CAIM Laboratory

Session 3: User Relevance Feedback

2019-2020 Q1

Ruben Martinez Escobar
Pol Renau Larrodé

UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Facultat d'Informàtica de Barcelona

1. Experimentació 1: variació de paràmetres

En aquest apartat el que farem serà, donat una query que ens dona un nombre reduït de documents com a resultat de la query, anar modificant els paràmetres per veure quina és la seva funcionalitat i quina importància tenen amb el resultat final obtingut.

Durant tot aquesta part de l'experimentació utilizarem els següents paràmetres:

--alpha 1, --beta 1 -R 4, --k 5, --nrounds 10 i

--query run^2 gun^3

No entrem en detalls en la implementació per què ha estat seguir el guió de la pràctica i reutilitzar funcions anteriorment tractades en pràctiques posteriors.
Cal remarcar l'ús de diccionaris en lloc de llistes per a un benefici en temps d'execució.

1.1 *Alpha*

En aquest subapartat variem el paràmetre *alpha*, anem incrementant el valor d'alpha de 1 fins a 10.

El resultat obtingut no ha estat diferent, sinó que ens ha mostrat els 4 mateixos documents tant amb *alpha* = 1 fins a *alpha* = 10, passant per totes i cada una de les iteracions, no obstant el que sí que hem vist que ha variat es la puntuació donada als documents obtinguts. El score ha passat de tenir un valor entre [140,160], mentre que al llarg de cambiar el paràmetre alpha els valors oscil·len [80,105].

També observem que a major alpha, els valors de la query inicial tenen major importància, el motiu és que el paràmetre alpha incrementa la importància de la query inicial. Això també pot explicar perquè els scores són més reduïts, atès que li dona major importància als valors de la query inicial i crea una major diferència amb els nous termes creats per *Rocchio*, de tal manera que el score es veu ressentit.

1.2 *Beta*

En aquest subapartat altarem el paràmetre beta, incrementarem progressivament el seu valor de 1 fins a 10.

Degut a que tenim *alpha* = 1 < *beta* = 10, conforme més iteracions transcorren, el pes de la query inicial baixa i les paraules afegides prenen més pes. Això també comporta un increment en l'score considerable:

Iteració 10: (*beta* = 1) = 155.76738

Iteració 10: (*beta* = 10) = 182.12238

Per tant podem concloure, que el paràmetre beta influeix directament sobre les paraules afegides per Rocchio (quan més alt més pes prenen), això també explica

que l'score augmenti, degut a que mentre *Rocchio* afegeixi nous termes i s'augmenti el seu pes, l'score també augmentarà.

En aquest cas si que obtindrem un resultat diferent, amb *beta* = 1, obtenim 4 documents, en canvi amb *beta* = 10 obtenim 2 documents, això es perfectament explicable, degut a que al donar més pes a les paraules afegides (que són més que les de la nostra pròpia query (4 paraules afegides per *Rocchio* > 2 paraules inicials)) significa que la query a buscar és més complexa i difícil de satisfer.

1.3 *R*

En aquest subapartat tractem el paràmetre *R*, incrementarem progressivament *R* des de el seu valor inicial 4 fins a 10.

Podem observar que conforme el valor de *R* és incrementat el valor del score final és reduït, això pot ser perfectament degut a que com la query estarà conformada amb més termes i tenim *beta*=1, satisfer tots els termes nous afegits per *Rocchio* serà més complicat i conseqüentment l'score serà més baix. Un raonament similar el podem aplicar en el resultat; amb *R*=10 obtenim com a resultat 0 documents, això és degut a que la nostre query restringeix més les seves possibles solucions, degut a esta conformada per més paraules.

1.4 *Nrounds*

Variant el paràmetre *nrounds*, no hem obtingut grans canvis en el resultat, el que si que es veritat, es que al agregar més iteracions veiem que els scores varien, ja que treballem sobre la mateixa query, però més iteracions, amb el que veiem que la query va degenerant bastant si el *nrounds* pren un valor bastant elevat. En el exemple que hem analitzat, veiem que a part de que els scores pugen més d'un 600% per a nrounds a valor 100, podem veure que el paràmetre run ha estat eliminat de la query i substituït per un altre valor de major rellevància.

1.5 *K*

En aquest subapartat hem variat el paràmetre *k* de 5 a 505 en increments de 10 en el valor de *k*, el resultat obtingut ha estat que a mesura que incrementavem el valor de la variable, més resultats obteniem, tot i que els resultats obtinguts cada cop eren menys rellevants.

En aquest experiment, hem passat de tenir 3 documents rellevants a tenir-ne 68, això es degut a que augmentant k permet que hi hagi major *Recall*, no obstant disminuim la *Precision* atès que els documents obtinguts no tots són realment rellevants.

2. Variació de la query inicial modificant a punts extrems

En aquest apartat modificarem els 2 valors principals que tenen a veure amb la nova query resultant, aquests són *alpha* i *beta*. Per veure quin és l'impacte d'aquests els forçarem a tenir uns valors elevats per veure en el extrem quina és la seva aportació a la nova query.
Per a totes les experimentacions, usarem la query del apartat d'experimentació anterior.

2.1 Alpha 100

Hem modificat el paràmetre alpha a valor 100, el resultat obtingut és el següent:
els mateixos 3 documents, però amb oscil·lacions del score [65,90], és a dir com hem vist anteriorment el score es veu reduït a causa de que se li dóna molta importància als valors de la query original. Ja que els paràmetres inicials han passat a tenir una rellevància de 0.88 gun i de 0.55 run, bastant més incrementat que a l'inici amb *alpha* igual a 1, 0.33 i 0.16 respectivament.

2.2 Beta 100

Modificant el paràmetre *beta* fins a 100, hem obtingut 2 documents, i respecte a la query, veiem que el valor gun baixa fins a 0.31 i run ha desaparegut de la query. Això és degut a que altres paraules han tingut major importància que run i per això ha estat desbancat.

2.3 *Beta* i *aplha* 100

Al haver pujat els dos valors, el resultat obtingut és que el score de tots els documents és més elevat, això es degut a que els valors la query tant els nus com els originals tenen major rellevància que els que tenen amb *alpha* i *beta* igual a 1.

També podem apreciar que els documents obtinguts són els mateixos això es deu a que al haver modificat de la mateixa manera, els dos paràmetres la variació de la query doncs ha estat constant.

3. Baixem *R* a valor 1

Al baixar el valor de *R* fins al valor de 1, veiem que els resultats obtinguts són bastant desastrosos, ja que tenim com a resultat 93 documents.

Analitzant una mica més el resultat, veiem que la query només té un valor i aquest és gun, que te tota la importància, és a dir que el resultat que obtindrem seràn aquells 93 textos on apareix la paraula gun. Per tant podem veure que hem perdut el paràmetre run, i la creació de nous paràmetres, ja que ha estat eclipsats per la paraula gun i per tant aquells paràmetres no tindràn un impacte en el score d'aquells documents.

4. Conclusions

4.1 *Alpha i Beta*

Per aquests dos paràmetres hem pogut concloure amb l'experimentació que *Alpha* afecta directament a la importància de la query inicial i que *Beta* afecta directament a la importància dels termes afegits. Conclusió trivial degut a la propia fórmula de *Rocchio* pero que ha estat possible observar-la a la pràctica. Podem extreure també que quan més elevat sigui *Alpha*, el nostre *Recall* també serà més elevat degut a que en la sortida tindrà més pes les pròpies aparicions de la query inicial. En canvi, quan més gran sigui *Beta*, la nostre *Precision* serà més alta, degut a que prendran més importància els termes afegits a la query que intentaràn eliminar falsos positius intentant donar un bon contexte a la paraula.

4.2 *R*

Per l'experimentació realitzada hem trobat que *R* té un efecte directe entre *Recall* i *Precision*. Quan més elevat tinguem el paràmetre *R*, més *Precision* obtindrem i conseqüentment menys *Recall*. Això te sentit, degut a que es podran afegir més paraules a la query i d'aquesta manera elevar el contexte de la query inicial, però al ser més estricta la cerca perdrem documents, tot i que una gran part d'ells seran falsos positius.

4.3 *nRouds*

Degut a la nostre experimentació podem extreure que aquest paràmetre ha de estar ben fitat degut a que pot arribar a donar resultats finals no corresponents. Si *nRounds* es adaptat correctament a la query inicial trobarem una gran millora en els resultats, degut a que utilitzarà un bon nombre de iteracions per a posar en contexte la paraula i alliberar-nos de molts falsos positius, però a la vegada mantenint un bon nivell de *Recall* per a no derivar en reultats que no tenen res a veure amb la nostre query inicial.

4.4 *K*

Per aquest paràmetre em obtingut que afecta directament al *Recall* degut a que quan més alt és *K* més resultats obtenim, conseqüentment tenim un *Recall* més alt i un *Precision* més baix, ja que molts dels resultats són falsos positius que poc tenen a veure amb la query inicial.