Taules de hash i filtres de Bloom

Grau A Curs 2019-2020*

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

Resum

Aquest projecte té com a objectiu l'aprenentatge de diferents implementacions d'estructures de dades basades en 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'ha de comunicar a alg@cs.upc.edu abans del* ***28 de Febrer de 2020****.*

El lliurament del projecte es farà en línia via Racó, i teniu temps fins las 10:59 hores del dia ***23 de Març de 2020****.*

Alguns grups podeu rebre preguntes per e-mail o ser convocats a una entrevista a finals d'Abril. Una setmana abans publicarem la planificació de dates al Racó.

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 amb dos tipus de estructures de dades per diccionaris:

- Taules de hash.
- Filtres de Bloom.

L'objectiu és veure experimentalment si les diferències entre els diferents mètodes i els pros i contres de cadascun d'ells així com la sensibilitat als seus paràmetres de definició. 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. Mes endavant podeu estendre els resultats experimentant amb altres tipus de text.

Aquest document és intencionadament vague. Per tant, a més d'estudiar, analitzar i documentar diferents versions, haureu de documentar el disseny d'experiments per contrastar les vostres hipòtesis i trobar les eines adients per mostrar els resultats obtinguts.

Com a referència inicial us suggerim la secció 13.6 de [1]. Una descripció curta d'alguns dels mètodes la podeu trobar a la Wikipedia (Hash table, Bloom filter). Podeu trobar més informació als fitxers que s'adjunten a aquesta documentació.

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

II. Programes

Implementeu un programa en C++ per a cada mètode. En la versió més senzilla (suficient per aprovar el projecte, si es complementa amb una bona experimentació i una documentació entenedora) heu d'implementar un diccionari amb

1. Taula de hash amb separate chaining
2. Taula de hash amb open addressing
3. Un filtre de Bloom

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 de alguna variació d'aquests mètodes o experimentar amb els mateixos algorismes amb funcions de hash diferents.

Tingueu en compte que haureu de fer un seguiment de diversos comptadors que reflecteixin la quantitat de treball que el programa fa en funció dels paràmetres seleccionats. Per exemple, el nombre total o esperat de col·lisions per taules de hash o de fals positius en el cas del filtre de Bloom, la mida de la taula de hash. Penseu, si es el cas, en altres mesures útils i documenteu-les. Per exemple, és possible que vulgueu calcular mitjanes separades per l'èxit i el fracàs, o estimar altres quantitats.

III. Dades

La idea general és que per experimentar amb les vostres implementacions 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`, etc., com a texts. Això us permetrà determinar els valors experimentals de les mesures de eficiència del mètode implementat en funció dels paràmetres seleccionats, tant per al mètode de hash com del text, la mida del diccionari, per tal de poder comparar-les amb els valors teòrics.

Assegureu-vos que a cadascun dels exemples n és prou gran per tal que pugui obtenir bons resultats experimentals. 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.

En una segona fase executeu els vostres algoritmes sobre altres tipus de diccionari i dissenyeu les cerques adients per poder comparar experimentalment les vostres implementacions amb el comportament teòric.

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 alguns aspectes rellevants relacionats amb aquesta competència: els algorismes per crear i consultar la estructura de dades, les funcions de hash utilitzades, i el disseny i l'anàlisi dels experiments. La qualificació final del projecte reflectirà la qualitat del vostre aprenentatge, de l'experimentació feta i de la documentació on es reflectirà tot. La qualitat del codi entregat (programes) es pressuposa i representarà una part petita de la qualificació final.
- La documentació ha de recollir i presentar la feina feta, les fonts que s'han consultat, el que heu après i els resultats de l'experimentació. En particular és molt important que reflecteixi de forma succinta el que heu après. Si no es compleix aquesta condició, la qualificació final del projecte reflectirà només la qualitat de la presentació.

Referències

[1] J. Kleinberg and E. Tardos. *Algorithm Design.* Pearson & Addison-Wesley, 2006.