ΑΡΙΣΤΟΤΕΛΕΙΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΘΕΣΣΑΛΟΝΙΚΗΣ

ΑΝΟΙΚΤΑ ΑΚΑΔΗΜΑΙΚΑ ΜΑΘΗΜΑΤΑ

Θεωρία Υπολογισμού

Ενότητα 7: Πεπερασμένη αναπαράσταση γλωσσών

Επ. Καθ. Π. Κατσαρός
Τμήμα Πληροφορικής

Ευρωπαϊκή Ένωση
Ευρωπαϊκό Κοινωνικό Ταμείο
ΕΠΙΧΕΙΡΗΣΙΑΚΟ ΠΡΟΓΡΑΜΜΑ
ΕΚΠΑΙΔΕΥΣΗ ΚΑΙ ΔΙΑ ΒΙΟΥ ΜΑΘΗΣΗ
επένδυση στην κοινωνία της γνώσης
ΥΠΟΥΡΓΕΙΟ ΠΑΙΔΕΙΑΣ ΚΑΙ ΘΡΗΣΚΕΥΜΑΤΩΝ
ΕΙΔΙΚΗ ΥΠΗΡΕΣΙΑ ΔΙΑΧΕΙΡΙΣΗΣ
ΕΣΠΑ 2007-2013
πρόγραμμα για την ανάπτυξη
ΕΥΡΩΠΑΪΚΟ ΚΟΙΝΩΝΙΚΟ ΤΑΜΕΙΟ
Με τη συγχρηματοδότηση της Ελλάδας και της Ευρωπαϊκής Ένωσης

Άδειες Χρήσης

Χρηματοδότηση

- Το παρόν εκπαιδευτικό υλικό έχει αναπτυχθεί στα πλαίσια του εκπαιδευτικού έργου του διδάσκοντα.
- Το έργο «Ανοικτά Ακαδημαϊκά Μαθήματα στο Αριστοτέλειο Πανεπιστήμιο Θεσσαλονίκης» έχει χρηματοδοτήσει μόνο την αναδιαμόρφωση του εκπαιδευτικού υλικού.
- Το έργο υλοποιείται στο πλαίσιο του Επιχειρησιακού Προγράμματος «Εκπαίδευση και Δια Βίου Μάθηση» και συγχρηματοδοτείται από την Ευρωπαϊκή Ένωση (Ευρωπαϊκό Κοινωνικό Ταμείο) και από εθνικούς πόρους.

Ευρωπαϊκή Ένωση
Ευρωπαϊκό Κοινωνικό Ταμείο
ΕΠΙΧΕΙΡΗΣΙΑΚΟ ΠΡΟΓΡΑΜΜΑ
ΕΚΠΑΙΔΕΥΣΗ ΚΑΙ ΔΙΑ ΒΙΟΥ ΜΑΘΗΣΗ
επένδυση στην κοινωνία της γνώσης
ΥΠΟΥΡΓΕΙΟ ΠΑΙΔΕΙΑΣ ΚΑΙ ΘΡΗΣΚΕΥΜΑΤΩΝ
ΕΙΔΙΚΗ ΥΠΗΡΕΣΙΑ ΔΙΑΧΕΙΡΙΣΗΣ
ΕΣΠΑ 2007-2013
πρόγραμμα για την ανάπτυξη
ΕΥΡΩΠΑΪΚΟ ΚΟΙΝΩΝΙΚΟ ΤΑΜΕΙΟ
Με τη συγχρηματοδότηση της Ελλάδας και της Ευρωπαϊκής Ένωσης

1. Κανονικές εκφράσεις
2. Κλάσεις γλωσσών
    - Κλάσεις και ιδιότητες γλωσσών
    - Κανονικές γλώσσες
    - Αναγνώριση γλωσσών
    - Παραγωγή γλωσσών
    - Εφαρμογές

Πεπερασμένη αναπαράσταση γλωσσας

- Κάθε πεπερασμένη γλώσσα μπορεί να περιγραφεί με *εξαντλητική απαρίθμηση* των συμβολοσειρών της. Υπάρχει δηλαδή τουλάχιστο μια *πεπερασμένη αναπαράσταση* της γλώσσας.
- Τι γίνεται με τις άπειρες γλώσσες;
- Κάθε πεπερασμένη αναπαράσταση μιας (πεπερασμένης ή άπειρης) γλώσσας πρέπει
  - να είναι η ίδια συμβολοσειρά ενός αλφαβήτου Σ και
  - να διαφέρει από την αναπαράσταση όλων των άλλων γλωσσών

Κανονικές εκφράσεις

Ορισμός 1 (Κανονική έκφραση)

Κάθε *κανονική έκφραση* περιγράφει μία γλώσσα με απλά σύμβολα, το $\emptyset$ και συνδυασμούς τους που προκύπτουν από τις πράξεις που ορίζουν τα σύμβολα $\cup$ και * και, ίσως, με παρενθέσεις.

Παράδειγμα: $L = \{w \in \{a, b\}^* :$ η w έχει μία ή δύο εμφανίσεις του b, αλλά όχι συνεχόμενες$\}$

$$L = a^* ba^* (aba^* \cup \emptyset^*)$$

Κανονικές εκφράσεις

- *Κανονικές εκφράσεις* επί του Σ^* είναι οι συμβολοσειρές στο αλφάβητο $\Sigma \cup \{(,),\cup,\emptyset,^*\}$ που σχηματίζονται ως εξής:
  - το $\emptyset$ και κάθε στοιχείο του Σ είναι κανονική έκφραση
  - αν α και β είναι κανονικές εκφράσεις, τότε
    - και η $(\alpha\beta)$ είναι κανονική έκφραση
    - και η $(\alpha\cup\beta)$ είναι κανονική έκφραση
  - αν η α είναι κανονική έκφραση, τότε και η α^* είναι κανονική έκφραση
  - μία συμβολοσειρά αν δεν ικανοποιεί κάποιο από τα παραπάνω δεν είναι κανονική έκφραση

Κανονικές εκφράσεις

- Η $\mathcal{L}$ είναι μία *συνάρτηση* από συμβολοσειρές σε γλώσσες: Μία κανονική έκφραση α αναπαριστά τη γλώσσα $\mathcal{L}(\alpha)$.
- Η συνάρτηση $\mathcal{L}$ ορίζεται ως εξής:
  1. $\mathcal{L}(\emptyset) = \emptyset$ και $\mathcal{L}(\alpha)=\{\alpha\}$ για κάθε $\alpha\in\Sigma$
  2. Αν α και β είναι κανονικές εκφράσεις, τότε
     - $\mathcal{L}(\alpha\beta) = \mathcal{L}(\alpha)\mathcal{L}(\beta)$
     - $\mathcal{L}(\alpha\cup\beta) = \mathcal{L}(\alpha)\cup\mathcal{L}(\beta)$
  3. Αν α είναι κανονική έκφραση, τότε
     - $\mathcal{L}(\alpha^*) = \mathcal{L}(\alpha)^*$

Αν $\mathcal{L}(\alpha)=\{\text{base,foot}\}$, $\mathcal{L}(\beta)=\{\text{ball}\}$,
$\mathcal{L}(\alpha\beta)=\{\text{baseball,football}\}$
$\mathcal{L}(\alpha\cup\beta)=\{\{\text{base,foot}\},\{\text{ball}\}\}$

- Ο τελεστής * έχει υψηλότερη προτεραιότητα από τον $\cup$.

Χαρακτηριστικά κανονικών εκφράσεων

- Αν μια γλώσσα μπορεί να περιγραφεί από μία κανονική έκφραση, μπορεί να περιγραφεί από άπειρες κανονικές εκφράσεις.
  Παράδειγμα:
  οι εκφράσεις α και (α∪∅) αναπαριστούν την ίδια γλώσσα.
- Αφού η ένωση και η παράθεση συνόλων είναι προσεταιριστικές, οι περιττές παρενθέσεις μπορούν να παραλείπονται.
  Παράδειγμα:
  - $(\mathcal{L}_1 \cup \mathcal{L}_2) \cup \mathcal{L}_3 = \mathcal{L}_1 \cup (\mathcal{L}_2 \cup \mathcal{L}_3) = \mathcal{L}_1 \cup \mathcal{L}_2 \cup \mathcal{L}_3$
  - $(\mathcal{L}_1\mathcal{L}_2)\mathcal{L}_3 = \mathcal{L}_1(\mathcal{L}_2\mathcal{L}_3) = \mathcal{L}_1\mathcal{L}_2\mathcal{L}_3$

Παράδειγμα γλώσσας κανονικής έκφρασης

$$
\begin{aligned}
\mathcal{L}((\alpha\cup\beta)^*\alpha) &= \mathcal{L}((\alpha\cup\beta)^*)\mathcal{L}(\alpha) && \text{από } \mathcal{L}(\alpha\beta) = \mathcal{L}(\alpha)\mathcal{L}(\beta) \\
&= \mathcal{L}((\alpha\cup\beta)^*)\{\alpha\} && \text{από } \mathcal{L}(\alpha)=\{\alpha\} \text{ για } \alpha\in\Sigma \\
&= \mathcal{L}(\alpha\cup\beta)^*\{\alpha\} && \text{από } \mathcal{L}(\alpha^*)=\mathcal{L}(\alpha)^* \\
&= (\mathcal{L}(\alpha)\cup \mathcal{L}(\beta))^*\{\alpha\} && \text{από } \mathcal{L}(\alpha\cup\beta)=\mathcal{L}(\alpha)\cup\mathcal{L}(\beta) \\
&= (\{\alpha\}\cup\{\beta\})^*\{\alpha\} && \text{από } \mathcal{L}(\alpha)=\{\alpha\} \text{ για } \alpha\in\Sigma \\
&= \{w \in\{\alpha,\beta\}^*\text{: όπου } w \text{ τελειώνει σε } \alpha\}
\end{aligned}
$$

Παραδείγματα γλωσσών κανονικών εκφράσεων

- $\mathcal{L}(01)=\{01\}$
- $\mathcal{L}(01\cup 0)=\{01,0\}$
- $\mathcal{L}(0(1\cup 0))= \{01,00\}$
- $\mathcal{L}(0^*)=\{\varepsilon,0,00,000,\ldots\}$
- $\mathcal{L}((0\cup 10)^*(\varepsilon\cup 1))$
  $=\{w \in \{0,1\}^*\colon$ η w δεν περιλαμβάνει 2 συνεχόμενα 1$\}$

Παραδείγματα γλωσσών κανονικών εκφράσεων

- Η κανονική έκφραση $(0^*1^*)^*(0110)(0^*1^*)^*$ αναπαριστά το σύνολο των συμβολοσειρών του $\{0,1\}$ που περιέχουν τουλάχιστο μία φορά τη συμβολοσειρά 0110.
  - 0^*1^* - κάποια (ίσως κανένα) 0 ακολουθούμενα από κάποια (ίσως κανένα) 1
  - $(0^*1^*)^*$ - τα 0 και 1 εναλλάσονται 0 ή περίσσότερες φορές
  - (0110) - η συμβολοσειρά 0110 εμφανίζεται αυτούσια
- Συμβάσεις:
  - $\alpha + = \alpha\alpha^*$ (παράθεση της α τουλάχιστο μία φορά)
  - $\alpha ? = \alpha \cup \emptyset$ (παράθεση της α το πολύ μία φορά)

Κλάσεις γλωσσών

Ορισμός 2 (Κλάση γλωσσών)

Κλάση γλωσσών είναι ένα σύνολο γλωσσών με κάποιες κοινές ιδιότητες.

- Παράδειγμα: Ή κλάση των κανονικών γλωσσών
- Ιδιότητες γλωσσών
  - Ιδιότητες απόφασης
  - Ιδιότητες κλειστότητας

Ιδιότητες των γλωσσών

Ορισμός 3 (Ιδιότητες κλειστότητας)

Μία *ιδιότητα κλειστότητας* μίας κλάσης γλωσσών δηλώνει ότι κάποια πράξη που εφαρμόζεται στις γλώσσες της κλάσης παράγει γλώσσα που βρίσκεται στην ίδια κλάση.

Ορισμός 4 (Ιδιότητες απόφασης)

Μία *ιδιότητα απόφασης* για μία κλάση γλωσσών είναι ένας αλγόριθμος που δέχεται ως είσοδο την τυπική αναπαράσταση μιας γλώσσας και δίνει ως έξοδο την απόφαση αν μια ιδιότητα ισχύει.

Κανονικές γλώσσες

Ορισμός 5 (Κανονικές γλώσσες)

Η κλάση των *κανονικών γλωσσών* ενός αλφαβήτου Σ αποτελείται από τις γλώσσες L που περιγράφονται από κάθε κανονική έκφραση που μπορεί να οριστεί στο Σ.

- Συνεπώς, κάθε γλώσσα που μπορεί να περιγραφεί από κανονικές εκφράσεις είναι κανονική γλώσσα.
- Η κλάση των κανονικών γλωσσών με αλφάβητο Σ είναι κλειστή ως προς την ένωση, παράθεση και Kleene star.
- Ωστόσο, οι κανονικές εκφράσεις δε μπορούν να αναπαραστήσουν όλες τις απλές γλώσσες.
  Παράδειγμα μη κανονικής γλώσσας: $\{0^n1^n : n \geq 0\}$
- Εναλλακτικά, χρησιμοποιούνται αναπαραστάσεις του τύπου
  $L = \{w \in \Sigma^* : w$ έχει την ιδιότητα $P\}$.

Μηχανή αναγνώρισης γλώσσας

Ορισμός 6 (Μηχανή αναγνώρισης γλώσσας)

Η *μηχανή αναγνώρισης γλώσσας* είναι ένας αλγόριθμος που «αποφασίζει» αν μία συμβολοσειρά w ανήκει στη γλώσσα L.

- Παράδειγμα: Μία μηχανή αναγνώρισης της γλώσσας $L = \{w \in \{a, b\}^*$: η w δεν περιέχει τη συμβολοσειρά ααα$\}$ μπορεί να λειτουργεί ως εξής:
    - Αρχικοποίησε ένα μετρητή με τιμή μηδέν.
    - Διάβαζε ένα σύμβολο σ τη φορά, από τα αριστερά προς τα δεξιά:
        - Αν σ=β, μηδένισε το μετρητή
        - Αν σ=α, πρόσθεσε 1 στο μετρητή.
        - Επέστρεψε 'ΟΧΙ' αν ο μετρητής είναι τρία.
    - Επέστρεψε 'ΝΑΙ' αν μετρητής έχει τιμή μικρότερη από τρία και δεν υπάρχουν άλλα σύμβολα.

Παράγωγοι γλωσσών

- Οι κανονικές εκφράσεις περιγράφουν έναν τρόπο παραγωγής των συμβολοσειρών μιας γλώσσας.
- Παράδειγμα: $(e \cup 1 \cup 11)(0 \cup 01 \cup 011)^*$
  «Γράψε τίποτα ή 1 ή 11. Έπειτα γράψε 0 ή 01 ή 011 οσες φορές θέλεις».
- Τέτοιες περιγραφές *παράγωγης γλωσσών* δεν είναι αλγόριθμοι γιατί
  - δεν έχουν σχεδιαστεί ώστε να δίνουν ένα μόνο αποτέλεσμα
  - δεν είναι σαφώς ορισμένες οι ενέργειες που εκτελούνται (πώς επιλέγουμε αν θα γράψουμε 0 ή 01 ή 011)

Θα εξετάσουμε για πολλές διαφορετικές κλάσεις γλωσσών (ξεκινώντας από τις κανονικές γλώσσες) το πώς σχετίζονται οι μηχανισμοί παραγωγής με τις μηχανές αναγνώρισης των γλωσσών αυτών.

Εφαρμογές των κανονικών εκφράσεων

- Οι κανονικές εκφράσεις χρησιμοποιούνται σε πολλά συστήματα που χρειάζεται μέσω μίας απλής γλώσσας να περιγραφούν ακολουθίες γεγονότων.
- Παραδείγματα:
  - Επεξεργασία κειμένου
  - Λεξική ανάλυση

Αναζήτηση συμβολοσειρών σε έγγραφα κειμένου

- Το λογισμικό Notepad διαθέτει μία γλώσσα κανονικών εκφράσεων για να μπορεί ο χρήστης να περιγράφει συμβολοσειρές που αναζητά σε ένα κείμενο.
- Η γλώσσα περιλαμβάνει τα σύμβολα
  - [ ] - δηλώνει ένα σύνολο ή εύρος από εναλλακτικούς χαρακτήρες
    - πχ Η [abc] ταιριάζει οποιονδήποτε από τους a, b ή c
    - πχ Η [a-z] ταιριάζει οποιονδήποτε χαρακτήρα στο a-z
  - ^ - εξαιρεί ένα σύνολο ή εύρος χαρακτήρων (αν χρησιμοποιείται μέσα σε σύνολο) ή δηλώνει την αρχή μιας γραμμής (αν δε χρησιμοποιείται μέσα σε σύνολο)
    - πχ Η by[^e] ταιριάζει τη by, αλλά όχι τη bye
    - πχ Η [^a-z] ταιριάζει την 8 ή την R αλλά όχι την r
  - $ - δηλώνει το τέλος μιας γραμμής
  - . - δηλώνει οποιοδήποτε σύμβολο
  - $\backslash s$ - δηλώνει ένα κενό
  - $*$ - δηλώνει την παράθεση 0 ή περισσότερες φορές
  - $+$ - δηλώνει την παράθεση 1 ή περισσότερες φορές

Τέλος ενότητας

Επεξεργασία: Εμμανουέλα Στάχτιαρη
Θεσσαλονίκη, 24/07/2014