Filtres de Bloom i Cryptohashing

Grau A QP Curs 2016-2017*

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

Resum

Aquest projecte té com a objectiu una validació experimental de l'efectivitat de diferents formes de generar filtres de Bloom per a representar un conjunt de claus.

El projecte es farà en grups de 2 o 3 persones. El lliurament del projecte es farà en línia via Racó, teniu temps fins les 23:59 hores del dia ***23 d'abril de 2017****.*

Alguns grups poden ser convocats per a una entrevista personal (data per decidir la setmana del 29 de maig) amb prova interactiva. ***És obligatori que a l'entrevista estiguin presents tots els membres del grup.*** *Alguns grups poden ser convocats a l'entrevista per aclarir dubtes relatius a la seva pràctica o bé rebre qüestions per e-mail. Altres grups poden ser escollits aleatòriament per tal d'explicar el seu treball.*

I. Objectius

L'objectiu d'aquesta pràctica és analitzar la probabilitat de fals positiu en un filtre de Bloom creat per a representar un conjunt de claus. Per aquesta raó us proposem que construïu filtres de Bloom amb diferents algorismes:

- Fent servir k funcions de hash tal i com s'ha explicat a classe.
- Fent servir primer la funció d'encriptació **SHA256** per encriptar les claus, i després creant un filtre de Bloom (com a l'apartat previ) sobre les claus encriptades.

L'objectiu és veure experimentalment si hi han diferències significatives (o no) a la probabilitat de fals positiu entre els diferents mètodes de representació del conjunt de claus. Juntament amb una anàlisi del cost computacional en funció dels valors dels paràmetres i de les decisions d'implementació i els resultats. Per centrar-nos en aquest aspecte, simplificarem una mica el context i assumirem que les claus estan formades per d caràcters alfanumèrics.

Aquest document és intencionadament vague. Per tant, a més de implementar i analitzar diferents versions d'algorismes, considerar diferents funcions de hash i/o encriptacions, haureu de documentar les fonts què heu fet servir en el seu disseny, els paràmetres i propietats rellevants, i el seu cost computacional. Per una altra part cal decidir, i inclou-re, el disseny d'experiments per contrastar les vostres hipòtesis i observacions, així com un estudi comparatiu dels resultats dels experiments.

*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 i/o variant. El nivell de sofisticació i esforç dedicat al projecte és opcional i es tindrà en compte a l'hora d'avaluar-ho. En la versió més senzilla (suficient per aprovar si està acompanyada d'un bon disseny d'experiments) implementeu programes en C++ seguint la versió més simple explicada a classe de teoria. Versions més sofisticades del projecte inclouran la implementació d'algorismes no bàsics o de variants dels algorismes bàsics. Cal que documenteu i justifiqueu les decisions de disseny que preneu.

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, a més de comptabilitzar els paràmetres d'interès. També haureu de mesurar el cost mitjà (en comparacions i crides a funcions o altres) necessaris per crear i cercar al filtre. Penseu (si cal) en altres mitjanes útils (i documenteu-les) i en quins són els paràmetres del vostre disseny per analitzar.

III. Dades

La idea general per un experiment és que primer creeu un fitxer, `claus`, amb n claus seleccionades d'acord amb algun criteri. Després, creeu un seguit de fitxers, `test1` ...`testT`. Feu servir `claus` com a base per a crear el filtre de Bloom i `test1` ...`testT` com a seqüencies de cerques per comptabilitzar els fals positius i altres mesures. Per suposat, una part de la vostra feina és avaluar la sensibilitat del vostre disseny als diferents paràmetres, decidir els criteris d'interès per omplir `claus` i com generar tests per reflectir diferents entorns de cerca. Cal que incloeu un experiment en què els arxius de test incloguin $c\,n$ claus aleatòries i una certa proporció de claus a `claus`.

Assegureu-vos que, a cadascun dels exemples, n és prou gran com per a poder obtenir bons resultats experimentals. Això no vol dir necessàriament que n hagi de ser el més gran possible. Una n 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 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 comprimida. El seu nom ha de començar amb l'identificador assignat al vostre grup. Cal 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 les 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 el vostre projecte. S'han d'incloure les instruccions per a la compilació i execució, així com per a la generació dels fitxers de dades.