Comparativa entre sistema de recuperación de información y CRAWLER

Autor: David Márquez Calavia

NIP: 700940

Autor: Víctor Labrador Ortiga

NIP: 701658

Comparación de resultados

Los resultados de las distintas necesidades de información son los siguientes (aunque solamente se han mostrado los cinco primeros resultados):

Para la necesidad **03-5**, relacionada con el análisis de la **crisis económica mundial**:

Sistema tradicional:

1	recordsdc\oai_zaguan.unizar.es_30821.xml	Relevante
2	recordsdc\oai_zaguan.unizar.es_11517.xml	Relevante
3	recordsdc\oai_zaguan.unizar.es_15623.xml	No relevante
4	recordsdc\oai_zaguan.unizar.es_17141.xml	Relevante
5	recordsdc\oai_zaguan.unizar.es_15416.xml	Relevante

Crawler:

1	recordsdc\oai_zaguan.unizar.es_30822.xml	Relevante
2	recordsdc\oai_zaguan.unizar.es_5738.xml	No relevante
3	recordsdc\oai_zaguan.unizar.es_6548.xml	No relevante
4	recordsdc\oai_zaguan.unizar.es_10620.xml	No relevante
5	recordsdc\oai_zaguan.unizar.es_63883.xml	Relevante

Que se ha obtenido mediante la siguiente consulta al crawler:

```
content:crisis AND content:econ?m* AND content:desarrollo AND content:empres* AND content:20?? AND content:análisis
```

Donde se usa el "?" para indicar que puede ser una letra cualquier para así poder coger "económ" como "econom" en este caso por si no se ha puesto la tilde, seguir encontrando el documento. Para el caso de la cerradura de Kleene es el "*" que se usa para así encontrar cualquier palabra que empiece por lo indicado. En este caso, cualquier palabra que empiece por "empres" como podría ser "empresa", "empresario", etc. En este caso, se ve que funciona mejor nuestro sistema tradicional ya que encuentra más documentos que cumplen las especificaciones de la consulta que el crawler que encuentra resultados que apenas tienen que ver con la crisis o la economía como puede ser el documento "6548" que trata sobre una central térmica de biomasa en Venezuela sin relevancia alguna sobre la crisis y la economía.

Para la necesidad **05-2**, relacionada con el análisis de los **estudios de feminismo**:

Sistema tradicional:

1	recordsdc\oai_zaguan.unizar.es_12350.xml	Relevante
2	recordsdc\oai_zaguan.unizar.es_64995.xml	Relevante
3	recordsdc\oai_zaguan.unizar.es_31517.xml	No relevante
4	recordsdc\oai_zaguan.unizar.es_56867.xml	Relevante
5	recordsdc\oai_zaguan.unizar.es_8095.xml	No relevante

Crawler:

1 recordsdc\oai_zaguan.unizar.es_ 47893.xml Relevante
2 recordsdc\oai_zaguan.unizar.es_ 64995.xml Relevante
3 recordsdc\oai_zaguan.unizar.es_ 56965.xml Relevante
4 recordsdc\oai_zaguan.unizar.es_ 56606.xml No relevante (No habla del feminismo en sí, sólo de la mujer y las revistas)
5 recordsdc\oai_zaguan.unizar.es_ 64008.xml No relevante

Que se ha obtenido mediante la siguiente consulta al crawler:

```
content:femin* OR content:mujer*) AND content:evolución AND content:papel
```

En este caso los dos sistemas funcionan de manera similar ya que, ambos devuelven 3 documentos relevantes sobre 5. Aunque hay un documento que tiene que ver un poco con las mujeres como puede ser el 56606 devuelto por el crawler, aunque no trate el tema principal como es el feminismo.

Para la necesidad **07-3**, relacionada con el análisis de las **construcciones españolas**:

Sistema tradicional:

1 recordsdc\oai_zaguan.unizar.es_13660.xml No relevante
2 recordsdc\oai_zaguan.unizar.es_13845.xml No relevante
3 recordsdc\oai_zaguan.unizar.es_12301.xml No relevante
4 recordsdc\oai_zaguan.unizar.es_13614.xml No relevante
5 recordsdc\oai_zaguan.unizar.es_9735.xml Relevante

Crawler:

1 recordsdc\oai_zaguan.unizar.es_ 36880.xml No relevante
2 recordsdc\oai_zaguan.unizar.es_ 47886.xml No relevante
3 recordsdc\oai_zaguan.unizar.es_ 62582.xml No relevante
4 recordsdc\oai_zaguan.unizar.es_ 47144.xml No relevante
5 recordsdc\oai_zaguan.unizar.es_ 32422.xml No relevante

Que se ha obtenido mediante la siguiente consulta al crawler:

```
(content:construc* AND (content:españ* OR content:Españ*) AND (content:19?? OR content:20??) AND content:María )OR (content:construc* AND (content:españ* OR content:Españ*) AND (content:19?? OR content:20??))
```

El problema con esta necesidad es que al pedir que el nombre de pila del autor o director sea María obliga a buscar la palabra María y se le da más peso en nuestro sistema tradicional a encontrar el nombre de María y al final acaba obteniendo resultados con la palabra María, pero no relevantes con las construcciones españolas. Aun así, ha encontrado un documento mientras que el crawler no ha encontrado ninguno relevante de los 5 posibles.

Para la necesidad **08-3**, relacionada con el análisis de **TFM o TFG de Profesorado, pero no de FP**:

Sistema tradicional:

1	recordsdc\oai_zaguan.unizar.es_16593.xml	Relevante (TFM y no FP)
2	recordsdc\oai_zaguan.unizar.es_8481.xml	No relevante (TFM y FP)
3	recordsdc\oai_zaguan.unizar.es_11343.xml	No relevante (TFM y FP)
4	recordsdc\oai_zaguan.unizar.es_11333.xml	Relevante (TFM y no FP)
5	recordsdc\oai_zaguan.unizar.es_8118.xml	No relevante (TFM y FP)

Crawler:

1	recordsdc\oai_zaguan.unizar.es_ 62404.xml	No relevante (No TFM o TFG)
2	recordsdc\oai_zaguan.unizar.es_ 61423.xml	No relevante (No TFM o TFG)
3	recordsdc\oai_zaguan.unizar.es_ 62483.xml	No relevante (TFM y FP)
4	recordsdc\oai_zaguan.unizar.es_ 58674.xml	Relevante (TFM y no FP)
5	recordsdc\oai_zaguan.unizar.es_ 62402.xml	No relevante (TFM y FP)

Que se ha obtenido mediante la siguiente consulta al crawler:

```
content:profesor* -content:"?ormación profesional" -content:FP AND
(content:?achillerato OR content:ESO OR content:"?ducación ?ecundaria")
AND content:20??
```

En este caso se encuentran muchos resultados que tienen que ver con TFM o TFG de Profesorado, pero para FP y en la consulta se pide que no sea para FP, entonces no se da por relevante. Sin embargo, aun así, nuestro sistema obtiene muchos mejores resultados ya que todas las consultas al menos tienen que ver con TFM's y no como en el crawler que hay 2 de 5 que no tienen que ver.

Para la necesidad **11-4**, relacionada con el análisis sobre **las energías renovables**:

Sistema tradicional:

1	recordsdc\oai_zaguan.unizar.es_31583.xml	Relevante (fotovoltaicas)
2	recordsdc\oai_zaguan.unizar.es_6217.xml	Relevante (fotovoltaicas)
3	recordsdc\oai_zaguan.unizar.es_31473.xml	Relevante (fotovoltaicas)
4	recordsdc\oai_zaguan.unizar.es_6383.xml	Relevante (no fotovoltaicas)
5	recordsdc\oai_zaguan.unizar.es_5618.xml	Relevante (fotovoltaicas)

Crawler:

1	recordsdc\oai_zaguan.unizar.es_ 31583.xml	Relevante (fotovoltaicas)
2	recordsdc\oai_zaguan.unizar.es_ 10186.xml	Relevante (no fotovoltaicas)
3	recordsdc\oai_zaguan.unizar.es_ 10213.xml	No relevante (no energías renov.)
4	recordsdc\oai_zaguan.unizar.es_ 10289.xml	Relevante (no fotovoltaicas)
5	recordsdc\oai_zaguan.unizar.es_ 10290.xml	Relevante (no fotovoltaicas)

Que se ha obtenido mediante la siguiente consulta al crawler:

```
(content:201? AND content:energ* AND content:renovable* OR
content:"placas fotovoltaicas") OR (content:201? AND
content:energ* AND content:renovable*)
```

En este caso se ve que nuestro sistema tradicional funciona mucho mejor ya que todos los resultados obtenidos son relevantes pese a que uno no contenga nada relacionado con las placas fotovoltaicas. Sin embargo, para el caso del crawler hay uno que ni siquiera habla de energías renovables y de los 4 restantes que se consideran relevantes porque hablan de energías renovables, sólo hay uno que habla de las placas fotovoltaicas.

Procesos de indexación

Para el caso del crawler se crea sólo un campo "content" como se ve en las consultas hechas en lo que respecta al documento y el proceso del sistema tradicional consiste en que se separa en distintos campos que componen todo el fichero como pueden ser los que ya hemos visto y trabajado con ellos como "title" o "creator".

Esto hace que para el sistema tradicional se permitan búsquedas más concretas como puede ser encontrar un "title" exacto en vez de buscar en todo el volcado del documento. Además, el sistema tradicional que se usó utiliza un analizador para tokenizar mediante Lucene y trabaja en el idioma del castellano, sin embargo, SOLR trabaja con uno genérico lo que hace que sea menos preciso.

Procesos de búsqueda

Debido a como se ha visto que se ha indexado para cada uno de los sistemas, ahora cambia a la hora de realizar las búsquedas. En el primer caso, SOLR, busca las palabras claves que se le han puesto en el campo "content" mirando por todo el documento indiferentemente. Se puede trabajar como se ha mencionado anteriormente con "?" y la cerradura de Kleene ("*") para así poder hacer búsquedas mejores que contengan partes de palabras, por ejemplo. Esto hace que haya consultas que pierdan mucha eficacia como puede ser el tema de rango de fechas. En este caso si se busca algo como "a partir del 1900", esto buscará directamente en el documento los tokes "a", "partir", "del" y "1900" pero sin embargo si nosotros con el sistema tradicional le decimos que puede encontrar "a partir del", "desde", etc (cualquier otro sinónimo para indicar un rango de fechas), lo encontrará y luego buscará un número que será lo que trate como fecha. Además, puede trabajar con los campos ya mencionados como podría ser el caso del campo "date" donde se marcan las fechas de los documentos.

Ocurriría lo mismo para el caso que se ha dicho de María como autora o directora de una construcción ya que, por ejemplo, para el caso del sistema tradicional se puede dar ciertos pesos para así darle más importancia a que sea hecho por María o no.

También es cierto que, el sistema tradicional, se implementó conociendo ya previamente las consultas que se iban a realizar a diferencia del crawler lo que da un resultado de documentos más relevante.