Teorie algoritmů — 5. týden

Marie Demlová
http://math.fel.cvut.cz/en/people/demlova

17. 3. 2026

Huffmanův kód

Je dán text t nad abecedou C a pro každé $c \in C$ je dána frekvence výskytu c v textu — $c.freq$.

Úkol: Najít binární kód textu t.

Jsou možnosti

- kód stejné délky – kódová slova mají délku k, kde k je nejmenší splňující $|C| \leq 2^k$.
- bezprefixový kód — ne všechna slova jsou stejné délky, ale žádné není prefixem jiného.

Huffmanův kód

Příklad.

V datech o 100 000 symbolech se vyskytují pouze symboly a až f a to s frekvencemi $a.freq = 45$, $b.freq = 13$, $c.freq = 12$, $d.freq = 16$, $e.freq = 9$ a $f.freq = 5$ (hodnoty jsou uváděny v tisících.

Huffmanův kód

Optimální kódy

Je dán text nad abecedou C s frekvencemi $c.\mathit{freq}$ pro $c \in C$. Pak délka kódu w je

$$B(w) = \sum_{c \in C} c.\mathit{freq} \cdot |w(c)|,$$

kde $w(c)$ je kódové slovo odpovídající c.

Strom binárního kódu

Kód w lze reprezentovat binárním stromem T, kde

- hrany jsou označeny 0 nebo 1,
- listy stromu jsou označeny $c \in C$,
- orientovaná cesta z kořene do c je označena $w(c)$.

Huffmanův kód

Platí

$$B(w) = B(T) = \sum_{c \in C} c.freq \cdot d_T(c),$$

kde $d_T(c)$ je hloubka c v T.

Je-li T strom optimálního kódu, pak každý vrchol, který není list, má dva následníky.

Huffmanův kód

Konstrukce Huffmanova kódu.

Vstup: Abeceda C spolu s $c.\mathit{freq}$
Výstup: Strom T optimálního binárního kódu w

1. $Q := C$; $\mathcal{T} := \{T_c \mid c \in C\}$, kde T_c je strom s jedním vrcholem označeným $c.\mathit{freq}$.
2. Dokud $|Q| \neq 1$ vybereme $x, y \in Q$ se dvěma nejmenšími $x.\mathit{freq}$, $y.\mathit{freq}$, $x.\mathit{freq} \leq y.\mathit{freq}$;
   nahradíme x, y v Q vrcholem z, položíme
   $z.\mathit{freq} := x.\mathit{freq} + y.\mathit{freq}$;
   Nahradíme T_x, T_y novým T_z s kořenem z označeným $z; z.\mathit{freq}$, kde levý podstrom z je T_x, pravý je T_y.
3. Je-li $Q = \{q\}$ (a $\mathcal{T} = \{T_q\}$), pak vrátíme T_q.

Huffmanův kód

Variant.

Po každém průchodu bodem 2 má množina Q o jeden prvek méně (totéž platí pro množinu $\mathcal{T}$). Tedy po $n-1$ průchodech bodem 2 algoritmus skončí.

Invariant

Předpokládejme, že T' je strom optimálního kódu pro C', kde T' jsme získali z T podle Huffmanova algoritmu. Pak T je optimální strom pro C.

Huffmanův kód

Tvrzení 1.

Nechť C je abeceda s $c.freq$, $c \in C$. Předpokládejme, že $x, y \in C$ jsou symboly s dvěma nejmenšími $c.freq$.
Pak existuje optimální strom T takový, že x a y mají kódová slova stejné délky a liší se pouze v posledním bitu.

Tvrzení 2.

Nechť C je abeceda s $c.freq$ a $x, y \in C$ jsou symboly s 2 nejmenšími $c.freq$. Označme $C' = (C \setminus \{x, y\}) \cup \{z\}$ a $z.freq = x.freq + y.freq$.

Předpokládejme, že T' je optimální strom C'. Pak T získané z T' nahrazením z následníky x a y je optimální strom C.

Huffmanův kód

Věta.

Huffmanův kód zkonstruovaný výše je optimální binární kód pro abecedu C.

Turingovy stroje

Neformální popis Turingova stroje.

Turingův stroj (zkráceně TM) má tyto části:

- *řídící jednotka*, která se může nacházet v jednom z konečně mnoha stavů,
- potenciálně nekonečná páska rozdělená na políčka, každé z nich obsahuje jeden páskový symbol a
- hlavu, která čte obsah políčka pásky, a přepisuje obsah políčka.

Na základě páskového symbolu X, stavu q řídící jednotky, TM změní svůj stav na p, zapíše páskový symbol Y do čteného pole pásky a posune hlavu doprava nebo doleva. Tato změna je určena přechodovou funkcí δ.

Turingovy stroje

Formální definice.

TM je sedmice $(Q, \Sigma, \Gamma, \delta, q_0, B, F)$, kde

- Q je konečná neprázdná množina stavů,
- Σ je konečná neprázdná množina vstupních symbolů (vstupů),
- Γ je konečná množina páskových symbolů, kde $\Sigma \subset \Gamma$,
- B je blank *blank*, speciální symbol znamenající, že políčko pásky je prázdné; $B \in \Gamma \setminus \Sigma$,
- δ je přechodová funkce, tj. parciální zobrazení $(Q \setminus F) \times \Gamma$ do $Q \times \Gamma \times \{L, R\}$, (zde L znamená pohyb hlavy doleva, R pohyb hlavy doprava),
- $q_0 \in Q$ je počáteční stav a
- $F \subseteq Q$ je množina koncových (akceptujících) stavů.

Turingovy stroje

Stavový diagram

Situace TM

Situace ID (také *konfigurace*) je

$$X_1\, X_2\, \ldots\, X_{i-1}\, q\, X_i\, X_{i+1}\, \ldots\, X_k,$$

kde

- $X_1, X_2, \ldots, X_k$ je obsah části pásky takový, že všechny ostatní políčka obsahují B,
- TM je ve stavu q,
- hlava čte symbol X_i.

Turingovy stroje

Počáteční ID

Na začátku práce TM nad vstupním slovem $w \in \Sigma^\star$, $w = a_1 \dots a_n$

- TM je ve stavu q_0,
- páska obsahuje $a_1 \dots a_n$ v n po sobě následujících políčkách, ostatní políčka obsahují B,
- hlava čte a_1.

Tedy počáteční ID je tedy

$$q_0\, a_1\, a_2 \dots a_n.$$

Turingovy stroje

Předpokládejme, že $\delta(q, X_i) = (p, Y, R)$, pak

$$X_1\, X_2\, \ldots\, X_{i-1}\, q\, X_i\, \ldots\, X_k \vdash X_1\, X_2\, \ldots\, X_{i-1}\, Y\, p\, X_{i+1}\, \ldots\, X_k. \quad (1)$$

Je-li $i = k$, pak

$$X_1\, \ldots\, X_{k-1}\, q\, X_k \vdash X_1\, \ldots\, X_{k-1}\, Y\, p\, B.$$

Předpokládejme, že $\delta(q, X_i) = (p, Y, L)$, pak

$$X_1\, X_2\, \ldots\, X_{i-1}\, q\, X_i\, \ldots\, X_k \vdash X_1\, \ldots\, X_{i-2}\, p\, X_{i-1}\, Y\, X_{i+1}\, \ldots\, X_k. \quad (2)$$

Je-li $i = 1$, pak

$$q\, X_1\, \ldots\, X_k \vdash p\, B\, Y\, \ldots\, X_k.$$

Turingovy stroje

Výpočet

$\vdash^\star$ je tranzitivní a reflexivní uzávěr relace $\vdash$.

Slovo $w = a_1 \ \ldots \ a_k$ je přijímáno TM, jestliže

$$q_0\, a_1 \ \ldots \ a_k \vdash^\star \alpha\, p\, \beta.$$

pro některé $p \in F$, $\alpha, \beta \in \Gamma^\star$.

Jazyk přijímaný TM je množina slov $L(M)$, kde

$$L(M) = \{w \in \Sigma^\star \mid \ w \text{ je přijímán } M\}.$$

Turingovy stroje

Také říkáme, že se Turingův stroj

- úspěšně zastaví nad slovem w, jestliže
$$q_0\, a_1\, \ldots\, a_k \vdash^\star \alpha\, p\, \beta,$$
pro $p \in F$;
- neúspěšně zastaví nad slovem w, jestliže
$$q_0\, a_1\, \ldots\, a_k \vdash^\star \alpha\, p\, X \beta,$$
a $\delta(p, X)$ není definováno a $p \notin F$.

Turingovy stroje

Je dána funkce $f\colon \Sigma^\star \to \Sigma^\star$. Řekneme, že TM realizuje f, jestliže

- pro každé w, pro které $f(w)$ není definováno, TM se zastaví neúspěšně;
- pro každé w, pro které $f(w)$ je definováno, TM se zastaví úspěšně a při zastavení je na pásce $f(w)$.

Turingovy stroje

Je dána funkce $f\colon \mathbb{N}^k \to \mathbb{N}$. Řekneme, že TM realizuje f, jestliže

- pro každé $(n_1, n_2, \ldots, n_k)$ utvoříme slovo $w = 0^{n_1}10^{n_2}1\ldots10^{n_k}$.
- pro každé $w \in \{0,1\}^\star$, které nemá tento tvar, TM se zastaví neúspěšně;
- pro každé $w = 0^{n_1}10^{n_2}1\ldots10^{n_k}$, TM se úspěšně zastaví a při zastavení páska obsahuje $0^{f(n_1,n_2,\ldots,n_k)}$.

Turingovy stroje

Jazyk L je přijímán/rozhodován TM

Řekneme, že $L \subseteq \Sigma^\star$ je přijímán TM, jestliže

- pro každé $w \in L$ se TM úspěšně zastaví;
- pro každé $w \notin L$ TM se nezastaví úspěšně, tj. buď se zastaví neúspěšně nebo se nezastaví.

Řekneme, že $L \subseteq \Sigma^\star$ je rozhodován TM, jestliže

- pro každé $w \in L$ se TM úspěšně zastaví;
- pro každé $w \notin L$ se TM neúspěšně zastaví.

Časová a paměťová složitost TM

Definice.

Časová složitost Turingova stroje je parciální zobrazení $T(n)$ z množiny všech přirozených čísel do reálných čísel definované:

- Jestliže pro nějaký vstup délky n se Turingův stroj nezastaví, $T(n)$ není definováno.
- V opačném případě je to maximální počet kroků Turingova stroje před jeho zastavením, kde maximum je přes všechna slova $w \in \Sigma^\star$ délky n.

Časová a paměťová složitost TM

Definice.

Paměťová složitost Turingova stroje je parciální zobrazení $S(n)$ z množiny všech přirozených čísel do reálných čísel definované:

- Jestliže pro některé slovo $w \in \Sigma^\star$ délky n vyžaduje TM nekonečně mnoho políček pásky, $S(n)$ není definováno.
- V opačném případě je $S(n)$ rovno největšímu rozdílu pořadových čísel polí, které byly během výpočtu použity, kde maximum se bere přes všechna slova $w \in \Sigma^\star$ délky n.