BibTeX Bibliography Index Maker: Meeting Notes

Ramon Xuriguera

24-02-2010

1 Multithreading

He estudiat la possibilitat d'utilitzar diferents threads i sembla una opció que pot ajudar a millorar els temps d'execució pels casos en que el número de fitxers pels quals hem d'obtenir referències és gran. Comparació del temps per obtenir múltiples pàgines aleatòries d'Internet de forma seqüencial i paral·lela:

2 pàgines		5 pàgines		10 pàgines		20 pàgines	
Seq.	5 Threads	Seq.	5 Threads	Seq.	5 Threads	Seq.	5 Threads
0.9010	0.5481	2.1830	0.6612	4.3153	1.5914	7.9295	2.5949
0.7467	0.3795	2.1558	0.7441	4.3186	1.2311	8.5483	2.1958
0.7678	0.5641	2.0645	0.5383	9.2930	1.4415	8.7202	2.5749
0.7421	0.3876	2.0684	0.8551	4.9859	1.5294	8.4732	2.2841
0.9674	0.5477	2.1510	0.8550	5.3600	1.3116	9.2901	2.2257
Mitjana:							
0.8250	0.4854	2.1246	0.7307	5.6546	1.4210	8.5923	2.3751
Guany:							
-0.34s		**-1.39s**		**-4.23s**		**-6.22s**	

Els temps de la taula tenen en compte el temps necessari per crear i finalitzar els diferents fils d'execució. Els resultats s'han obtingut al fer consultes aleatòries a Google. El temps necessari per obtenir altres pàgines seran més alts. Hem utilitzat consultes aleatòries per evitar l'efecte dels proxys i cachés.

Algunes idees:

- Utilitzar múltiples fils o no depenent del número de fitxers pels quals hem d'obtenir la referència
- Tenir un pool de fils d'execució que consumeixen fitxers d'una cua d'entrada i emmagatzemen els resultats en una cua de sortida. Els fils es reutilitzen.
- Ajustar el número màxim de fils segons la velocitat de la xarxa

2 SQLite

Segons la documentació és perfectament adequat per aplicacions que tenen un número inferior a 100K hits diaris, un número molt superior del que nosaltres necessitem.

3 Wrapper induction

Una idea:

1. A l'obtenir una referència emmagatzemar les dades a BD
   Dades que podríem guardar: ruta del pdf al sistema, url a partir de la qual hem obtingut la referència, camps principals
2. Per generar un wrapper, identifiquem on es troba cada peça d'informació dins de la pàgina html
   Podem comprovar què falla abans de tornar a generar tot el wrapper.
3. Generem un fitxer de configuració del wrapper
4. Configuration-driven wrappers:
   Generalitzem el codi dels wrappers per tal que tinguin un comportament diferent per configuracions diferents.

A considerar:

- La solució només funciona per *field wrappers*
- Si una pàgina passa a canviar el format de les seves urls (poc habitual), no es podrà regenerar el wrapper:
  Podem tornar a cercar-la amb google i actualitzar la url.