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 llegeix el contingut dels documents.

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. *CiteSeer* és un projecte *open source* de característiques similars, però que també té limitacions. El sistema funciona analitzant la bibliografia dels articles, 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

Phasellus eu ante diam, eu euismod nunc. Vivamus non dolor sem. Sed id metus enim. Curabitur consectetur eleifend quam porta sagittis. Mauris sed augue fermentum leo pharetra posuere nec euismod risus. In dui elit, iaculis eget vestibulum eu, suscipit at purus. Mauris hendrerit condimentum velit, in facilisis dui consectetur non. Quisque tristique velit vitae enim posuere suscipit. Integer condimentum rutrum accumsan. Suspendisse bibendum urna eget orci aliquam faucibus congue urna consequat. Nam elementum, lectus a volutpat gravida, felis nibh faucibus nibh, id fringilla arcu purus sed orci.

2.2 El format BibTeX

- Com podem distingir entre diferents tipus d'entrada (article, book, inproceedings, etc.) a partir del fitxer?
- Format dels noms. Un nom consisteix de diferents parts: First, von, Last, Jr. El token *von* o *de la* cal posar-los en minúscules. Per tal que el nom es reconegui, cal que tingui el format: von Last, Jr, First. D'aquesta manera, si hi ha més d'un cognom no passa res.
- Caràcters Unicode entre claus per poder ser utilitzats correctament amb l'estil *alpha*. Per exemple: `Jos{\'{e}}`
- Per prevenir que BibTeX canviï un text a minúscules, cal posar el text entre claus.
- Si hi ha massa autors, truncar la llista amb *et al.*
- Utilitzar abreviatures de tres lletres per als mesos
- Utilitzar el camp `key` per a organitzacions amb un nom llarg, de manera que s'utilitzin les inicials de l'organització al fer una cita.

2.3 Característiques

At vero eos et accusamus et iusto odio dignissimos ducimus qui blanditiis praesentium voluptatum deleniti atque corrupti quos dolores et quas molestias excepturi sint occaecati cupiditate non

provident, similique sunt in culpa qui officia deserunt mollitia animi, id est laborum et dolorum fuga. Et harum quidem rerum facilis est et expedita distinctio.

2.4 Planificació Temporal

Capítol 3

Disseny del sistema

3.1 Mòduls

Capítol 4

Extracció dels continguts d'un 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 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 texts extrets de diferents documents PDF.
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 Primera idea: *Google Scholar*

5.2 Resta de cercadors

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.

5.3 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.

5.4 *Multithreading*

Un dels inconvenients que suposa el fet d'haver d'accedir a Internet, és la latència. Per reduir el temps que es perd esperant les dades, hem fet que l'aplicació creï diferents fils d'execució.

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.

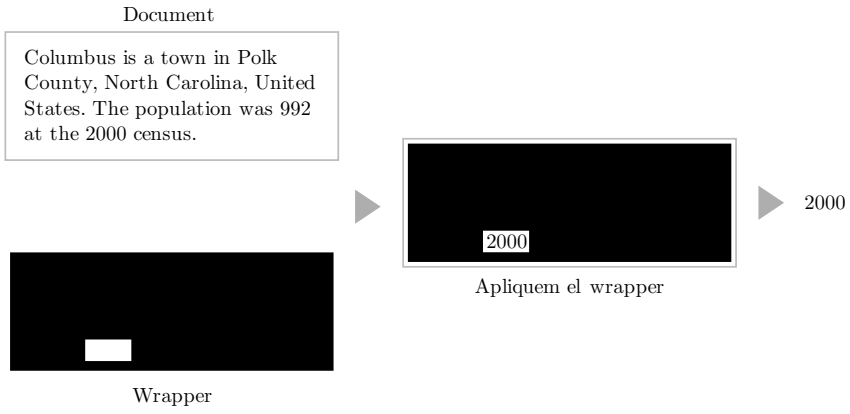

Figura 6.1: Funció d'un *wrapper*

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
-