BibTeX Bibliography Index Maker

Ramon Xuriguera Albareda

Primavera 2010

Índex

Motivació

Guió típic en un treball de recerca:

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

Motivació

Guió típic en un treball de recerca:

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

Motivació

Guió típic en un treball de recerca:

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

Motivació

Guió típic en un treball de recerca:

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

Objectiu

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

Objectiu

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

Objectiu

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

Objectiu

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

Objectiu

Un únic objectiu principal:

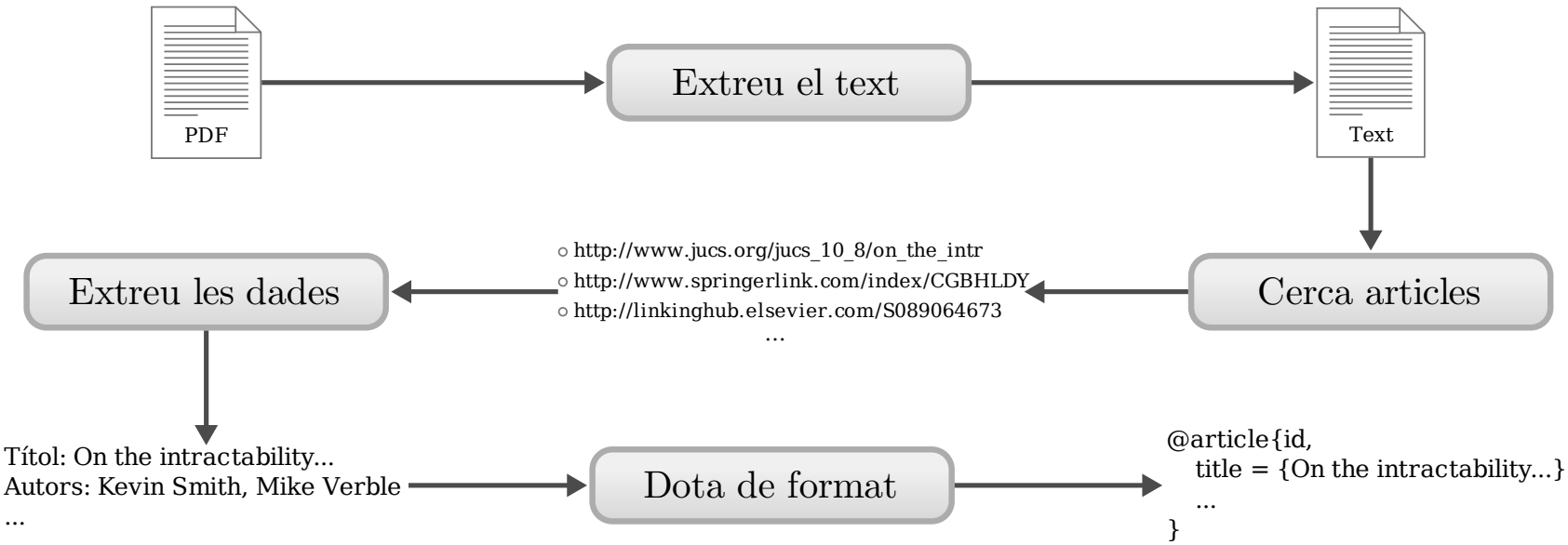
- La creació d'índexs bibliogràfics de forma automàtica

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

Objectiu

Esquema d'extracció



Cerca de referències

Passos a seguir:

- Extreu el contingut dels fitxers PDF
- Genera consultes a partir del text
- Obté els resultats de les consultes (*Google*, *Bing*, *Yahoo*)
- Filtra i ordena els resultats

Extracció d'informació

Dos mètodes d'extracció:

- *Reference Wrappers*
- *Field Wrappers*

Extracció d'informació

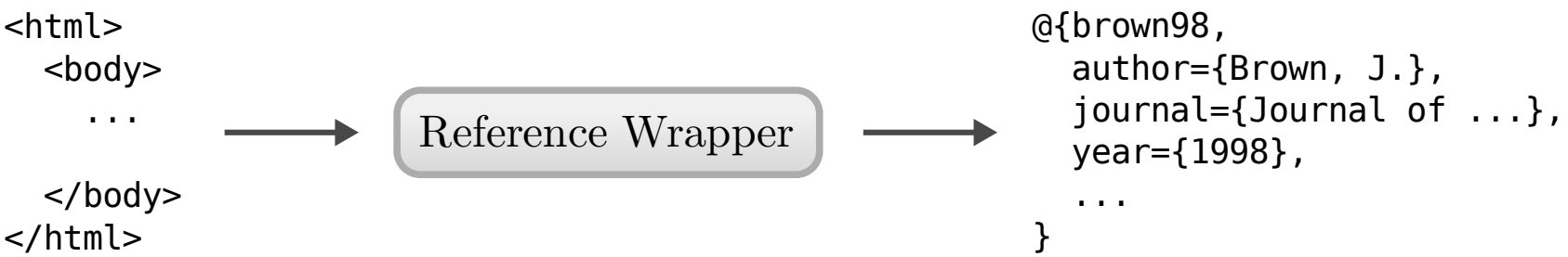
Reference Wrappers

Característiques:

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

Extracció d'informació

Reference Wrappers



Extracció d'informació

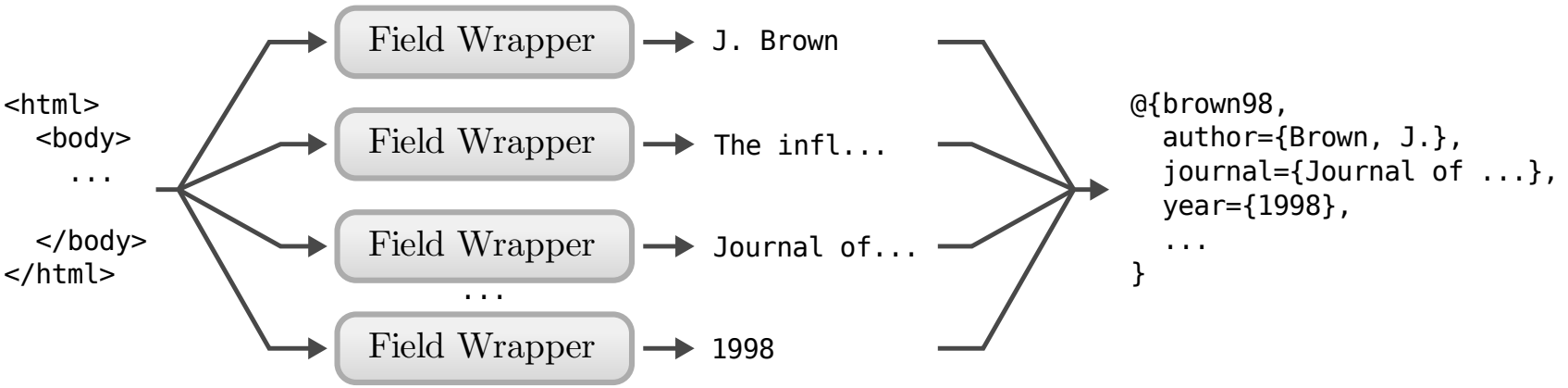
Field Wrappers

Característiques:

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

Extracció d'informació

Field Wrappers



Generació de regles

Només es generen *Field Wrappers*
Consisteixen en regles
Trobar l'element que conté la informació
Generar una expressió regular per extreure-la
Avaluar el *wrapper*

Demostració

Conclusions i Treball Futur