



Recuperación de Información con Terrier

Recuperación de Información



Contenido

- Introducción
- Componentes
- Instalación
- Configuración / Preparación
- Ejecución



Introducción

Experimentación IR

Motores de búsqueda disponibles

No-Académicos

- Lucene (Apache)
- Minion (Oracle)
- Xapian (Cambridge)
- Sphinx (Sphinx Inc.)

Académicos

- **Terrier** (Glasgow)
- Lemur/Indri (CMU/UMass)
- Zettair
- PISA engine



Introducción

“**Terrier** is a highly flexible, efficient, and effective open source search engine, readily deployable on large-scale collections of documents”

Eficiente

Extensible

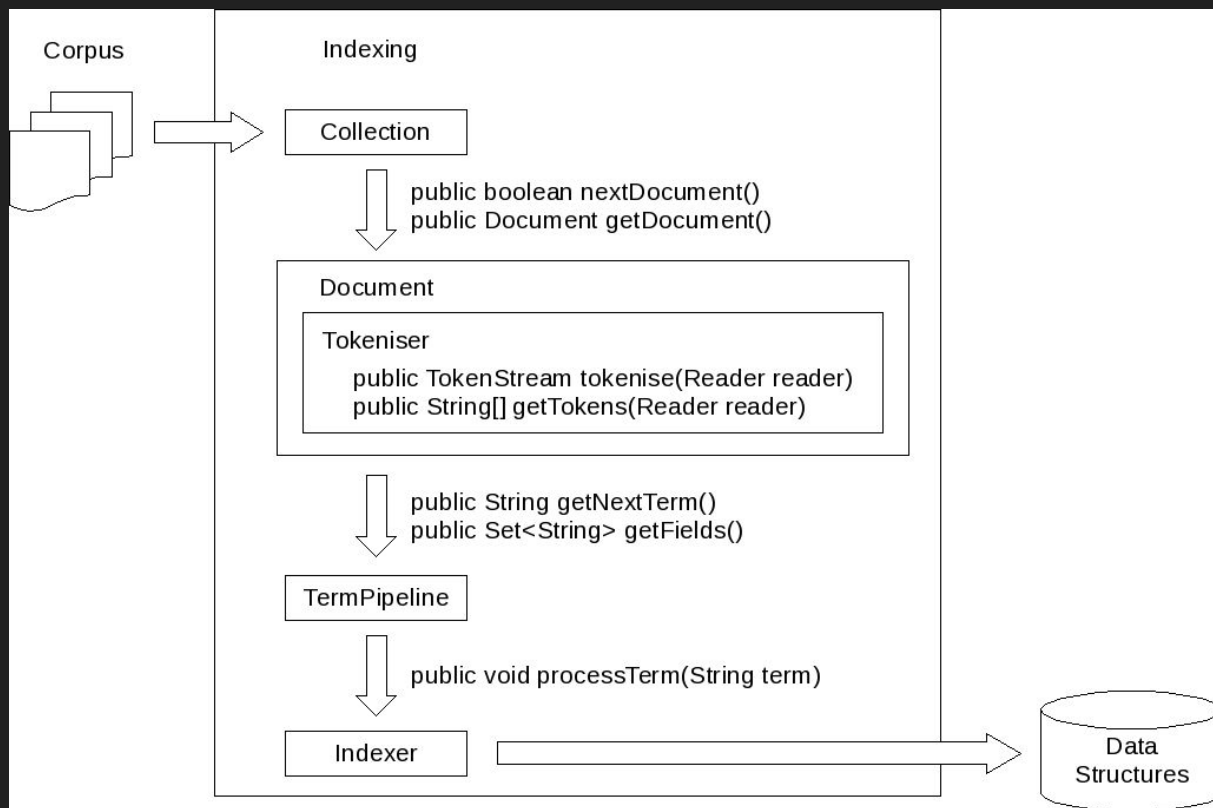
Efectivo

Multi-lenguaje

Flexible

Interactivo

Componentes - Indexación

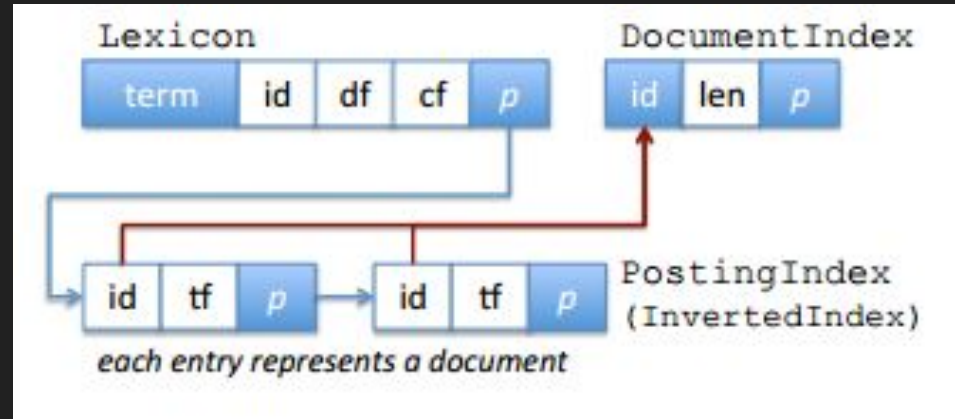




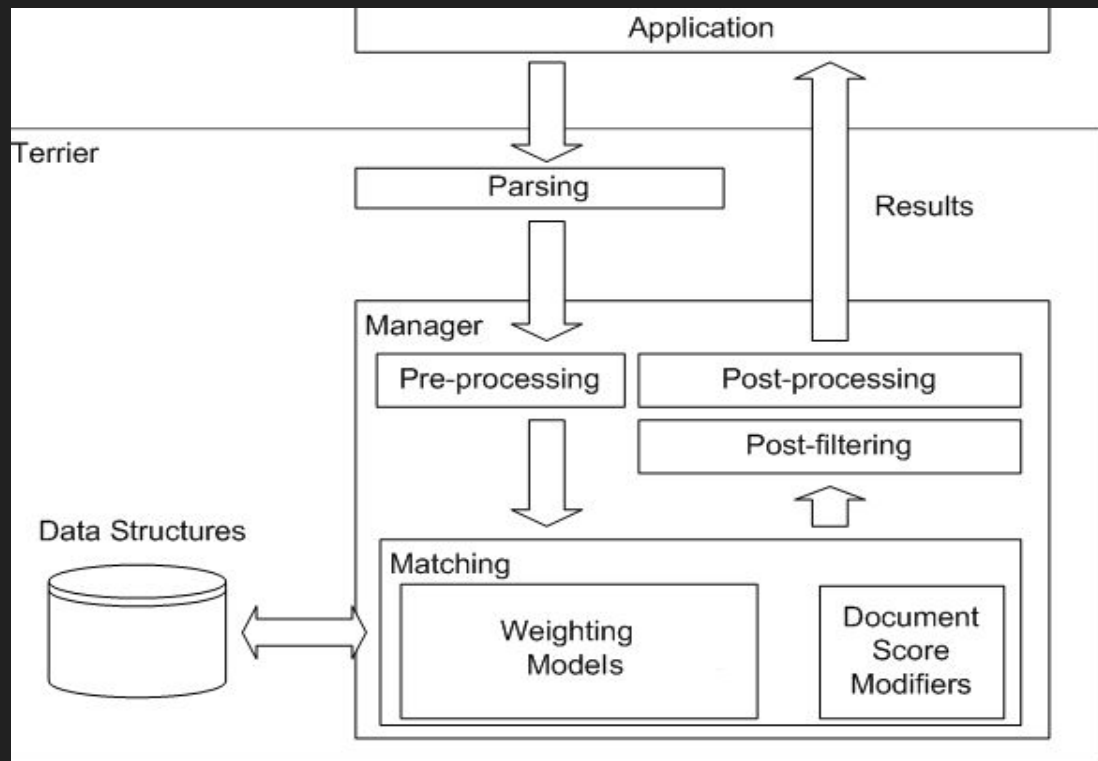
Componentes - Indexación

Estructuras generadas

- Lexicon
- Inverted index
- Document index (id doc, long.)
- Direct index (términos de cada doc y frecuencia)



Componentes - Recuperación





Instalación

- Descargar desde terrier.org (v5.x)
- Requerimientos
 - Java JDK 1.8+
- Ya compilado! Listo para usar
- Estructura de directorios

/bin - scripts para ejecutar Terrier

/etc - archivos de configuración

/share - archivos de utilidad (ej: stopwords)

/src - código fuente

/var - índices y resultados



Configuración / Preparación

- Configuración por defecto (/etc/terrier.properties)
- Soporte para configurar propiedades a través de la línea de comandos
- Documentación de todas las propiedades configurables terrier.org/docs/v4.0/properties.html

```
#default controls for query expansion
querying.postprocesses.order=QueryExpansion
querying.postprocesses.controls=qe:QueryExpansion
#default controls for the web-based interface. SimpleDecorate
#is the simplest metadata decorator. For more control, see Decorate.
querying.postfilters.order=SimpleDecorate,SiteFilter,Scope
querying.postfilters.controls=decorate:SimpleDecorate,site:SiteFilter,scope:Scope

#default and allowed controls
querying.default.controls=
querying.allowed.controls=scope,qe,qemodel,start,end,site,scope

#document tags specification
#for processing the contents of
#the documents, ignoring DOCHDR
TrecDocTags.doctag=DOC
TrecDocTags.idtag=DOCNO
TrecDocTags.skip=DOCHDR
#set to true if the tags can be of various case
TrecDocTags.casesensitive=false

#query tags specification
TrecQueryTags.doctag=TOP
TrecQueryTags.idtag=NUM
TrecQueryTags.process=TOP,NUM,TITLE
TrecQueryTags.skip=DESC,NARR

#stop-words file
stopwords.filename=stopword-list.txt

#the processing stages a term goes through
termpipelines=Stopwords,PorterStemmer|
```



Configuración / Preparación

- Formatos de corpus
 - Archivos simples (SimpleFileCollection)
 - XML (SimpleXMLCollection)
 - Tweets (TwitterJSONCollection)
 - PDF, Excel, Word, ...
- **TREC** (TRECCollection)

```
1 <DOC>
2 <DOCNO> 1 </DOCNO>
3 Contenido del documento uno
4 </DOC>
5 <DOC>
6 <DOC>
7 <DOCNO> 2 </DOCNO>
8 Contenido del documento dos
9 </DOC>
10 <DOC>
11 ...|
```



```
#document tags specification
#for processing the contents of
#the documents, ignoring DOCHDR
TrecDocTags.doctag=DOC
TrecDocTags.idtag=DOCNO
TrecDocTags.skip=DOCHDR
#set to true if the tags can be of various case
TrecDocTags.casesensitive=false
```



Configuración / Preparación

- Formato de Query's (Topics)

```
1 <TOP>
2 <NUM>1<NUM>
3 <TITLE>house dog
4 <DESC>description
5 <NARR>narrative
6 </TOP>
7 <TOP>
8 <NUM>2<NUM>
9 <TITLE>chair
10 <DESC>description
11 <NARR>narrative
12 </TOP>
13 ...|
```



```
#query tags specification
TrecQueryTags.doctag=TOP
TrecQueryTags.idtag=NUM
TrecQueryTags.process=TOP,NUM,TITLE
TrecQueryTags.skip=DESC,NARR
```

- Lenguaje Terrier

term1 term2 (term1 o term2)
+term1 +term2 (term1 y term2)
+term1 -term2 (term1 y no term2)
"term1 term2" (term1 y term2 como frase)



Configuración / Preparación

- Modelos de recuperación

- **TF-IDF**
- BM25 (modelo probabilístico)
- Hiemstra_LM (modelos de lenguaje)
- ... (lista completa en terrier.org/docs/v5.x/configure_retrieval.html)

```
#configure retrieval model
trec.model=BM25
```

- Juicios de relevancia

- Archivo de 4 columnas, en el que se debe especificar id query, id doc y relevancia.

id query

1	0	1	1
1	0	2	1
1	0	3	1
1	0	4	0
2	0	1	0
2	0	2	0
2	0	3	1
2	0	4	1
3	0	1	0
3	0	2	1
3	0	3	0
3	0	4	1
.			
.			
.			

id doc

relevancia (0-1)



Ejecución

1) Setup

```
>> bin/trec_setup.sh ruta/absoluta/coleccion
```

En /etc se crea un archivo collection.spec con los archivos que se van a indexar

2) Indexación

```
>> bin/terrier batchindexing
```

```
>> bin/terrier batchindexing -Dtrec.collection.class=SimpleFileCollection
```

Se crean las estructuras de datos dentro de /var/index. Se pueden ver estadísticas de las mismas ejecutando:

```
>> bin/terrier indexstats
```

```
>> bin/terrier indexutil --printlex lexicon
```



Ejecución

3) Recuperación

Modo interactivo

```
>> bin/terrier interactive
```

Batch

```
>> bin/terrier batchretrieve
```

```
>> bin/terrier batchretrieve -t ruta/absoluta/topics_file.txt
```

```
>> bin/terrier batchretrieve -t ruta/absoluta/topics_file.txt -Dtrec.model=BM25
```

Dentro de /var/results se crea un archivo .res con los resultados de la recuperación

```
1 Q0 4 0 2.102432 TF_IDF
1 Q0 2 1 1.832432 TF_IDF
1 Q0 1 2 0.432344 TF_IDF
.
.
.
```

id_query

id_doc

similitud

Ejecución

4) Evaluación

```
>>bin/terrier batchevaluate -q /ruta/absoluta/qrels_file.txt
```

Dentro de /var/results se genera un archivo .eval con

los resultados de la evaluación

```
Number of queries = 35
Retrieved       = 25055
Relevant        = 1742
Relevant retrieved = 1043
```

```
Average Precision: 0.0300
R Precision       : 0.0479
```

```
Precision at 1 : 0.0571
Precision at 2 : 0.0571
Precision at 3 : 0.0667
Precision at 4 : 0.0714
Precision at 5 : 0.0686
Precision at 10 : 0.0486
Precision at 15 : 0.0495
Precision at 20 : 0.0514
Precision at 30 : 0.0476
Precision at 50 : 0.0457
Precision at 100 : 0.0440
Precision at 200 : 0.0401
Precision at 500 : 0.0397
Precision at 1000 : 0.0298
```

```
Precision at 0%: 0.2484
Precision at 10%: 0.0917
Precision at 20%: 0.0816
Precision at 30%: 0.0784
Precision at 40%: 0.0698
Precision at 50%: 0.0610
Precision at 60%: 0.0370
Precision at 70%: 0.0122
Precision at 80%: 0.0000
Precision at 90%: 0.0000
Precision at 100%: 0.0000
```

```
Average Precision: 0.0300
```

The background of the slide features a large, faint, circular seal of the Universidad Nacional de Luján. The seal contains a central shield with a bird (phoenix) rising from flames, flanked by two hands holding a banner. The text "UNIVERSIDAD NACIONAL DE LUJÁN" is inscribed around the perimeter of the seal.

FIN