Βελτιώνοντας την επίδοση του packrat parsing

Νίκος Μαυρογεώργης

Σχολή Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών
Εθνικό Μετσόβειο Πολυτεχνείο

Παρουσίαση Διπλωματικής
Ιούνιος 2020

Πίνακας Περιεχομένων

1 Εισαγωγή

2 Parsing Expression Grammars

3 Packrat Parsing

4 Γεννήτορας συντακτικών αναλυτών packrat

5 Packrat Parsing με ελαστικό κυλιόμενο παράθυρο

6 Παράλληλο Packrat Parsing

7 Πειραματικά Αποτελέσματα

8 Συμπεράσματα

Πίνακας Περιεχομένων

1 Εισαγωγή

2 Parsing Expression Grammars

3 Packrat Parsing

4 Γεννήτορας συντακτικών αναλυτών packrat

5 Packrat Parsing με ελαστικό κυλιόμενο παράθυρο

6 Παράλληλο Packrat Parsing

7 Πειραματικά Αποτελέσματα

8 Συμπεράσματα

Συντακτική Ανάλυση

- Πρακτικά όλες οι γλώσσες, είτε φυσικές είτε γλώσσες μηχανής, βασίζονται στην έκφραση της πληροφορίας με γραμμικό τρόπο
- Συνήθως η αναπαράστηση γίνεται με τη μορφή μίας *συμβολοσειράς*, που είναι μια ακολουθία χαρακτήρων από ένα τυποποιημένο σύνολο
- Οποιαδήποτε εφαρμογή επεξεργασίας γλώσσας πρέπει να μετατρέψει τις συμβολοσειρές σε πιο αφηρημένες δομές όπως λέξεις, φράσεις, προτάσεις, εκφράσεις ή εντολές

Συντακτική Ανάλυση

- Πρακτικά όλες οι γλώσσες, είτε φυσικές είτε γλώσσες μηχανής, βασίζονται στην έκφραση της πληροφορίας με γραμμικό τρόπο
- Συνήθως η αναπαράσταση γίνεται με τη μορφή μίας *συμβολοσειράς*, που είναι μια ακολουθία χαρακτήρων από ένα τυποποιημένο σύνολο
- Οποιαδήποτε εφαρμογή επεξεργασίας γλώσσας πρέπει να μετατρέψει τις συμβολοσειρές σε πιο αφηρημένες δομές όπως λέξεις, φράσεις, προτάσεις, εκφράσεις ή εντολές

Ορισμός

Συντακτική ανάλυση (parsing) είναι η διαδικασία που εξάγει χρήσιμη δομημένη πληροφορία από γραμμικό κείμενο.

Πόσο κοστίζει η συντακτική ανάλυση?

Πόσο κοστίζει η συντακτική ανάλυση?

- Αποτελεί σημαντικό κομμάτι της εκτέλεσης προγραμμάτων, ειδικά στις διερμηνευόμενες γλώσσες όπου οι εντολές δεν μετατρέπονται σε ένα εκτελέσιμο, αλλά εκτελούνται διαρκώς εκ νέου:
    - Γλώσσες Σεναρίων: Python, Javascript
    - Γλώσσες Σήμανσης: HTML, CSS, Postscript
    - Γλώσσες ανταλλαγής δεδομένων: XML, JSON

Πόσο κοστίζει η συντακτική ανάλυση?

- Αποτελεί σημαντικό κομμάτι της εκτέλεσης προγραμμάτων, ειδικά στις διερμηνευόμενες γλώσσες όπου οι εντολές δεν μετατρέπονται σε ένα εκτελέσιμο, αλλά εκτελούνται διαρκώς εκ νέου:
    - Γλώσσες Σεναρίων: Python, Javascript
    - Γλώσσες Σήμανσης: HTML, CSS, Postscript
    - Γλώσσες ανταλλαγής δεδομένων: XML, JSON
- Κατά το rendering ιστοσελίδων, η συντακτική ανάλυση των HTML, CSS και Javascript καταναλώνει έως και το 40% της διαδικασίας.

Πόσο κοστίζει η συντακτική ανάλυση?

- Αποτελεί σημαντικό κομμάτι της εκτέλεσης προγραμμάτων, ειδικά στις διερμηνευόμενες γλώσσες όπου οι εντολές δεν μετατρέπονται σε ένα εκτελέσιμο, αλλά εκτελούνται διαρκώς εκ νέου:
    - Γλώσσες Σεναρίων: Python, Javascript
    - Γλώσσες Σήμανσης: HTML, CSS, Postscript
    - Γλώσσες ανταλλαγής δεδομένων: XML, JSON
- Κατά το rendering ιστοσελίδων, η συντακτική ανάλυση των HTML, CSS και Javascript καταναλώνει έως και το 40% της διαδικασίας.

Συμπέρασμα

Θα άξιζε να μειώναμε το χρόνο εκτέλεσής της, ιδιαίτερα αν αξιοποιούσαμε και τα πολυπύρηνα συστήματα που είναι σχεδόν πάντα διαθέσιμα.

Σε ποιες γραμματικές απευθύνεται το packrat?

Πίνακας Περιεχομένων

1 Εισαγωγή
2 Parsing Expression Grammars
3 Packrat Parsing
4 Γεννήτορας συντακτικών αναλυτών packrat
5 Packrat Parsing με ελαστικό κυλιόμενο παράθυρο
6 Παράλληλο Packrat Parsing
7 Πειραματικά Αποτελέσματα
8 Συμπεράσματα

Parsing Expression Grammars - Κίνητρο

- Οι δύο πιο συνηθισμένες μέθοδοι για να περιγραφεί η σύνταξη μίας γλώσσας: οι κανονικές εκφράσεις και οι γραμματικές χωρίς συμφραζόμενα (CFGs)

Parsing Expression Grammars - Κίνητρο

- Οι δύο πιο συνηθισμένες μέθοδοι για να περιγραφεί η σύνταξη μίας γλώσσας: οι κανονικές εκφράσεις και οι γραμματικές χωρίς συμφραζόμενα (CFGs)
- Ένα ακόμη χρήσιμο πρότυπο περιγραφής της σύνταξης είναι οι *Parsing Expression Grammars (PEGs)*
- Μοιάζουν με τις γραμματικές χωρίς συμφραζόμενα, αλλά έχουν και ορισμένες θεμελιώδεις διαφορές

Parsing Expression Grammars - Κίνητρο

- Οι δύο πιο συνηθισμένες μέθοδοι για να περιγραφεί η σύνταξη μίας γλώσσας: οι κανονικές εκφράσεις και οι γραμματικές χωρίς συμφραζόμενα (CFGs)
- Ένα ακόμη χρήσιμο πρότυπο περιγραφής της σύνταξης είναι οι *Parsing Expression Grammars (PEGs)*
- Μοιάζουν με τις γραμματικές χωρίς συμφραζόμενα, αλλά έχουν και ορισμένες θεμελιώδεις διαφορές
- Δαισθητικά μια CFG μας περιγράφει το πώς *κατασκευάζεται* μία συμβολοσειρά που ανήκει σε κάποια γλώσσα, ενώ οι PEGs το πώς *αναλύεται* η συμβολοσειρά ώστε να προκύψει δομική πληροφορία για αυτή

Parsing Expression Grammars - Κίνητρο

Example

Γλώσσα από τη συνένωση ζευγών **a**

- Παραγωγικός ορισμός: $\{s \in \mathbf{a}^* | s = (\mathbf{aa})^n\}$ δηλαδή μια γλώσσα με ένα μόνο γράμμα στο λεξιλόγιό της της οποίας οι συμβολοσειρές κατασκευάζονται συνενώνοντας ζεύγη από **a**
- Αναγνωριστικός ορισμός: $\{s \in \mathbf{a}^* | (|s| mod 2 = 0)\}$ δηλαδή μία συμβολοσειρά από **a**'s γίνεται αποδεκτή μόνο αν το μήκος της είναι άρτιο

Parsing Expression Grammars - Κίνητρο

Example

Γλώσσα από τη συνένωση ζευγών **a**

- Παραγωγικός ορισμός: $\{s \in \mathbf{a}^* | s = (\mathbf{aa})^n\}$ δηλαδή μια γλώσσα με ένα μόνο γράμμα στο λεξιλόγιό της της οποίας οι συμβολοσειρές κατασκευάζονται συνενώνοντας ζεύγη από **a**
- Αναγνωριστικός ορισμός: $\{s \in \mathbf{a}^* | (|s| mod 2 = 0)\}$ δηλαδή μία συμβολοσειρά από **a**'s γίνεται αποδεκτή μόνο αν το μήκος της είναι άρτιο

Ο σχεδιαστής της γραμματικής είναι ευκολότερο να σκέφτεται πώς αναλύεται μία δοσμένη συμβολοσειρά στα συστατικά της, παρά πώς θα γεννηθεί (generated) η συμβολοσειρά μέσα από τους κανόνες της γραμματικής.

Parsing Expression Grammars - Ορισμοί

- Κανόνες της μορφής `$n \leftarrow e$', όπου n μη τερματικό και e έκφραση ("για να αναγνωρίσεις το n, αναγνώρισε πρώτα το e")

Parsing Expression Grammars - Ορισμοί

- Κανόνες της μορφής `$n \leftarrow e$', όπου n μη τερματικό και e έκφραση ("για να αναγνωρίσεις το n, αναγνώρισε πρώτα το e")
- Αριστερό βέλος αντί για δεξί: διασθητική διαφορά στην "ροή της πληροφορίας"

Parsing Expression Grammars - Ορισμοί

- Κανόνες της μορφής \`$n \leftarrow e$', όπου n μη τερματικό και e έκφραση ("για να αναγνωρίσεις το n, αναγνώρισε πρώτα το e")
- Αριστερό βέλος αντί για δεξί: διασθητική διαφορά στην "ροή της πληροφορίας"
- Οι κανόνες των CFGs εκφράζουν "παραγωγές" από μη τερματικά στις αντίστοιχες εκφράσεις τους ενώ των PEGs αναπαριστούν "αφαιρέσεις" από τις εκφράσεις στους αντίστοιχους κανόνες

Parsing Expression Grammars - Εκφράσεις

Κενή συμβολοσειρά `()' : "Μην προσπαθήσεις να διαβάσεις τίποτα: απλά επίστρεψε επιτυχώς χωρίς να καταναλώσεις τίποτα από την είσοδο."

Parsing Expression Grammars - Εκφράσεις

Κενή συμβολοσειρά \`()' : "Μην προσπαθήσεις να διαβάσεις τίποτα: απλά επίστρεψε επιτυχώς χωρίς να καταναλώσεις τίποτα από την είσοδο."

Τερματικό \`α': "Αν το επόμενο τερματικό στην είσοδο είναι α τότε κατανάλωσε ένα τερματικό και επίστρεψε επιτυχώς. αλλιώς, απότυχε και μην καταναλώσεις τίποτα."

Parsing Expression Grammars - Εκφράσεις

Κενή συμβολοσειρά ‘()' : "Μην προσπαθήσεις να διαβάσεις τίποτα: απλά επίστρεψε επιτυχώς χωρίς να καταναλώσεις τίποτα από την είσοδο."

Τερματικό ‘α': "Αν το επόμενο τερματικό στην είσοδο είναι α τότε κατανάλωσε ένα τερματικό και επίστρεψε επιτυχώς. αλλιώς, απότυχε και μην καταναλώσεις τίποτα."

Μη Τερματικό ‘A': "Προσπάθησε να διαβάσεις την είσοδο με βάση τον κανόνα που αντιστοιχεί στο A και επίστρεψε επιτυχώς ή απότυχε αντίστοιχα."

Parsing Expression Grammars - Εκφράσεις

Ακολουθία \`$(e_1 e_2 \ldots e_n)$': "Προσπάθησε να διαβάσεις μία συμβολοσειρά ώστε να επιτύχει η e_1. Αν η e_1 επιτύχει, κάνε το ίδιο με την e_2, ξεκινώντας από το σημείο της εισόδου που δεν κατανάλωσε η e_1 κ.ό.κ. Αν και οι n εκφράσεις αναγνωριστούν επίστρεψε επιτυχώς και κατανάλωσε τα αντίστοιχα κομμάτια της εισόδου. Αν οποιαδήποτε υποέκφραση αποτύχει, απότυχε χωρίς να καταναλώσεις τίποτα."

Parsing Expression Grammars - Εκφράσεις

Διατεταγμένη Επιλογή `$(e_1/e_2/\ldots/e_n)$': "Προσπάθησε να διαβάσεις μία συμβολοσειρά ώστε να επιτύχει η e_1. Αν επιτύχει τότε η επιλογή επιστρέφει επιτυχώς καταναλώνοντας το αντίστοιχο κομμάτι της εισόδου. Αλλιώς, προσπάθησε με την e_2 και την αρχική είσοδο κ.ό.κ, μέχρις ότου να επιτύχει κάποια από τις υποεκφράσεις. Αν καμία από τις n εναλλακτικές δεν πετύχουν, τότε απότυχε χωρίς να καταναλώσεις τίποτα."

Parsing Expression Grammars - Εκφράσεις

Διατεταγμένη Επιλογή `$(e_1/e_2/\ldots/e_n)$': "Προσπάθησε να διαβάσεις μία συμβολοσειρά ώστε να επιτύχει η e_1. Αν επιτύχει τότε η επιλογή επιστρέφει επιτυχώς καταναλώνοντας το αντίστοιχο κομμάτι της εισόδου. Αλλιώς, προσπάθησε με την e_2 και την αρχική είσοδο κ.ό.κ, μέχρις ότου να επιτύχει κάποια από τις υποεκφράσεις. Αν καμία από τις n εναλλακτικές δεν πετύχουν, τότε απότυχε χωρίς να καταναλώσεις τίποτα."

Example

Έστω ο κανόνας $Number \rightarrow Digit\ Number/Digit$. Η σειρά έχει σημασία, διότι αν ήταν ανάποδα και θέλαμε να αναλύσουμε τον αριθμό 12, θα πηγαίναμε πρώτα στην εναλλακτική $Digit$, θα αναγνωρίζαμε το 1 και θα επιστρέφαμε χωρίς να πάμε στο 2.

Πίνακας Περιεχομένων

1 Εισαγωγή
2 Parsing Expression Grammars
3 Packrat Parsing
4 Γεννήτορας συντακτικών αναλυτών packrat
5 Packrat Parsing με ελαστικό κυλιόμενο παράθυρο
6 Παράλληλο Packrat Parsing
7 Πειραματικά Αποτελέσματα
8 Συμπεράσματα

Ορισμοί

- Ο απλούστερος και διαισθητικά προφανής τρόπος να σχεδιάσουμε έναν συντακτικό αναλυτή είναι η από πάνω προς τα κάτω ανάλυση ή ανάλυση αναδρομικής κατάβασης.
- *Προβλέποντες (predictive) συντακτικοί αναλυτές*: επιχειρούν να προβλέψουν ποιο στοιχείο της γλώσσας ακολουθεί, βλέποντας ορισμένα από τα προπορευόμενα σύμβολα στην είσοδο.
- *Συντακτικοί αναλυτές με οπισθαναχώρηση (backtracking)*: παίρνουν αποφάσεις υποθετικά (speculatively) και δοκιμάζουν διαδοχικά διάφορες εναλλακτικές. Αν μία αποτύχει, τότε ο αναλυτής οπισθαναχωρεί στη θέση της εισόδου που ήταν προτού δοκιμάσει την εναλλακτική και μετά εξετάζει την επόμενη εναλλακτική.

Πίνακας Περιεχομένων

1 Εισαγωγή

2 Parsing Expression Grammars

3 Packrat Parsing

4 Γεννήτορας συντακτικών αναλυτών packrat

5 Packrat Parsing με ελαστικό κυλιόμενο παράθυρο

6 Παράλληλο Packrat Parsing

7 Πειραματικά Αποτελέσματα

8 Συμπεράσματα

Πίνακας Περιεχομένων

1 Εισαγωγή

2 Parsing Expression Grammars

3 Packrat Parsing

4 Γεννήτορας συντακτικών αναλυτών packrat

5 Packrat Parsing με ελαστικό κυλιόμενο παράθυρο

6 Παράλληλο Packrat Parsing

7 Πειραματικά Αποτελέσματα

8 Συμπεράσματα

Πίνακας Περιεχομένων

1 Εισαγωγή

2 Parsing Expression Grammars

3 Packrat Parsing

4 Γεννήτορας συντακτικών αναλυτών packrat

5 Packrat Parsing με ελαστικό κυλιόμενο παράθυρο

6 Παράλληλο Packrat Parsing

7 Πειραματικά Αποτελέσματα

8 Συμπεράσματα

Πίνακας Περιεχομένων

1 Εισαγωγή

2 Parsing Expression Grammars

3 Packrat Parsing

4 Γεννήτορας συντακτικών αναλυτών packrat

5 Packrat Parsing με ελαστικό κυλιόμενο παράθυρο

6 Παράλληλο Packrat Parsing

7 Πειραματικά Αποτελέσματα

8 Συμπεράσματα

Πίνακας Περιεχομένων

1 Εισαγωγή

2 Parsing Expression Grammars

3 Packrat Parsing

4 Γεννήτορας συντακτικών αναλυτών packrat

5 Packrat Parsing με ελαστικό κυλιόμενο παράθυρο

6 Παράλληλο Packrat Parsing

7 Πειραματικά Αποτελέσματα

8 Συμπεράσματα