

Inverzna metoda generisanja Huffmanovog stabla za kodiranje

Izložena je jedna ideja za algoritam generisanja Huffmanovog stabla. Za razliku od klasične metode generisanja Huffmanovog stabla za kodiranje, koja stablo gradi počinjući od karaktera najmanje učestalosti, izložena metoda gradi stablo polazeći od karaktera najveće učestalosti.

Ključne reči. kompresovanje podataka; algoritam za kompresovanje.

Kompresovanje podataka predstavlja transformaciju ulaznog niza podataka u izlazni niz manje dužine. Posmatrajmo neki nekompresovani niz karaktera. Na primer, u običnom tekstu na srpskom jeziku razmak i samoglasnici se pojavljuju često, dok se recimo 'đ', 'ž' i slična slova pojavljuju ređe, a svi se zapisuju nizom od 8 bita. Vidimo da bi se moglo uštedeti dosta prostora ako bi karakteri sa većom učestalošću imali kraću bit reprezentaciju, dok bi oni koji se ređe pojavljuju imali bit reprezentaciju dužu od 8. Karakteri koji se uopšte ne pojavljuju ne bi imali bit reprezentaciju. Metoda kodiranja zasniva se upravo na tom principu.

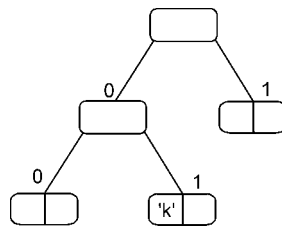
Jedan od poznatijih algoritama je Huffmanovo kodiranje, koje je zasnovano na tablici učestanosti pojavljivanja različitih karaktera u ulaznom nizu koja je sortirana u opadajućem redosledu. U standardnom slučaju, algoritam formira stablo počinjući od karaktera najmanje učestalosti.

Opis inverzne metode

Inverzna metoda, ovde predložena, takođe je zasnovana na tablici učestalosti. Radi lakše analize, uvešćemo pojam učestalosti podstabla koja predstavlja zbir učestalosti svih listova tog podstabla. Početno stablo se sastoji od prva dva karaktera iz tablice. Dokle god tabela nije prazna, iz nje se uzima karakter sa najvećom učestalošću i rekursivno prosleđuje podstablu sa manjom učestalošću. Kada algoritam prosledi karakter podstablu koje je list, taj list postaje čvor. Jedan list tako nastalog čvora sadrži karakter iz njega samog, a drugi sadrži karakter mu je prosleđen.

*Jovan Brakus (1982),
Zrenjanin, Vojvode
Putnika 20, učenik 1.
razreda Matematičke
gimnazije u Beogradu*

U toku izvršavanja, algoritam nastoji da napravi drvo sa što bližim vrednostima učestalosti podstabala, tako što karaktere uvek stavlja u podstablo sa manjom učestalošću. Nakon formiranja stabla na opisani način, pristupa se kodiranju ulaznog niza. Bit reprezentacija određenog karaktera određuje se na osnovu putanje kojom se kroz stablo do njega dolazi. Levo podstablo se predstavlja sa 0, a desno sa 1, pa ako se do traženog karaktera "stiže" putanjom levo-levo-desno tada će bit reprezentacija karaktera biti 001. Na primer, bit reprezentacija karaktera 'k' je (01) izgleda:




Primer

Neka je ulazni niz: 'karakter'. Prvo treba formirati tablicu učestalosti. Nju ćemo jednostavno formirati prebrojavanjem elemenata iz ulaznog niza. Za dati ulazni niz ona će izgledati ovako:

Karakter	Učestalost karaktera
k	2
a	2
r	2
e	1
t	1

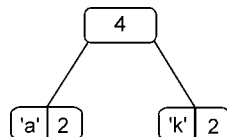
U daljem tekstu, tokom opisivanja binarnog drveta, list će biti predstavljen u obliku:

 – gde levi deo figure predstavlja karakter, a desni njegovu učestalost.

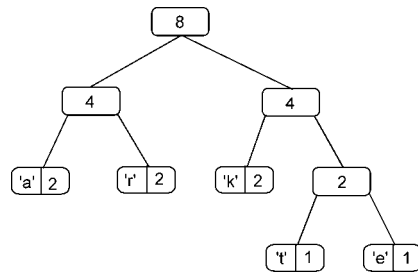
Čvor drveta će biti predstavljen u obliku:

 – u figuri će biti upisan učestalost drveta čiji je koren dati čvor.

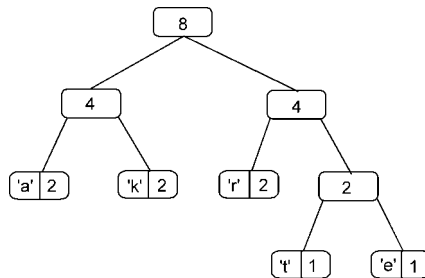
U početku imaćemo drvo koje se sastoji samo od dva prva karaktera iz tablice učestalosti. To drvo će izgledati:



Primenom inverzne metode dobija se sledeće stablo:



U slučaju klasične metode generisanja Huffmanovog stabla imamo:



U prvom slučaju kôd tabela će izgledati:

Karakter	Bit prezentacija karaktera
a	00
r	01
k	10
t	110
e	111

dok će u drugom slučaju tabela biti:

Karakter	Bit prezentacija karaktera
a	00
r	10
k	01
t	110
e	111

Vidimo da su bit reprezentacije svih karaktera jednake dužne, pa će i izlazni niz biti jednake dužne u oba slučaja:

:: 10 00 01 00 10 110 111 01 (u prvom slučaju)

:: 01 00 10 00 01 110 111 10 (u drugom slučaju)

biće dugačak 19 bita odnosno 3 bajta (karaktera).

Rezultat uporednih testova inverzne i Huffmanove metode

Uporednim testom ova dva algoritma pokazalo se da inverzni metod generisanja stabla za kodiranje daje u dosta slučajeva bolje rezultate kompresije od klasičnog Huffmanovog metoda, ali je vremenski zahtevniji. U najprostijim testovima, kada su ulazni fajlovi imali svega nekoliko različitih elemenata, algoritmi su generisali dva identična stabla, i dali izlazne fajlove iste dužine. U testovima u kojima su ulazni fajlovi imali manje različitih karaktera (do 50), klasičan algoritam je imao bolje rezultate, dok je u fajlovima u kojima se pojavljuje više od 50 različitih karaktera inverzni algoritam imao nešto bolje rezultate od klasičnog.

Zahvalnost. Zoranu Rilaku na pruženoj pomoći i savetima.

Literatura

Data Compression Reference Center (dostupno na www.rasip.fer.hr)

Jovan Brakus

Generating the Huffman Coding B-Tree Using the “Inverse Method”

This paper introduces the new algorithm which generates the Huffman B-tree. This B-tree is later used to compress strings (text). While the standard method generates the tree beginning with the minimum frequency character, method presented herein begins with the character having maximum frequency.

The Inverse method also uses the frequency table. During its operation, this algorithm is trying to generate the B-tree containing sub-trees with equal or almost equal frequencies. After that, the input string is coded. Bit-mapped representation of the single character is determined by using the path found in the B-tree. Left branch is coded by 0 and right by 1, so left-left-right path is coded like '001'.

In comparison to the standard algorithm, the Inverse method is in most cases more time-consuming, but it offers better compression for input strings containing more than 50 different characters.

