algoritmi e strutture di dati

hashing

m.patrignani

nota di copyright

- queste slides sono protette dalle leggi sul copyright
- il titolo ed il copyright relativi alle slides (inclusi, ma non limitatamente, immagini, foto, animazioni, video, audio, musica e testo) sono di proprietà degli autori indicati sulla prima pagina
- le slides possono essere riprodotte ed utilizzate liberamente, non a fini di lucro, da università e scuole pubbliche e da istituti pubblici di ricerca
- ogni altro uso o riproduzione è vietata, se non esplicitamente autorizzata per iscritto, a priori, da parte degli autori
- gli autori non si assumono nessuna responsabilità per il contenuto delle slides, che sono comunque soggette a cambiamento
- questa nota di copyright non deve essere mai rimossa e deve essere riportata anche in casi di uso parziale

sommario

- tipi astratti di dato
  - array associativo
  - insieme
- tabelle hash
  - collisioni e liste di trabocco
  - uso per la realizzazione di tipi astratti di dato
- funzioni hash
  - per interi, per stringhe, per oggetti

il tipo astratto di dato array associativo

- domini
  - il dominio di interesse è l'insieme A degli array associativi
  - dominio di supporto: le chiavi K dell'array associativo
  - dominio di supporto: i valori V dell'array associativo
    - comprensivo della costante "valore nullo"
  - dominio di supporto: i booleani B = {true, false}
- costanti
  - l'array associativo vuoto
- operazioni
  - aggiunge una coppia ⟨chiave,valore⟩ : PUT: $A \times K \times V \rightarrow A$
  - restituisce il valore associato ad una chiave: GET: $A \times K \rightarrow V$
    - può restituire il valore nullo se nessun elemento è associato alla chiave
  - rimuove la coppia ⟨chiave,valore⟩: DELETE: $A \times K \rightarrow A$
  - verifica che un a chiave sia utilizzata: EXISTS: $A \times K \rightarrow B$
  - …

realizzazioni di array associativi

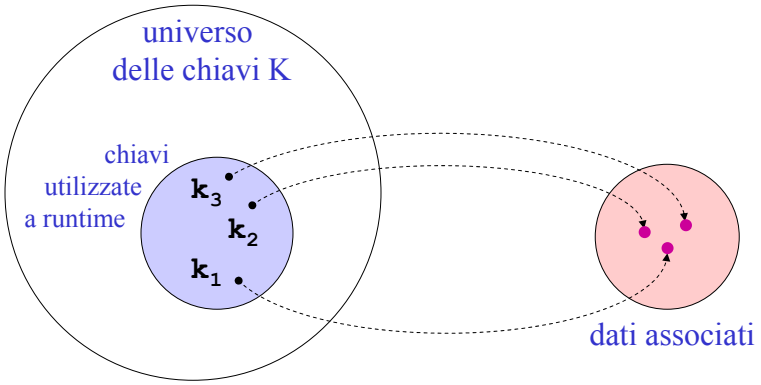
- le realizzazioni degli array associativi
  - non sono ovvie come quelle degli array tradizionali
    - alcuni linguaggi di programmazione li supportano solo tramite librerie aggiuntive
      - esempi: Pascal, C, C++, Java, ecc.
    - altri linguaggi li supportano nativamente
      - esempi: PHP, Python, Ruby, ecc.
  - non garantiscono tempi accettabili nel caso peggiore, ma solamente nel caso medio

difficoltà da superare nella realizzazione

- due difficoltà principali
  1. la tipologia delle chiavi
     - le chiavi non sono necessariamente degli interi
     - non possiamo confidare nelle chiavi per indicizzare direttamente un array
  2. la numerosità delle possibili chiavi
     - anche se le chiavi fossero degli interi, un array in grado di contenere tutte le chiavi sarebbe troppo grande e troppo sparso
       - per esempio se la chiave fosse un numero di matricola di sei cifre dovrei allocare un array con un milione di posizioni anche se gli studenti del corso che voglio considerare sono solo qualche centinaio

array associativi: osservazione

- il numero di chiavi effettivamente utilizzate dal programma in esecuzione è molto minore del numero delle chiavi possibili
  - quest'ultimo è chiamato "universo" delle chiavi K

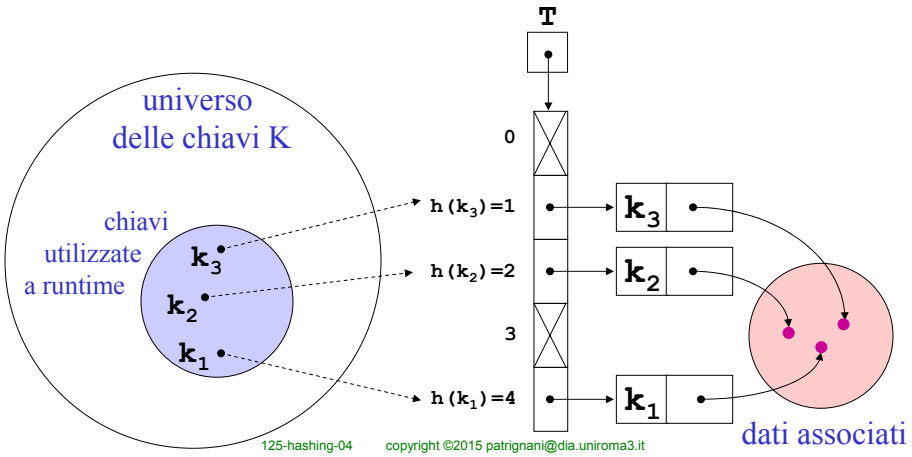


realizzazione tramite tabelle hash

- utilizzo un array `T` per memorizzare i dati associati
  - la dimensione *m* dell'array `T`
    - è molto minore della dimensione dell'universo K
    - è molto vicina al numero delle chiavi effettivamente utilizzate dal programma in esecuzione
  - l'array `T`, come tutti gli array, può essere indicizzato solo da un intero
- definisco una funzione hash *h* che trasforma le chiavi di K negli interi nel range [0…*m*-1]

tabella hash

- l'array **T** indicizzato tramite la funzione hash *h* è chiamato *tabella hash* (oppure *hashtable*)



funzione hash

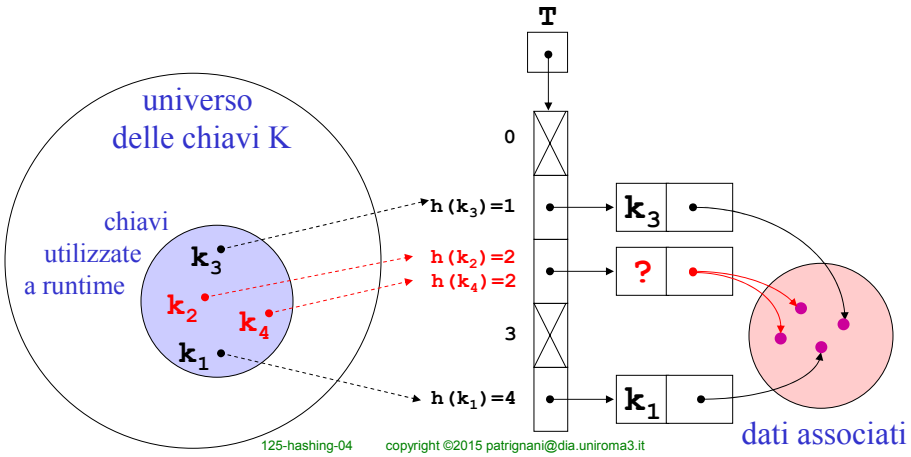
- la funzione hash *h*
  - definisce una corrispondenza tra l'universo K delle chiavi e gli indici della tabella hash **T** [0…*m*-1]

$$h: K \rightarrow \{0, 1, \ldots, m-1\}$$

  - deve essere deterministica
    - altrimenti dopo aver messo i valori nell'array non riesco più a ritrovare la loro posizione
  - si richiede che sia calcolabile in tempo costante
    - per contenere i tempi di calcolo
- l'elemento con chiave $k \in K$ si troverà nella posizione $h(k)$ nella tabella **T**

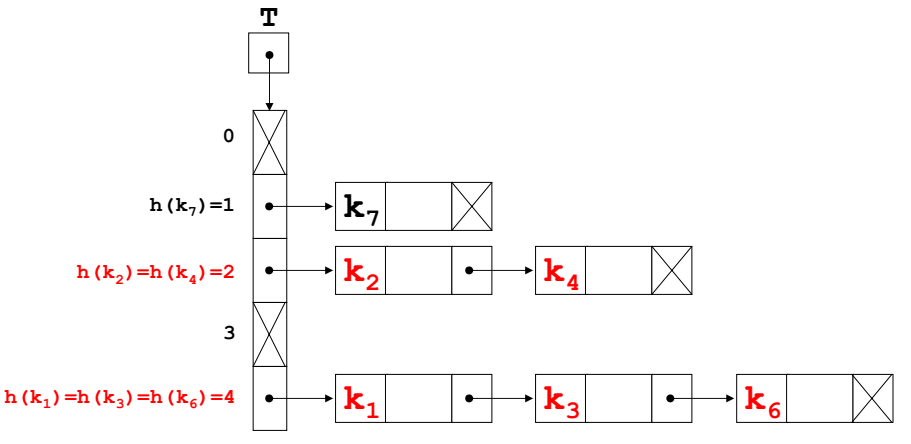
il problema della collisione

- la funzione *h* ha un codominio (gli *m* indici di **T**) molto più piccolo del dominio (tutti gli elementi di K)
  - è inevitabile che si generino collisioni



gestione delle collisioni con liste di trabocco

- ogni posizione di **T** è un riferimento al primo elemento di una lista detta "di trabocco"



liste di trabocco

- la lista di trabocco è una lista semplicemente concatenata
- ogni nodo della lista è un oggetto con tre campi
  - `key`: valore della chiave
    - può essere un riferimento ad oggetto
  - `info`: valore associato alla chiave
    - può essere un riferimento ad oggetto
  - `next`: riferimento al prossimo nodo
    - è `NULL` per l'ultimo nodo della lista

ricerca di un elemento in base alla chiave

- per ricercare un elemento devo scorrere la lista opportuna

```
GET(T,k) ▷ ritorna il valore associato alla chiave (o NULL)
1.  i = HASH(k)     ▷ devo guardare la lista i-esima
2.  x = T[i]        ▷ iteratore per elementi della lista T[i]
3.  while x != NULL
4.      if EQUAL(k,x.key)
5.          return x.info     ▷ l'ho trovato!
6.      x = x.next
7.  return NULL               ▷ non l'ho trovato
```


inserimento di una coppia ⟨chiave, valore⟩

- anche l'inserimento prevede una ricerca

```
PUT(T,k,v) ▷ inserisce <k,v> (eventualmente sovrascrive)
1.  i = HASH(k)     ▷ devo cercare nella lista i-esima di T
2.  x = T[i]        ▷ iteratore per elementi della lista T[i]
3.  while x != NULL
4.      if EQUAL(x.key,k)
5.          x.info = v     ▷ sovrascrivo il vecchio valore
6.          return         ▷ ho finito ed esco
7.      x = x.next
8.  y.key = k              ▷ y nuovo elemento della lista
9.  y.info = v
10. y.next = T[i]
11. T[i] = y               ▷ inserimento in testa
```


cancellazione di un elemento

- l'elemento viene rimosso in base alla chiave

```
DELETE(T,k) ▷ rimuove l'elemento (se esistente)
1.  i = HASH(k)     ▷ devo cercare nella lista i-esima di T
2.  x = T[i]        ▷ iteratore per elementi della lista T[i]
3.  prev = NULL     ▷ punterà all'elemento che precede x
4.  while x != NULL
5.      if EQUAL(x.key,k)        ▷ l'ho trovato
6.          if prev == NULL      ▷ x è il primo della lista
7.              T[i] = x.next
8.          else                 ▷ non è il primo della lista
9.              prev.next = x.next      ▷ lo saltiamo
10.         return               ▷ ho finito ed esco
11.     prev = x      ▷ ancora non trovato, provo il prossimo
12.     x = x.next
```

funzione EQUAL e funzione HASH

- una precondizione perché si possa realizzare un array associativo con hashtable è che siano definite
  - una funzione EQUAL
  - una funzione HASH
- entrambe le funzioni devono essere definite in base al contesto applicativo
- per motivi di efficienza si richiede generalmene che entrambe le funzioni siano calcolabili in tempo costante

una funzione EQUAL errata

- supponi che le chiavi siano stringhe
  - per esempio realizzate tramite array di caratteri
- la funzione EQUAL seguente è errata:

```
EQUAL-WRONG(A,B)      ▷ A e B sono due array di caratteri
1.  return A == B
```

  - in questo modo non vengono confrontati i valori contenuti negli array A e B, ma i loro riferimenti
    - cioè gli indirizzi, che sono necessariamente diversi anche quando le due stringhe sono uguali

una funzione EQUAL corretta

- la funzione seguente è una funzione corretta per questo contesto applicativo

```
EQUAL(A,B) ▷ A e B sono due array di caratteri
1.  if A.length != B.length
2.      return FALSE          ▷ lunghezza diversa
3.  for i=0 to A.length-1
4.      if A[i] != B[i]
5.          return FALSE      ▷ almeno un carattere diverso
6.  return TRUE
```

  - questa volta vengono confrontati tutti i caratteri delle due stringhe
    - la complessità della procedura è ancora $\Theta(1)$ se le stringhe hanno una dimensione massima nota

HASH: requisiti

- requisiti funzionali
  - è deterministica
    - data una chiave k, dà sempre lo stesso risultato HASH(k)
- requisiti prestazionali
  - è calcolabile in tempo costante
  - distribuisce le chiavi utilizzate in esecuzione in maniera pseudocasuale nell'intervallo [0…*m*-1]
    - questo requisito potrà solo essere soddisfatto solo in modo probabilistico
      - le chiavi che saranno utilizzate dall'utente in esecuzione non sono note a priori
      - comunque si scelga la funzione `HASH` esisterà sempre un insieme di chiavi che corrispondono alla stessa casella di T

HASH: ipotesi di distribuzione uniforme

- è soddisfatta dalla funzione `HASH` quando, data una chiave $k \in K$, la probabilità che `HASH`(k)=c sia la stessa per ogni casella $c \in [0 \ldots m\text{-}1]$ di `T`
  - indipendentemente da quali altre chiavi siano state già inserite in `T`
- implicazioni dell'ipotesi di distribuzione uniforme
  - per chiavi simili vengono generati hash diversi
    - spesso le chiavi utilizzate sono molto simili
  - quali che siano le chiavi utilizzate, queste vengono con alta probabilità distribuite uniformemente nell'intervallo $[0 \ldots m\text{-}1]$

fattore di carico

- si definisce *fattore di carico* il rapporto α tra il numero n di elementi memorizzati e il numero m di posizioni disponibili

$$\alpha = \frac{n}{m}$$

- α è il numero medio di elementi memorizzati in ogni lista concatenata
- a seconda del valore di α abbiamo
  - $\alpha < 1$ molte posizioni disponibili rispetto agli elementi memorizzati
  - $\alpha = 1$ il numero di elementi corrisponde al numero delle posizioni disponibili
  - $\alpha > 1$ molti elementi da memorizzare rispetto al numero delle posizioni disponibili

complessità delle operazioni

- caso migliore
  - l'operazione è eseguita su una lista di trabocco vuota o con un solo elemento
  - la complessità dell'operazione è data dalla complessità di `HASH` oppure di `HASH` + `EQUAL`
    - complessità $\Theta(1)$
- caso peggiore
  - tutte le chiavi utilizzate corrispondono alla stessa posizione
  - la complessità coincide con quella che si ha per il calcolo di `HASH`(k) + la ricerca in una lista con n posizioni + il calcolo di `EQUAL` per n volte
    - complessità $\Theta(1) + \Theta(n) + \Theta(n) = \Theta(n)$

complessità nel caso medio

- caso medio
  - se la funzione `HASH` distribuisce le chiavi in modo uniforme nell'intervallo [0…m-1]
    - la lunghezza attesa delle liste di trabocco coincide con la lunghezza media α
    - le operazioni hanno complessità $\Theta(\alpha)$
    - se α non supera mai una soglia fissata α_{max} la complessità di ogni operazione è $\Theta(1)$
  - se la funzione `HASH` non dà garanzie rispetto alla distribuzione delle chiavi
    - la complessità è la stessa del caso peggiore

funzioni hash

- considereremo delle funzioni hash per le seguenti tipologie di chiavi
  - funzioni hash per interi
    - metodo della divisione
      - veloce ma raramente adottato
    - metodo della moltiplicazione
  - funzioni hash per stringhe
  - funzioni hash per oggetti arbitrari

metodo della divisione: MOD-H

- utilizza il resto di una divisione intera

$$h(k) = k \bmod m$$

- in pseudocodice:

```
MOD-H(k,m)    ▷ k ed m sono interi
1.  return k mod m
```

- è un metodo molto veloce
- se le chiavi sono già degli interi pseudocasuali la funzione MOD-H viene utilizzata per riportare le chiavi nell'intervallo $[0 \ldots m\text{-}1]$

metodo della divisione: MOD-H

- se le chiavi non sono pseudocasuali
  - la praticità del metodo è compromessa
  - MOD-H ha delle forti proprietà di località
    - con altissima probabilità MOD-H(k+1) = MOD-H(k)+1
    - se m è una potenza di 2 o di 10, MOD-H(k) produce la parte meno significativa del numero k espresso in quella base
  - se si vuole usare questo metodo, è raccomandabile adottare come m un numero primo lontano da una potenza di due per limitare le collisioni
    - per esempio 701

metodo della moltiplicazione: osservazione 1

- supponiamo che le chiavi siano numeri reali pseudocasuali nell'intervallo (0,1)
- la funzione

$$h(k) = \lfloor m \cdot k \rfloor$$

è una buona funzione hash

  - $h(k)$ è deterministica
  - $h(k)$ può essere calcolata in $\Theta(1)$
  - $h(k)$ distribuisce uniformemente le chiavi nell'intervallo $[0 \ldots m\text{-}1]$
    - in quanto le chiavi erano già uniformemente distribuite!

metodo della moltiplicazione: osservazione 2

- sia *irr* un numero irrazionale
  - *irr* ha infinite cifre dopo la virgola, ma non è periodico
- date delle chiavi intere qualsiasi, le cifre decimali dopo la virgola del prodotto $k \cdot irr$ si possono assumere uniformemente distribuite nell'intervallo (0,1)
  - questo valore coincide con $k \cdot irr - \lfloor k \cdot irr \rfloor$
  - Knuth propone

$$irr = \frac{\sqrt{5}-1}{2} = 0.6180339...$$

    - è la parte dopo la virgola della sezione aurea

$$\varphi = \frac{\sqrt{5}+1}{2} = 1.6180339...$$

    - è un numero irrazionale

metodo della moltiplicazione

- si utilizza come hash la parte intera di un prodotto

$$h(k) = \lfloor\ m \cdot (k \cdot irr - \lfloor k \cdot irr \rfloor)\ \rfloor$$

- dove
  - *irr* è un numero irrazionale in (0,1)
    - per esempio la parte dopo la virgola della sezione aurea φ definita nella slide precedente
  - *m* può essere scelto arbitrariamente
    - di solito si usa una potenza di due: $m=2^p$, dove p è un intero

metodo della moltiplicazione: MUL-H

- la costante *irr* viene calcolata una volta sola

```
1.  a.irr = (SQRT(5)-1)*0.5     ▷ a.irr costante irrazionale
```

- la funzione MUL-H riceve in input l'array associativo a (per avere *m* ed *irr*) e la chiave *k*

```
MUL-H(a,k)                  ▷ a array associativo, k intero
1.  m = a.T.length          ▷ m è un numero intero
2.  prod = k * a.irr        ▷ prod è un numero reale
3.  prod = m * (prod - INT(prod))     ▷ reale in (0,m)
4.  out = INT(prod)                   ▷ intero in [0,m-1]
5.  return out
```

  - la funzione INT tronca un reale all'intero inferiore

HASH per stringhe: SIMPLE-H

```
SIMPLE-H(S)   ▷ S è una stringa (array di caratteri)
1.  hash = 0
2.  for i = 0 to S.length-1
3.      hash = hash + ASCII(S[i])
4.  return hash
```

- ritorna un numero intero che poi deve essere ridotto nell'intervallo [0…*m*-1] con la funzione MOD-H
- introdotta nella prima edizione del Kernigham-Ritchie
- veloce ma generalmente considerata poco efficace nel distribuire i valori in modo pseudocausuale
  - permutazioni di caratteri hanno lo stesso hash!

HASH per stringhe: DJB2-H

```
DJB2-H(S)      ▷ S è una stringa (array di caratteri)
1.  hash = 5381
2.  for i = 0 to S.length-1
3.      hash = hash*33 + ASCII(S[i])
4.  return hash
```

- introdotta da Daniel J. Bernstein (donde il nome)
- il numero 5381 è un numero primo
- il "magic numer" 33 non è giustificato teoricamente
  - ma dà ottimi risultati nella pratica
  - corrisponde al prodotto * 32 (traslazione di cinque caselle della rappresentazione) + una somma
    - può essere realizzato velocemente

funzioni hash per oggetti

- supponiamo che
  - l'oggetto abbia h campi $c_1, c_2, \ldots, c_h$
  - siano già definite opportune funzioni hash $\mathtt{HASH}_1(c_1), \mathtt{HASH}_2(c_2), \ldots \mathtt{HASH}_h(c_h)$
- una funzione hash si può ottenere facilmente con la loro somma

$$\mathtt{HASH}(o) = \mathtt{HASH}_1(o.c_1)+\mathtt{HASH}_2(o.c_2)+\ldots+\mathtt{HASH}_h(o.c_h)$$

- il risultato può essere riportato nell'intervallo opportuno tramite MOD-H

insiemi

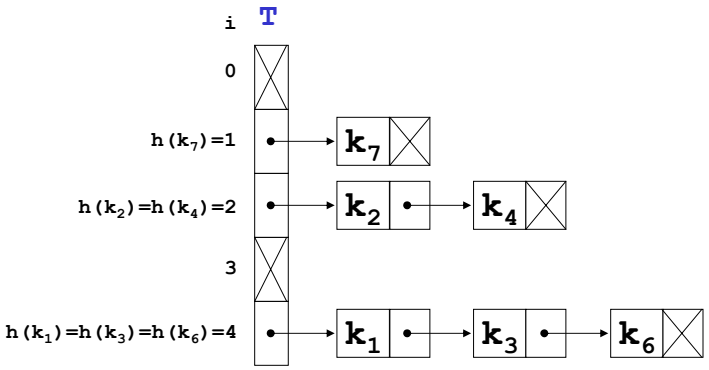
- un insieme è una collezione di elementi omogenei
- esempi di insieme
  - l'insieme degli studenti
  - l'insieme degli oggetti creati da un programma
  - l'insieme delle variabili utilizzate da un programma

il tipo astratto di dato insieme

- domini
  - il dominio di interesse è l'insieme I degli insiemi
  - dominio di supporto: gli elementi E dell'insieme
  - dominio di supporto: i booleani B = {true, false}
- costanti
  - l'insieme vuoto
- operazioni
  - aggiunge un elemento: ADD: $I \times E \rightarrow I$
  - elimina un elemento dall'insieme: REMOVE: $I \times E \rightarrow I$
  - verifica l'appartenenza: CONTAINS: $I \times E \rightarrow B$
  - …

realizzazione di un insieme

- si può usare una hashtable in cui lo stesso elemento funge da valore e da chiave



problemi

1. illustra l'inserimento in una tabella hash di di dimensione m=10 gestita con liste di trabocco delle chiavi 32, 17, 19, 31, 33, 15, 38, 46, utilizzando la funzione hash `MOD-H`
2. scrivi lo pseudocodice delle funzioni `ADD`(I,e), `REMOVE`(I,e) e `CONTAINS`(I,e) dove I è un insieme realizzato tramite una hashtable ed e è un elemento dell'insieme
   - assumi che siano definite opportune funzioni `HASH`(e) e `EQUAL`(e_1,e_2)

