Gramatická indukcia a jej využitie

Michal Malý

KAI FMFI UK

29. Marec 2010

Prehľad

Na počiatku bolo slovo. A slovo...

Definícia (Slovo)

Slovo je konečná postupnosť symbolov z abecedy.

Definícia (Abeceda)

Abeceda – množina znakov (symbolov, tokenov)

Definícia (Jazyk)

Jazyk – množina slov.

Príklad:
abeceda: $\Sigma = \{a, b\}$
slová: *abba*,*aaaabbbb*
jazyk: $L = \{\varepsilon, ab, ba, aabb, abba, baba, bbaa, ...\}$ = tie slová, kde počet *a*-čok a *b*-čok v slove je rovnaký

Gramatika

Definícia (Terminálny symbol)

Terminály – symboly, z ktorých sa skladajú slová (na výstupe).

Definícia (neterminálny symbol)

Neterminály – pomocné symboly pri odvodzovaní, nesmú sa objaviť na výstupe.

Definícia (Gramatika)

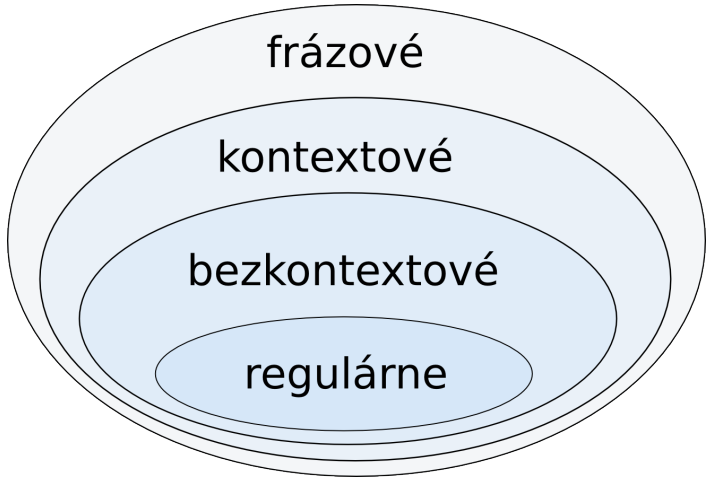
Gramatika je určená množinou neterminálov, množinou terminálov, množinou odvodzovacích pravidiel a štartovacím symbolom.

Gramatika – príklad

terminály: $T = \{1, 2, 3, \ldots 9, +, -, (,)\}$
neterminály: $N = \{V, C\}$
štartovací symbol: V
pravidlá:
$V \rightarrow C$
$V \rightarrow V + V$
$V \rightarrow V - V$
$V \rightarrow (V)$
$C \rightarrow 1,\ C \rightarrow 2,\ \ldots,\ C \rightarrow 9$

Vieme vyrobiť slovo $1 - (4 + 5)$ napríklad takto:
$V \Rightarrow V-V \Rightarrow C-V \Rightarrow C-(V) \Rightarrow C-(V+V) \Rightarrow C-(C+V) \Rightarrow$
$C - (C + C) \Rightarrow 1 - (C + C) \Rightarrow 1 - (4 + C) \Rightarrow 1 - (4 + 5)$

Chomského hierarchia



1. regulárne gramatiky $\equiv$ konečný automat (nevedia $a^n b^n$)
2. bezkontextové gramatiky $\equiv$ nedeterministický zásobníkový automat (nevedia $a^n b^n c^n$)
3. kontextové gramatiky $\equiv$ nedeterministický lineárny automat (nevie EXPSPACE-hard problémy)
4. frázové gramatiky (neobmedzené, typu 0) $\equiv$ Turingov stroj (vie všetko, čo je ,,intuitívne vypočítateľné")

Krátka charakterizácia gramatík

- regulárna gramatika: vľavo iba neterminál, vpravo sa neterminál môže vyskytovať iba na konci: $S \rightarrow aX$
- bezkontextová gramatika: vľavo iba neterminál: $S \rightarrow XaYbb$
- kontextové gramatiky: môže sa vyskytovať kontext, ktorý ale ostáva: $\alpha S \beta \rightarrow \alpha XYa\beta$
- frázové gramatiky: neobmedzené

Gramatická indukcia

- Jazyk – zápis (formalizácia) prakticky ľubovoľného problému
- Príklady: pozitívne, negatívne
- Riešenie problému?

Definícia

Gramatická indukcia je spôsob, ako odvodiť formálnu gramatiku z množiny vzoriek – pozorovaní.

- využitie – spracovanie prirodzeného jazyka, kompresia, ...

Gramatická indukcia – metódy

- pokus-omyl
- greedy (LZW,Sequitur)
- genetické algoritmy (John R. Koza)
- formálne metódy

Gramatická indukcia – pokus-omyl

pozitívne príklady: $\mathcal{D}^+ = \{a, aaa, aaab, aab\}$
negatívne príklady: $\mathcal{D}^- = \{ab, abc, abb, aabb\}$

i	$\mathbf{x}_i^+$	$\mathcal{P}$	$\mathcal{P}$ produces $\mathcal{D}^-$?
1	a	$S \to A$ $A \to \mathtt{a}$	No
2	aaa	$S \to A$ $A \to \mathtt{a}$ $A \to \mathtt{a}A$	No
3	aaab	$S \to A$ $A \to \mathtt{a}$ $A \to \mathtt{a}A$ $A \to \mathtt{ab}$	Yes: $\mathtt{ab} \in \mathcal{D}^-$
3	aaab	$S \to A$ $A \to \mathtt{a}$ $A \to \mathtt{a}A$ $A \to \mathtt{aab}$	No
4	aab	$S \to A$ $A \to \mathtt{a}$ $A \to \mathtt{a}A$ $A \to \mathtt{aab}$	No

Gramatická indukcia – stemming cez Myhill-Nerodovu ekvivalenciu

Veta (Myhill-Nerodova veta)

Jazyk L je regulárny práve vtedy, ak relácia R definovaná ako

$$u \; R \; v \iff_{def} \forall x \in \Sigma^*(ux \in L \Leftrightarrow vx \in L)$$

je reláciou ekvivalencie konečného indexu.

- dve slová sú v relácii, ak pridaním nejekého suffixu (koncovky) dostanem pre obe buď slovo z jazyka, alebo pre obe slovo nie z jazyka
- konštrukcia minimálneho deterministického automatu

Výsledná gramatika pre slovo mesto

pravidlá pre neterminál číslo...	použití
$0 \rightarrow \ldots \mid mesta\ 56 \mid meste \mid mesteck\ 61 \mid mesto\ 48 \mid$ $mestsk\ 455 \mid mestu \mid \ldots$	(štart)
$3 \rightarrow i \mid \varepsilon$	111
$15 \rightarrow ho \mid j \mid \varepsilon$	179
$48 \rightarrow m \mid \varepsilon$	101
$49 \rightarrow ch \mid m \mid \varepsilon$	111
$56 \rightarrow ch \mid m\ 3 \mid \varepsilon$	121
$61 \rightarrow a \mid o \mid u$	7
$83 \rightarrow m \mid u$	35
$455 \rightarrow a \mid e\ 15 \mid i \mid o\ 83 \mid u \mid y\ 49$	2

Výsledný stemovací slovník pre slovo mesto

...
mestach mestami mestam mesta
meste
mestecka mestecko mestecku
mestom mesto
mestska mestskeho mestskej mestske mestski mestskom mestskou
mestsku mestskych mestskym mestsky
...

Koniec

Ďakujem za pozornosť.
Otázky?