Búsqueda y Minería de Información

Enrique Cabrerizo Fernández
Guillermo Ruiz Álvarez

1 ANÁLISIS MÉTRICO

A continuación se realiza un análisis métrico comparativo para las tres colecciones de documentos: ClueWeb-1K, ClueWeb-10K y ClueWeb-100K. El código utilizado para obtener los resultados es reutilizado de la práctica anterior, añadiendo el nuevo buscador a la clase SearcherTest implementada.

1.1 RESULTADOS OBTENIDOS

1.1.1 ClueWeb-1K

En la siguiente tabla se muestran los valores promedio de las métricas de precisión P@5 y P@10 para la colección de 1K documentos.

Valores promedio									
Buscador	**TF-IDF Searcher**			**Literal Searcher**			**Proximal Searcher**		
Índice	**Basic**	**Stop**	**Stem**	**Basic**	**Stop**	**Stem**	**Basic**	**Stop**	**Stem**
P@5	0,2	0,16	0,12	0,36	0,28	0,28	0,36	0,36	0,32
P@10	0,16	0,14	0,12	0,34	0,34	0,36	0,4	0,4	0,36

1.1.2 ClueWeb-10K

En la siguiente tabla se muestran los valores promedio de las métricas de precisión P@5 y P@10 para la colección de 10K documentos.

Valores promedio									
Buscador	**TF-IDF Searcher**			**Literal Searcher**			**Proximal Searcher**		
Índice	**Basic**	**Stop**	**Stem**	**Basic**	**Stop**	**Stem**	**Basic**	**Stop**	**Stem**
P@5	0,06	0,04	0	0,14	0,1	0,08	0,22	0,22	0,16
P@10	0,07	0,07	0,04	0,12	0,13	0,14	0,2	0,21	0,2

1.1.3 ClueWeb-100K

En la siguiente tabla se muestran los valores promedio de las métricas de precisión P@5 y P@10 para la colección de 100K documentos.

Valores promedio									
Buscador	**TF-IDF Searcher**			**Literal Searcher**			**Proximal Searcher**		
Índice	**Basic**	**Stop**	**Stem**	**Basic**	**Stop**	**Stem**	**Basic**	**Stop**	**Stem**
P@5	0,16	0,15	0,13	0,20	0,18	0,17	0,21	0,19	0,18
P@10	0,15	0,14	0,13	0,18	0,18	0,17	0,23	0,21	0,20

1.2 ANÁLISIS DE LOS RESULTADOS

Para las tres colecciones de documentos, los resultados obtenidos para el buscador proximal mejoran con respecto al resto de buscadores considerablemente, llegando a mejorar incluso un 15% en algunos casos (colección de 10K documentos).

El motivo es que, como ya sabemos, este buscador le da más puntuación a los documentos que con mayor proximidad contiene los términos de la búsqueda.

Por el contrario, el buscador literal es más restrictivo, obligando a que los términos aparezcan consecutivamente. Por otro lado, el buscador TF-IDF no pone estas restricciones, sin embargo, la puntuación que otorga a los documento es independiente de la relación entre los términos, ya que sólo tiene en cuenta la frecuencia de los términos y el tamaño de los documentos.

2 ALGUNOS EJEMPLOS

A continuación se muestran ejemplos de casos de búsqueda. Todos ellos se hacen sobre el conjunto de mil documentos, debido a que el funcionamiento es igual sobre todas las colecciones.

Para la consulta "obama family tree" obtenemos:

```
Search time: 204.97235 milliseconds

Showing top 5 documents:

ID: 319     Name: clueweb09-en0010-79-2218.html Score: 30.09

ID: 121     Name: clueweb09-enwp00-04-9625.html Score: 8.81

ID: 16      Name: clueweb09-enwp00-39-9864.html Score: 8.81

ID: 159     Name: clueweb09-enwp00-00-9498.html Score: 8.81

ID: 539     Name: clueweb09-enwp00-41-6215.html Score: 8.81
```

Se puede observar que las puntuaciones son bastante altas debido a que en el sumatorio del algoritmo de puntuación se incluye una cantidad muy alta de intervalos.

Si acudimos cualquiera de los documentos, vemos que la cantidad de términos que contienen es muy alta, así como la proximidad de los mismos, produciendo una puntuación como las obtenidas. Además, no se incluye la restricción de que los términos estén colocados de forma consecutiva, tal y como aparecen en la consulta, solamente se pide que aparezcan en el documento y que su proximidad sea notable, obteniendo de esta forma una cantidad considerable de documentos relevantes.

El buscador se ha implementado de tal manera que, dado que la búsqueda proximal con un sólo término no tiene sentido, en caso de que se realice una consulta de un

sólo término, se crea una instancia del buscador TF-IDF para resolver la búsqueda. Esto se puede observar en el siguiente ejemplo.

Para la consulta "obama" obtenemos:

```
Search time: 20.468608 milliseconds

Showing top 5 documents:

ID: 786     Name: clueweb09-en0001-02-21241.html     Score: 0.16

ID: 319     Name: clueweb09-en0010-79-2218.html      Score: 0.16

ID: 959     Name: clueweb09-en0011-02-12744.html     Score: 0.14

ID: 846     Name: clueweb09-en0010-57-32937.html     Score: 0.13

ID: 969     Name: clueweb09-enwp01-39-17275.html     Score: 0.12
```

Se puede observar que se obtienen los mismos resultados que con el buscador TF-IDF.

En caso de introducir una consulta más larga, se puede observar que las puntuaciones disminuyen, al estar añadiendo la siguiente restricción: la cantidad de términos que han de encontrarse en los documentos es mayor, y además buscamos proximidad entre una cantidad mayor de términos. Esto se puede observar en el siguiente ejemplo:

```
Search time: 163.750155 milliseconds

Showing top 5 documents:

ID: 881     Name: clueweb09-en0006-85-33176.html     Score: 7.60

ID: 865     Name: clueweb09-enwp00-73-20643.html     Score: 3.12

ID: 600     Name: clueweb09-enwp00-62-20587.html     Score: 3.12

ID: 409     Name: clueweb09-enwp00-67-20889.html     Score: 3.12

ID: 4       Name: clueweb09-enwp03-36-9316.html      Score: 3.12
```