PRÁCTICA III:

...

Implementación de un Sistema de Recuperación de Información utilizando Lucene

...

Parte D: Indexación de Facetas

11 de noviembre de 2019

ÍNDICE	2

Índice

1.	. Objetivo		3 3 5	
2.	Búsqueda por Facetas			
3.	. Facetas en Lucene . Búsqueda y Facetas			
4.			7	
	4.1.	Mostra	amos los resultados obtenidos para las facetas	10
4.2. Otros mecanismos de agregación para las facetas			necanismos de agregación para las facetas	11
		4.2.1.	Acumulando la relevancia	12
		4.2.2.	Agrupando valores por rango definidos por el usuario	13
5.	. Filtrando la búsqueda		14	
6.	. Gestionando los árboles de facetas		16	

1 OBJETIVO 3

1. Objetivo

El objetivo de esta práctica es profundizar en el uso de facetas en un SRI. Para ello, el alumno deberá de incluir dicha estrategia de búsqueda en la aplicación que esta diseñando. Aunque la búsqueda por facetas es mas lenta que una búsqueda normal (son muchos los cálculos internos que tenemos que hacer), su eficiencia se mantiene proporcional al número de hits (documentos que emparejan con la consulta).

El paquete facetas de Lucene hace fácil añadir categorías a una aplicación de búsqueda, dotándolo de características interesantes como muestro aleatorio (para el caso en que son muchos los documentos que emparejan con la consulta, distintos mecanismos de agregación, permitir tener las facetas en RAM, interfaces personalizados, etc.).

2. Búsqueda por Facetas

Una búsqueda por facetas nos permite acceder a la información refinando la búsqueda de acuerdo a una clasificación por categorías, filtrando los datos teniendo en cuenta las categorías a las que pertenecen. Para ello, es necesario que cada documento pueda ser clasificado a lo largo de múltiples dimensiones (llamadas facetas) como por ejemplo el autor, el idioma o en un sitio de comercio electrónica cada una de las posibles categorías bajo las que podemos clasificar un producto (marca, modelo, características, etc.).

Un atractivo de la mezcla de la búsqueda con el uso de la navegación por facetas es que, ante una consulta, podemos mostrar el número de elementos recuperados en cada una de las categorías. Así, por ejemplo, podemos saber cuantos trabajos, de entre los relevantes a la consulta, han sido publicados en el año 2015 o el 2016, o cuántos de ellos han sido escritos por un determinado autor. Además, cuando el usuario selecciona una de ellas podemos restringir la búsqueda (drill down) entre los documentos que pertenecen a dicha categoría. Esta información hace fácil la búsqueda de los elementos de interés, ya que el usuario puede navegar fácilmente por los resultados, facilitando las siguientes interacciones con el sistema para refinar la búsqueda.

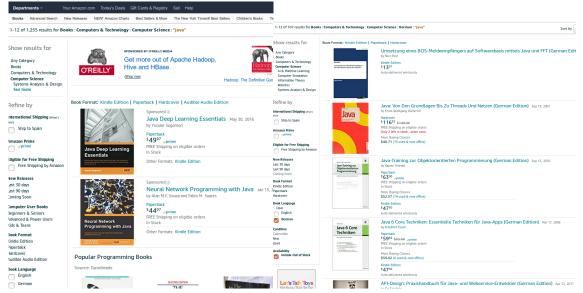
Esta peculiaridad ha hecho que la búsqueda por facetas sea muy común en sitios de comercio electrónico, como por ejemplo Amazon. En la Figura 1 podemos ver cómo ante la consulta "Java" encontramos un total de 22075 libros en el portal de ventas Amazon. A la izquierda de la misma encontramos un frame en el que se permite mostrar los resultados por categorías (aunque Amazon, por motivos internos, ha decidido no mostrar cuántos libros hay en cada una de las categorías). Entre las categorías que considera Amazon encontramos el tipo de libro, lenguaje, autores, formato, etc.

Figura 1: Búsqueda de libros en Amazon, consulta: "Java". Se encuentran un total de 22.075 libros.



Así, podemos centrar la búsqueda dentro de la categoría Computer-Science (imagen a la izquierda de la Figura 2), encontrando un total de 1255 libros y dentro de ella, podemos de nuevo restringirnos a los libros que han sido editados en Alemán (imagen a la derecha de la Figura 2), encontrando un total de 107 libros.

Figura 2: Consulta: "Java", restringimos la búsqueda a los libros que se encuadran dentro de la categoría computer-science (izq.) o computer-science ->Alemán (dcha.) Se encuentran un total de 22.075 libros.



3. Facetas en Lucene

La facetas fueron incluidas en Lucene en la versión 3.4, y podemos encontrar una descripción detallada en http://lucene.apache.org/core/4_2_0/facet/org/apache/lucene/facet/doc-files/userguide.html. Aunque dicha guía ha sido borrada de la documentación de Lucene hace años, y por tanto no se encuentra actualizadas, es posible encontrar ejemplos de su uso en la versión 7.1.0 de Lucene en https://lucene.apache.org/core/7_1_0/demo/src-html/org/apache/lucene/demo/facet/.

Para poder realizar una búsqueda con facetas es necesario añadirlas en tiempo de indexación. Para ello, Lucene utiliza dos "índices" distintos: uno que almacena la información sobre las categorías (facetas) y otro, el índice normal, que almacena los documentos Lucene. Por tanto, necesitaremos de dos Directory distintos, que nos permitirán almacenar cada uno de estos índices por separado y dos writers diferentes, IndexWriter y DirectoryTaxonomyWriter, que se encargarán de crear las estructuras necesarias para los índices. El DirectoryTaxonomyWriter es un TaxonomyWriter que utiliza un Directory para almacenar la información de las

categorías (taxonomía) en disco.

```
FSDirectory indexDir = new FSDirectory.open(Paths.get(
       INDEX_DIR));
2
   FSDirectory taxoDir = new FSDirectory.open(Paths.get(
       FACET_DIR));
3
4
   IndexWriterConfig iwconfig = new IndexWriterConfig(new
       WhitespaceAnalyzer());
5
   iwconfig . setOpenMode (OpenMode .CREATE);
6
   FacetsConfig fconfig = new FacetsConfig();
7
8
   IndexWriter indexWriter = new IndexWriter(indexDir, iwconfig);
    DirectoryTaxonomyWriter taxoWriter = new
       DirectoryTaxonomyWriter(taxoDir);
```

Al igual que podemos configurar el IndexWriter, podremos configurar las facetas utilizando la clase FacetsConfig, que permite almacenar la configuración para las distintas facetas. Por defecto, si no se indica nada, la configuración de una faceta tendrá

- un único valor (por ejemplo, si consideramos el año de nacimiento como faceta estamos hablando de valor único, pero si consideramos que una asignatura esta puede ser impartida por varios profesores podemos hablar de una faceta multivaluada)
- una única dimensión (el año tendrá una única dimensión, pero el género puede tener dos o mas como cuando consideramos un segundo nivel dentro de la categoría novela, como por ejemplo romántica, aventuras, etc.)

Para ello, podemos utilizar los "setters" de FacetsConfig, como por ejemplo:

```
fconfig.setMultiValued("Profesor", true);
fconfig.setHierarchical("Titulacion", true);
```

Una vez que hemos configurado las facetas, podemos pasar a indexarlas.

Cuando construimos las facetas debemos de crearnos un FacetFields, que permite asignar pares categoría-valor a cada documento. El siguiente código muestra cómo podemos añadir distintos elementos (facetas o no) a un documento Lucene, el cual puede pasar a ser indexado (línea 20). Es importante indicar que en

este caso no indexamos el documento Lucene tal cual lo hemos creados, sino el documento Lucene que se construye cuando consideramos toda la información asociada a las facetas, mediante el método build de FacetsConfig.

```
2 Document doc = new Document();
4 // Incluimos los campos de indexación
6 doc.add(new StringField("Codigo", "123456", Field. Store.YES));
7 doc.add(new TextField("Nombre", "Nombre de la Asignatura ....",
      Field. Store. YES));
8 doc.add(new TextField("Resumen", "Esta asignatura presenta ....",
      Field. Store.NO));
9
10 // Incluimos las facetas
11
12 doc.add(new FacetField("Profesor", "nombrel apellido1"));
13 doc.add(new FacetField("Profesor", "nombre2 apellido2"));
14 doc.add(new FacetField("Profesor", "nombre3 apellido3"));
15
16 doc.add(new FacetField("Titulacion", "Grado en Informatica", "
      Sistemas de Informacion"));
17
18 // Indexamos el documento.
19
20 index Writer.addDocument(fconfig.build(taxoWriter,doc));
```

Finalmente, tendremos que cerrar los writers que hemos utilizado

```
1 indexWriter.close();
2 taxoWriter.close();
```

4. Búsqueda y Facetas

El módulo de facetas en Lucene se caracteriza por realizar la mayor parte del trabajo en tiempo de indexación. Para cada uno de los documentos indexados, Lucene analiza todas sus facetas, algunas de las cuales pueden ser jerárquicas, y mapea cada una de ellas con un único identificador (entero), los cuales se almacenan dentro de un campo del tipo docValues, que en resumen permiten obtener eficientemente todos los valores de un campo en distintos documentos. Un índice separado (TaxonomyWriter) almacena estos emparejamientos asegurándose que siempre la misma etiqueta sea asociada al mismo identificador.

En tiempo de búsqueda el costo es mínimo: Para cada uno de los documentos en la lista de resultados, se consultan todos los identificadores y se agregan los conteos en un array, resumiendo los resultados al final si es necesario quedarnos con las top N facetas. En el caso de no hacerlo así, como por ejemplo lo hacen las implementaciones de Solr o ElasticSearch, el obtener las facetas asociadas a un documento hay que hacerlo en tiempo de búsqueda, siendo por tanto menos eficiente. En cualquier caso, esta segunda aproximación permitirá una mayor flexibilidad.

Como es lógico, para realizar una búsqueda por facetas debemos de inicializar algunas componentes que nos permitirán leer los dos índices, el índice invertido y el índice que contiene los árboles de categorías. Ambos deben ser abierto es modo lectura.

```
DirectoryReader indexReader = DirectoryReader.open(indexDir);
IndexSearcher searcher = new IndexSearcher(indexReader);
TaxonomyReader taxoReader = new DirectoryTaxonomyReader(taxoDir);
```

DirectoryReader e indexReader se utilizan para realizar las búsquedas sobre el índice (como hemos visto en el apartado de búsquedas de las prácticas). TaxonomyReader es el encargado de realizar la búsqueda sobre las facetas.

En general, la búsqueda por facetas implica contar, de entre los documentos recuperados, cuántos documentos emparejan con una determinada faceta o categoría. Por tanto, requiere el uso de colectores específicos que permitan agilizar dicha tarea. Tanto es así, que es el colector el que realiza la búsqueda para asegurar que obtiene de forma eficiente la información que necesita.

Para ejecutar las búsqueda con facetas y recuperar las categorías para un determinada consulta debemos utilizar la clase FacetCollector https://lucene.apache.org/core/7_1_0/facet/org/apache/lucene/facet/FacetsCollector.html. Recibe una lista de nodos de un árbol de categorías y devuelve los valores (número de hits) para cada uno de los descendientes en el árbol. Un nodo puede

ocupar cualquier posición dentro del árbol de categoría (por ejemplo puede ser la raíz como "/Titulación/" o un descendiente como "/Titulacion/Grado en Informática/".

```
FacetsCollector fc = new FacetsCollector();
TopDocs tdc = FacetsCollector.search(searcher, query, 10, fc);
```

Para realizar la búsqueda por facetas debemos utilizar uno de los distintos métodos search* de la clase FacetCollector. Los parámetros del método search son el IndexSearcher, la consulta Lucene, el número de documentos recuperados y el objeto de la clase FacetCollector encargado de recoger las facetas asociadas a los documentos que emparejan con la consulta. Este objeto nos permitirá realizar la agrupación de resultados en las distintas categorías a las que pertenecen (faceting).

Tras ejecutar la búsqueda se obtienen dos salidas distintas:

 Los documentos que emparejan con la consulta, los resultados de la búsqueda Lucene, son devueltos en un objeto de la clase TopDocs o TopFieldDocs.
 Con este objeto podemos iterar sobre todos los documentos recuperados de la forma usual y recuperar los campos que nos interesen almacenados en el documento Lucene.

■ Las facetas son recogidas dentro del collector que se pasa como parámetro. Una vez que hemos recogido las facetas, podemos instanciar una de las distintas subclases de Facets para realizar el conteo de las mismas (lo normal es contar cuantos documentos se asocian a una determinada faceta, pero también podemos estar interesados en asociar a cada faceta la suma de los scores de los documentos recuperados).

Para realizar el conteo se utiliza la clase FastTaxonomyFacetCounts que recupera el número de documentos que emparejan con cada faceta y las devuelve en un objeto de tipo Facets que podrá ser utilizado para mostrar los elementos..

```
Facets facetas = new FastTaxonomyFacetCounts(taxoReader, fconfig, fc);
```

4.1. Mostramos los resultados obtenidos para las facetas

Una vez que tenemos el objeto de tipo Facets podemos realizar las siguientes acciones:

- List<FacetResults> getAllDims (int topN) para devolver una lista con los topN FacetResults (estructura que entre otros contiene las etiquetas y los contadores -valores- asociados) de cada dimensión o categoría, ordenadas por el número de hits con los que emparejan. Se suele utilizar cuando no todos los documentos son clasificados bajo las mismas categoría, como por ejemplo en e-comercio donde se dispongan de múltiples tipos de productos.
- FacetResults getTopChildren(int topN, String dim, String... path) Devuelve los topN FacetResults de una determinada categoría (podemos dar el camino hasta la misma). Por ejemplo, clasificamos las asignaturas bajo la dimensión "Titulacion", y permitimos la jerarquía "Grado/Especialidad/" las siguientes llamadas nos permiten recuperar las facetas para todas las titulaciones, o sólo para el "Grado en Informatica".

```
getTopChildren(10, "Titulacion");
getTopChildren(10, "Titulacion", "Grado en Informatica");
```

Number getSpecificValue(String dim, String... path)
 Devuelve el número (o valor numérico computado) para una categoría concreta.

El siguiente ejemplo nos muestra como mostrar los distintos resultados obtenidos. Para ello, podemos recorrer la lista de FacetsResult pudiendo obtener entre otros datos, el nombre de la categoría o dimensión (por ejemplo, Profesorado) y para cada dimensión consultar para cada una de sus etiquetas (nombre del profesor) y valor (número de asignaturas que emparejan con la consulta)

```
1 Facets facetas = new FastTaxonomyFacetCounts(taxoReader, fconfig
      , fc);
3 // Iterando sobre las facetas obtenidas //
5 List < FacetResult > TodasDims = facetas.getAllDims(100);
6 System.out.println("Categorias totales" + TodasDims.size());
7 for (FacetResult fr: TodasDims) {
               System.out.println("Categoria " + fr.dim);
               for (LabelAndValue lv : fr.labelValues){
9
10
                   System.out.println("
                                            Etiq: " + lv.label + ",
                      valor (#n)->" + lv.value);
11
               }
12
           }
13
14 FacetResult fresult = facetas.getTopChildren(10, "Profesores"));
```

Así, una posible salida de este código ante una consulta sería:

```
1 Categorias totales 3
2 Categoria Centro
       Etiq: ETSII, valor (\#n)->4
      Etiq: FacDoc, valor (#n)->1
  Categoria cuatrimestre
      Etiq: segundo, valor (\#n) -> 3
6
      Etiq: primero, valor (#n)->2
7
  Categoria Profesores
9
       Etiq: Silvia Acid, valor (#n)->4
10
      Etiq: Juan Huete, valor (#n)->2
      Etiq: Nicolas Perez, valor (#n)->1
11
12
       Etiq: Andres Cano, valor (#n)->1
       Etiq: Francisco Cortijo, valor (#n)->1
```

4.2. Otros mecanismos de agregación para las facetas

Como hemos comentado, en algunas aplicaciones es posible que deseemos considerar otros mecanismos -mas allá del simple conteo- para cuantificar la importancia de una faceta ante la consulta. Veremos algunos ejemplos:

12

4.2.1. Acumulando la relevancia

El primero de ellos nos permite acumular los scores de los documentos recuperados por la consulta que emparejan con las facetas. Para ello, en primer lugar debemos de indicarle al FacetCollector que también almacene los valores de los scores, pasándolo el valor true al constructor. Una vez que lo tenemos, debemos de utilizar la clase TaxonomyFacetSumValueSource que se encarga de sumar los valores indicados en el último parámetro, de tipo DoubleValuesSource. Finalmente las etiquetas se ordenan por dicho valor.

Obteniendo la siguiente salida, donde podemos ver el cambio en el orden con respecto al ejemplo anterior:

```
1 Sumamos scores
2 Etiq: Silvia Acid, valor (#n) ->1.4176474
3 Etiq: Juan Huete, valor (#n) ->0.76920474
4 Etiq: Andres Cano, valor (#n) ->0.3643428
5 Etiq: Francisco Cortijo, valor (#n) ->0.3643428
6 Etiq: Nicolas Perez, valor (#n) ->0.35077688
```

También podemos utilizar como información para acumular aquellos NumericDocValuesField¹ (https://lucene.apache.org/core/7_1_0/core/

¹Incluso podríamos utilizar string indexando como SortedDocValuesField o SortedSetDocValuesField, pero necesitariamos de un mecanismo para convertir ese string a un valor numérico.

org/apache/lucene/document/Field.html) como muestra el siguiente ejemplo:

```
1 // En indexacion ....
2 Document doc = new Document();
3
4 doc.add(new NumericDocValuesField("nota",10));
6 indexWriter.addDocument(doc);
8 // En consulta
10 facetas=new TaxonomyFacetSumValueSource(taxoReader, fconfig,
      fc_Score , DoubleValuesSource.fromIntField("nota"));
11 System.out.println("Sumamos nota");
12 FacetResult frsuma = facetas.getTopChildren(10, "Profesores");
13 for (LabelAndValue lv: frsuma.labelValues) {
14
        System.out.println("
                                 Etiq: " + lv.label + ", valor (#n)
            ->" + lv.value);
15 }
```

con la salida

```
Sumamos nota

Etiq: Silvia Acid, valor (#n) ->25.0

Etiq: Juan Huete, valor (#n) ->12.0

Etiq: Andres Cano, valor (#n) ->8.0

Etiq: Francisco Cortijo, valor (#n) ->8.0

Etiq: Nicolas Perez, valor (#n) ->5.0
```

4.2.2. Agrupando valores por rango definidos por el usuario

Otra potencialidad que nos da Lucene es agrupar los valores almacenados en los DocValues por rangos definidos por el usuario como nos muestra el siguiente ejemplo.

```
LongRange[] ranges = new LongRange[3];
ranges[0] = new LongRange("10-30", 10L, true, 30L, false);
ranges[1] = new LongRange("21-59", 31L, true, 60L, false);
ranges[2] = new LongRange("60-100", 60L, true, 100L, true);
```

5. Filtrando la búsqueda

En una aplicación típica que use facetas, podemos ver una columna a la izquierda con los conteos del número de documentos recuperados que empareja con cada faceta (como hemos ilustrado en los casos anteriores). Esto nos indica cuántos documentos podemos recuperar si restringimos (DrillDownQuery) la búsqueda a esa faceta, tras añadirla como restricción a la búsqueda actual. Por tanto, DrillDown añade filtros a la búsqueda, pudiendo utilizar múltiples categorías (OR lógico)

```
1 Query q1 = new MatchAllDocsQuery();
3 DrillDownQuery ddq = new DrillDownQuery (fconfig, q1);
4 ddq.add("Centro", "ETSII");
5 //ddq.add("Profesores", "Juan Huete"); //Icluimos una segunda
      restriccion OR
7 | System.out.println(" Filtramos query[ "+ ddq.toString() +" ]");
9 FacetsCollector fc1 = new FacetsCollector();
10 TopDocs td2 = FacetsCollector.search(searcher, ddq, 10, fc1);
                            Total hits = " + td2.totalHits);
11 System.out.println("
12 Facets fcCount2 = new FastTaxonomyFacetCounts(taxoReader,
      fconfig, fc1);
13 List < Facets Results > all Dims = fcCount2.getAllDims(100);
14
15
16 System.out.println("Categorias "+ allDims.size());
17 for (FacetResult fr : allDims) {
```

```
18
19
      System.out.println("Dimension " + fr.dim);
      for (LabelAndValue lv: fr.labelValues){
20
          System.out.println("
                                   "+lv.label + ":: #->" + lv.value)
21
22
      }
23 }
24
25 for (ScoreDoc scoreDoc : td2.scoreDocs) {
            doc = searcher.doc(scoreDoc.doc);
26
            System.out.println(" Docs Score -> " + scoreDoc.score +
27
                       + doc.get("asignatura"));
28
           }
```

Si hacemos esto aquellas facetas que no concuerdan dentro de la opción de búsqueda desaparecen, por ejemplo se eliminan del árbol los resultados asociados a centros distintos de la ETSII, lo cual puede parecer normal. Pero si este proceso lo repetimos reiteradamente sobre las distintas opciones de búsqueda, al final si queremos deshacer algunos de los cambios nos resultará imposible, teniendo que volver a empezar la navegación desde el principio o utilizar el botón de vuelta atrás.

Una alternativa es utilizar DrillSideWays, que nos asegura no perder los conteos de otras facetas al restringir la búsqueda sobre una de ellas.

```
System.out.println("Ahora con DrillSideWays");
q1 = new MatchAllDocsQuery();
System.out.println("QUERY "+q1.toString());

dq = new DrillDownQuery(fconfig,q1);
dq.add("Centro", "ETSII");
System.out.println(" Filtramos query["+ dq.toString() +" ]");

prillSideways ds = new DrillSideways(searcher, fconfig, taxoReader);
DrillSidewaysResult dsresult = ds.search(dq,10);
System.out.println("dsw hits "+ dsresult.hits.totalHits);
System.out.println(dsresult.facets.getAllDims(10).toString());
```

6. Gestionando los árboles de facetas

En principio, todas las facetas son añadidas a un único índice, con nombre por defecto \$facets. Cada faceta de las añadidas es un nodo en dicho árbol. En cualquier caso, Lucene permite separar las facetas en distintos árboles, pudiendo tratar cada uno de ellos por separado, lo que permite que en tiempo de búsqueda nos podamos centrar un un subconjunto de las facetas de nuestro interés. Para ello, debemos de configurar las facetas de forma adecuada utilizando el método setIndexFieldName que recibe el nombre de la faceta y el nombre del índice (para las facetas que no se indica nada, se ubican dentro del índice \$facets)

```
1 FacetsConfig fconfig = new FacetsConfig();
2 | fconfig.setIndexFieldName("Titulacion", "facet_tit"); // Faceta
      Titulacion la gestiona dentro del factet_tit
3 | fconfig.setIndexFieldName("Alumnos", "facet_alumnos"); // Faceta
       alumnos la gestiona facet_alumnos
5 fconfig.setHierarchical("Titulacion", true);
6 . . .
7
8 Document doc = new Document();
9 doc.add(new FacetField("Centro", "ETSII")); //a $facets
10 doc.add(new FacetField("Cuatrimestre", "primero")); //a $facets
11 doc.add(new FacetField("Titulacion", "Grado en Informatica",
12
                                         "Sistemas de Informacian",
13
                                         "Obligatoria")); //a
                                            facet_tit
14 doc.add(new NumericDocValuesFacetField("alumnos", 35L)); // a
      facet_alumnos
```

En este caso, en tiempo de búsqueda podemos restringir el cálculo de las facetas a uno de los índices, indicándole al método que realiza los conteos con el que se desea trabajar, por defecto, sin no se indica nada, es \$facets.

```
// Para buscar en $facets
Facets facets = new FastTaxonomyFacetCounts(taxoReader, fconfig, fcollector);

// Para buscar en "facet_tit"
Facets ftit = new FastTaxonomyFacetCounts("facet_tit", taxoReader, fconfig, fcollector);

// Listamos categorias de nivel Titulacion
System.out.println(ftit.getTopChildren(3, "Titulacion").toString());

// Listamos categorias de nivel Titulacion/GradoEnInformatica
System.out.println(ftit.getTopChildren(3, "Titulacion", "Grado en Informatica").toString());
```