BibTeX Bibliography Index Maker: Meeting Notes

Ramon Xuriguera

21-04-2010

1 Generació de regles

Tal i com vam comentar a la sessió anterior, el principal problema de la generació de regles és la informació repetida dins de la mateixa pàgina: com més vegades apareix la informació que busquem dins de la pàgina, més probable és agafar una de les etiquetes que no ens convenen. Per solucionar-ho, **generem múltiples wrappers per cadascun dels camps de la referència** i considerem que un *wrapper* només s'encarrega d'extreure la informació corresponent a un únic camp.

1.1 Rating

Per decidir quins *wrappers* són els que realment funcionen, els ordenem segons un sistema de valoració que depèn dels encerts i errors. La idea és molt similar a la que s'utilitza en molts llocs web per ordenar comemantaris o recomanacions de productes:

- Per cada *wrapper* tindrem dos comptadors: **vots positius i vots negatius**.
- Cada vegada que utilitzem un *wrapper* per extreure informació, donem un vot positiu o negatiu segons si l'extracció ha estat vàlida o no.
- Això farà que la llista de *wrappers* s'ordeni automàticament. Aquells que funcionen millor sempre es provaran primer.

Aquest sistema també ens permet veure quan cal generar nous *wrappers*: només cal donar un cop d'ull a la llista per veure si encara en queda algun amb una valoració positiva.

Possible millora: Si un *wrapper* no funciona una vegada, és probable que deixi de fer-ho d'aquí en endavant. Per aquest motiu, potser seria bo accelerar el procés fent que els vots negatius tinguin més pes que els positius. Per exemple: donant més d'un vot negatiu per cada referència que s'ha extret malament.

Problemes: Els camps pels quals no es pot comprovar la validesa com ara el número de pàgines. Potser es pot votar positiu/negatiu segons la validesa de la resta de camps de la referència.

1.2 Algorismes per valorar *wrappers*

- Resta: $score = positius - negatius$

- Mitjana: $score = \frac{positius}{votstotals}$
  Els percentatges ens poden fer triar un *wrapper* que no és el que volem. Això pot passar sobretot quan hi ha casos amb pocs vots.
- Wilson score interval: és robust per a casos on tenim pocs vots.
- Bayesian Average

2 Wrappers multi valor (e.g. autors)

Pel que hem pogut comprovar, en la majoria de pàgines, els camps de més d'un valor estan estructurats com:

- Elements germans:

```
<ul>
<li>Element 01</li>
<li>Element 02</li>
...
</ul>
```

- Dins del mateix element i amb algun separador:

```
Autors: Sisco Solsona, Jordi Carles, Miquel Rius
```

Primera idea: utilitzar una expressió regular que permeti qualsevol ordre dels autors dins del document. Problema: el parser HTML que fem servir (BeautifulSoup), ens complica la feina en aquells casos en que els autors es troben en elements HTML diferents.

Una solució més naïve, seria abordar els dos casos que hem vist de forma diferent:

- Elements germans: Agafar el text de tots els germans
- Mateix element: Construir una expressió regular que detecti els separadors entre el diferents valors. Agafar tots els *matches*.

3 Índex de la memòria

1. Introducció
   - Descripció
   - Treball existent
2. Definició del projecte
   - Context
   - El format BibTeX
   - Característiques

- Planificació temporal

3. Extracció de continguts PDF
    - Dificultats
    - Software existent: Per què utilitzem xPDF
4. Cerca a la web
    - Primera idea: Google Scholar. Problemes
    - Resta de cercadors: Google, Bing i Yahoo
    - Ajustaments (Paràmetres de la cerca, llista negra , etc.)
    - *Multithreading*
5. Extracció d'Informació
    - *Wrappers* a mà
    - Inducció de *Wrappers*
        - Generació automàtica de regles:
          Rutes i expressions regulars
        - Valoració dels *wrappers*
        - Reaprenentatge
6. Anàlisi de resultats
    - Proves realitzades i resultats obtinguts
7. Conclusions i Treball Futur
    - Objectius assolits
    - Possibles millores
8. Annexos
    - Resultats dels tests
        - Comparació resultats *Multithreading*
        - Comparació dels cercadors
        - etc.
    - Biblioteques utilitzades
    - Diagrames ?
    - Llicència ?

4 Tasques pendents

Llistat de tasques pendents a realitzar:

- Netejar, encara més, l'HTML abans de fer l'extracció: treure comentaris i etiquetes `script`, `style`, etc.
- Valoració dels *wrappers*
- *Wrappers* per a camps multi-valor com ara els autors
- Comprovació de l'estat dels wrappers actuals. (utilitzant els *ratings* dels *wrappers*)
- Interfície