Estensione dei metodi di ranking mediante analisi dell'interspaziatura fra occorrenze

Maria C. Daniele, Claudio Carpineto, and Andrea Bernardini

Fondazione Ugo Bordoni, Rome, Italy
mariac.daniele@gmail.com, carpinet@fub.it, aberna@fub.it

Abstract. L'analisi frequentistica delle occorrenze, tipica dei modelli di ranking di information retrieval, può essere integrata con l'analisi della spaziatura fra le occorrenze di una singola parola, mutuata dallo studio dei livelli di energia dei sistemi statistici di quanti disordinati. Queste due aree di ricerca sono fortemente interrelate, perché entrambe hanno l'obiettivo di assegnare dei pesi di rilevanza alle singole parole di un documento, e sembrano complementari, perché si basano su metodologie differenti. Tuttavia finora esse sono progredite in modo separato. L'obiettivo di questa ricerca è di favorire una loro riconciliazione. I contributi principali del lavoro sono tre: (a) estensione del metodo basato sull'interspaziatura mediante analisi di corpora, (b) verifica sperimentale che la pesatura quantistica è scorrelata da quella frequentistica, (c) studio della combinazione ottimale dei pesi quantistici e frequentistici ai fini del miglioramento delle prestazioni del ranking. Il risultato principale dei nostri esperimenti è che il metodo quantistico da solo non funziona bene, ma che il metodo combinato consente di migliorare in modo significativo le prestazioni del metodo classico frequentistico. Un ulteriore risultato riguarda le potenzialità di applicazione selettiva dei due metodi di pesatura: buone in funzione della lunghezza dei documenti recuperati, modeste rispetto alla difficoltà stimata delle interrogazioni.[1]

1 Introduzione

Ordinare i documenti di una collezione per pertinenza a fronte di una richiesta d'utente è il problema chiave dell'Information Retrieval. Nel corso degli ultimi decenni sono stati ideati numerosi modelli di ranking (vettoriale, probabilistico, basato sulla modellazione del linguaggio, o sullo scostamento dalla casualità), che tipicamente assegnano un punteggio o una probabilità a ciascun documento basandosi su una valutazione dell'importanza che i singoli termini dell'interrogazione rivestono nei documenti che li contiene. Le grandezze sulle quali si basano la maggior parte di questi modelli dipendono dalle frequenze con le quali i termini compaiono nei singoli documenti e nell'intera collezione. Coi

[1] Questo lavoro è basato sulla tesi di laurea magistrale in ingegneria informatica di Maria Daniele "Sperimentazione di tecniche d'Information Retrieval basate sulla Fisica dei Quanti", svolta presso la Fondazione Ugo Bordoni e discussa all'Università Roma Tre nel luglio 2011.

progressi degli ultimi anni però, i margini per ulteriori miglioramenti nelle tecniche tradizionali di ranking si sono ridotti: un avanzamento sostanziale ormai sarà difficile che avvenga senza un vero e proprio cambiamento di paradigma.

Parallelamente, nell'ultimo decennio si è sviluppato un ramo della ricerca riguardante l'estrazione delle parole rilevanti di un testo che prescinde dalla frequenza delle parole. Tale approccio, nato da studi sui livelli di energia dei sistemi statistici di quanti disordinati, si basa sull'analisi dell'interspaziatura fra le occorrenze di uno stesso termine. Un ruolo fondamentale è giocato dalle forze di attrazione e repulsione cui sono soggette le singole occorrenze di un termine. Più il termine è rilevante, maggiore è lattrazione fra sue occorrenze, quindi più tali parole si concentrano in aree determinate del documento, generando la formazione di clusters; viceversa, più un termine è comune e poco rilevante, più deboli sono queste forze, per cui il termine si distribuisce uniformemente lungo tutto il testo.

Ortuño et al. [6] sono stati i primi a mostrare che in un testo la distribuzione spaziale di una parola rilevante è molto diversa da quella corrispondente a una non rilevante, postulando un'analogia tra il linguaggio naturale e il linguaggio del DNA. In seguito, ci sono state altre proposte derivate da quella pioneristica di Ortuño et al., ad esempio [9], [5], e [1]. In [9] vengono accertate alcune limitazioni dell'indice di pesatura ideato da Ortuño et al. sulle quali ritorneremo in seguito. Una caratteristica importante di tutte queste tecniche quantistiche è che non serve una collezione esterna da analizzare: esse si basano esclusivamente sul contenuto dei singoli documenti.

Fra queste due aree di ricerca, quella frequentistica e quella quantistica, esiste una forte connessione, perché entrambe puntano ad assegnare un peso di rilevanza ai singoli termini di un documento. Tuttavia, esse sono state portate avanti in modo esclusivo nelle due comunità, di information retrieval e di fisica dei quanti, senza cercare di analizzare i rispettivi vantaggi e svantaggi o di combinarle per trovare un approccio più potente di quelli singoli. Da questa osservazione è scaturita la nostra ricerca. L'obiettivo è il tentativo di cominciare a riconciliare questi due approcci.

La prima area di intervento è stata l'estensione della pesatura quantistica con statistiche estratte da un corpus (considerando in particolare le variazioni della frequenza di ciascun termine rispetto all'insieme dei documenti), ai fini di premiare la capacità di discriminazione di un termine. Il secondo tema che è stato studiato è la complementarietà dei ranking prodotti dalle metriche quantistiche e da quelle frequentistiche (in particolare quelle basate su tf-idf). Infine, dati gli esiti deludenti dell'applicazione diretta delle metriche quantistiche, con o senza estensione, al ranking dei documenti, abbiamo fatto una serie di esperimenti per valutare l'efficacia di una combinazione dei due metodi. I risultati sono stati incoraggianti, con prestazioni migliori di quelle ottenibili con i metodi convenzionali di information retrieval (in particolare BM25).

Il seguito di questo articolo è strutturato nel seguente modo. Dopo avere ricapitolato l'approccio quantistico alla pesatura dei termini, così come presentato in letteratura, introduciamo la sua estensione basata sulle variazioni di frequenza

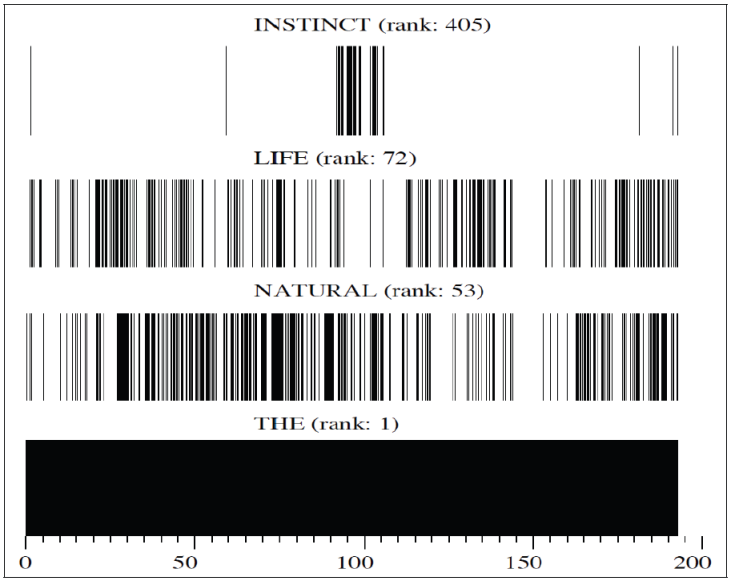

Fig. 1. Analisi spettrale e rank dei termini istinct, life, natural e the, estratte dal testo di Charles Darwin The Origin of Species" [6].

nel corpus. Successivamente, viene discussa la combinazione di pesatura quantistica estesa e pesatura tradizionale ai fini del ranking, presentando una serie di esperimenti su due collezioni campione. Infine, viene discussa la possibilità di un uso selettivo delle due metriche di pesatura guidato da lunghezza dei documenti e difficoltà delle interrogazioni.

2 Pesatura delle parole basata su interspaziatura fra occorrenze: σ_p

Il fenomeno della diversa distribuzione spaziale di parole rilevanti e non rilevanti è illustrato in Figura 1. I grafici sono relativi al testo The Origin of Species di Charles Darwin. Le occorrenze di parole rilevanti, come istinct, natural, e life, hanno distribuzione non omogenea e tendono a unirsi (fenomeno d'attrazione) formando dei clusters. Ciò accade indipendentemente dal numero di occorrenze, perché queste parole hanno nel ranking delle frequenze posizioni differenti. Simmetricamente, le occorrenze di parole non rilevanti, quali the (che è il termine con maggiore frequenza nel testo), sono equidistribuite.

Da un punto di vista fisico, nel caso di una parola chiave ogni livello d'energia attrae se stesso. La controparte linguistica di questo comportamento è che un termine rilevante è di solito il soggetto principale in un contesto locale di un documento, perciò occorre con maggiore frequenza in qualche area del testo e minore

in altre, generando il fenomeno di clustering. Invece, nel caso di parole non rilevanti tali livelli d'energia risultano scorrelati, corrispondentemente al fatto che tali parole si distribuiscono attraverso l'intero documento senza caratterizzarne in modo specifico nessuna parte.

Per quantificare questo fenomeno si utilizza il seguente approccio. Ogni occorrenza di un termine è considerata come un livello di energia che si trova all'interno di uno spettro energetico formato da tutte le occorrenze della data parola nel testo che si sta analizzando. Ogni valore del livello di energia è dato semplicemente dalla posizione che il termine ha nel documento. In pratica, per una data parola w, si estraggono le posizioni corrispondenti, creando il vettore x(w) = x_1 ,..., x_n (ogni x_i corrisponde ad un livello di energia). Ad esempio, nella frase "a great scientist must be a good teacher and a good researcher", per la parola "a" si estrae il vettore di posizioni x(a) =1,6,10. Si considera, poi, il vettore delle distanze d_i, dist(w) = d_1,...,d_n, con $d_i = x_{i+1}$ - x_i, tra le occorrenze consecutive della parola w e si calcola la corrispondente media delle distanze μ:

$$\mu \;=\; \frac{1}{n+1} \;\cdot\; \sum_{i=0}^{n} (x_{i+1} - x_i) \;=\; \frac{x_{n+1} - x_0}{n+1} \tag{1}$$

Denotando con p(x) la frequenza relativa di occorrenza di una data distanza x, la sua funzione di distribuzione integrata P_1(x) è:

$$P_1(x) \;=\; \sum_{x' < x} p(x') \tag{2}$$

Se la parola è distribuita in modo casuale (random) lungo il testo, la distribuzione P_1, nel limite continuo, sarà una distribuzione Poissoniana:

$$P_1(\mu) \;=\; 1 - exp(-\mu) \tag{3}$$

Se invece il termine respinge se stesso (quindi è distribuito uniformemente lungo tutto il testo) allora la sua distribuzione P_1 sarà più piccola di quella di Poisson per $\mu < 1$. Viceversa, se il termine attrae se stesso, P_1 sarà più grande della distribuzione di Poisson per brevi distanze (per un trattamento probabilistico piu approfondito si rimanda a [6]). Questo perché, come già osservato, le parole rilevanti di un testo compaiono generalmente in un ambito specifico, con oscillazioni apprezzabili fra i diversi ambiti.

Il calcolo della funzione di distribuzione P_1 per tutte le parole di un testo è molto oneroso dal punto di vista computazionale. Per questo motivo, al posto di P_1, viene utilizzata la deviazione standard s:

$$s \;=\; \frac{1}{n-1} \;\cdot\; \sum_{i=0}^{n} ((x_{i+1} - x_i) - \mu)^2 \tag{4}$$

Per eliminare la dipendenza dalla frequenza per differenti parole, la deviazione standard viene normalizzata rispetto al corrispondente valore medio delle distanze moltiplicato per $\sqrt{1-p}$:

$$\sigma_p \;=\; \frac{s}{\mu} \;\cdot\; \frac{1}{\sqrt{1-p}} \tag{5}$$

dove n è il numero di occorrenze della parola w all'interno del documento e N è il numero totale di parole nel testo. Questa funzione è molto semplice da calcolare e si dimostra robusta contro le oscillazioni. Le parole con il valore di σ_p più elevato saranno quelle più importanti.

3 Estensione della pesatura quantistica mediante analisi di corpora: σ^*

Numerose analisi e modifiche di σ_p sono state proposte. Uno dei lavori piu importanti è [9], dove sono evidenziati vari problemi. Il primo è che può accadere che parole comuni (rilevanti) abbiano alto (basso) valore di σ_p. Ad esempio, la parola you, che è indubbiamente un termine con scarso valore informativo, nella Bibbia ha valore 2,71 ed è classificata in posizione 550, che è molto elevata considerando che ci sono 12.910 parole distinte all'interno del libro; inoltre, la parola Sirach rispetto a you è più rilevante, ma ha solo un valore pari a 0,24 con corrispondente ranking di 9543. In secondo luogo, il metodo è alquanto instabile perché il valore di σ_p puo' essere influenzato fortemente dal cambio di una delle posizioni, specialmente in testi molto grandi. Ancora, ad alti valori non sempre corrisponde una distribuzione concentrata localmente. Ad esempio, la distribuzione 3,5,7,20 è clusterizzata nella regione [3,7], mentre per 3,5,18,20 si trovano due piccoli cluster in [3,5] e [18,20]; la metrica non fa distinzione tra questi due insiemi, a cui corrisponde lo stesso valore di σ_p. Un altro problema evidenziato, particolarmente importante per la nostra applicazione, è che la dimensione di un testo ha un forte impatto sulle prestazioni generali del sistema. Più il testo è breve, più l'indice classifica male le parole, collocando fra le prime posizioni quelle parole con frequenze molto basse, che all'interno del documento compaiono solamente pochissime volte e in posizioni molto ravvicinate tra loro (che in testi corti può accedere anche ad articoli o preposizioni).

I tentativi presenti in letteratura hanno cercato di presentare dei correttivi alla funzione σ_p senza però abbandonare l'assunzione di base, e cioè che l'ordinamento dei termini viene costruito soltanto analizzando il particolare testo che si sta considerando. Mentre questa assunzione può essere utile in determinate situazioni, sembra ragionevole cercare di estendere l'approccio quantistico utilizzando informazioni aggiuntive sulla importanza dei singoli termini basate sull'analisi di corpora, la disponibilità di corpora essendo oggigiorno vasta.

In particolare, noi proponiamo di correggere la metrica originaria con un fattore che abbia un duplice obiettivo: penalizzare le parole rare, perché in collezioni reali queste spesso costituiscono "rumore", e premiare le parole che riescono a discriminare meglio il testo in osservazione da altri testi, capacità questa che manca completamente nella pesatura quantistica. Il nostro approccio prende lo spunto da una metrica ben nota in information retrieval, la deviazione standard delle frequenze dei termini [7] Essa è definita nel seguente modo. Si consideri il

vettore delle frequenze f_i, freq(w) = $f_1, ..., f_{ND}$ relativo a una ad una parola w negli ND documenti della collezione., La media delle frequenze μ_f è:

$$\mu_f(w) \ = \ \frac{1}{ND} \ \cdot \ \sum_{i=1}^{n} \ f_i(w) \tag{6}$$

Si noti che ND è il numero totale di documenti della collezione: vengono considerate quindi anche le frequenze nulle, cioè i documenti in cui la parola non compare. La deviazione standard delle frequenze s_f sara' data da:

$$s_f(w) \ = \ \frac{1}{ND} \ \cdot \ \sum_{i=1}^{n} \ (f_i(w) - \mu_f)^2 \tag{7}$$

Chiaramente, s_f assumerà valori piccoli nel caso in cui la distribuzione di frequenza è uniforme (con f_i circa uguale a μ_f) o il termine appare in pochissimi documenti (essendo la maggior parte degli f_i uguali a zero e μ_f circa uguale a zero). Viceversa, s_f sarà grande quando la distribuzione di frequenza presenta forti variazioni a fronte di una frequenza media apprezzabile. Queste caratteristiche sembrano in grado di compensare i limiti di σ_p.

La deviazione standard può essere poi normalizzata rispetto al corrispondente valore medio delle frequenze μ_f, come visto in precedenza nel caso della pesatura quantistica:

$$\sigma_f \ = \ \frac{s_f}{\mu_f} \tag{8}$$

Nel complesso, questo approccio ha l'ulteriore vantaggio che il suo razionale è analogo a quello impiegato per sviluppare la funzione di pesatura originale σ_p. In questo caso i livelli di energia di una parola non corrispondono più alla posizione delle sue occorrenze in un testo, bensì alle frequenze in ciascun documento della collezione. Pertanto, l'analogia in questo caso è fra lo spettro di energia dei sistemi di quanti disordinati e l'insieme delle frequenze che una certa parola assume nella collezione.

La funzione di pesatura quantistica estesa σ^*, relativa ad una singola parola, è data dal prodotto di σ_p e σ_f:

$$\sigma^*(w) \ = \ \sigma_p(w) \ \cdot \ \sigma_f(w) \tag{9}$$

Per farsi un'idea più precisa delle caratteristiche dei termini estratti da testi lunghi mediante metriche frequentistiche e quantistiche, nonché del loro grado di complementarietà, abbiamo svolto il seguente esperimento. Come metrica di pesatura frequentistica abbiamo scelto tf-idf, che è semplice ed ha una valenza paradigmatica in information retrieval, nelle due versioni con e senza stop words (denotate rispettivamente tf-idf e tf-idf*), e come metriche quantistiche σ_p e σ^*. Abbiamo utilizzato come testo The Bible[2] e come corpus di riferimento per

[2] http://www.gutenberg.org/ebooks/10

calcolare i valori tf-idf e σ_f la collezione TREC WT10g, pre-elaborarata secondo quanto descritto nella Sezione 5.

I risultati sono mostrati in Tabella 1. La metrica tf-idf ha riportato nelle prime posizioni molte stop words arcaiche, poiché queste parole, oltre ad avere un valore elevato di tf nel testo originario, hanno conseguito anche un alto valore di idf nella collezione di riferimento (costituita da testi moderni). La metrica tf-idf* (cioè con rimozione di stop words) ha funzionato molto meglio, anche se ha restituito diversi termini generici nelle prime dieci posizioni, quali ad esempio "son", "king", "man", "land", "men". Le parole estratte da σ_p sembrano invece piú precise nel descrivere il contenuto della Bibbia, e consentono di identificare molti concetti e nomi propri importanti. Passando a σ^*, si nota che le parole diventano ancora piú specifiche (anche se non si tratta di termini rari in un testo come la Bibbia) e corrispondono a brani piú circoscritti all'interno del libro. Alcuni di questi termini hanno conseguito un alto valore di σ^* non solo in virtú della loro elevata concentrazione nella Bibbia ma anche per l'infrequenza con la quale appaiono nel corpus, secondo quanto giá evidenziato nella discussione di tf-idf. Nel complesso le parole estratte da σ^* sono meno caratterizzanti al livello del testo globale ma hanno sicuramente una maggiore capacità di discriminazione (ad esempio rispetto ad altri testi di carattere religioso).

rank	tf-idf	tf-idf*	σ_p	σ^*
1	unto (1,14)	lord (6,64)	jesus (24,35)	jesus (7,89)
2	shall (0,82)	god (3,12)	christ (18,31)	saul (4,97)
3	lord (0,81)	absalom (2,287)	paul (11,74)	absalom (4,97)
4	thou (0,71)	son (1,74)	peter (9,91)	jephthah (2,08)
5	thy (0,60)	king (1,55)	disciples (9,64)	jubile (2,08)
6	thee (0,50)	behold (1,46)	faith (9,39)	ascendeth (2,07)
7	him (0,42)	man (0,40)	john (9,14)	abimelech (1,96)
8	god (0,38)	judah (1,10)	david (8,75)	elias (1,95)
9	his (0,38)	land (1,05)	saul (8,70)	joab (1,86)
10	hath (0,31)	men (1,02)	gospel (8,01)	haman (1,82)

Table 1. Ordinamento e punteggi dei primi dieci termini della Bibbia secondo le metriche tf-idf (con e senza stop words), σ_p, e σ^*.

Se poi confrontiamo la somiglianza dei ranking prodotti dalle diverse metriche, ci accorgiamo che metriche frequentistiche e quantistiche restituiscono termini molto differenti. Considerando i primi 100 termini, ci sono 15 termini in comune fra σ_p e i due tf-idf, che scendono a due con σ^*, precisamente "jesus" e "saul". Inoltre, i pochi termini in comune hanno posizioni molto differenti. Ad esempio, la parola "jesus", che usando σ_p e σ^* compare nella prima posizione, viene invece classificata rispettivamente in quarantesima e quindicesima posizione da tf-idf e tf-idf*. Questi risultati indicano chiaramente che i ranking prodotti dai due tipi di ordinamento sono completamente scorrelati, in particolar modo quando si considera σ^* invece di σ_p, anche se bisogna sottolineare che i

nostri esperimenti sono stati effettuati su un testo lungo che non contiene errori. I testi che vengono tipicamente considerati nelle applicazioni di information retrieval sono invece brevi e rumorosi. Nelle prossime sezioni verranno presentati una serie di esperimenti con la seconda tipologia di dati.

4 Applicazione di σ^* al ranking

La metrica σ^* puo essere adoperata per fare il ranking di una collezione di documenti rispetto ad una interrogazione q, semplicemente sommando i valori relativi a tutti i termini di q presenti nel documento. Il punteggio $\sigma^*(d, q)$ conferito al generico documento d sarà dato da:

$$\sigma^*(d, q) \; = \; \sum_{w \in q} \sigma^*(w) \tag{10}$$

Vista la complementarietà delle metriche di pesatura quantistica e frequentistica, un approccio naturale è quello di cercare di integrare le due tecniche. Uno dei modi più intuitivi è fare una combinazione lineare dei punteggi assegnati dalle due tecniche a ciascun documento, preceduta da una normalizzazione degli stessi. Lo schema di normalizzazione adoperato è stato il seguente:

$$weight_{NORM} \; = \; \frac{weight - weight_{Min}}{weight_{Max} - weight_{Min}} \tag{11}$$

Il punteggio finale è dato da:

$$score \; = \; \alpha \cdot score_{BM25} \; + \; (1 - \alpha) \cdot score_{\sigma^*} \tag{12}$$

5 Esperimenti

Come collezioni di prova abbiamo utilizzato la WT10g e la Robust, due collezioni sviluppate in ambito TREC. La prima contiene oltre un milione e mezzo di pagine web, la seconda circa 500 mila documenti estratti da varie sorgenti informative. Per WT10g sono state utilizzate le 50 topics 501-550, mentre per la collezione Robust sono state usate 250 queries, le topics 301-450 che sono quelle del track "ad hoc" delle TREC 6-8, e le topics 601-700 del track " robust" delle TREC 2003-2004. Su queste collezioni è stata applicata una riduzione dello spazio dei termini, sia per rendere pi ù efficiente l'esecuzione degli esperimenti sia per cercare di migliorare l'efficacia attraverso una riduzione del rumore insito nei testi (abbreviazioni, refusi, ecc.). In particolare sono state rimosse le parole contenute in meno di dieci documenti, e quelle che contenevano pi ù di tre caratteri consecutivi uguali o che erano lunghe pi ù di venti caratteri. Tale procedimento ha portato l'insieme di documenti WT10g ad avere 435.744 invece di 5.167.898 di termini distinti (considerando anche i numeri interi), mentre per la Robust siamo passati da 1.178.484 a 485.326. Per quest'ultima collezione però abbiamo notato che per alcune topics c'era soltanto un documento che conteneva i

termini corrispondenti; eliminando la restrizione sulla frequenza dei documenti siamo passati a 835.760 termini.

Come sistema di indicizzazione e ricerca è stato utilizzato Lucene,[3] con l'estensione a BM25 fornita da Perez-Iglesias[4]. Lucene è stato adoperato sia per calcolare il ranking secondo BM25, sia per fornire i documenti di input (tutti quelli che contenevano almeno una parola dell'interrogazione) alle routine sviluppate per calcolare il ranking secondo σ^* e il successivo ranking integrato σ^* + BM25. Il valore di α usato negli esperimenti (= 0,8) è stato determinato utilizzando le topics 451-500, viste come al training set di WT10g.

In Tabella 2 sono riportate le prestazioni dei tre metodi di ranking, cioè BM25, σ^* e la loro combinazione BM25+σ^*, su ciascuna delle due collezioni.[5] BM25 va molto meglio di σ^*, probabilmente a causa del fatto che i documenti rilevanti sono di lunghezza ridotta, ma il metodo combinato ha ottenuto le prestazioni migliori in tutti e due i casi, con un miglioramento piuttosto netto anche rispetto a BM25. La differenza fra le prestazioni del metodo integrato e di BM25 sono statisticamente significative utilizzando il T-Test.

Collezione	Topics	BM25	σ^*	BM25 + σ^*
WT10g	501-550	0.143	0.057	0.153
Robust	301-450, 601-700	0.195	0.089	0.203

Table 2. MAP (mean average precision) medio dei metodi di ordinamento singoli e combinati sulle collezioni WT10g e Robust.

Per esaminare meglio le prestazioni relative dei tre metodi, abbiamo calcolato il valore di MAP sulle singole interrogazioni. In Figura 2 abbiamo graficato i risultati per le interrogazioni di WT10g. In questo caso il metodo combinato migliora in 28 casi e peggiora nei rimanenti 22, rispetto a BM25. I risultati per Robust sono leggermente differenti, perché a fronte di un miglioramento medio percentuale piu contenuto, la robustezza rispetto alle singole interrogazioni aumenta: 197 i miglioramenti, 53 i peggioramenti.

Per valutare la robustezza del metodo rispetto al parametro α abbiamo ricalcolato le prestazioni facendo variare il valore di α nell'intervallo fra uno e zero, i due estremi coincidendo rispettivamente con BM25 e σ^*. I risultati, mostrati in Tabella 3, suggeriscono chiaramente che il metodo è sufficientemente robusto, perché c'è un intervallo di valori per i quali le prestazioni si mantengono elevate, e questo comportamento è riscontrabile su entrambe le collezioni.

[3] http://lucene.apache.org/

[4] http://nlp.uned.es/ jperezi/Lucene-BM25/

[5] Abbiamo fatto una serie di esperimenti per valutare le potenzialità per il ranking anche della metrica σ_p, sia da sola, sia in combinazione con BM25, sia infine come riordinamento del ranking prodotto da BM25. I risultati però sono stati insoddisfacenti.

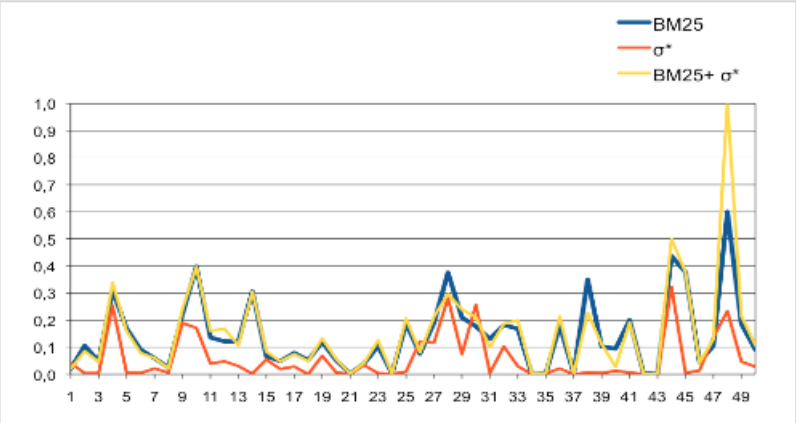

Fig. 2. Analisi delle prestazioni sulle singole topics di WT10g.

α	1	0.9	0.8	0.7	0.6	0.5	0.4	0.3	0.2	0.1	0
WT10g	0.143	0.146	0.153	0.153	0.150	0.137	0.122	0.096	0.081	0.067	0.054
Robust	0.195	0.203	0.203	0.198	0.167	0.154	0.142	0.120	0.107	0.096	0.089

Table 3. MAP medio del metodo di ranking combinato BM25+σ^* sulle due collezioni, al variare del parametro α.

6 Applicazione selettiva delle metriche frequentistiche e quantistiche

Finora abbiamo considerato l'ipotesi di combinare la pesatura frequentistica e quantistica in modo sistematico, per ciascuna interrogazione e su tutta la collezione. Poiche però pesatura quantistica e frequentistica hanno caratteristiche e requisiti differenti, ci siamo chiesti se è possibile prevedere una utilizzazione selettiva dei due paradigmi di ranking in funzione di determinate caratteristiche dei documenti e dell'interrogazione. La prima variabile che abbiamo considerato è stata la lunghezza dei documenti, perché il metodo quantistico dovrebbe andare meglio sui testi lunghi. Vogliamo valutare se effettivamente la metrica quantistica è più efficace nel recuperare i documenti lunghi e quella frequentistica i documenti brevi.

A questo scopo abbiamo riportato due grafici relativi a WT10g, uno per BM25 e uno per σ^*, in cui sull'asse x ci sono i valori della lunghezza del documento in numero di parole, mentre sull'asse y è riportata la percentuale di documenti rilevanti (nei due casi in cui vengano ritrovati o non ritrovati) che hanno meno del corrispondente numero di parole dell'asse x. Ad esempio, il grafico di sinistra mostra che per i documenti rilevanti di lunghezza < 2000, i ritrovati da BM25 sono l80% del totale dei rilevanti ritrovati e solo il 60% dei rilevanti non ritrovati. Risulta quindi confermato che gli andamenti sono opposti

a seconda della metrica che si considera. Questi risultati sono incoraggianti dal punto di vista di un'applicazione selettiva guidata dalla lunghezza dei documenti. Lo sviluppo e la sperimentazione di un metodo di pesatura basato su queste osservazioni è stato lasciato come lavoro futuro.

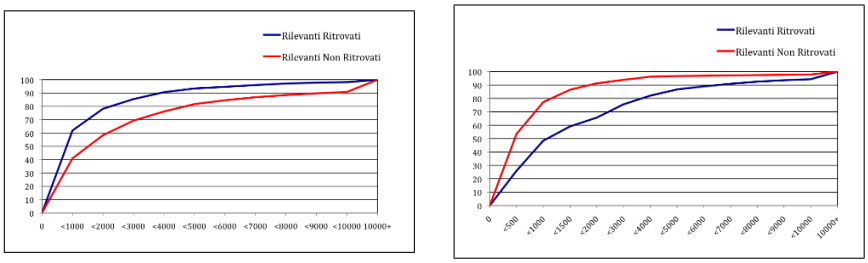

Fig. 3. Percentuali cumulative dei documenti rilevanti ritrovati e rilevanti non ritrovati da BM25 (sinistra) e σ^* (destra), in funzione della lunghezza dei documenti.

La seconda variabile per l'applicazione selettiva che abbiamo considerato è stata la difficoltà stimata delle interrogazioni. La speranza era che le metriche fossero efficaci in modo inverso rispetto a quest'ultima, in particolare che la pesatura quantistica conseguisse buone prestazioni sulle topics ritenute piu difficili. Abbiamo utilizzato due noti predittori pre-retrieval: Simplified Clarity Score [4] e σ_1 [8]. In Figura 4 abbiamo riportato due grafici, uno per WT10g con predittore σ_1 e uno per Robust con predittore Simplified Clarity Score, in cui ciascuna topic viene rappresentata con il valore restituito dal predittore (asse x) e con il suo valore di MAP (asse y), quest'ultimo calcolato utilizzando sia BM25 sia a σ^*. Nelle figure sono graficate anche le rispettive regressioni lineari. Risulta chiaro che le due metriche hanno un comportamento simile. In questo caso quindi, non sembrano esserci i presupposti per un'applicazione selettiva delle due tecniche.

7 Conclusioni

In questo lavoro abbiamo cercato di riconciliare la pesatura quantistica delle parole, basata sull'interspaziatura delle occorrenze e sviluppata prevalentemente nell'ambito della fisica, e la pesatura frequentistica adottata in information retrieval. Abbiamo visto che le due tecniche sono essenzialmente complementari e che la loro combinazione può migliorare sia la pesatura quantistica, incorporando statistiche legate all'analisi di corpus, sia quella frequentistica, per trovare termini rilevanti che sfuggono ai normali criteri basati su tf-idf. In una serie di esperimenti preliminari abbiamo dimostrato che è possibile migliorare il ranking attraverso una semplice combinazione delle due metriche, anche se le potenzialità di questo approccio sono ancora in gran parte da investigare. Oltre al ranking, questa tecnica può essere utilizzata per migliorare altri classici compiti di information retrieval nei quali l'individuazione delle parole chiave presenti in uno

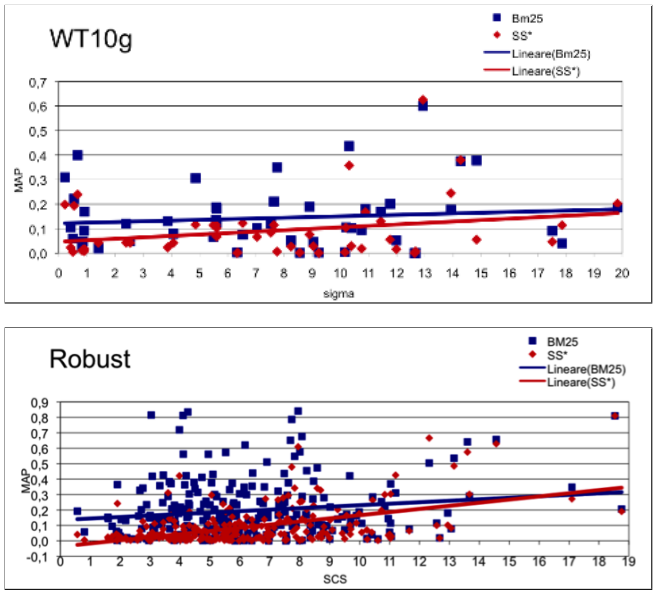

Fig. 4. MAP delle singole topics in funzione della loro difficolta' stimata

o più documenti è cruciale ed è stata finora affrontata con tecniche frequentistiche, in particolare la diversificazione e il clustering dei risultati [2] e l'espansione automatica delle interrogazioni [3].

References

1. P. Carpena, P. Bernaola-Galvă, M. Hackenberg, A. V. Coronado, and J. L. Oliver. Level statistics of words: Finding keywords in literary texts and symbolic sequences. *Physical Review E 79:035102*, 2009.
2. C. Carpineto, M. D'Amico, and G. Romano. Evaluating Subtopic Retrieval Methods: Clustering Versus Diversification of Search Results. Information Processing and Management, in press, 2012.
3. C. Carpineto and G. Romano. A Survey of Automatic Query Expansion in Information Retrieval . ACM Computing Surveys, in press, 2012.
4. B. He and I. Ounis. Query performance prediction. *Inf. Sys.*, 31(7):585–594, 2006.
5. J. P. Herrera and P. A. Pury. Statistical keyword detection in literary corpora. *European Physical Journal B*, 63:135–146, 2008.
6. M. Ortuño, P. Carpena, P. Bernaola-Galvă, E. Muñoz, and M. Somoza. Keyword detection in natural languages and dna. *Europhysics Letters*, 57(5):759–764, 2002.
7. G. Salton. *A Theory of indexing*. Society for Industrial and Applied Mathematics, 1975.
8. Y. Zhao, F. Scholer, and Y. Tsegay. Effective pre-retrieval query performance prediction using similarity and variability evidence. In *ECIR '08*, pages 52–64, 2008.
9. H. Zhou and G. W. Slater. A metric to search for relevant words. *Phyisica A 329*, pages 309–327, 2003.