MergeSort és un 13% més ràpid que QuickSort

Roser González Valeri, José Antonio Rius Cobo i Erik Carrasco Alastrué.

Novembre 2021

Resum

Objectiu: Comparar l'eficiència temporal de dos algorismes d'ordenació d'ordre O(*nlog(n)*): MergeSort i QuickSort per decidir si canvien el segon, ja implementat, pel primer.

Mètodes: Hem generat 50 vectors de grandàries distribuïdes uniformement entre 0 i 10^6 contenint números enters distribuïts també uniformement entre 0 i deu vegades la grandària del vector. Hem aplicat els dos algorismes d'ordenació a tots els vector amb un ordre aleatori de execució. Com la diferència depenia de la magnitud del vector, però no així el rati, hem transformat les dades amb el seu logaritme natural. Hem fet el CH de igualtat de mitjanes amb la t-Student per dades aparellades.

Resultats: El temps mig (DE[1]) d'execució en ms ha estat de 2777 (1901) pel MergeSort i de 3123 (2039). Rebutgem H_0.

Conclusió: L'algorisme de MergeSort és un 13% ($IC_{95\%}$: de 5 a 20%) més eficient que el de QuickSort. Si altres consideracions (espai, etc.) fossin equivalents, convé substituir el segon pel primer.

Introducció

En l'àmbit de la informàtica, el disseny de programes té una importància cabdal. Aquests han de complir unes determinades característiques: llegibilitat, robustesa, reusabilitat, eficiència etc...

Volem veure si podem fer més ràpida certa aplicació. Ara hi ha implementat l'algoritme Quicksort (Q) d'ordenació de vectors, però volem decidir si convé canviar-ho per Mergesort (M) –en cas de que tingui major eficiència temporal. Per això, intentarem rebutjar la hipòtesi nul·la de igualtat de temps a favor de la alternativa unilateral de que M és més ràpid que Q.

[1] DE: Desviació Estàndar, o Tipus

Material y mètodes

Hem generat al atzar 50 vectors de grandària entre 0 i 10^6 amb números d'una D. Uniforme i els hem ordenat amb els dos algorismes, alternant aleatòriament l'ordre d'execució.

Hem empleat el programari R v.2.11.1 amb un Intel Core 2 de 2.66GHz i 3.46 Gb de memòria RAM; i amb Microsoft Windows XP.

Hem recollit la grandària del vector i el temps d'execució d'ambdós algorismes. Com el seu temps d'execució depèn de la grandària [‘O(*nlog(n)*)’] hem estudiat si era més simple resumir amb una diferència o un rati –que hem obtingut a partir de la diferència dels logaritmes naturals de les variables.

Hem fet la descriptiva amb mitjanes i desviacions estàndard (DE); i la comparació amb la t-Student per a mostres aparellades.

I. Càlcul de la diferència dels temps d'execució

Hem creat la variable (D) com diferència o bé dels temps d'execució (M–Q) o bé dels seus logaritmes.

II. Premisses convenients (sigui pels temps o pel seus logs).

1. Normalitat per la variable diferència.
2. Homocedasticitat. Estudiarem si la diferència és raonablement similar per qualsevol grandària. Si depengués de la longitud de la llista de números a ordenar, estudiarien si el seu rati es pot considerar constant..
3. Obtenció a l’atzar de les grandàries i assignació a l’atzar dels ordres de execució.

III. Càlcul de l'estadístic de la comparació:

Les dades no són independents, per tant s’ha de realitzar un test aparellat.

$T_D = \dfrac{\bar{d}}{s\sqrt{\dfrac{1}{n}}} \sim t_{n-1}$ on *s* és la estimació de la DE de les diferències M–Q i *n* és la grandària mostral.

IV. Construcció de l'interval de confiança per la diferència

$$IC(D, 1-\alpha) = \left[\bar{d} \mp t_{n-1,\ 0.975} \cdot \frac{s}{\sqrt{n}} \right]$$

Resultats

Descriptiva

La taula 1 conté la mitjana i la desviació tipus dels temps emprats pels dos algorismes en els 50 vectors i la seva diferència. El mètode d'ordenació per mescla inverteix en la mostra un temps inferior

Algorisme	Mitjana (DE)
Merge	2777 (1901)
Quick	3123 (2039)
Diferència	346 (1198)

Taula 1: Mitjana i desviació tipus dels temps d'execució del 50 vectors.

En l'annex I es mostren els resultats per les variables sense transformar. S'observa que les diferències no semblen seguir la D. Normal i que són més amples a mesura que s'augmenta la longitud del vector. Fem la transformació logaritme natural dels temps per intentar oferir un únic resultat que apliqui a qualsevol longitud.

En la figura 1 mostra un millor ajustament a la D. Normal i en la figura 2 observem una distribució de la diferències dels logaritmes raonablement homogènia per totes les grandàries.

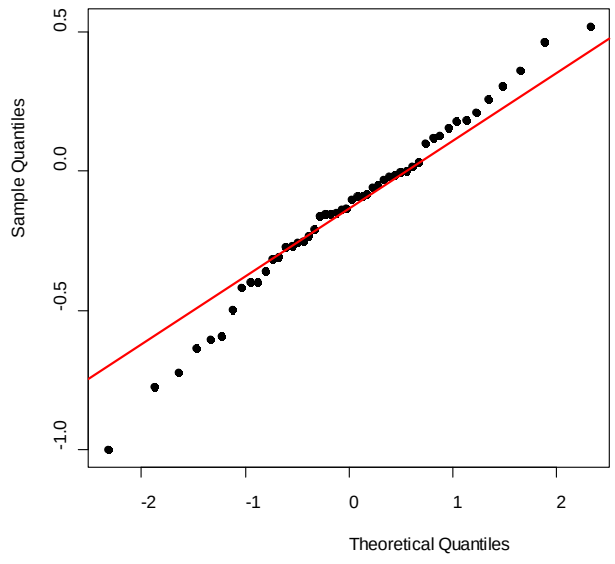


Figura 1. QQ-norm de la diferència dels logaritmes del temps (ln{ms}).

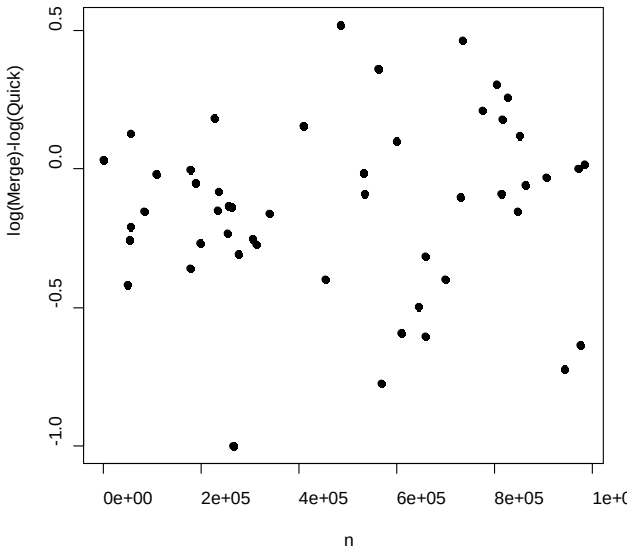


Figura 2: Representació de les diferències dels logaritmes del temps (ln{ms}).) en funció de la mida del vector.

Per comparar la velocitat d'ambdós algorismes d'ordenació fem un contrast de hipòtesi d'igualtat de mitjanes amb dades aparellades. La sortida proporcionada per R és la següent[2]:

```
        Paired t-test

data:  LogMerge and LogQuick
to 0
95 percent confidence interval:
 -0.22774047 -0.04882862
sample estimates:
mean of the differences
              -0.1382845
```

Els graus de llibertat son 49, que corresponen al nombre d'observacions menys un. La mitjana del logaritme del temps de M esta 0.14 log{ms} per sota de Q amb un $IC_{95\%}$ de [0.05 a 0.23]. Si desfem els logaritmes:

$$\log(Merge) - \log(Quick) = -0.1382845 \Rightarrow \log\left(\frac{Merge}{Quick}\right) = -0.1382845 \Rightarrow \frac{Merge}{Quick} = e^{-0.1382845} = 0.8708509$$

$$Eficiencia\ relativa = \left(1 - \frac{Merge}{Quick}\right) \cdot 100 = 12.91491\%$$

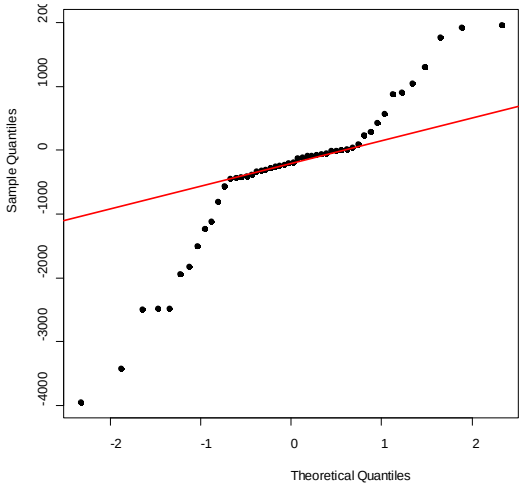
equival a dir que el algorisme M és un 12.9% ($IC_{95\%}$: de 4.8 a 20.4%) més eficient que Q.

Discussió

Sense tenir en compte altre consideracions (espai, per exemple), aquest resultats aconsellen substituir Q per M en el nostre sistema. Els resultats observats ens han aconsellat canviar la variable principal de l'estudi per afavorir la seva interpretació i assolir les premisses, però cal dir que l'anàlisi amb les dades no transformades també rebutja la hipòtesi de igualtat.

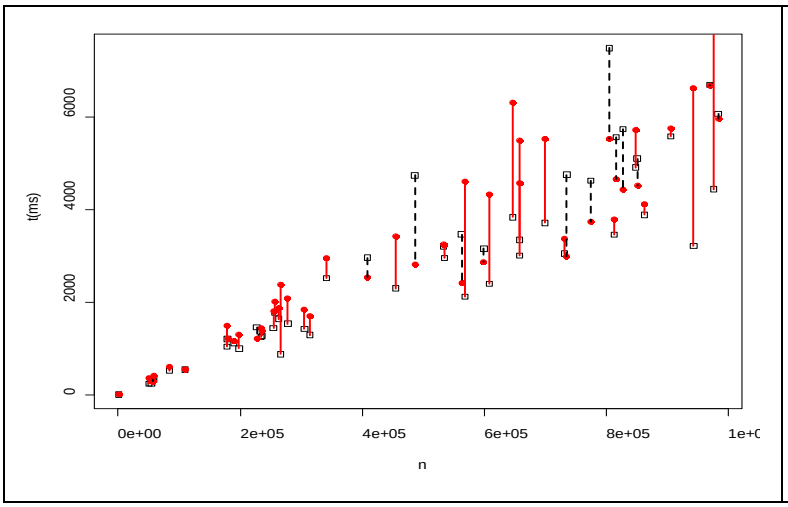
Encara que hem obtingut aquests resultats amb un ordinador concret (microprocessador Intel Core 2de 2.66GHz i 3.46 Gb de memòria RAM), creiem que poden aplicar-los a qualsevol tipus. En canvi, el sistema operatiu emprat (Microsoft Windows XP) podria condicionar els temps obtinguts i aconsellem repetir l'estudi en altres SO abans de estendre aquesta conclusió. Els nostres resultats estan comprovats amb grandàries de vectors per sota de 10^6. Si es preveuen grandàries superiors convindria repetir l'estudi.

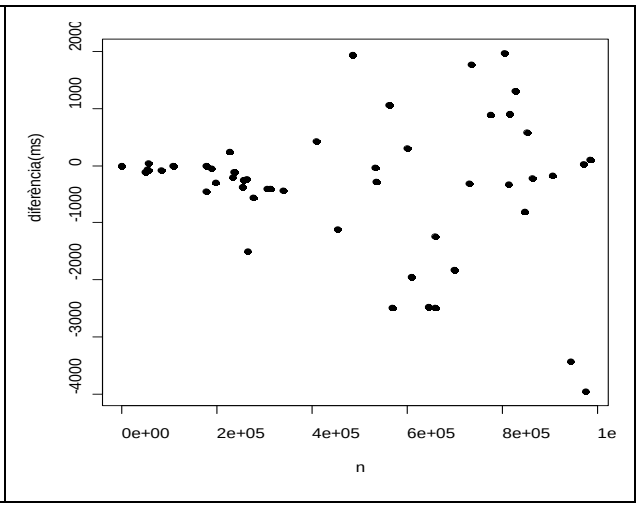
ANNEX I. Test aparellat de les diferències sense transformar



El QQ-plot adjunt mostra un desviament important de la Normalitat.

La següent figura mostra que els temps d'execució dels dos programes i la seva diferència augmenten segons les grandàries mostrals dels vectors a ordenar. La ultima figura ressalta que les diferències augmenten amb la grandària. Aquesta situació té dues implicacions pràctiques: (1) no té sentit estimar un valor comú per totes les grandaries; i (2) no es pot aplicar el test per comparació de mitjanes amb dades aparellades ja que totes les observacions no tenen la mateixa variància i l'estadístic no s'ajustaria a la distribució de la t-Student.






```
        Paired t-test

data:  dades$Merge and dades$Quick
95 percent confidence interval:
 -686.489824   -5.819776
sample estimates:
mean of the differences
              -346.1548
```

ANNEX II. Script en R

```
### Lectura de les dades
dades <- read.table('SortTimes.txt',header=TRUE,sep='\t')

### Variable diferència
dif <-  dades$Merge-dades$Quick

### Descriptiva

apply(dades[,2:3],2,mean)
apply(dades[,2:3],2,sd)

mean(dif)
sd(dif)

### Normalitat
qqnorm(dif,pch=19)
qqline(dif,lwd=2,col=2)

### boxplot
boxplot(dades[,2:3])

### Gràfic diferències
plot(dades$n,dades$Merge,pch=22,cex=1,xlab="n",ylab="t(ms)")
points(dades$n,dades$Quick,pch=19,col=2,cex=1)
for (i in 1:n){
     if(dades$Merge[i]>dades$Quick[i]){co <- 1 ; lt <- 2}
     else{co <- 2 ; lt <- 1}

segments(dades$n[i],dades$Merge[i],dades$n[i],dades$Quick[i],col=co,lty=lt,lwd=2
)
}

# Heterocedasticitat
plot(dades$n,dif,xlab="n",ylab="diferència(ms)",pch=19)

# boxplot diferencies
boxplot(dif)

### Test aparellat pels temps sense transformar
t.test(dades$Merge,dades$Quick,paired=TRUE)


# Treure logaritmes
LogMerge <- log(dades$Merge)
LogQuick <- log(dades$Quick)
dif <-  LogMerge-LogQuick

plot(dades$n,LogMerge,pch=22)
points(dades$n,LogQuick,pch=19,col=2)

qqnorm(dif,pch=19)
qqline(dif,lwd=2,col=2)

plot(dades$n,dif,pch=19,xlab="n",ylab="log(Merge)-log(Quick)")
boxplot(dif)

### Test aparellat pels temps transformats
t.test(LogMerge,LogQuick,paired=TRUE)
```

La velocitat real de telefònica sembla ser més baixa.

Olaguer Mas Maragall, Encarna Rajoy Zapatero i Olivier Sarcozy Miterrand.

Maig 2021

Resum

Objectiu: estimar per interval la diferència entre el temps de baixada real de Telefònica i la seva competència.

Mètodes: 41 voluntaris han proporcionat els seus proveïdors i els temps de baixada d'acord amb un mateix procediment.

Resultats: Amb una confiança del 95%, la diferència del temps de baixada es troba entre 14.1 i 3510.7 a favor de la competència de telefònica.

Discussió: Els nostres resultats son favorables a la competència, però els voluntaris estudiats no eren una mostra aleatòria i podrien ser els més molestes amb el principal proveïdor. A més, les premisses de Normalitat i de igualtat de variàncies no son clares. Abans de treure una conclusió definitiva, haurien de repetir el estudi en una mostra aleatòria i, possiblement, tenint en compte els temps contractats.

Introducció

La diferència entre la velocitat contractada i la velocitat real en les connexions a Internet és una de les queixes més freqüents entre els usuaris de les línies ADSL. A més hi ha una gran competència entre els diferents proveïdors per proporcionar un bon servei a un preu assequible. En el cas de l'operador de Telefònica, la OCDE destaca que "*El operador dominant ha estat capaç de mantenir una elevada quota de mercat en els serveis d'ADSL, però la diferència entre els seus preus i els dels competidors per l'accés superràpid a Internet segueix sent inusualment elevada*"

El que no es sap és si aquesta diferència de preu està justificada per la obtenció d'un millor servei. En aquest cas, ens centrarem en avaluar la diferència en la velocitat de descàrrega dels usuaris de Telefònica i els altres operadors.

Objectiu

L'objectiu d'aquest estudi és comparar la velocitat de baixada real d'aquells que tenen com a proveïdor Telefònica respecte aquells que tenen un altre proveïdor i obtenir una estimació d'aquesta diferència amb la seva incertesa.

Material y mètodes

Varen proposar als 164 estudiants de Estadística d'Informàtica (FIB) recollir informació sobre la seva connexió d'ADSL. D'aquests, 41 (25%) van fer una única prova de connexió durant els mesos de març i abril de 2010. No hem inclòs les mesures realitzades des de mòdems amb velocitats de 56 kb/s o inferiors, des de dispositius d'Internet mòbils o de des xarxes locals. Varen recollit les variables sobre el proveïdor i les velocitats de baixada de dades: la contractada amb el proveïdor del servei; i la real, mesurada amb el test de la plana web http://www.internautas.org/testvelocidad/.

Hem descrit les variables numèriques amb la mitjana i la desviació estàndard (DE) i la de les variables categòriques amb els percentatges. Hem comparat els dos grups amb la t-Student per a mostres independents segons el següent procediment.

I. Premisses

4. Normalitat per cadascun dels grups de comparació.
5. Igualtat de variàncies entre ambdós grups.
6. Mostra aleatòria.

II. Estadístic i distribució de referència:

$$T = \frac{\bar{x}_T - \bar{x}_O}{s_p\sqrt{\frac{1}{n_T}+\frac{1}{n_O}}} \sim t_{n_T+n_O-2}$$

on s_p és la estimació de desviació comuna a ambdues mostres.

III. Construcció de l'interval de confiança per la diferència

$$IC(\mu_T-\mu_O,1-\alpha)=\left[\bar{x}_T-\bar{x}_O \mp t_{n_T+n_O-2,\ 0.975}\cdot s_p\sqrt{\frac{1}{n_T}+\frac{1}{n_O}}\right]$$

Resultats

Descriptiva

La taula 1 mostra que les velocitats reals són inferiors a les contractades, i que les dues son superiors per la competència.

	Global		Operador			
			Telefònica		Altres	
	n	Mitjana (DE)	n	Mitjana (DE)	n	Mitjana (DE)
Velocitat de baixada real	41	4198.5(2677.3)	26	3563.4(1837.6)	14	5325.8(3653.0)
Velocitat de baixada contractada	39	6406.6(5004.0)	26	5238.2(2958.3)	12	8618.7(7530.6)

Taula 1: Descriptiva global i desglossada per proveïdor. Les velocitats estan expressades en Kb/s.

		Global
		n (%)
Proveïdor	Jazztel	3(7.5%)
	Ono	5(12.5%)
	Orange	2(5.0%)
	Telefónica	26(65.0%)
	Vodafone	1(2.5%)
	Ya.com	3(7.5%)

La taula 2 mostra que la companyia amb més usuaris és Telefònica

.

Taula 2: Descriptiva del operador.

La figura 1 mostra que els usuaris de Telefònica tenen una velocitat de baixada real menor que la resta. El fet que la mediana estigui situada en la part inferior de la capsa i que la longitud dels bigotis superiors sigui més gran indica una asimetria amb cues a la dreta que posen en dubte la premissa de Normalitat de les dades. També mostra una lleugera major dispersió entre les velocitats dels participants que tenen altres proveïdors.

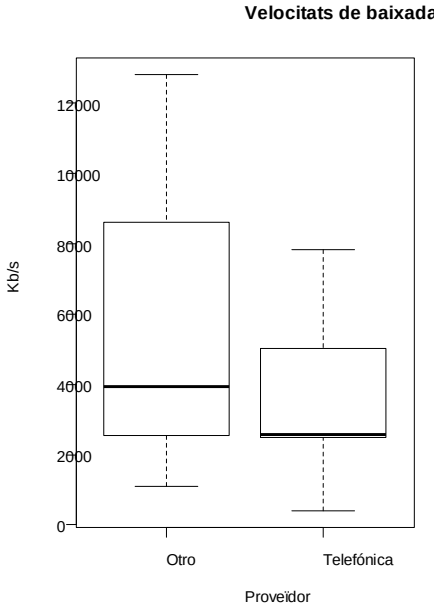


Figura 3: Boxplot de les velocitats de baixada segons el proveïdor

La figura 2 no recolza la premissa de Normalitat.

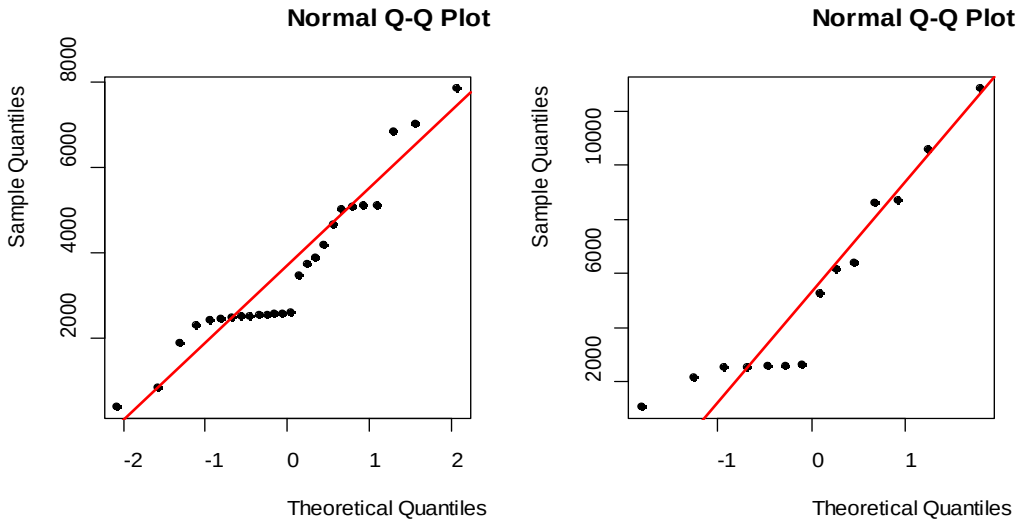


Figura 4: QQ- Norm de la variable principal temps de baixada real amb l'operador de Telefònica (esquerra) i amb un altre operador (dreta)

Els resultat del test en R és:

```
        Two Sample t-test

data:  adsl$down.speed by adsl$proveTel
95 percent confidence interval:
   14.13955 3510.66265
sample estimates:
      mean in group Otro mean in group Telefónica
                5325.786                  3563.385
```

Els graus de llibertat son 38, que són el resultat de restar 2 a totes les observacions vàlides (hi ha una dada mancant). L'estimació de la diferència és 1762.4 [$IC_{95\%}$ de 14.1 a 3510.7] a favor de la competència.

Discussió

(Conclusió principal:) Aquests resultats recolzen als competidors de Telefònica, (Limitacions:) però no són una prova definitiva al no tractar-se ni d'un disseny d'experiments, ni de mostres aleatòries. Per altra banda, les dades provenen de un 25% de voluntaris, el que provoca una incertesa addicional que no està contemplada en les mesures estadístiques de l'error aleatori. (Extrapolació:) A més a més, podria ser que els estudiants de informàtica no representin a poblacions més generals. Finalment, les premisses de Normalitat i igualtat de variàncies tampoc son clares del tot. (Feina futura:) Els fets de que les desviacions siguin més grans quan més gran és la mitjana (taula 1) i de que les velocitats contractades semblin també diferents aconsella estudiar si una transformació logarítmica del temps de baixada o el seu quocient amb el temps contractat porten a les mateixes conclusions.