

Kontekstno-neodvisne gramatike za kodiranje in stiskanje podatkov

Janez Podlogar

Univerza v Ljubljani, Fakulteta za matematiko in fiziko

16. 9. 2024

2024-09-14

Gramatike za stiskanje podatkov

Kontekstno-neodvisne gramatike
za kodiranje in stiskanje podatkov

Janez Podlogar

Univerza v Ljubljani, Fakulteta za matematiko in fiziko

16. 9. 2024

Razdeljamo pojme v naslovu.

- Kodiranje je spreminjanje zapisa sporočila.
- Stiskanje podatkov je kodiranje, katerega cilj je zapisati sporočilo v zgoščeni obliki.

Kontekstno-neodvisna gramatika

Definicija

Kontekstno-neodvisna gramatika je četverica $G = (V, \Sigma, P, S)$, kjer je

- V abeceda *nekončnih simbolov*;
- Σ abeceda *končnih simbolov* taka, da $\Sigma \cap V = \emptyset$;
- $P \subseteq V \times (V \cup \Sigma)^*$ celovita relacija;
- $S \in V$ *začetni simbol*.

Gramatike za stiskanje podatkov

└ Kontekstno-neodvisna gramatika

└ Kontekstno-neodvisna gramatika

2024-09-14

Definicija

Kontekstno-neodvisna gramatika je četverica $G = (V, \Sigma, P, S)$, kjer je

- V abeceda *nekončnih simbolov*;
- Σ abeceda *končnih simbolov* taka, da $\Sigma \cap V = \emptyset$;
- $P \subseteq V \times (V \cup \Sigma)^*$ celovita relacija;
- $S \in V$ *začetni simbol*.

- Abeceda je končna neprazna množica.
- Relacija $R \subseteq A \times B$ je *celovita*, če velja $\forall x \in A \exists y \in B: (x, y) \in R$.
- $(V \cup \Sigma)^*$ je množica vseh nizov končne dolžine končnih in nekončnih simbolov. Vsebuje tudi prazen niz ε , saj je to niz dolžine 0.
- Prazen niz ε je enota za operacijo stik.
- Okrajšamo z KNG.

Primer

Naj bo $V = \{S\}$, $\Sigma = \{a, b, c\}$, $P = \{(S, aSb), (S, \varepsilon)\}$ in $S = S$.

Gramatike za stiskanje podatkov

└ Kontekstno-neodvisna gramatika

- Elementom relacije P pravimo *prepisovalna pravila*.
- Prepisovalno pravilo $(A, \beta) \in P$ pišemo $A \rightarrow \beta$.
- Niz $\alpha A \gamma$ se *prepiše* s pravilom $A \rightarrow \beta$ v $\alpha \beta \gamma$, pišemo $\alpha A \gamma \Rightarrow \alpha \beta \gamma$.
- Niz α *izpelje* niz β , če lahko α prepišemo v β z uporabo končno mnogo prepisovalnih pravil.

- Elementom relacije P pravimo *prepisovalna pravila*.
- Prepisovalno pravilo $(A, \beta) \in P$ pišemo $A \rightarrow \beta$.
- Niz $\alpha A \gamma$ se *prepiše* s pravilom $A \rightarrow \beta$ v $\alpha \beta \gamma$, pišemo $\alpha A \gamma \Rightarrow \alpha \beta \gamma$.
- Niz α *izpelje* niz β , če lahko α prepišemo v β z uporabo končno mnogo prepisovalnih pravil.

- *Leva stran prepisovalnega pravila $A \rightarrow \beta$ je A in desna stran prepisovalnega pravila je β .*

Primer

prepisovalna pravila zapišemo kot $P = \{S \rightarrow aSb, S \rightarrow \epsilon\}$. Izpeljemo nize

$$S \Rightarrow \epsilon, S \Rightarrow aSb \Rightarrow ab, S \Rightarrow aSb \Rightarrow aaSbb \Rightarrow aabb, \dots$$

Jezik je $\{a^n b^n \mid n \geq 0\}$.

Gramatike za stiskanje podatkov

Kontekstno-neodvisna gramatika

- Elementom relacije P pravimo *prepisovalna pravila*.
- Prepisovalno pravilo $(A, \beta) \in P$ pišemo $A \rightarrow \beta$.
- Niz $\alpha A \gamma$ se *prepiše* s pravilom $A \rightarrow \beta$ v $\alpha\beta\gamma$, pišemo $\alpha A \gamma \Rightarrow \alpha\beta\gamma$.
- Niz α *izpelje* niz β , če lahko α prepíšemo v β z uporabo končno mnogo prepisovalnih pravil.

Definicija
Jezik KNG je množica nizov, ki jih lahko izpeljemo iz začetnega simbola in ne vsebujejo nekončnih simbolov.

- Elementom relacije P pravimo *prepisovalna pravila*.
- Prepisovalno pravilo $(A, \beta) \in P$ pišemo $A \rightarrow \beta$.
- Niz $\alpha A \gamma$ se *prepiše* s pravilom $A \rightarrow \beta$ v $\alpha\beta\gamma$, pišemo $\alpha A \gamma \Rightarrow \alpha\beta\gamma$.
- Niz α *izpelje* niz β , če lahko α prepíšemo v β z uporabo končno mnogo prepisovalnih pravil.

Definicija

Jezik KNG je množica nizov, ki jih lahko izpeljemo iz začetnega simbola in ne vsebujejo nekončnih simbolov.

- *Leva stran prepisovalnega pravila* $A \rightarrow \beta$ je A in *desna stran prepisovalnega pravila* je β .

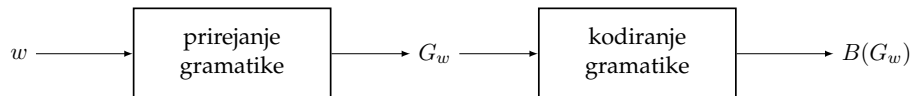
Primer

prepisovalna pravila zapišemo kot $P = \{S \rightarrow aSb, S \rightarrow \epsilon\}$. Izpeljemo nize

$$S \Rightarrow \epsilon, S \Rightarrow aSb \Rightarrow ab, S \Rightarrow aSb \Rightarrow aaSbb \Rightarrow aabb, \dots$$

Jezik je $\{a^n b^n \mid n \geq 0\}$.

Stiskanje niza

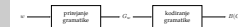


2024-09-14

Gramatike za stiskanje podatkov

- └ Kontekstno-neodvisna gramatika
 - └ Stiskanje niza

Stiskanje niza



- Glavna ideja je, da je jezik G_w je $\{w\}$, saj lahko enolično rekonstruiramo niz w .
- Alternativni pristop: KNG G je poznana tako kodirniku kot dekodirniku in jezik KNG G vsebuje vse nize, ki jih želimo stisniti. Posamezni niz stisnemo tako, da stisnemo izpeljavo niza iz začetnega simbola.

Dopustna gramatika

Definicija

KNG je *deterministična*, če vsak nekončen simbol $A \in V$ nastopa natanko enkrat kot leva stran nekega prepisovalnega pravila.

2024-09-14

Gramatike za stiskanje podatkov

- └ Kontekstno-neodvisna gramatika
 - └ Dopustna gramatika
 - └ Dopustna gramatika

- Prejšnji primer ni deterministična gramatika.
- Če odstranimo odstarnimo pravilo $S \rightarrow aSb$, postane deterministična, a je jezik $\{\varepsilon\}$.
- Če odstranimo $S \rightarrow \varepsilon$ postane deterministična, a je njen jezik prazna množica.
- Determinizem sam po sebi ni dovolj močan, da prepreči praznost jezika. Zato bomo od dopustnih gramatik zahtevali, da je njihov jezik neprazen in da ne vsebuje pravila oblike $A \rightarrow \varepsilon$.

Dopustna gramatika

Definicija

KNG je *deterministična*, če vsak nekončen simbol $A \in V$ nastopa natanko enkrat kot leva stran nekega prepisovalnega pravila.

Trditev

Jezik deterministične KNG je enojec ali prazna množica.

2024-09-14

Gramatike za stiskanje podatkov

- └ Kontekstno-neodvisna gramatika
 - └ Dopustna gramatika
 - └ Dopustna gramatika

Definicija

KNG je *deterministična*, če vsak nekončen simbol $A \in V$ nastopa natanko enkrat kot leva stran nekega prepisovalnega pravila.

Trditev

Jezik deterministične KNG je enojec ali prazna množica.

- Prejšnji primer ni deterministična gramatika.
- Če odstranimo odstarnimo pravilo $S \rightarrow aSb$, postane deterministična, a je jezik $\{\varepsilon\}$.
- Če odstranimo $S \rightarrow \varepsilon$ postane deterministična, a je njen jezik prazna množica.
- Determinizem sam po sebi ni dovolj močan, da prepreči praznost jezika. Zato bomo od dopustnih gramatik zahtevali, da je njihov jezik neprazen in da ne vsebuje pravila oblike $A \rightarrow \varepsilon$.

2024-09-14

Gramatike za stiskanje podatkov

└ Kontekstno-neodvisna gramatika

└ Dopustna gramatika

Definicija

KNG ne vsebuje neuporabnih simbolov, če se vsak simbol $y \in V \cup \Sigma$, $y \neq S$ pojavi vsaj enkrat v izpeljavi niza, ki je v jeziku KNG.

Definicija

KNG ne vsebuje neuporabnih simbolov, če se vsak simbol $y \in V \cup \Sigma$, $y \neq S$ pojavi vsaj enkrat v izpeljavi niza, ki je v jeziku KNG.

- Povedano drugače: simbol je neupraven, če se ne pojavi v nobeni izpeljavi niza, ki je v jeziku KNG.

Definicija

KNG je *dopustna gramatika*, če je:

- deterministična,
- ne vsebuje neuporabnih simbolov,
- ima neprazen jezik,
- prazen niz ne nastopa kot desna stran kateregakoli prepisovalnega pravila.

2024-09-14

Gramatike za stiskanje podatkov

└ Kontekstno-neodvisna gramatika

└ Dopustna gramatika

Definicija

KNG je *dopustna gramatika*, če je:

- deterministična,
- ne vsebuje neuporabnih simbolov,
- ima neprazen jezik,
- prazen niz ne nastopa kot desna stran kateregakoli prepisovalnega pravila.

Definicija

KNG je *dopustna gramatika*, če je:

- deterministična,
- ne vsebuje neuporabnih simbolov,
- ima neprazen jezik,
- prazen niz ne nastopa kot desna stran kateregakoli prepisovalnega pravila.

Posledica

Jezik dopustne gramatike je enojec.

2024-09-14

Gramatike za stiskanje podatkov

└ Kontekstno-neodvisna gramatika

└ Dopustna gramatika

Definicija

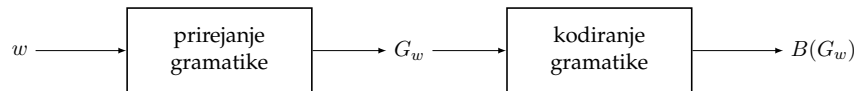
KNG je *dopustna gramatika*, če je:

- deterministična,
- ne vsebuje neuporabnih simbolov,
- ima neprazen jezik,
- prazen niz ne nastopa kot desna stran kateregakoli prepisovalnega pravila.

Posledica

Jezik dopustne gramatike je enojec.

Prirejanje gramatike



Od tu naprej naj bo:

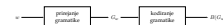
- \mathcal{A} poljubna abeceda, $|\mathcal{A}| \geq 2$;
- $\{A_0, A_1, \dots\}$ končna množica, $\mathcal{A} \cap \{A_0, A_1, \dots\} = \emptyset$.

2024-09-14

Gramatike za stiskanje podatkov

└ Prirejanj gramatike

└ Prirejanje gramatike



Od tu naprej naj bo:

- \mathcal{A} poljubna abeceda, $|\mathcal{A}| \geq 2$;
- $\{A_0, A_1, \dots\}$ končna množica, $\mathcal{A} \cap \{A_0, A_1, \dots\} = \emptyset$.

Definicija

Naj bo $\mathcal{G}(\mathcal{A})$ množica vseh KNG G , ki zadostujejo:

- 1 G je dopustna gramatika;
- 2 $\Sigma \subseteq \mathcal{A}$;
- 3 $V = \{A_0, A_1, \dots, A_{|V|-1}\}$;
- 4 $S = A_0$;
- 5 Če naštejemo nekončne simbole V v vrstnem redu prve pojavitve pri izpeljavi niza gramatike, dobimo zaporedje $A_0, A_1, A_2, \dots, A_{|V|-1}$.

2024-09-14

Gramatike za stiskanje podatkov

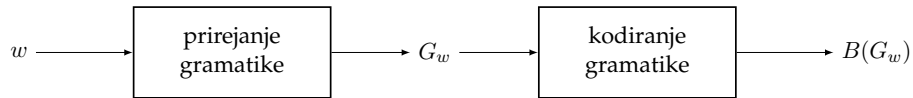
└ Prirejanj gramatike

Definicija

Naj bo $\mathcal{G}(\mathcal{A})$ množica vseh KNG G , ki zadostujejo:

- G je dopustna gramatika;
- $\Sigma \subseteq \mathcal{A}$;
- $V = \{A_0, A_1, \dots, A_{|V|-1}\}$;
- $S = A_0$;
- Če naštejemo nekončne simbole V v vrstnem redu prve pojavitve pri izpeljavi niza gramatike, dobimo zaporedje $A_0, A_1, A_2, \dots, A_{|V|-1}$.

- Z zahtevo 5. so nekončni simboli poimenovani po edinstvem vrstnem redu. Ta vrstni red omogoča pravilno dekodiranje.
- KNG $G \notin \mathcal{G}(\mathcal{A})$, ki izpolnjuje zahtevi 1. in 2., preimenujmo nekončne simbole tako, da izpolnimo 5. zahtevo. Potem izpolnimo tudi 3. in 4. zahtevi. Dobimo $[G] \in \mathcal{G}(\mathcal{A})$, ki jo imenujemo *kanonično oblika* G , in velja da sta jezika $[G]$ in G enaka.



Definicija

Prirejanje gramatike nizu abecede \mathcal{A} je preslikava

$$\pi: \mathcal{A}^+ \rightarrow \mathcal{G}(\mathcal{A}),$$

$$w \mapsto G_w.$$



Definicija

Z $\mathcal{G}^*(\mathcal{A})$ označimo pravo podmnožico množice $\mathcal{G}(\mathcal{A})$, da za vsak $G \in \mathcal{G}^*(\mathcal{A})$ velja

$$\forall A, B \in V, A \neq B: f_G^\infty(A) \neq f_G^\infty(B).$$

- Za prirejanja, ki jih predstavimo potrebujemo še eno omejitev s katero se v predstavitvi ne obremenjujemo.

Definicija

Z $\mathcal{G}^*(\mathcal{A})$ označimo pravo podmnožico množice $\mathcal{G}(\mathcal{A})$, da za vsak $G \in \mathcal{G}^*(\mathcal{A})$ velja

$$\forall A, B \in V, A \neq B: f_G^\infty(A) \neq f_G^\infty(B).$$

Definicija

Z $|G|$ označimo vsoto dolžin desnih strani prepisovalnih pravil KNG G .

- Za prirejanja, ki jih predstavimo potrebujemo še eno omejitev s katero se v predstavitvi ne obremenjujemo.

Asimptotsko kompaktno prirejanje gramatike

Definicija

Prirejanje gramatike nizu abecede \mathcal{A} je *asimptotsko kompaktno*, če za vsak niz $w \in \mathcal{A}^+$ velja $G_w \in \mathcal{G}^*(\mathcal{A})$ in je

$$\lim_{n \rightarrow \infty} \max_{w \in \mathcal{A}^n} \frac{|G_w|}{|w|} = 0.$$

Gramatike za stiskanje podatkov

└ Prirejanj gramatike

└ Asimptotsko kompaktno prirejanje gramatike

└ Asimptotsko kompaktno prirejanje gramatike

2024-09-14

Definicija

Prirejanje gramatike nizu abecede \mathcal{A} je *asimptotsko kompaktno*, če za vsak niz $w \in \mathcal{A}^+$ velja $G_w \in \mathcal{G}^*(\mathcal{A})$ in je

$$\lim_{n \rightarrow \infty} \max_{w \in \mathcal{A}^n} \frac{|G_w|}{|w|} = 0.$$

- Primer je bisekcijsko prirejanje gramatike.

Neskrčljivo prirejanje gramatike

Definicija

Pravimo, da je $G \in \mathcal{G}^*(\mathcal{A})$ *neskrčljiva gramatika*, če:

- ❶ Vsak $A \in V$, $A \neq S$ nastopa vsaj dvakrat v desni strani prepisovalnih pravil;
- ❷ Ne obstajata $y_1, y_2 \in V \cup \Sigma$, da niz $y_1 y_2$ nastopa kot podniz desne strani kateregali prepisovalnega pravila več kot enkrat na neprekrivajočih se mestih.

Gramatike za stiskanje podatkov

└ Prirejanj gramatike

└ Neskrčljivo prirejanje gramatike

└ Neskrčljivo prirejanje gramatike

2024-09-14

Neskrčljivo prirejanje gramatike

Definicija

Pravimo, da je $G \in \mathcal{G}^*(\mathcal{A})$ *neskrčljiva gramatika*, če:

- ❶ Vsak $A \in V$, $A \neq S$ nastopa vsaj dvakrat v desni strani prepisovalnih pravil;
- ❷ Ne obstajata $y_1, y_2 \in V \cup \Sigma$, da niz $y_1 y_2$ nastopa kot podniz desne strani kateregali prepisovalnega pravila več kot enkrat na neprekrivajočih se mestih.

- Primer prekrivajočih mest je 111 in 111.
- Primer neprekriavjočih mest je 11 11.
- Različna neskrčjiva prirejanja gramatike dobimo tako, da izvajamo različne nabore skritvenih pravil.
- Primer je metodo najdaljšega ujemačega podniza.

Neskrčljivo prirejanje gramatike

Definicija

Pravimo, da je $G \in \mathcal{G}^*(\mathcal{A})$ *neskrčljiva gramatika*, če:

- ❶ Vsak $A \in V$, $A \neq S$ nastopa vsaj dvakrat v desni strani prepisovalnih pravil;
- ❷ Ne obstajata $y_1, y_2 \in V \cup \Sigma$, da niz $y_1 y_2$ nastopa kot podniz desne strani kateregali prepisovalnega pravila več kot enkrat na neprekrivajočih se mestih.

Definicija

Neskrčljivo prirejanje gramatike nizu abecede \mathcal{A} vsakemu nizu abecede \mathcal{A} priredi neskrčljivo gramatiko.

Gramatike za stiskanje podatkov

└ Prirejanj gramatike

└ Neskrčljivo prirejanje gramatike

└ Neskrčljivo prirejanje gramatike

2024-09-14

Definicija

Pravimo, da je $G \in \mathcal{G}^*(\mathcal{A})$ *neskrčljiva gramatika*, če:

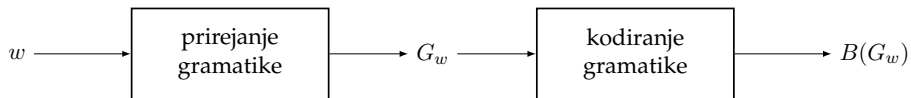
- ❶ Vsak $A \in V$, $A \neq S$ nastopa vsaj dvakrat v desni strani prepisovalnih pravil;
- ❷ Ne obstajata $y_1, y_2 \in V \cup \Sigma$, da niz $y_1 y_2$ nastopa kot podniz desne strani kateregali prepisovalnega pravila več kot enkrat na neprekrivajočih se mestih.

Definicija

Neskrčljivo prirejanje gramatike nizu abecede \mathcal{A} vsakemu nizu abecede \mathcal{A} priredi neskrčljivo gramatiko.

- Primer prekrivajočih mest je 111 in 111.
- Primer neprekriavjočih mest je 11 11.
- Različna neskrčljiva prirejanja gramatike dobimo tako, da izvajamo različne nabore skritvenih pravil.
- Primer je metodo najdaljšega ujemajočega podniza.

Binarno kodiranje gramatike



Definicija

Binarno kodiranje dopustne gramatike je preslikava

$$B: \mathcal{G}(\mathcal{A}) \rightarrow \{0, 1\}^+,$$

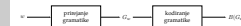
$$G \mapsto B(G).$$

2024-09-14

Gramatike za stiskanje podatkov

└ Binarno kodiranje gramatike

└ Binarno kodiranje gramatike



Definicija

Binarno kodiranje dopustne gramatike je preslikava

$$B: \mathcal{G}(\mathcal{A}) \rightarrow \{0, 1\}^+,$$

$$G \mapsto B(G).$$

Gramatike za stiskanje podatkov

└ Binarno kodiranje gramatike

Izrek

Obstaja bijektivno binarno kodiranje dopustne gramatike, da

- $\forall G_1, G_2 \in \mathcal{G}(\mathcal{A}), G_1 \neq G_2$ niz $B(G_1)$ ni predpona niza $B(G_2)$,
- $\forall G \in \mathcal{G}(\mathcal{A}): |B(G)| \leq |\mathcal{A}| + 4|G| + \lceil H(G) \rceil$.

Izrek

Obstaja bijektivno binarno kodiranje dopustne gramatike, da

- 1 $\forall G_1, G_2 \in \mathcal{G}(\mathcal{A}), G_1 \neq G_2$ niz $B(G_1)$ ni predpona niza $B(G_2)$,
- 2 $\forall G \in \mathcal{G}(\mathcal{A}): |B(G)| \leq |\mathcal{A}| + 4|G| + \lceil H(G) \rceil$.

- Abeceda je poznana tako kodirniku kot dekodirniku.

$$\kappa: \mathcal{A}^+ \rightarrow \{0, 1\}^+, \\ w \mapsto B(\pi(w)),$$

Definicija

Stiskanje niza abecede \mathcal{A} z gramatikami $\mathcal{G}(\mathcal{A})$ je par preslikav kodne in dekodne preslikave $\Phi = (\kappa, \delta)$. Kodna preslikava je

$$\kappa: \mathcal{A}^+ \rightarrow \{0, 1\}^+, \\ w \mapsto B(\pi(w)),$$

kjer je π prirejanje gramatike nizu abecede \mathcal{A} in B binarno kodiranje dopustne gramatike.

- Odvečnost meri količino ponavljajočih se ali predvidljivih podatkov znotraj sporočila, ki jih je mogoče odstraniti, da se prihrani prostor, brez izgube informacije.
- Odvečnost stiskanja z asimptotsko kompaktnim prirejanjem konvergira proti 0 v odvisnosti od izbire kodiranja znotraj razreda.
- Stiskanje z neskrčljivim prirejanjem tudi stiskanje z asimptotsko kompaktnim prirejanjem in odvečnost konvergira enakomerno proti 0 za vsa stiskanja z neskrčljivim prirejanjem, vsaj tako hitro kot $\frac{\log_2 \log_2(n)}{\log_2(n)}$ pomnoženo z neko konstanto.