Universitat Politècnica de Catalunya

Facultad de Informática de Barcelona

User Relevance Feedback

Búsqueda y Análisis de Información Masiva

Autores

Nicolas Llorens

Ricardo Lopez

10/11/2024

Contents

1 Implentación

En este proyecto, hemos implementado un sistema de retroalimentación de relevancia sobre Elasticsearch. La principal función de la modificación es modificar la consulta basa en la regla de Rocchio, esta regla ajusta los pesos de los terminos según la formula para mejorar la consulta.

$$\mathbf{Q}' = \alpha\mathbf{Q} + \frac{\beta}{k}\sum_{i=1}^{k}\mathbf{D}_i, \tag{1}$$

donde:

- $\mathbf{Q}'$: es la consulta actualizada.
- α: es el peso de la consulta original.
- $\mathbf{Q}$: query Original
- β: Peso de los documentos relevantes.
- k es el número de documentos relevantes considerados.
- $\mathbf{D}_i$ es el vector de términos del documento i.

La implementación esta hecha en python de manera que los valores de manera que los parámetros α, β, el número de documentos relevantes (k) y el número máximo de términos (R) y el **nrounds** el número total de iteraciones del bucle. Son parámetros que los podemos modificar para comprobar cual es el impacto de cada uno de los parámetros en la formual y/o en la respuesta de los documentos.

Poca dificultad a la hora de implementar el algoritmo, la mínima dificultad es entender la sintaxis de python

2 Experimentacion

Aqui modificaremos los diferentes valores para observar los diferentes resultados en la obtención de los documentos según la formula inicial y los parámetros de entrada de la función. Y tambien veremos como se modifica la query inicial según los parametros de α, β,

2.1 Valores de nRounds

En este caso modificaremos el numero de iteraciones del algoritmo. La hipótesis que plantemos es comprobar como no es necesario un número muy alto de iteraciones para producir la query. Los valores fijos son $\mathbf{k} = 1000$, $\mathbf{R} = 4$, $\alpha = 1$, $\beta = 0.1$ y nround $= 100$

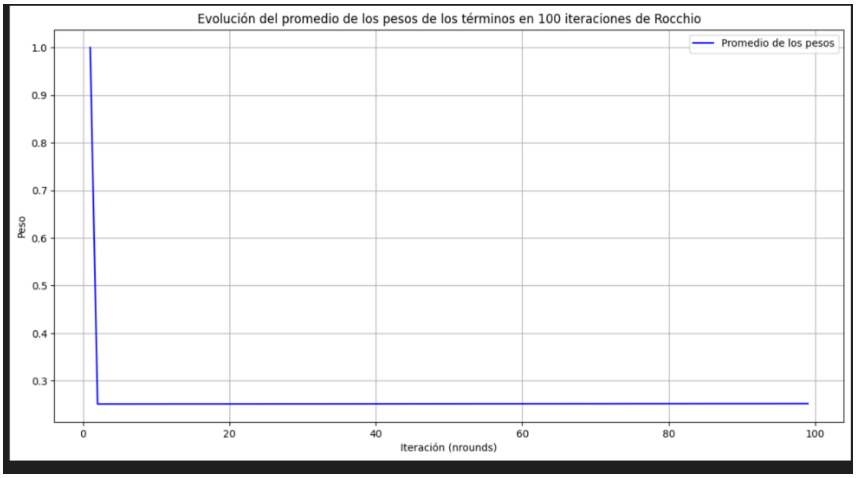

Figure 1: Promeedio de los pesos

Tal como vemos en la imagen anterior, podemos corroborar lo que planteábamos en la hipótesis, no es necesario tener un alto número de iteraciones dado que llega un punto que los valores se repiten.

2.2 Valor de R

En este caso modificaremos el numero de terminos de la query. La hipotesis que plantemos es que a medida que el numero es más alto el número de documentos disminuye.

Los valores fijos son $\mathbf{k} = 1000$, $\mathbf{nrounds} = 1$, $\alpha = 1$, $\beta = 0.1$

Tal como hemos dicho en la hipótesis, la gráfica nos confirma que a medida que el valor de R crece, el número de documentos que obtenemos es bajo o casi nulo. Podemos observar que siendo R ¡ 10 el número de documentos es 0. Esto es bastante lógico dado que a medida que la query crece se vuelve mucho más estricta por lo tanto es más difícil obtener un documento que coincida con la query generada.

2.3 Valor de α

En este caso miraremos para los valores 0.5,1.0,2.0,5.0,10.0 para α. Dado que este parámetros impulsa los terminos de la query, por lo tanto lo que queremos comprobar

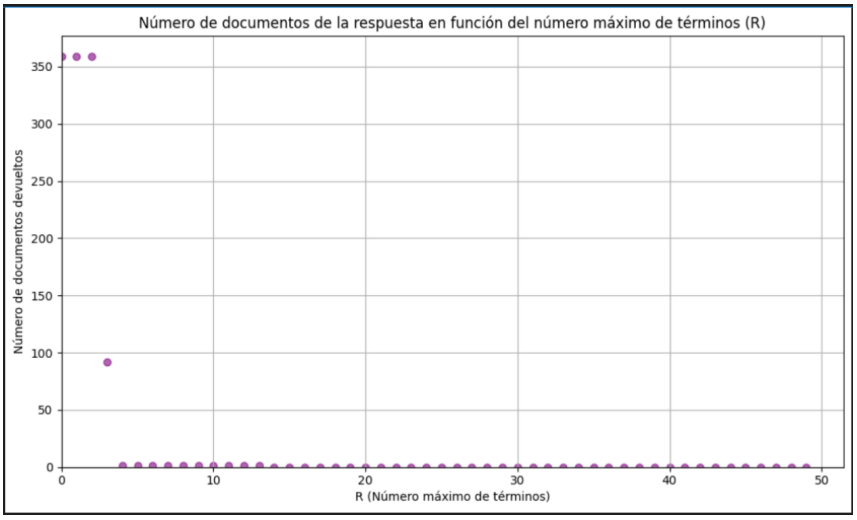

Figure 2: Documetos por Valores de R

es como los termintos principales de la query tienen un valor más alto en la query.

Los valores fijos son $\mathbf{k} = 1000$, $\mathbf{R} = 4$, $\mathbf{k} = 1000$, $\beta = 0.1$ y nround $= 5$

Table 1: Términos de la consulta actualizada para diferentes valores de alpha.

Alpha	**Términos de la consulta actualizada**
0.0	Toronto^{3}, detroit^{5}
0.5	detroit$^{0.4304}$, Toronto$^{0.2560}$, toronto$^{0.0030}$, leafs$^{0.0021}$
1.0	detroit$^{0.8592}$, Toronto$^{0.5133}$, toronto$^{0.0030}$, leafs$^{0.0022}$
2.0	detroit$^{1.7167}$, Toronto$^{1.0278}$, toronto$^{0.0030}$, leafs$^{0.0022}$
5.0	detroit$^{4.2891}$, Toronto$^{2.5712}$, toronto$^{0.0030}$, leafs$^{0.0022}$
10.0	detroit$^{8.5766}$, Toronto$^{5.1437}$, toronto$^{0.0030}$, leafs$^{0.0022}$

Tal como vemos en la tabla anterior, los valores de la query original tiene unos valores demasiado altos.Esto era algo esperable dado que el peso de los terminos de la query orginial es directamente proporcional al valor de alpha

2.4 Valores de β

En este caso miraremos para los valores 0.5,1.0,2.0,5.0,10.0 para β. Dado que este parámetros impulsa los terminos de los documentos, por lo tanto lo que queremos comprobar es como los termintos de los documentos tienen un valor más alto en la query.

Los valores fijos son $\mathbf{k} = 1000$, $\mathbf{R} = 5$, $\alpha = 0.1$, $\mathbf{k} = 1000$ y nround $= 5$

Table 2: Términos de la consulta actualizada para diferentes valores de beta.

Beta	**Términos de la consulta actualizada**
0.0	Toronto, nyc, water
0.5	water$^{0.0588}$, nyc$^{0.0574}$, Toronto$^{0.0573}$, mwra$^{0.0016}$, dept$^{0.0007}$
1.0	water$^{0.0598}$, nyc$^{0.0570}$, Toronto$^{0.0569}$, mwra$^{0.0032}$, dept$^{0.0013}$
2.0	water$^{0.0617}$, nyc$^{0.0562}$, Toronto$^{0.0561}$, mwra$^{0.0064}$, dept$^{0.0027}$
5.0	water$^{0.0670}$, nyc$^{0.0536}$, Toronto$^{0.0533}$, mwra$^{0.0156}$, dept$^{0.0065}$
10.0	water$^{0.0741}$, nyc$^{0.0487}$, Toronto$^{0.0483}$, mwra$^{0.0295}$, dept$^{0.0123}$

En este caso no obtenemos el valor esperdao, dado quue los valores de la query siguen teniendo un valor más algo que los nuevos valores(los terminos de los documentos). En este caso creemos que es que los terminos de la query deben ser muy repetidos en los documentos y por este motivo no han sido remplazados.

2.5 Valores de K

Los valores fijos son **nround** $= 5$, $\mathbf{R} = 4$, $\alpha = 1$, $\beta = 0.1$

La k nos indica los documentos relevantes que consideremos a la hora de aplicar la query. Plateamos que por normal general el número de documentos que obtenemos dependerá siempre de este número.

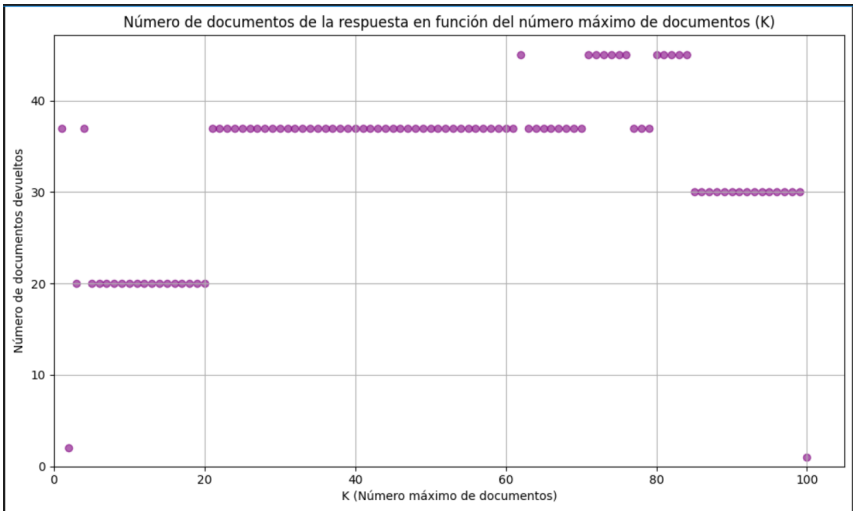

Figure 3: Documetos por Valores de K

Tal como vemos en la imagen a medida que el valor de K aumenta el numero de documentos que obtenemos también aumenta. Otra observación es que los valores se aplanan a partir de un cierto valor de K y luego decae en a un nivel que ya no se obtienen resultados

3 Conclusiones

A partir de los experimentos realizados, se pueden extraer las siguientes conclusiones clave sobre la influencia de los parámetros en la aplicación de la regla de Rocchio:

- **Iteraciones (nRounds)**: Se observó que, al aumentar el número de iteraciones, los pesos de los términos de la consulta tienden a estabilizarse. Esto confirma la hipótesis de que no es necesario un número alto de iteraciones para obtener una consulta refinada y estable. Más allá de un umbral (alrededor de 20-30 iteraciones), el proceso de actualización alcanza un estado de equilibrio, donde los cambios en los pesos son mínimos o inexistentes.
- **Número de términos (R)**: Los resultados mostraron que un mayor número de términos en la consulta actualizada conduce a una disminución en el número de documentos devueltos. Esto es consistente con la hipótesis de que una consulta más larga y específica hace que la búsqueda sea más restrictiva. La gráfica confirma que a partir de un cierto valor de R, el número de documentos devueltos se reduce drásticamente, destacando la necesidad de encontrar un balance óptimo entre el tamaño de la consulta y la cantidad de resultados.
- **Peso de la consulta original (α)**: Los experimentos demostraron que a medida que el valor de α aumenta, los términos de la consulta original mantienen un peso considerablemente alto en la consulta actualizada. Esto era esperado, ya que α influye directamente en la relevancia de los términos originales. La tabla de resultados muestra cómo, al incrementar α, los términos de la consulta inicial dominan la actualización, confirmando su impacto en mantener la dirección original de la búsqueda.
- **Peso de los documentos relevantes (β)**: Aunque se esperaba que un aumento en β diera mayor importancia a los términos de los documentos relevantes, los resultados indicaron que los términos de la consulta original aún mantenían un peso relativamente alto. Esto sugiere que los términos originales tienen una presencia significativa en los documentos relevantes y que β debe ajustarse cuidadosamente para lograr un balance adecuado. La repetitividad de los términos de la consulta en los documentos puede ser un factor importante en este comportamiento.
- **Número de documentos relevantes (K)**: Se observó que al incrementar K, el número de documentos devueltos inicialmente aumentaba, pero a partir de un cierto punto, la cantidad de documentos se estabilizaba o incluso disminuía. Esto sugiere que, aunque agregar más documentos puede enriquecer la actualización de la consulta, existe un límite en el que la relevancia marginal de nuevos documentos se reduce, y la consulta se vuelve demasiado general o incorpora ruido.

En resumen, la experimentación ha permitido validar las hipótesis iniciales y destacar la importancia de una parametrización adecuada en la aplicación de la regla de Rocchio. Cada parámetro tiene un papel fundamental en el comportamiento de la actualización de la consulta, y un ajuste cuidadoso es esencial para maximizar la relevancia y precisión de los resultados obtenidos.