BibTeX Bibliography Index Maker: Meeting Notes

Ramon Xuriguera

14-04-2010

1 Generació de regles

El principal problema que he trobat un cop implementada la generació de regles 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 exemple, en el cas de la pàgina *Science Direct*, el títol de l'article apareix en el títol de la pàgina, juntament amb troços d'informació que van canviant. Això fa que la regla resultant no sigui tant bona com la que es podria haver obtingut fent servir una de les altres aparicions del títol dins la pàgina:

```
Path: [[[u'title', {}, 5]]]
Regex: u'ScienceDirect\ \-\ (?:.*)e\ (?:.*)\ \:\ (.*)(?:.*)
    \ '
```

Enlloc de:

```
Path: [[[u'div', {u'id':'articleTitle'}, 0]]]
Regex: u'(.*)'
```

Aquest problema afecta molt en el cas dels anys, sobretot en articles nous on l'any també pot aparéixer al copyright de la pàgina, etc.

2 Emmagatzemar wrappers a la base de dades

Fet

3 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.
- Afegir els camps especials com ara autors a la generació de wrappers.
- Comprovació de l'estat dels wrappers actuals. (Comparar timestamps)
- Interfície