

# Hashing

GRAU A      CURS 2015-2016\*

Departament de Ciències de la Computació  
alg@cs.upc.edu

## Resum

*Aquest projecte té com objectiu una validació experimental de l'efectivitat de diferents algorismes de hash comparant-los també amb l'eficàcia de la cerca binària.*

*La pràctica es farà en grups de 2 o 3 persones (excepcionalment 1, sota autorització expressa). La composició dels grups s'haurà de comunicar a alg@cs.upc.edu abans del 15 de desembre.*

*El lliurament de la pràctica es farà en línia via Racó, teniu temps fins las 23:59 hores del dia 11 de gener de 2016. Alguns grups poden ser convocats per a una entrevista personal (data per decidir la setmana del 18de gener) amb prova interactiva. Alguns grups poden ser convocats a l'entrevista per aclarir dubtes relatius a la seva pràctica. Altres grups poden ser escollits aleatòriament. És obligatori que a l'entrevista estiguin presents tots els membres del grup.*

## I. OBJECTIUS

L'objectiu d'aquesta pràctica és analitzar el cost de cerca dels mots d'un text en un diccionari. Per això us proposem fer la cerca de tres maneres:

- Cerca binària
- Fent servir una algorisme de hash amb taula.
- Fent servir un filtre de Bloom.

L'objectiu és veure experimentalment si hi ha diferència entre els temps dels diferents mètodes. Per centrar-nos en aquest aspecte simplifiquem una mica el context i assumirem que tant el diccionari com el text són un seguit de nombres enters no negatius.

Aquest document és intencionadament vague. Per tant, a més de analitzar diferents versions d'algorismes de hash haureu de documentar el disseny d'experiments per contrastar les vostres hipòtesis.

## II. PROGRAMES

Implementeu un Programa en C++ per a cada mètode. En la versió més senzilla (suficient per aprovar) podeu implementar els algorismes més bàsics vistos a classe.

Versions més sofisticades de la Pràctica (el nivell de sofisticació i esforç dedicat és opcional i es tindrà en compte a l'hora d'avaluar la Pràctica) inclouran la implementació d'altres algorismes amb cost teòric millor que els bàsics.

Teniu en compte que haureu de mesurar el temps dels algorismes. A més, haureu de fer un seguiment de diversos comptadors que reflecteixen la quantitat de treball que el programa fa. Per exemple, per a la cerca binària: nombre total de comparacions de claus necessàries per ordenar les

---

\*La versió més actualitzada d'aquest document, així com qualsevol material adicional relacionat es publicarà al Racó.

dades del diccionari, nombre total de comparacions de claus per a totes les cerques reeixides/fallides, etc.

També haureu de mesurar el cost mitjà (en comparacions i crides a funcions o altres) necessaris per crear i cercar a la taula de hash. Penseu en altres mitjanes útils i documenteu-les. Per exemple, és possible que vulgueu calcular mitjanes separades per l'èxit i el fracàs, o estimar quant més cara és una crida a una funció de hash que una comparació d'enters i utilitzar aquesta estimació per calcular una mitjana ponderada.

### III. DADES

La idea general és que primer creeu un fitxer, arxiu1, amb  $n$  nombres enters seleccionats al atzar. Després, creeu un segon (o un seguit de) fitxer, arxiu2, que contingui com a mínim  $2n$  nombres i una certa proporció de nombres de l'arxiu1. Feu servir arxiu1 com a diccionari i arxiu2 com a text.

Assegureu-vos que en cadascun dels exemples  $n$  és prou gran com per què pugui obtenir bons resultats experimentals. La comparació de dos programes que s'executen en 0,05 segons i 0,07 segons i concloent que el primer és més ràpid no és bona tècnica experimental - no res hauria pogut contribuir a la diferència! Això no vol dir necessàriament que  $n$  tingui que ser el més gran possible. Una  $n$  de moderadament gran amb múltiples assajos en més d'un conjunt de dades pot ser revelador.

Assegureu-vos de mantenir  $n$  petita, mentre esteu provant el programa. Per garantir la reproductibilitat dels experiments haureu de lliurar també els algorismes de generació dels arxius de dades que feu servir.

### IV. QUÈ CAL LLIURAR

Cal lliurar una carpeta que contingui:

- Una breu descripció del que heu implementat, les proves que heu fet i la comparació dels resultats que heu obtingut. També és interessant que indiqueu idees que heu provat encara que no hagin donat bons resultats. El document en format pdf ha d'incloure les referències adients.
- Una carpeta amb totes les fonts necessàries per compilar i executar la vostra pràctica. S'han d'incloure les instruccions per a la compilació i execució, així com per a la generació dels fitxers de dades.