BibTeX Bibliography Index Maker: Meeting Notes

Ramon Xuriguera

16-03-2010

1 Comparació entre motors de cerca

Per tal d'evitar que Google ens bloquegi he estat provant altres motors de cerca. A continuació es mostren els resultats obtinguts:
La següent taula mostra els resultats obtinguts al processar un directori amb 47 fitxers PDF. D'aquests 47, n'hi ha 5 dels quals no es pot extreure el contingut. Els percentatges estan donats en funció d'aquests $47 - 5 = 42$ fitxers:

	Results		Extracted		Valid		Invalid	
	Total	%	Total	%	Total	%	Total	%
Google								
{6,10} skip 3	35	83.33%	33	78.57%	23	54.76%	10	23.81%
{8,12} skip 3	34	80.95%	30	71.43%	21	50.00%	9	21.43%
{10,15} skip 0	35	83.33%	34	80.95%	23	54.76%	11	26.19%
Scholar								
{6,10} skip 3	33	78.57%	28	66.67%	20	47.62%	8	19.05%
{8,12} skip 3	35	83.33%	31	73.81%	23	54.76%	8	19.05%
{10,15} skip 0	34	80.95%	28	66.67%	21	50.00%	7	16.67%
Bing								
{6,10} skip 3	35	83.33%	34	**80.95%**	16	38.10%	18	42.86%
{8,12} skip 3	36	**85.71%**	28	66.67%	16	38.10%	12	28.57%
{10,15} skip 0	35	83.33%	30	71.43%	17	40.48%	13	30.95%
Yahoo								
{6,10} skip 3	28	66.67%	25	59.52%	14	33.33%	11	26.19%
{8,12} skip 3	24	57.14%	23	54.76%	10	23.81%	13	30.95%
{10,15} skip 0	29	69.05%	25	59.52%	14	33.33%	11	26.19%
Max		**85.71%**		80.95%		54.76%		16.67%

La resta de paràmetres per fer la cerca s'ha deixat fixa: `max_queries_to_try = 5` i `too_many_results = 25`. Modificant aquests paràmetres també s'obtenen variacions, però les proves mostren que 5 i 25 permeten obtenir prou bons resultats.

El temps d'execució mig per processar aquests 47 fitxers és d'uns 40 segons.

Un inconvenient d'utilitzar altres motors de cerca enlloc d'*Scholar* són:

- No hi han indexades les pàgines d'SpringerLink.

- En alguns casos (e.g. Bing), els resultats de portal.acm sobre els autors tenen prioritat per sobre le pàgines de le referències. (Moltes vegades només surt la de l'autor)

1.1 Es poden millorar aquests resultats?

Sí. Els principals problemes:

- Extracció de text del PDF: Es pot provar d'utilitzar el nom del fitxer o bé les metadades per fer la cerca.
- No es troben resultats a Internet: La solució de reduir la mida de les consultes ja queda contemplada amb els límits inferior i superior de paraules de la consulta. Es pot mirar de tornar a provar totes les *queries* a altres cercadors.
- Extraccció de la informació: només cal afegir més *wrappers* i ordenar millor els resultats.

Un altre problema que s'ha presentat és que dels 42 fitxers dels que es pot extreure el contingut del PDF, n'hi ha 3 pels quals el contingut s'extreu, però no té sentit. Per exemple, tenim les següents consultes:

```
Query: "rgbGphihgfBc q ed H a CFQ SV S P 6 H C8"
Query: "BA I8GF6D3 C  2 (c) A 3  (c) 5"
Query: "h c P u X hT Qy hT XH h H"
```

Caldrà comptar amb una reducció del percentatge d'extraccions deguda a aquest fenòmen.

El factor més influent per comparar els resultats dels diferents cercadors és el número de fitxers pels quals s'han obtingut resultats. El número de referències (vàlides i no vàlides) també és important, però depèn molt dels *wrappers* dels que es disposa.

1.2 BlackList

Algunes pàgines com ara Microsoft Academic Search no arriben a oferir la informació obligatòria per construir la referència BibTeX. En aquests casos, es pot afegir l'adreça arrel de la pàgina al fitxer de configuració per tal que s'ometin.

2 Base de Dades

Finalment s'utilitza SQLite amb SQLAlchemy, un toolkit SQL i ORM per a Python que ofereix un nivell més d'abstracció per poder treballar amb bases de dades diferents sense haver de canviar el codi.

El mòdul corresponent a la base de dades necessita una mica de refactoring.

3 Bloqueig ACM

Moltes consultes provoquen que portal.acm bloqui la ip i retorni l'error 403 (Forbidden).

4 Wrapper induction

Després de llegir uns quants articles sembla que extreure informació estructurada no hauria de ser molt complicat.

Tenir com a mínim dos exemples de la pàgina etiquetats. Per nosaltres, les etiquetes seran aquells trossos d'informació que ens interessa extreure. Utilitzarem un dels exemples per generar el wrapper i en comprovarem la correctesa amb la resta. Tindrem en compte:

- Posició relativa dins del document:
  Calcularem la posició *bottom-up*, des de les fulles i fins on sigui necessari. Si no necessitem arribar a l'arrel per identificar l'element, no ho farem.

  Haurem de tenir en compte, també, l'ordre de l'element respecte els seus *germans*.

  Utilitzarem tant etiquetes com atributs. Els principals atributs que mirarem seran `id` i `class` ja que no variaran entre document i document del mateix lloc web i són molt comuns ja que avui en dia gairebé tots els llocs web utilitzen CSS.

- Posició relativa dins de l'element:
  De la mateixa manera que WHISK (Stephen Soderland 1999), si el contingut de l'element no només conté la peça d'informació que estem buscant, l'haurem d'extreure utilitzant expresssions regulars.
  Podem crear aquestes expressions regulars mirant la resta de contingut de l'etiqueta. Es pot comparar amb la resta d'exemples disponibles per decidir quina és la informació rellevant.

5 Informe del projecte

No més tard de tres mesos (dos mesos) abans de la defensa del projecte, l'estudiant ha de presentar un informe del projecte als membres del tribunal, tal i com s'especifica a la normativa. Aquest informe ha de portar el vist-i-plau del director/ponent i és molt convenient que l'estudiant el presenti personalment i en mà als membres del tribunal.