Hashing

Grau A Curs 2017-2018*

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

Resum

Aquest projecte té com a objectiu l'aprenentatge d'algorismes de Hashing, per una part, juntament amb una validació experimental de la seva efectivitat en l'aplicació a un problema de cerca amb diccionari.

El projecte es farà en grups de 3 persones (excepcionalment 2, sota autorització expressa). La ***composició dels grups*** *s'haurà de comunicar a alg@cs.upc.edu abans del* ***6 d'Abril de 2018****.*

El lliurament de la pràctica es farà en línia via Racó, i teniu temps fins las 23:59 hores del dia ***13 de Maig de 2018****. Alguns grups poden ser convocats per a una entrevista personal (data per decidir la setmana del 28 de Maig) amb prova interactiva.* ***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 diferències entre els temps dels diferents mètodes i els pros i contres de cadascun dels mètodes. Per centrar-nos en aquest aspecte simplificarem 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 d'estudiar, analitzar i documentar 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 si es complementa amb una bona experimentació i una documentació entenedora) podeu implementar a més de la cerca binària un dels algorismes de hash amb taula i un filtre de Bloom com els descrits als surveys que s'adjunten a aquesta documentació.

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

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

Tingueu en compte que haureu de mesurar el temps dels algorismes. A més, haureu de fer un seguiment de diversos comptadors que reflecteixin la quantitat de treball que el programa fa. Per exemple, per a la cerca binària serien el nombre total de comparacions de claus necessàries per ordenar les dades del diccionari, el 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 o al filtre de Bloom. 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 a l'atzar. Després, creeu un segon (o un seguit de) fitxer(s), `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 a cadascun dels exemples n és prou gran per tal que 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 hagi de 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, però, de mantenir n petita, mentre esteu provant el programa.

Per tal de 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 documentació adequada del algorismes i mètodes que heu implementat, les proves que heu fet i la comparació dels resultats que heu obtingut. També és interessant que indiqueu altres idees que hagueu provat, encara que no hagin donat bons resultats. El document en format PDF ha d'incloure les referències adients.

- Una carpeta amb tots els programes font necessaris per a compilar i executar la vostra pràctica. S'han d'incloure les instruccions per a la compilació i l'execució, així com per a la generació dels fitxers de dades.

- Tingueu en compte que la documentació entregada ens ha de permetre valorar el nivell d'assoliment de la competència transversal que hem d'avaluar: Capacitat d'autoaprenentatge. En el context del projecte hi han dos aspectes rellevants relacionats amb aquesta competència: els algorismes per crear i consultar taules de hash, i el disseny i anàlisi dels experiments.