Títol: **BibTEX Bibliography Index Maker**

Volum: **1/1**
Alumne: **Ramon Xuriguera Albareda**

Director/Ponent: **Marta Arias**
Departament: **LSI**
Data: **Primavera 2010**

DADES DEL PROJECTE

Títol del Projecte:

Nom de l'estudiant: Ramon Xuriguera Albareda

Titulació: Enginyeria Informàtica

Crèdits: 37,5

Director/Ponent: Marta Arias

Departament: LSI

MEMBRES DEL TRIBUNAL *(nom i signatura)*

President:

Vocal:

Secretari:

QUALIFICACIÓ

Qualificació numèrica:

Qualificació descriptiva:

Data:

Índex

Capítol 1

Introducció

1.1 Descripció

BibTEX Bibliography Index Maker és una eina d'ajuda a la creació d'índexs bibliogràfics pensada com un complement a aplicacions de maneig de referències ja existents com poden ser *JabRef*[1] o *Mendeley*[2].

La principal funcionalitat consisteix en escanejar un directori que conté articles científics en PDF i generar un índex bibliogràfic en BibTEX amb les referències d'aquests fitxers. Aquest índex es pot importar des de les aplicacions esmentades o bé pot ser referenciat directament des d'un nou document TEX.

1.2 Treball Existent

Actualment existeixen nombroses aplicacions dedicades al maneig de referències. Algunes d'elles utilitzen les meta-dades dels fitxers per tal de trobar informació com ara el títol o l'autor, però cap de les eines que hem trobat aprofita el contingut dels documents per generar la referència.

Empreses com ara *Google* o *Microsoft* agafen la informació de documents PDF per oferir serveis com ara *Scholar* o *Academic Search*, però no ofereixen el codi font i per tant, no sabem com funcionen. Per una altra banda, *CiteSeer* és un projecte *open source* de característiques similars, però que també té limitacions. El sistema funciona analitzant la bibliografia dels articles, informació que acostuma a ser bastant estructurada, però té problemes per obtenir els camps de la capçalera del propi fitxer, que és el que ens interessa.

[1] http://jabref.sourceforge.net
[2] http://www.mendeley.com

Capítol 2

Definició del Projecte

2.1 Context

2.2 BibTEX

Per poder entendre el context del projecte cal que descrivim l'eina de maneig de referències BibTEXi la sintaxi del llenguatge que utilitza. En el nostre cas farem servir aquest llenguatge com a format de sortida al generar els índexos bibliogràfics. Al llistat 2.1 es mostra un exemple d'una referència d'un article científic expressat en el format BibTEX:

```
@article{MoSh:27,
  title = {Size direction games over the real line},
  author = {Moran, Gadi and Shelah, M., Saharon},
  journal = {Israel Journal of Mathematics},
  pages = {442--449},
  volume = {14},
  year = {1973},
}
```

Llistat 2.1: Referència expressada en BibTEX

Alguns aspectes a comentar sobre l'exemple anterior:

- La primera línia conté el tipus de document i un identificador. El primer defineix els camps obligatoris que s'han d'especificar, i el segon ens permetrà citar a la referència des d'un document. En el nostre cas només ens interessen les referències de tipus *article* i haurem de definir, com a mínim, els camps: *author*, *title*, *journal* i *year*.

- Es considera que el nom d'un autor o editor pot constar de quatre parts diferents: *First*, *von*, *Last*, *Jr.*. Es poden ordenar de diverses maneres, però nosaltres ho farem amb `<von> <last>, <middle>, <first>`. Cal separar múltiples noms amb la paraula `and`.

- L'últim camp d'una referència pot acabar o no amb una coma.

2.3 Característiques

2.4 Planificació Temporal

Capítol 3

Disseny del sistema

3.1 Mòduls

Hem organitzat el codi del sistema en els següents mòduls:

- *Raw Content Extraction* (rce): Agrupa totes les classes encarregades d'extreure el contingut dels documents PDF.
- *Information Retrieval* (ir): Encarregat de comunicar-se amb els diferents cercadors disponibles a Internet per obtenir pàgines que contenen informació de la referència que volem extreure.
- *Information Extraction* (ie): Conté tot el codi que permet obtenir la referència a partir d'una pàgina HTML. A més, també és l'encarregat de generar nous *wrappers*.
- *References*: Per una banda fa un anàlisis sintàctic de les referències extretes per poder-les validar. Per l'altra, transforma a BibTeXles referències extretes.
- Base de dades (db): Tal i com indica el seu nom, duu a terme els accessos la base de dades.
- *Main*: Enllaça tots els mòduls anteriors i proporciona punts d'entrada a la interfície d'usuari. Fa de façana del sistema.
- *Graphical User Interface* (gui): Interfície d'usuari més o menys amigable.

La figura 3.1 mostra com interaccionen entre ells.

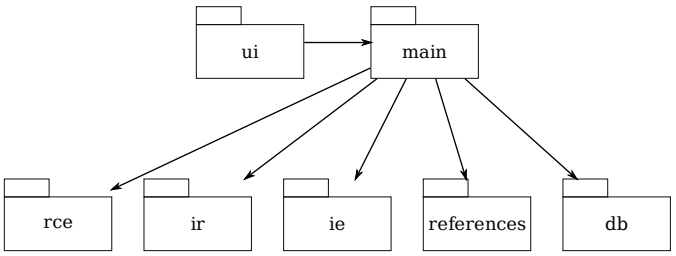

Figura 3.1: Mòduls del sistema

Capítol 4

Extracció dels continguts d'un PDF

Un dels aspectes que han influït més en l'enfocament que hem donat al sistema ha estat la dificultat d'extreure el text dels documents PDF. La primera idea a l'hora d'abordar el nostre projecte va ser intentar extreure informació directament dels fitxers PDF dels quals es disposa.

4.1 Dificultats

Les principals dificultats que es troben a l'hora d'obtenir el text d'un fitxer PDF són:

- Caràcters especials: com Unicode o lligadures
- Flux del text dins del fitxer

4.2 Programari existent

Tot hi haver-hi diverses utilitats que permeten l'extracció del contingut d'un fitxer PDF en forma de text pla o HTML, totes presenten problemes similars en els punts comentats a la secció anterior.
A l'apèndix A hi ha exemples de com queden els continguts de diferents documents PDF després d'extreure'ls.
xPDF proporciona eines executables des de la línia de comandes per extreure text i altres elements dels fitxers PDF. Es distribueixen binaris de la utilitat tant per Windows com per Linux (que també funcionen per MAC OS). El principal motiu pel qual hem escollit xPDF és la qualitat dels resultats, en especial, el fet que no separa els paràgrams en diferents línies i que obté el text segons l'ordre de lectura i no l'ordre en que es troben en el document (e.g. dues columnes). També serà útil la possibilitat d'extreure les metadades del fitxer de forma fàcil.
Altres opcions que s'han tingut en compte:

- PyPDF
- PDFMiner
- PDFBox

Capítol 5

Cerca de referències a Internet

5.1 Cercadors

La primera idea per cercar pàgines que contenen referències d'artícles va ser utilitzar *Google Scholar*. La falta d'APIs i el bloqueig periòdic de les cerques automàtiques van fer Hem preparat el nostre cercador per tal d'utilitzar les APIs dels cercadors *Google*, *Yahoo* i *Bing* i hem

El principal avantatge és la

Un inconvenient, hi ha biblioteques virtuals que no estan indexades en aquests serveis.

5.2 Ajustaments

Podem ajustar la manera com es fan les cerques a partir de certs paràmetres que es detallen a continuació.

En moltes ocasions, el cercador *Bing* mostra resultats corresponents a *Microsoft Academic Search* (un projecte molt similar a *Google Scholar*). Aquestes pàgines, però, no mostren prou informació com per generar referències. Per tant, les hem d'ometre.

5.3 *Multithreading*

Un dels inconvenients més grans que implica el fet d'haver d'accedir a Internet, és que el temps perdut esperant dades és molt alt. Per reduir-lo, s'ha estudiat la possibilitat d'utilitzar diferents fils d'execució per fer més d'una consulta de forma més o menys simultània. La taula següent mostra una comparativa del temps necessari per obtenir mútliples pàgines web de forma seqüencial o bé utilitzant fins a cinc fils d'execució diferents.

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:							
-44.96%		**-65.6%**		**-74.87%**		**-72.35%**	

Les pàgines pàgines corresponen a consultes aleatòries a Google per evitar l'efecte dels *proxies* i la memòria *cache*. Com a conclusió, tot i que es dades obtingudes no són riguroses, ens donen una idea força clara de la millora que s'obté utilitzant aquesta tècnica.

Al nostre sistema hem implementat un *pool* amb un número variable de fils d'execució que es van reutilitzant mentre queden referències per extreure.

Capítol 6

Extracció d'Informació

En aquest capítol tractarem

En el nostre context, anomenarem *wrapper* a un troç de codi que podem utilitzar per extreure una peça d'informació concreta d'un document.

Podem imaginar-ho com filtre que només ens deixa veure una part del document que ens interessa.

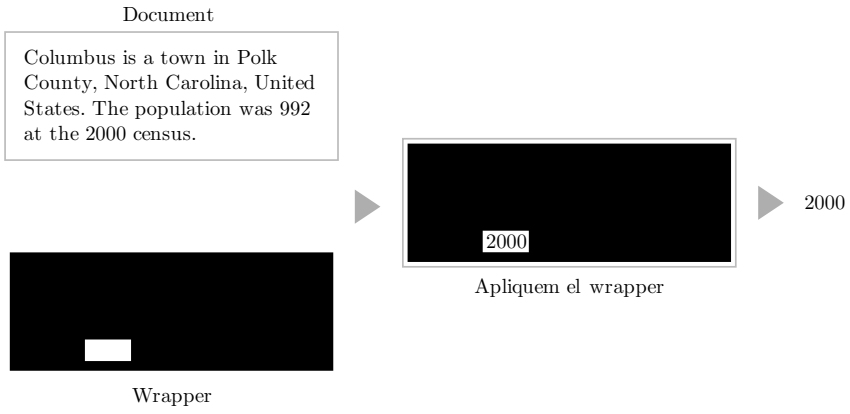

Figura 6.1: Funció d'un *wrapper*

6.1 *Wrappers* a mà

6.2 Inducció de *wrappers*

6.2.1 Generació automàtica de regles

Ruta d'un element HTML

Expressió regular

6.2.2 Avaluació dels *wrappers*

Una vegada hem generat el conjunt dels *wrappers* possibles per a un conjunt de documents, cal que avaluem quins d'ells funcionen millor. Utilitzem un sistema de vots positius i negatius i en calculem la mitjana amb la següent fórmula:

$$score = \frac{vots\ positius}{vots\ totals}$$

6.2.3 Reaprenentatge

El sistema està dissenyat per tal que, quan hi ha una davallada en el nombre de referències extretes correctament, provi de reaprendre els *wrappers* automàticament a partir dels exemples que té emmagatzemats d'execucions passades.

Capítol 7

Anàlisi de resultats

En aquest capítol es mostren les principals proves realitzades amb la nostra aplicació. Per cada prova s'explica el perquè dels resultats obtinguts.
A l'apèndix B es mostren tots els test que s'han dut a terme.

7.1 Només amb *wrappers* induïts

7.2 Utilitzant *wrappers* de referència

Capítol 8

Conclusions i Treball Futur

8.1 Objectius Assolits

8.2 Possibles Millores

Bibliografia

[Jr06] Nobody Jr. My article, 2006.

Apèndix A

Extracció Contingut PDF

Apèndix B

Resultats dels tests

A continuació es mostren els resultats complets de totes les proves realitzades a la nostra aplicació. L'explicació d'aquests números s'explica al capítol 7.

Apèndix C

Biblioteques utilitzades

A continuació es llisten les diferents biblioteques i mòdules *Python* que s'han utilitzat en l'aplicació

- SimpleJSON
- DiffLib
-