STRUKTURE PODATAKA LETNJI SEMESTAR 2015/2016

HASH TABLICE

Prof. Dr Leonid Stoimenov Katedra za računarstvo Elektronski fakultet u Nišu

Pregled Predavanja

- o Uvod − šta su Hash tablice
- Osnovni pojmovi
- Hash funkcija
- Kolizija i način rešavanja

Uvod - šta je Hash tablica

 Hash tablica je struktura podataka koja omogućava

• direktan pristup podacima ...

• ... izračunavanjem njihovog položaja u tablici ...

 ... na osnovu vrednosti ključa.

	ključ	vrednost	Ost.pod
0			
1			
2			
3			
741			
5			
•			
•			
•			
N-1			

Uvod - šta je Hash tablica

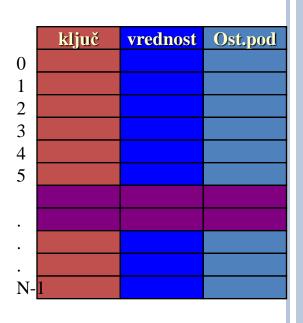
Hash tablica

o generalizacija običnog polja

indeksira se celobrojnim indeksima

okao polje indeksirano proizvoljnim ključevima mogu biti:

- ⇒ brojevi,
- ⇒ nizovi znakova, i sl.



Uvod - šta je Hash tablica

o Cilj:

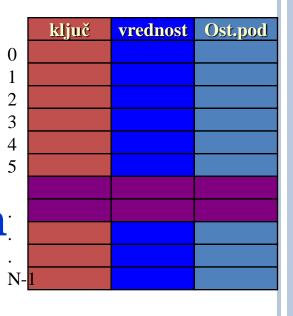
- smeštanje/dodavanje i
- brzo pretraživanje velike količine podataka.

o Osnovna ideja:

da se ulazni skup veličine N
 (koji može biti jako veliki;
 veći od raspoložive memorije)

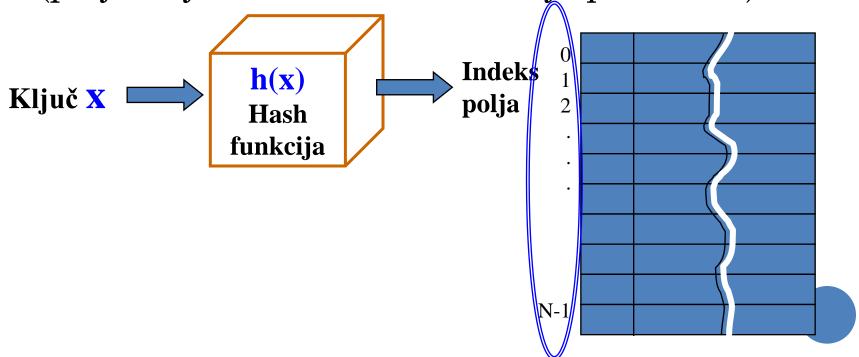
• svede na manji broj, *n*, stavki.

Takvu "kompresiju" radi hash
 funkcija



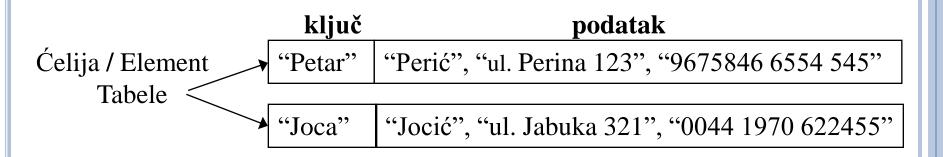
ILUSTRACIJA RADA HASH FUNKCIJE

 Hash funkcija h(x) vrši preslikavanje vrednosti ključa x u adrese (indekse) tablice (polja koje se koristi za čuvanje podataka)



Element tabele — ključ i podatak

- Zbog traženja, u svakoj ćeliji tabele odnosno u svakom elementu tabele, čuva se
 - ključ i odgovarajući podatak,
 - ... ali **odvojeno**, osim kada je vrednost ključa deo podatka.



OSNOVNI POJMOVI

(detaljnije objašnjenje sledi nadalje)

- Hash tablica struktura podataka
- Hash funkcija funkcija za preslikavanje ključa u adresu (indeks) tablice
- Ključ vrednost na osnovu koje se određuje adresa podatka
- Kolizija problem kada se dva ključa preslikavaju u istu adresu u tablici
- Sinonimi podaci sa različitim ključevima koji se preslikavaju u istu adresu u tablici
- Faktor popunjenosti odnos broja elemenata i veličine tablice

HASH FUNKCIJA

- **Ključ** hash tablice vrednost podatka na osnovu koga se vrši smeštanje u hash tablicu.
- *Hash* funkcija Funkcija koja preslikava vrednosti ključeva (vrednosti na osnovu kojih se smeštaju podaci) u cele brojeve, tj. vrednosti indeksa polja.
- Hash funkcija omogućava mehanizam pristupa koji **izbegava pretraživanje** strukture radi nalaženja elementa.
- Na osnovu **vrednosti podataka** vrši se izbor funkcije za preslikavanje njihove vrednosti u adrese, odnosno indekse polja.
- Njen izbor zahteva izvesno predznanje o vrednostima podataka koje se smeštaju u hash tablicu

FORMALNA DEFINICIJA HASH FUNKCIJE

Hash funkcija h je preslikavanje iz skupa vrednosti ključa U (|U| = N, N veličina hash tablice) u skup indeksa hash tablice:

$$h: U \to \{0, 1, \dots, N-1\}$$

- Za funkcionisanje ove ideje važno je da hash funkcija h(*el_type*) **ravnomerno raspoređuje** vrednosti iz skupa el_type na indekse tablice 0,1,2,...,N-1.
- To ne može da bude uvek slučaj!!
- Svrha hash funkcije je da preslika ključeve iz nekog opsega u vrednosti indeksa ali tako da se vrednosti ključeva distribuiraju slučajno u celom opsegu indeksa tablice.
- Ključevi mogu biti kompletno slučajni ili delimično slučajni.

ZAHTEVI KOD IZBORA HASH FUNKCIJE

- •Perfektna *hash* funkcija??
- <u>Cilj</u>: Izabrati dobru, odnosno što je moguće bolju hash funkciju
- Osnovni zahtevi koje treba da ispuni dobra hash funkcija su:
 - da izbegava kolizije,
 - da rasipa vrednosti ravnomerno po čitavoj tablici i
 - da se jednostavno (tj. brzo) izračunava.

ZAHTEVI KOD IZBORA HASH FUNKCIJE

- Od hash funkcije jako zavisi brzina umetanja i traženja elemenata u hash tablici.
- O Dobra hash funkcija treba da izvrši preslikavanje ključeva u indekse što slučajnije i uniformnije jer u tom slučaju ima najmanje kolizija.
- \circ U idealnom slučaju su ubacivanje i traženje elementa složenosti O(1).
- Ovo važi posle računanja Hash funkcije složenost njenog računanja može da zavisi od dužine ključa.

Metode za izračunavanje hash **FUNKCIJE**

- Najčešće metode za izračunavanje hash funkcija su sledeće:
 - metod **deljenja**, $h(k) = k \mod M$
 - $h(k) = \left| \frac{M}{W} (k^2 \mod W) \right|$ $h(k) = \left| \frac{M}{W} (ak \mod W) \right|$ • metod sredine kvadrata,
 - metod množenja,
 - Fibonačijev metod,

- metod presavijanja,
- metod **ekstrakcije** (izvlačenja) i
- metod transformacije osnove.

Više informacija: PRAKTIKUM!!

PRIMER RADA HASH FUNKCIJE

Hash funkcijah=(key) mod 10

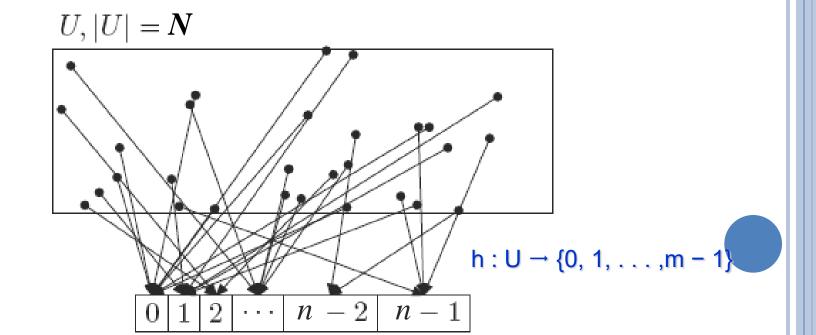
Ključ	\mathbf{h}
-------	--------------

- **o** 81 1
- **o** 64 4
- **o** 36
- **o** 49 9
- Nema kolizije !!
- **o** 61 ?
- **o** 34
- **o** 14

	IDStud	lme	BrTel	
0				
1	81	Pera	12345	
2				
3				
4	64	Sonja	23456	
5				
6	36	Mara	34567	
7				
8				
9	49	Fića	87654	
Indeks	KLJUČ			

KOLIZIJA KOD HASH FUNKCIJE

- Kolizija: Ako dva ključa imaju istu vrednost hash funkcije (različiti ključevi preslikavaju se u istu adresu), kažemo da je došlo do kolizije.
- Ključevi koji izazivaju koliziju nazivaju se sinonimi.



KOLIZIJA KOD HASH FUNKCIJE (2)

- Osnovni problem prilikom izbora *hash* funkcije je:
 - Veliki skup vrednosti ključeva, i
 - Daleko manji broj podataka koji se zaista smeštaju u tablicu.
- Za dimenziju polja gde se smeštaju podaci biramo M, tako da je $M \ge N$.
- Ako je | K | kardinalnost skupa ključeva, tada je | K |
 M,
- Posledica: hash funkcija, označena sa h, nema jednoznačno preslikavanje, odnosno. važi
 (∃(i, j) ∈ K) (i ≠ j => h(i) = h(j))

Hash funkcija – dodatni pojmovi

• Hash funkcija koja različite ključeve mapira u različite adrese naziva se perfektna hash funkcija, za koju važi:

$$(\forall (i, j) \in K) (h(i) = h(j) \Rightarrow i = j)$$

- o Perfektna hash funkcija ne generiše kolizije.
- Mogućnost realizacije?

FAKTOR POPUNJENOSTI TABLICE

- Faktor popunjenosti tablice (engl.load factor), FP
 - predstavlja odnos broja elemenata u tablici i broja rezervisanih lokacija za elemente tablice (veličine tablice):

$$FP =$$

veličina tablice

• "Idealna" vrednost $\mathbf{FP} = \mathbf{0.8}$

IMPLEMENTACIJA HASH TABLICE

- Za Hash tablicu se rezerviše polje od N ćelija ili elemenata tablice
- o Indeksi elemenata su 0,1,2,...,N-1 (ili 1,2,...,N).
- o Hash funkcija h(el_type) se implementira kao potprogram koji sukcesvino preslikava skup elemenata određenog tipa na skup celih brojeva izmedu 0 i N−1, gde je N celobrojna konstanta.
- Implementacija funkcije dodavanja elementa u hash tablicu se svodi na to da element x smestimo u ćeliju tablice sa indeksom h(x).
- Implementacija funkcije čitanja/traženja je obrnuta element sa ključem x tražimo u ćeliji sa indeksom h(x).

OSNOVNE OPERACIJE

- Osnovne operacije za rad sa *hash* tablicama su:
 - insertItem dodaje novi element u tablicu,
 - find pronalazi element tablice sa zadatim ključem
 - removeElement uklanja element sa zadatim ključem iz tablice.
- Dodatne funkcije:
 - getLength vraća veličinu tablice i
 - getLoadFactor- vraća faktor popunjenosti tablice.
- Implemenacija **primarne** transformacije (*hash* funkcija), i
- Implementacija sekundarne transformacije.

OPŠTI POSTUPAK KOD DODAVANJA PODATKA SA KLJUČEM K

Algoritam:

- 1. Izračunavanje hash funkcije h(K) na osnovu vrednosti ključa;
- 2. Ako je pozicija h(K) slobodna, podatak se upisuje na to mesto.
- 3. Inače, lokacija je zauzeta i došlo je do kolizije:
 - 1) vrši se rešavanje kolizije nekom od metoda
 - 2) Menja se vrednost primarne adrese u sekundarne adrese i pokušava upis sve dok se ne nađe slobodna lokacija
 - 3) Ako slobodna lokacija postoji, vrši se upis podatka,
 - 4) Inače, tablica je puna.

OPŠTI POSTUPAK KOD BRISANJA PODATKA SA KLJUČEM K

Algoritam:

- 1. Izračunavanje hash funkcije h(K) na osnovu vrednosti ključa;
- 2. Ako je pozicija h(K) zauzeta,
 - 1. Ako je to podatak sa ključem K, obrisati ga.
 - 2. Inače, nastavi pretragu po sinonimima
 - Menja se vrednost primarne adrese u sekundarne adrese i pokušava traženje sve dok se ne nađe traženi podatak
 - Ako podatak postoji, vrši se brisanje podatka,
 - Inače, podatak koji se traži ne postoji, nemoguće brisanje
- 3. Inače, podatak sa ključem K ne postoji

Brisanje iz hash tablice

- Ne može se jednostavno samo obrisati element iz tablice na koji ukazuje h(k). Zašto?
- Koristi se specijalna vrednost DELETED za markiranje ćelije koja je sadržala obrisani element.
 - *Traženje* tretira vrednost DELETED kao da ćelija sadrži element sa ključem koji se ne poklapa sa traženim.
 - *Dodavanje* tretira vrednost DELETED kao da je ćelija prazna tako da se može koristiti za smeštanje novog podatka.
- o Nedostatak: Vreme traženja
 - Ulančavanje sinonima kao rešenje ovog problema!

REŠAVANJE KOLIZIJE

- Kolizija dva podatka sa različitim ključevima se preko hash funkcije preslikavaju u istu adresu
- Bez obzira kako izabrali *hash* funkciju, postoji velika verovatnoća da će prilikom unosa podataka doći do kolizije.
- Kolizije produžavaju vreme pristupa elementima u tablicama, tako da je potrebno što efikasnije rešiti problem smeštanja sinonima.

Rešavanje kolizije (2)

Metode za rešavanje kolizije:

Otvoreno adresiranje – kad dođe do kolizije, polje se pretražuje na izabrani sistematski način sve dok se ne nađe prazna ćelija za upis podatka.

o Ulančavanje sinonima

- spoljašnje ulančavanje sinonima kreira se lančana lista tako da se kod kolizije u nju dodaju sinonimi
- unutrašnje ulančavanje sinonima ulančavanje se vrši unutar tablice (praktično odgovara definiciji otvorenog adresiranja, s tim da se elementima tablice dodaje još jedno polje link na sinonime).

OTVORENO ADRESIRANJE

• Kod kolizije traženje se nastavlja traženjem slobodne lokacije korišćenjem sekundarne transformacije:

$$h_i(k) = (h(k) + c(i)) \mod M, \quad i = 0, 1, ..., M - 1$$

- Funkcija c(i) naziva se sekundarna transfomacija i treba da ima sledeća dva svojstva:
 - c(0) = 0, čime se obezbeđuje da se prvi pokušaj poklopi sa primarnom (hash) transformacijom, i
 - skup { c(0) mod M, c(1) mod M, ... c(M-1) mod M } mora sadržati sve cele brojeve iz intervala [0, M-1], kako bi se obezbedilo adresiranje čitavog adresnog prostora tablice.

OTVORENO ADRESIRANJE

Najčešće korišćene sekundarne transformacije su:

- Linearno traženje
- Modifikovano linearno traženje
- Kvadratno traženje
- Sekundarna hash funkcija
- Broj pristupa tablici prilikom čitanja zavisi od popunjenosti tablice.
- Što je tablica punija, to je veći broj koraka potreban da bi se našlo mesto za novi podatak, pa time i broj koraka za njegovo kasnije nalaženje.

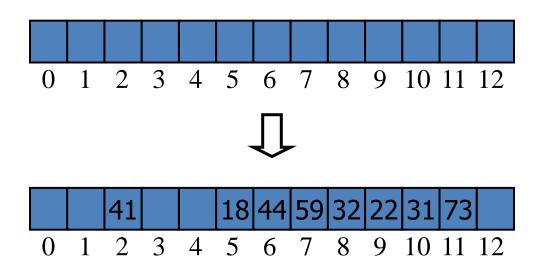
LINEARNO TRAŽENJE

- Traženje slobodne ćelije se vrši sekvencijalno (sekvencijalna promena indeksa u tablici) dok se ne dođe do prve slobodne lokacije
 - $c(i) = \alpha \cdot s$, gde je a uzajamno prost broj sa M.
- Problem: klasterovanje podataka u nekim delovima tablice
- o U tom slučaju, ako je tablica puna, spor pristup
- Treba u proseku manje od 5 pokušaja ako je tablica manje od 2/3 puna
- o U slučaju kada je α ≠ 1 ova transformacija se naziva modifikovano linearno traženje.

PRIMER LINEARNOG TRAŽENJA

• Dato je:

- $\bullet h(x) = x \bmod 13$
- Treba upisati: 18, 41, 22, 44, 59, 32, 31, 73, u navedenom redosledu



Traženje elementa sa Linearnim traženjem mesta za sinonime

```
Algoritam HT.1: Traženje u
   tablici
1. find(T,N,k)
2. \quad i \leftarrow h(k)
3. p \leftarrow 0
   repeat
5. c \leftarrow T[i]
6. if c = \emptyset
     return null
8. else if c.key = k
9.
         return T[i]
10.
     else
11. i \leftarrow (i+1) \mod N
12.
          p \leftarrow p + 1
13.
       until p = N
14. return null
```

- Traženje u tablici T počinje od lokacije h(k)
- Vrši se sukcesivni obilazak lokacija sve dok se ne desi:
 - Našli smo element sa ključem k,
 ILI
 - Našli smo praznu ćeliju, ILI
 - N ćelija je neuspešno obiđeno (N je veličina T)

AŽURIRANJE SA LINEARNIM TRAŽENJEM

Kod operacija dodavanja i brisanja, uveden je specijalni objekat, nazvan *DELETED*, koji zamenjuje obrisane elemente

Algoritam HT.2: Brisanje u tablici removeElement(T,N,k)

- 1. Traženje elementa sa ključem k (kao kod **find**)
- 2. Ako je element (k, o) nađen,
 - **Tada** zamenjujemo njegovu vrednost specijalnom vrednošću *DELETED* i vraćamo poziciju tog elementa
- 1. Inače, vraća se null pozicija

Ažuriranje sa linearnim traženjem

Kod operacija dodavanja i brisanja, uveden je specijalni objekat, nazvan *DELETED*, koji zamenjuje obrisane elemente

Algoritam HT.3: Dodavanje

insertItem(T, N, k, o)

- Ako je tabela puna Tada Prekoračenje Kraj
- 2. Početna ćelija za obradu je h(k)
- Obilazimo konsekventne ćelije sve dok se ne desi nešto od sledećeg (petlja)
 - Ćelija *i* je prazna ili sadrži *DELETED*, ili
 - N ćelija je neuspešno obiđeno
- 4. **Ako** je nađena prazna ćelija **tada** Smestimo element (**k**, **o**) u ćeliju **i**

Problem brisanja kod Linearnog traženja

• h=key mod 10

oinsertElement:

47, 57, 68, 18, 67

• find: 68

• find: 10

• removeElement: 47

• find: 57

0	
1	
2	
2 3 4 5 6 7	
4	
5	
6	
8 9	
9	

Problem brisanja kod Linearnog traženja (2)

Pregled:

- Lažno brisanje elemenata specijalna vrednost DELETED
- Za svaku ćeliju definišemo 3 moguća stanja:
 - aktivna
 - prazna
 - obrisana
- •Za operacije find ili removeElement
 - Pretraživanje se zaustavlja jedino ako je detektovana prazna ćelija (ne i za obrisanu!!)

Problem brisanja kod linearnog traženja (3)

o InsertElement

- Ćelija prazna ili obrisana
- Ćelija aktivna

insert na poz. H, cell = active

$$H = (H + 1) \mod TS$$

Find

- Ćelija prazna
- Ćelija obrisana
- Ćelija aktivna

NOT found

$$H = (H + 1) \mod TS$$

else
$$H = (H + 1) \mod TS$$

o removeElement

- Ćelija aktivna; key <> key u ćeliji
- Ćelija aktivna; key = key u ćeliji
- Ćelija obrisana
- Ćelija prazna

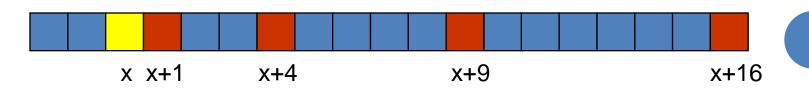
$$H = (H + 1) \mod TS$$

$$H = (H + 1) \bmod TS$$

NOT found

KVADRATNO TRAŽENJE

- Pokušaj da se izbegne problem klasterovanja
- Slobodna lokacija se traži:
 - $c(i) = \alpha \cdot i^2$
- Vrš se distribucija podataka u izdvojene ćelije u tablici
- Ako je indeks x, sledeće ćelije koje se obilaze su x+1, x+4, x+9, x+16 itd.
- Eliminiše primarno klasterovanje
- Problem ješto se svi ključevi koji se posle heširanja preslikavaju u istu ćeliju slede istu sekvencu kod traženja slobodne (sekundarno klasterovanje).



SEKUNDARNA HASH FUNKCIJA

- Dvostruko heširanje
- Vrši se transformacija ključa novom hash funkcijom, tzv. Sekundarna hash funkcija, koja se koristi i u svakom narednom koraku za dobijeni ključ:
 - $c(i) = i \cdot h'(k)$, gde je h'(k) *hash* funkcija različita od primarne transformacije.

h(k) k

h(i)

0

h1(j)

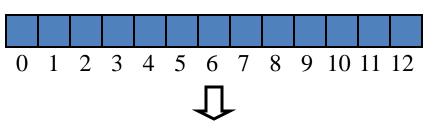
h2(j)

- Bolje rešenje u odnosu na prethodna, manji broj pokušaja u odnosu na linearno traženje
- Generiše sekvence koje zavise od novo generisanog ključa, i nisu iste za svaki ključ.
- Promena koraka je konstantna, ali se razlikuje za različite ključeve.
- Ne sme nikad da vrati 0.
- Zahteva da veličina tablice bude prost broj.

Primer dvostrukog heširanja

- Hash tablica sadrži cele brojeve
- Kolizija dvostruko heširanje
 - N = 13
 - $h(k) = k \mod 13$
 - $d(k) = 7 k \mod 7$
- Dodati:18, 41, 22, 44, 59,32, 31, 73, unavedenomredosledu

k	h(k)	d(k)	Adr	esa	
18	5	Ω	5		
41	2	1	2		
22	9	6	9		
44	5	5	5	10	
59	7	4	7		
32	6	3	6		
31	5	4	5	9	0
73	8	4	8		



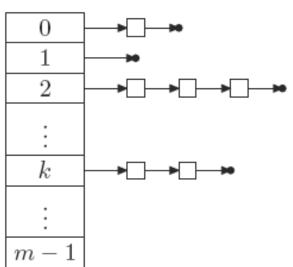
31		41			18	32	59	73	22	44		
0	1	2	3	4	5	6	7	8	9	10	11	12

Rešavanje kolizije — ilustracija ulančavanja sinonima

Ulančavanje sinonima

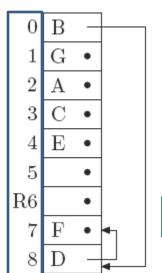
Spoljno ulančavanje

 Svaki element tablice sadrži element, ukazatelj na početak l.liste koja sadrži sinonime



Unutrašnje ulančavanje

- implementirano po ugledu na statičku implementaciju lančanih listi
- 2D polje, gde je jedna kolona za podatke a druga kolona za "link" unutar tablice



SPOLJAŠNJE ULANČAVANJE SINONIMA

- *Hash* tablica je organizovana kao vektor lančanih listi.
- Hash funkcija određuje u kojoj od lančanih listi se nalazi traženi podatak
- Lista obilazi standardnim metodom praćenja linkova.
- Broj elemenata liste je broj sinonima M
- Taj broj može da utiče na brzinu pristupa
- Redukuje se broj poređenja za faktor M, ali se koristi dodatni memorijski prostor za M linkova
- Brisanje elemenata iz tablice ne predstavlja problem
- Veličina tablice ne mora biti prost broj
- Mogu se koristiti i polja tzv buckets na svakoj lokaciji u hash tablici umesto ulančavanja.

Unutrašnje ulančavanje sinonima

- Svaka vrsta u tablici mora sadržati bar dva polja:
 - ključ
 - pokazivač na sledeći sinonim.
- *Hash* funkcijom određuje se pozicija prvog sinonima, a ukoliko traženi podatak nije na toj poziciji, traženje se nastavlja sleđenjem "linkova"
- Sinonimi mogu biti postavljeni bilo gde u okviru tablice
- Upis se vrši u dva prolaza
 - Sinonimi se smeju postavljati tek nakon zauzimanja svih lokacija dobijenih transformacijama ključa
 - Smeštanje se najčešće obavlja linearnim traženjem slobodne lokacije, počev od pozicije koju generiše *hash* funkcija. Kasnije dodavanje novih podataka u tablicu zahteva restrukturiranje tablice.
- Za sinonime se može rezervisati poseban prostor.

OSOBINE HASH TABLICA

- Struktura podataka koja obezbeđuje brz pristup i dodavanje podataka, *uglavnom* reda **O(1)**.
- Relativno se lako programira (u odnosu na druge složene strukture kao što su stabla ili grafovi napr.)
- Implementira se preko polja, s tim da pozicija elementa u polju zavisi od aritmetičke transformacije ključa.
- Pošto se implementacija zasniva na poljima, postoji poznati problem ekspanzije polja.
- Ne postoji podesan način za obilazak elemenata hash tablice u bilo kom redosledu.

EFIKASNOST HASHING-A

- Dodavanje i traženje je reda O(1).
- Kod kolizije, vreme pristupa zavisi od dužine traženja mesta za sinonime.
- Pojedinačna operacija upisa ima složenost koja je proporcionalna dužini traženja mesta za sinonime.
- Kvadratno traženje i dvostruko heširanje su istih performansi.
 - Kod uspešnog traženja: log₂(1 loadFactor) / loadFactor
 - Neuspešno traženje: 1 / (1 loadFactor)
- Traženje sa spoljašnjim ulančavanjem: 1 + loadFactor /2
 - Kod uspešnog traženja: 1 + loadFactor
 - Kod umetanja: 1 + loadfactor² za uređene liste i 1 za neuređene.
- Ako se koristi otvoreno adresiranje, dvostruko heširanje ima prednost u odnosu na kvadratno traženje.
- Ako imate na raspolaganju veliku količinu memorije, i ne zahteva se povećanje količine podataka, linearno traženje je vrlo jednostavno za implementaciju.
- Ako je količina podataka nepoznata, spoljašnje ulančavanje ima prednost
- Ako ste u dilemi šta da koristite, implementirajte spoljašnje ulančavanje

PITANJA, IDEJE, KOMENTARI

