

BIBTEX Bibliography Index Maker

Primavera 2010

Índex

Objectius

Procés d'extracció

PDF to text

Cerca de referències

Extracció d'informació

Generació de regles

Tipus de Regles

Avaluació

Demo

Conclusions

Motivació

Objectius

Guió típic d'un treball de recerca:

1. Lectura d'articles
2. S'acumulen els fitxers (PDF) en un directori
3. A l'hora d'escriure, per poder citar, fa falta tenir un índex bibliogràfic

Objectiu

Objectius

Un únic objectiu principal:

- ▶ La creació d'índexs bibliogràfics de forma automàtica

Però per aconseguir-lo necessitem:

- ▶ Cercar referències a Internet pels PDFs que tenim
- ▶ Extreure la informació de les pàgines HTML
- ▶ Crear regles d'extracció automàticament

Objectiu

Objectius

Un únic objectiu principal:

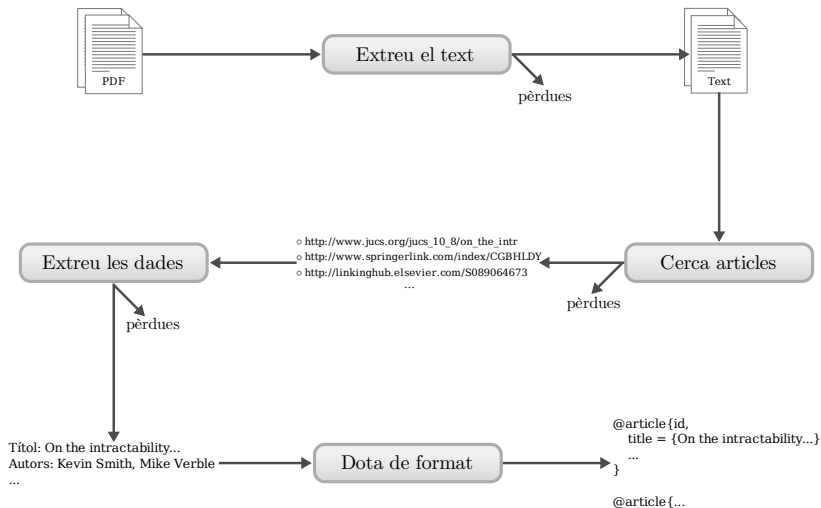
- ▶ La creació d'índexs bibliogràfics de forma automàtica

Però per aconseguir-lo necessitarem:

- ▶ Cercar referències a Internet pels PDFs que tenim
- ▶ Extreure la informació de les pàgines HTML
- ▶ Crear regles d'extracció automàticament

Procés d'extracció

Esquema d'extracció



Procés d'extracció

PDF to text

Quant a l'extracció de text dels fitxers PDFs:

- ▶ S'utilitza l'eina *xPDF*
- ▶ Resultats relativament bons...
- ▶ ...però continua tenint força problemes

Procés d'extracció

Cerca de referències

Passos a seguir per cercar referències:

- ▶ Genera consultes a partir del text
`([\w () ? !] + \) {min,max}`
“and slurry methods and have been tested in the selective”
- ▶ Obté els resultats de les consultes (*Google, Bing, Yahoo*)
 - ▶ `www.springerlink.com/index/G4588X...`
 - ▶ `www.ingentaconnect.com/content/klu/ca...`
- ▶ Filtra i ordena els resultats

Procés d'extracció

Extracció d'informació

Dos tipus de regles d'extracció d'informació:

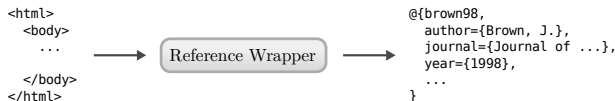
- ▶ *Reference Wrappers*
- ▶ *Field Wrappers*

Extracció d'informació

Reference Wrappers

Característiques:

- ▶ Extreuen referències senceres
- ▶ Només se'n necessita un per cada biblioteca digital
- ▶ Implementats manualment
- ▶ Els resultats solen ser bons

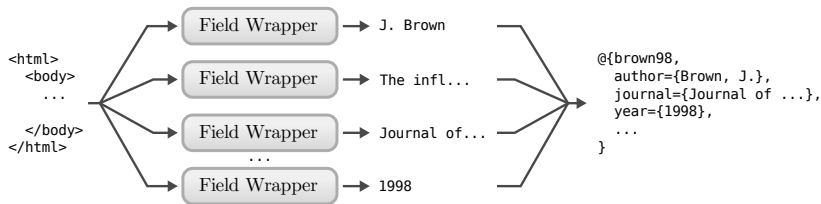


Extracció d'informació

Field Wrappers

Característiques:

- ▶ Especialitzats en treure un sol camp
- ▶ Se'n necessita un per cada camp i per cada biblioteca
- ▶ Implementats automàticament



Extracció d'informació

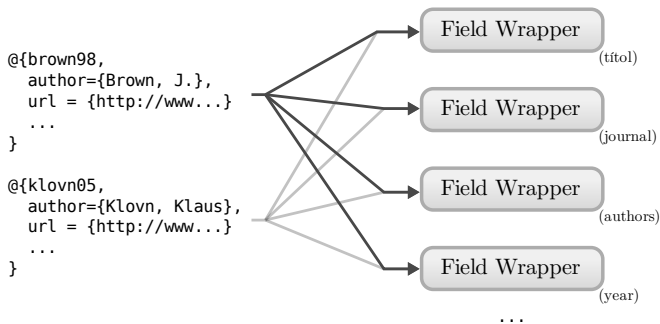
Validació

Característiques de la validació:

- ▶ Camps com ara el títol i autors: es comprova que es troben dins del document PDF
- ▶ Camps com ara el número de pàgines o any: es mira que compleixin una expressió regular.
- ▶ Cada camp té un pes sobre la validesa final

Generació de regles

Esquema



- ▶ S'utilitzen (poques) referències com a exemples.
- ▶ Corresponen a la mateixa biblioteca digital

Generació de regles

Tipus de regles

Path Rule:

- ▶ Localitza l'element del document HTML que conté el valor que busquem
- ▶ El patró consisteix en una ruta dins de l'arbre que representa el document

Regex Rule:

- ▶ Localitza el valor dins del contingut de l'element HTML obtingut per una *Path Rule*
- ▶ El patró consisteix en una expressió regular per extreure el valor desitjat

Generació de regles

Tipus de regles

Path Rule:

- ▶ Localitza l'element del document HTML que conté el valor que busquem
- ▶ El patró consisteix en una ruta dins de l'arbre que representa el document

Regex Rule:

- ▶ Localitza el valor dins del contingut de l'element HTML obtingut per una *Path Rule*
- ▶ El patró consisteix en una expressió regular per extreure el valor desitjat

Generació de regles

Altres regles per a camps multivalor

Separators Regex Rule:

- ▶ Separa els diferents valors continguts en un mateix element HTML
- ▶ El patró consisteix en un conjunt de cadenes de caràcters que separen els valors desitjats

MultiValue Regex Rule:

- ▶ Consisteix en una *regex rule* que actua sobre diferents valors

Person Rule:

- ▶ Separen noms en els camps: `last_name`, `middle_name` i `first_name`

Generació de regles

Altres regles per a camps multivalor

Separators Regex Rule:

- ▶ Separa els diferents valors continguts en un mateix element HTML
- ▶ El patró consisteix en un conjunt de cadenes de caràcters que separen els valors desitjats

MultiValue Regex Rule:

- ▶ Consisteix en una *regex rule* que actua sobre diferents valors

Person Rule:

- ▶ Separen noms en els camps: `last_name`, `middle_name` i `first_name`

Generació de regles

Altres regles per a camps multivalor

Separators Regex Rule:

- ▶ Separa els diferents valors continguts en un mateix element HTML
- ▶ El patró consisteix en un conjunt de cadenes de caràcters que separen els valors desitjats

MultiValue Regex Rule:

- ▶ Consisteix en una *regex rule* que actua sobre diferents valors

Person Rule:

- ▶ Separen noms en els camps: `last_name`, `middle_name` i `first_name`

Generació de regles

Avaluació de *Wrappers*

Quant a l'avaluació de *wrappers*:

- ▶ Els millors *wrappers* s'han de provar primer
- ▶ Per escollir un ordre inicial, s'apliquen sobre els mateixos exemples
- ▶ Cada vegada que s'utilitza un *wrapper* se li dóna un vot
- ▶ Els bons pugen i els dolents baixen

Demo

Conclusions

- ▶ Hi ha pèrdues a cadascun dels passos
- ▶ L'èxit dels resultats depèn completament de la qualitat dels *wrappers* disponibles
- ▶ Errors fàcilment corregibles manualment
- ▶ La part de generació i extracció, aplicable a qualsevol altre domini
- ▶ Molt per fer i millorar

Q&A