Ivan Cala Mesa

Informe CAIM sesión 3

El siguiente informe trata sobre un estudio realizado a una implementación propia del algoritmo de Rocchio, formulado como método basado en **relevance feedback**, y/o suposición de que la mayoría de los usuarios tienen una concepción general de lo que es la 'relevancia' para los documentos. El algoritmo original utiliza por un lado, una lista de K documentos más relevantes, y otra para los menos relevantes, siendo ambos parte del cálculo. Para esta implementación sin embargo, nos remitiremos a solo utilizar los relevantes. El objetivo es fundamentalmente conseguir, iteración tras iteración (nrows veces) una query cada vez más y más específica y en teoría, relevante para el contenido que creemos que el usuario quiere y busca.

Dicho lo cual, en la experimentación del siguiente informe trabajaremos viendo el comportamiento del algoritmo modificando los diferentes parámetros modificables que tenemos dentro de la implementación original y en la propia. Estos son nhits(o k), nrows, R, α i ß.

Fase de pruebas

Realizaremos una primera query con los siguientes parámetros:

```
ß = 1, α = 2, R = 4, nrows = 10, nhits = 5
Initial query: soccer
Result query: ['soccer^0.6679510968514077', 'volleyball^0.3935251045598352',
'gameboy^0.40236525028286974', 'genesis^0.2667626217529411']
1 Documents
```

A continuación, valoramos y decimos, qué ocurre si modificamos el parámetro alfa para que este sea inferior? El comportamiento esperado es una reducción del peso de las palabras que formaban parte de la query inicial.

```
ß = 1, α = 1, R = 4, nrows = 10, nhits = 5
Initial query: soccer
Result query: ['soccer^0.23788836121049203', 'gameboy^0.37204929049227886',
'volleyball^0.3638752000563102', 'genesis^0.23610766255059568']
1 Documents
```

¿Qué ocurre respecto a las iteraciones? ¿De qué manera afectan al resultado? Probamos a ver los resultados incrementando a 100 el número de iteraciones, y viendo la fluctuación de los resultados en el proceso.

```
ß = 1, α = 2, R = 4, nrows = 10, nhits = 100
Initial query: soccer

Result query iteración 10: ['soccer^0.6679510968514077',
'volleyball^0.3935251045598352', 'gameboy^0.40236525028286974',
'genesis^0.2667626217529411']
Result query iteración 13: ['soccer^0.5266101596557938',
'volleyball^0.4488961508718953', 'genesis^0.29611821197913984',
'gameboy^0.45898015146610255']
```

```
Result query iteración 25: ['gameboy^0.552984697012075',
'genesis^0.34486080406662517', 'volleyball^0.5408353742244879',
'soccer^0.29192380509923155']
Result query iteración 50: ['gameboy^0.5771748887265228',
'genesis^0.35740373627466937', 'volleyball^0.5644940965347723',
'soccer^0.23153195886645522']
Result query iteración 100: ['gameboy^0.5781604184127542',
'genesis^0.35791474636459925', 'volleyball^0.5654579736900399',
'soccer^0.22907154193502616']
```

Probamos con otra palabra bien conocida en el imaginario popular, y realizamos una primera prueba para ver cómo de similares son para nosotros los términos adicionales obtenidos en la query resultante.

```
ß = 1, α = 2, R = 4, nrows = 10, nhits = 5
Initial query: nintendo
Result query: ['tetris^0.3757067313081599', 'nintendo^0.6080897149457993',
'castlevania^0.43518343228368517', 'nemesis^0.40657581068660864']
1 Documents
```

Muy bien, pero qué ocurre pues si aumentamos la proporción de peso del parámetro ß con respecto al α?

```
ß = 3, α = 1, R = 4, nrows = 10, nhits = 5
Initial query: nintendo
Result query: ['tetris^0.42994770016047956', 'nemesis^0.46909652935430995',
'nintendo^0.4007171454095585', 'castlevania^0.4980110822258524']
1 Documents
```

Y hasta ahora hemos mantenido constante el parámetro hits. Este parámetro aumenta el tamaño del subconjunto de documentos que utilizamos como ‘relevantes’ a partir de cada consulta. De qué manera nos afecta a las palabras resultantes?

```
ß = 1, α = 2, R = 4, nrows = 10, nhits = 20
Initial query: nintendo
Result query: ['game^0.06845651584364312', 'games^0.0709551109402931',
'nintendo^1.2497938052477002', 'asking^0.2352496731133377']
3 Documents
```

Por último, queda ver qué ocurre si modificamos el parámetro R. Dentro de nuestra implementación, es el número de palabras ‘relevantes’ que permitimos que tenga la query resultante. Qué resultado tiene ese comportamiento?

```
ß = 1, α = 2, R = 10, nrows = 10, nhits = 20
Initial query: nintendo
Result query: ['asking^0.03588840462775546', 'super^0.03018636370124934',
'tetris^0.02926369208991048', 'game^0.04208637942352312',
'shipping^0.028464414580921384', 'games^0.03999915381847896',
'offer^0.028234304748237345', 'castlevania^0.03389631567323369',
'sony^0.028044206496917627', 'nintendo^1.2913385858145054']
0 Documents
```

Ivan Cala Mesa

Conclusiones

Relativo al **parámetro α**, ha quedado demostrado que su incremento o aumento del peso relativo al ß, provoca un aumento del peso asociado en los términos que existían previamente en la query inicial, no por definición pero en general, en ‘detrimento’ del resto.

Asimismo, similar ocurre con el **parámetro ß**. Si este gana peso relativo respecto al α, los términos añadidos a partir de los documentos relevantes tendrán un peso relativo en la query aumentado.

En cuanto al **número de iteraciones**, observamos que se produce una fluctuación de los resultados significativa cuanto más anteriores sean estas, y se estabilizan los resultados (es decir, los pesos) conforme aumenta el número de las iteraciones, habiendo diferencias mínimas y poco visibles entre ellas.

Sobre la **R**, no hay demasiado que añadir además del hecho de que incrementa el número de palabras posibles contenidas en la query inicial. Hemos observado que el peso de cada una de estas palabras ‘adicionales’ es muy pequeño a nivel unitario, posiblemente por producirse esa partición sin incrementar el peso relativo de los documentos relevantes.

Para **nhits o k** hemos visto que se produce, en efecto, un cambio en las palabras resultantes entre querys de mismos parámetros. Al incrementar el rango de documentos que consideramos relevantes nos exponemos a que la terminología de entre todos ellos abarque más (mayor recall o exhaustividad) pero una precisión inferior.
Para el caso de la nintendo, vemos como las palabras fueron de ser juegos de por ejemplo la NES, a terminología genérica como es ‘game’ o ‘games’.