονική γλώσσα. Τότε υπάρχουν συμβολοσειρές x, y, z τέτοιες ώστε $y \neq e$ και $xy^n z \in L$ για κάθε $n \geq 0$.
- **Απόδειξη.** Εφόσον η L είναι κανονική, γίνεται δεκτή από ένα πεπερασμένο αυτόματο M. Ας υποθέσουμε ότι το M έχει n καταστάσεις. Καθώς η L είναι άπειρη το L θα δέχεται και μία λέξη w μήκους n ή μεγαλύτερη.

Συνέχεια της Απόδειξης

- Έστω $l = |w|$, $w = \sigma_1\sigma_2...\sigma_l$ όπου κάθε $\sigma_i \in \Sigma$ και θεωρείστε τον υπολογισμό του M με είσοδο w:

$$(q_0, \sigma_1...\sigma_l) \vdash_M (q_1, \sigma_2...\sigma_l) \vdash_M \ldots \vdash_M (q_{l-1}, \sigma_l) \vdash_M (q_l, e)$$

  όπου q_0 είναι η αρχική κατάσταση του M και q_l μία τελική κατάσταση του M.
  Αφού $l \geq n$ και το M έχει μόνο n καταστάσεις, από την Αρχή του Περιστεριώνα υπάρχουν i και j τέτοια ώστε $0 \leq i < j \leq l$ έτσι ώστε $q_i = q_j$.

Συνέχεια της Απόδειξης

- Δηλαδή η συμβολοσειρά $\sigma_{i+1} \cdots \sigma_j$ οδηγεί το M από τη κατάσταση q_i πίσω στη κατάσταση q_i, και η συμβολοσειρά αυτή δεν είναι κενή καθώς $i + 1 \leq j$. Τότε όμως η κατάσταση αυτή μπορεί να αφαιρεθεί από τη w, ή θα μπορούσε να προστεθεί οποιοσδήποτε αριθμός επαναλήψεων της μέσα στη w αμέσως μετά το j-ιοστό σύμβολό της, και το M θα δεχόταν τη w.
- Δηλαδή το M δέχεται τη συμβολοσειρά $\sigma_1 \cdots \sigma_i (\sigma_{i+1} \cdots \sigma_j)^n (\sigma_{j+1} \cdots \sigma_l)$ για κάθε $n \geq 0$. Γιατί αν $x = \sigma_1 \cdots \sigma_i, y = \sigma_{i+1} \cdots \sigma_j, z = \sigma_{j+1} \cdots \sigma_l$ τότε

$$(q_0, xy^n z) \vdash^*_M (q_i, y^n z) \vdash^*_M (q_i, y^{n-1} z) \cdots \vdash^*_M (q_i, z) \vdash^*_M (q_l, e).$$

Παράδειγμα

- Η γλώσσα

$$L = \{a^n b^n : n \geq 0\}$$

  δεν είναι κανονική.
- **Απόδειξη:** Από το θεώρημα άντλησης έχουμε ότι υπάρχουν συμβολοσειρές x, y, z τέτοιες ώστε $y \neq e$ και $xy^n z \in L$ για κάθε $n \geq 0$.
  Αρα υπάρχουν τρεις δυνατότητες για το y, και θα δείξουμε για την κάθε περίπτωση ότι η L πρέπει να περιέχει κάποια συμβολοσειρά που δεν έχει τη σωστή μορφή.

Συνέχεια της Απόδειξης

- Το y να αποτελείται εξ ολοκλήρου από a. Τότε $x = a^p, y = a^q$ και $z = a^r b^s$ όπου $p, r, s \geq 0$ και $q > 0$.
  Όμως τότε η L πρέπει να περιέχει τη συμβολοσειρά $xy^n z = a^{p+nq+r} b^s$ για κάθε $n \geq 0$ και μόνο μία το πολύ συμβολοσειρά αυτής της μορφής έχει ίσο αριθμό a και b.
- Το y να αποτελείται εξ ολοκλήρου από b. Τότε έχουμε μία περίπτωση όμοια με την περίπτωση 1.
- Το y να περιέχει a και b. Τότε για κάποιο $n > 1$ η συμβολοσειρά $xy^n z$ περιέχει εμφάνιση του b πριν από εμφάνιση του a και έτσι δεν μπορεί να ανήκει στην L.

Παράδειγμα

- Η γλώσσα

$$L = \{a^n : n \text{ είναι πρώτος αριθμός}\}$$

  δεν είναι κανονική.
- **Απόδειξη:** Χρησιμοποιείστε το θεώρημα άντλησης.

Εφαρμογές Αυτομάτων και Κανονικών Γλωσσών

- LEX
  - Το πρόγραμμα lex του Unix χρησιμοποιείται για την αυτόματη παραγωγή προγραμμάτων **λεκτικής ανάλυσης** (σε C).
  - Η λεκτική ανάλυση είναι το πρώτο στάδιο της **μεταγλώττισης** ενός προγράμματος. Σ' αυτό το στάδιο ένας μεταγλωττιστής διαβάζει μια ακολουθία χαρακτήρων και τους ταξινομεί σε συμβολοσειρές που αναλογούν σε κάποιες δεδομένες κανονικές εκφράσεις.
  - Το πρόγραμμα lex χρησιμοποιείται συνήθως μαζί με το πρόγραμμα yacc που παράγει προγράμματα **συντακτικής ανάλυσης**.
  - Περισσότερες πληροφορίες μπορείτε να βρείτε στο βιβλίο "Compilers - Principles, Techniques and Tools" των A.V. Aho, R. Sethi and J.D. Ullman.

Εφαρμογές Αυτομάτων και Κανονικών Γλωσσών

- Κειμενογράφοι (π.χ. vi στο Unix)
  - Εύρεση συμβολοσειρών που αντιστοιχούν σε κανονικές εκφράσεις και αντικατάσταση τους από άλλες συμβολοσειρές.
  - Παράδειγμα: Εντολή *s/ */ /g*

Μελέτη

- Κεφάλαια 2.4-2.6 από το βιβλίο:
  Harry R. Lewis και Χρίστος Χ. Παπαδημητρίου, Στοιχεία Θεωρίας Υπολογισμού, Τεχνικό Επιμελητήριο Ελλάδας, 1992.