Sull'uso di meno topic nelle iniziative di valutazione per l'information retrieval

Andrea Berto and Stefano Mizzaro

Department of Mathematics and Computer Science
University of Udine
Udine, Italy
andrea@andreaberto.it, mizzaro@uniud.it

Sommario La possibilità di ridurre il numero di topic usati in TREC e in analoghe iniziative di valutazione è stata studiata di recente, con risultati incoraggianti: anche diminuendo di molto il numero di topic (ad esempio usandone solo 10 invece di 50) è possibile, almeno potenzialmente, ottenere risultati molto simili in termini di valutazione dei sistemi. La generalità di questo approccio è però in discussione, in quanto sembra che il sottoinsieme di topic selezionato su una popolazione di sistemi sia poi non adeguato a valutare altri sistemi. In questo lavoro riconsideriamo la questione della generalità: evidenziamo alcune limitazioni dei lavori precedenti e riportiamo alcuni risultati sperimentali che sono invece più positivi. I risultati supportano l'ipotesi che con opportuni accorgimenti, i pochi topic selezionati sulla base di una popolazione di sistemi possono poi essere adeguati a valutare anche una popolazione di sistemi differente.

Keywords: TREC, valutazione, test collection, meno topic

1 Introduzione

La valutazione dei sistemi d'Information Retrieval (IR) viene spesso effettuata tramite *test collections*: questa metodologia prevede che più gruppi di ricerca partecipino ad una competizione internazionale e cerchino di reperire in modo automatico i documenti *relevant* per alcuni *topic* (ossia, descrizioni testuali di bisogni informativi). La relevance dei documenti viene decisa da giudici umani. Esistono alcune varianti di questo processo, ma le maggiori iniziative di valutazione attive oggi (TREC, NTCIR, CLEF, INEX, FIRE) lo seguono in modo abbastanza preciso.

Uno dei costi maggiori di questa metodologia è l'espressione dei giudizi di relevance, e infatti vi sono state varie proposte per cercare di diminuire questi costi [1, 2, 4, 8, 9, 11, 12, 13]. Una possibilità è quella di usare meno topic: in [3] viene evidenziato sperimentalmente che questa strada è, almeno potenzialmente, promettente; però in [7] viene invece sollevato un dubbio sulla generalità di tale risultato.

Il nostro lavoro si basa sui due lavori [3,7] appena citati. Nel paragrafo 2 i due lavori vengono descritti più in dettaglio, e ne vengono evidenziate le limitazioni

APs	t_1	$\cdots$	t_n	MAP
s_1	AP(s_1, t_1)	$\cdots$	AP(s_1, t_n)	MAP(s_1)
s_2	AP(s_2, t_1)	$\cdots$	AP(s_2, t_n)	MAP(s_2)
$\vdots$		$\ddots$		$\vdots$
s_m	AP(s_m, t_1)	$\cdots$	AP(s_m, t_n)	MAP(s_m)

Tabella 1. AP e MAP, per n topic e m sistemi (run) (da [3, pag. 21:4]).

e le domande senza risposta che motivano la necessità di continuare le ricerche in questa direzione. Nei paragrafi 3 e 4 vengono descritti alcuni ulteriori esperimenti e vengono presentati i risultati che abbiamo ottenuto, che effettivamente mitigano i problemi sulla generalità sollevati in [7].

2 I due studi

2.1 Meno topic!

Il punto di partenza del lavoro [3] è illustrato in tabella 1: ogni riga fa riferimento ad un sistema[1] ed ogni colonna ad un topic. Ogni cella della matrice AP(s_i, t_j) misura la prestazione del sistema s_i sul topic t_j; la metrica standard utilizzata in TREC è Average Precision (AP). La prestazione di un sistema s_i, solitamente, è ottenuta calcolando la media aritmetica di *tutti* i valori AP(s_i, t_j) (una riga della tabella). Questa metrica è chiamata Mean Average Precision (MAP).

Il metodo utilizzato in [3] è il seguente. Partendo dall'insieme di n topic, si considera per ogni cardinalità $c \in \{1, \ldots, n\}$ e per ogni sottoinsieme di topic di cardinalità c il corrispondente valore di MAP per ogni sistema calcolato solo su questo sottoinsieme di topic: in altri termini, si fa la media delle c (e non n) colonne in tabella 1 relative al solo sottoinsieme di topic di cardinalità c selezionato. Per ogni sottoinsieme viene poi calcolata la correlazione di questi valori di MAP con i valori di MAP dell'intero insieme di n topic. Questa correlazione misura quanto bene il sottoinsieme considerato predice le prestazioni dei sistemi in relazione all'intero insieme di topic. Per ogni cardinalità c, vengono poi selezionati i *migliori* sottoinsiemi di topic, ossia quelli con i valori di correlazione più alti. Si selezionano anche i *peggiori* sottoinsiemi e si calcola poi la correlazione *media* su tutti i sottoinsiemi di cardinalità c.

In [3] vengono usati dati di TREC 8 [10] (da cui sono stati eliminati il 25% dei sistemi peggiori: tabella 1 con $n = 50$ e $m = 96$) ed NTCIR 6 (tabella 1 con $n = 50$ e $m = 74 - 25\% = 56$), varie metriche di efficacia (oltre a MAP, anche RPrec, P@10, GMAP, ed NDCG) e varie misure di bontà dei sottoinsiemi di topic (oltre alla Correlazione, anche Tau di Kendall e Tasso d'errore).

Il grafico in figura 1 riassume il risultato principale: i valori di correlazione per ogni cardinalità. Esso mostra che il miglior sottoinsieme di cardinalità, ad

[1] Anche se sarebbe più corretto, in terminologia TREC, usare *run*.

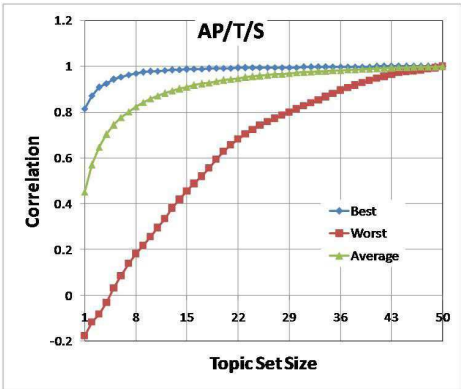


Figura 1. Correlazioni massima, media e minima per cardinalità. Misura MAP (da [3, pag. 21:5]).

esempio, $c = 5$ o $c = 10$ è decisamente migliore nel prevedere le prestazioni sull'intero insieme di 50 topic rispetto ad un sottoinsieme di pari cardinalità scelto a caso, il quale a sua volta si comporta molto meglio del peggior sottoinsieme. Interpretando la figura orizzontalmente, se l'obiettivo è una correlazione di 0.95 rispetto all'intero insieme, la scelta del miglior sottoinsieme permette di poter utilizzare soltanto 6 topic, rispetto ai 22 necessari se si sceglie un sottoinsieme casuale ed ai 41 se la scelta ricade sul peggior sottoinsieme. Risultati simili vengono riportati per le altre metriche di efficacia e misure di bontà.

In [3] sono studiati anche altri sottoinsiemi di topic con buona correlazione, i cosiddetti "best set": analizzando i 10 migliori sottoinsiemi per ogni cardinalità c, risulta che questi sono abbastanza differenti fra di loro. Inoltre viene analizzato anche il problema della generalizzazione, ossia di quanto i sottoinsiemi di buoni topic trovati sulla base di una certa popolazione di sistemi risultino buoni topic anche quando si misurano le prestazioni di un'altra popolazione di sistemi. Questo studio viene effettuato spezzando in due la popolazione dei sistemi partecipanti a TREC 8, ma lascia il dubbio che i run multipli effettuati con un unico sistema inficino in qualche modo l'esperimento.

2.2 Meno topic?

In [7] la generalizzazione viene ulteriormente studiata. Per fare ciò, oltre ai dati sui 96 sistemi di TREC 8 usati in [3] (denominati TREC96), vengono usate due nuove popolazioni di sistemi: TREC87 (TREC 8 senza i sistemi manual, per avere una popolazione di sistemi più omogenea) e Terrier (20 run di differenti varianti del sistema Terrier [5, 6]) per avere una popolazione di sistemi completamente differente seppure sugli stessi topic.

L'obiettivo principale di [7] è di capire se i migliori sottoinsiemi di topic selezionati per le varie cardinalità $c \in \{1, \ldots, n\}$ su una popolazione di sistemi

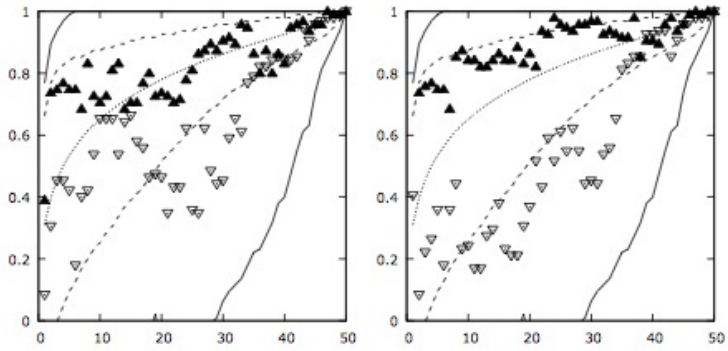


Figura 2. Tau di Kendall del migliore sottoinsieme di TREC96 (sinistra) e TREC87 (destra) applicati su Terrier (da [7, pag. 138]).

(vengono usate TREC96 e TREC87) risultano essere dei buoni sottoinsiemi di topic anche per valutare un'altra popolazione di sistemi (Terrier).

La figura 2 mostra il risultato ottenuto. Le cinque linee rappresentano rispettivamente i valori di correlazione massimi, il 95esimo percentile, medi (ossia, quelli attesi selezionando un sottoinsieme casuale di topic), il 5o percentile e peggiori ottenuti per Terrier; i triangoli pieni con punta verso l'alto sono i valori di correlazione dei sottoinsiemi migliori, ricavati su TREC96 o TREC87 e applicati a Terrier. Il risultato è piuttosto negativo, soprattutto per TREC96: il miglior sottoinsieme di topic, per ciascuna cardinalità, tende a comportarsi sempre meno bene del 95esimo percentile, e spesso anche peggio di un sottoinsieme di topic casuale della stessa cardinalità. I migliori sottoinsiemi selezionati su TREC87 sembrano comportarsi meglio: quando usati su Terrier portano a correlazioni vicine al 95esimo percentile e quasi sempre hanno una correlazione maggiore di un sottoinsieme casuale di topic.

2.3 Limitazioni e motivazioni

Il lavoro [7] mette quindi in discussione il risultato almeno potenzialmente positivo di [3]: sembra che i sottoinsiemi di topic adeguati per valutare una popolazione di sistemi non siano poi adeguati per valutare una popolazione di sistemi differente. Si possono però evidenziare alcune alcune limitazioni:

- L'analisi viene effettuata usando solo il singolo "best set"; resta in dubbio se vi siano altri sottoinsiemi di topic che siano buoni quasi quanto il migliore sottoinsieme di topic sulla popolazione di partenza, e che altresí generalizzino bene, ossia presentino una buona correlazione anche su una popolazione di sistemi differente.
- Vengono usate soltanto Tau di Kendall (non le altre misure di bontà) e GMAP e logit(AP) (e non le altre metriche di efficacia). I risultati potrebbero essere differenti per altre combinazioni di misure/metriche.

– Inoltre in nessuno dei due lavori [3, 7] viene detto nulla sul *numero* di best set: ossia, non è chiaro se vi siano molti o pochi sottoinsiemi di topic buoni (che consentono di valutare essenzialmente in modo analogo i sistemi).

Ha quindi senso continuare questa linea di ricerca. In questo lavoro ci chiediamo:

D1. Quanti "best set" ci sono?
D2. Se invece di considerare il singolo "best set" come fatto in [7] se ne considerano di più, i risultati sulla generalizzazione sono più positivi? In altri termini, se si considerano i 10 best set, quanto questi sono generali?

3 Esperimento 1: quanti "good subset"?

Per poter rispondere a **D1**, ossia sapere quanti "good subset" esistono, è stato condotto l'esperimento seguente. Per ogni cardinalità abbiamo usato l'euristica presentata in [3] per selezionare 10 milioni di sottoinsiemi di topic[2] e per ognuno di essi è stata calcolata la MAP parziale e la correlazione lineare di quest'ultima con la MAP dell'intero insieme di topic. Considerando 0.96 come soglia di correlazione oltre la quale un sottoinsieme predice bene i risultati finali, abbiamo contato il numero di sottoinsiemi che superano tale soglia. L'esperimento è stato condotto su tutte e tre le collezioni (TREC96, TREC87, Terrier) e abbiamo preso in esame, oltre alla correlazione, anche la Tau di Kendall (con soglia 0.85 anziché 0.96).

La figura 3 riporta i risultati sulle collezioni TREC96, TREC87 e Terrier. Essa mostra, per ogni cardinalità di ogni collezione, il numero di sottoinsiemi, tra i 10 milioni considerati, che hanno un valore di correlazione superiore a 0.96 e di Tau superiore a 0.85. Analizziamo prima le curve relative alla correlazione. Per quanto riguarda TREC96, si nota come il numero di "good subset" cresca velocemente: a cardinalità 25, ad esempio, più della metà dei sottoinsiemi calcolati è costituita da buoni sottoinsiemi e dalla cardinalità 35 si supera il 99% di "good subset".

In TREC87 le quantità di "good subset" sono simili anche se leggermente inferiori; questo è probabilmente dovuto all'assenza dei run manuali, notoriamente più efficaci e tali da esercitare una forte influenza nel calcolo dei risultati finali. Per Terrier i valori sono invece leggermente superiori, specie a cardinalità basse, dove si registrano già numerosi buoni sottoinsiemi (ad esempio a cardinalità 9 oltre il 20% dei sottoinsiemi risulta essere un "good subset").

Considerando la Tau di Kendall, i risultati ottenuti sono leggermente più bassi per ognuna delle tre collezioni analizzate: TREC96 riporta comunque un numero di "good subset" maggiore di TREC87 (i run manuali sono influenti indifferentemente dalla misura considerata), ma minore di Terrier (collezione formata da pochi run e molto simili tra loro).

L'esistenza di un numero così alto di sottoinsiemi di topic con buona correlazione fa pensare che sia effettivamente possibile trovarne di generali. Per studiare questo aspetto abbiamo eseguito un secondo esperimento.

[2] Per le cardinalità da 1 a 5 si sono analizzati tutti i possibili sottoinsiemi.

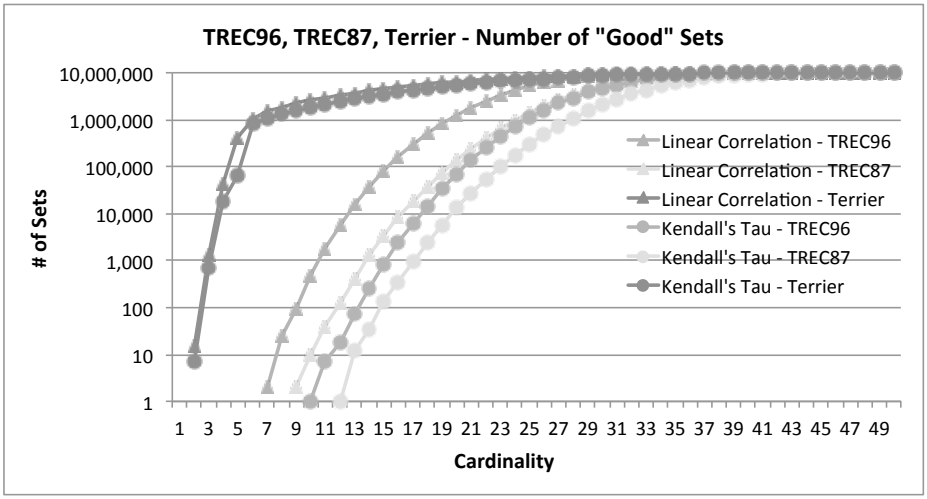


Figura 3. Il numero di "buoni" sottoinsiemi di topic alle varie cardinalità (scala semilogaritmica) per le 3 collezioni.

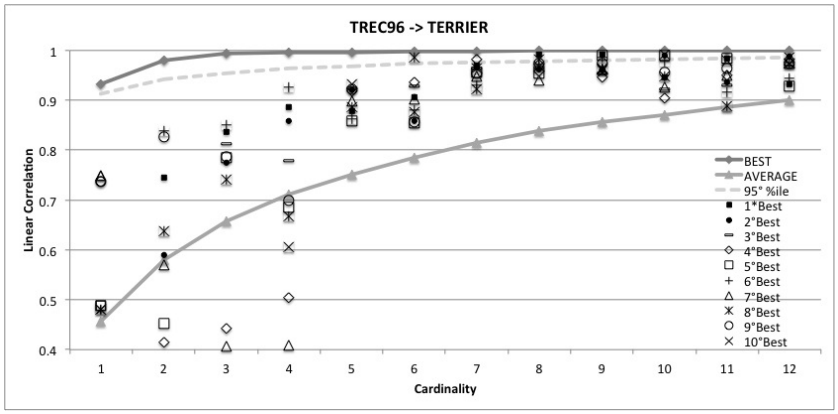
4 Esperimento 2: generalizzazione

Per poter rispondere alla seconda domanda **D2** è stato condotto un esperimento di generalizzazione prendendo da TREC96 e TREC87, per ogni cardinalità, i migliori 10 sottoinsiemi di topic e usandoli per valutare Terrier. L'obiettivo è di capire se fra i 10 migliori sottoinsiemi di topic ce ne sono alcuni che generalizzano (mentre in [7] si è guardato solo il migliore). In questo modo viene effettuato un test di generalità sulla capacità di valutazione dei migliori sottoinsiemi su una collezione diversa da quella da cui sono ricavati.

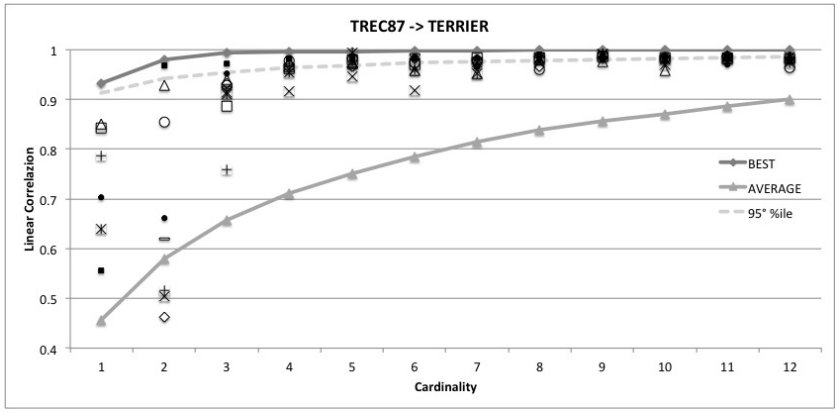
L'esperimento è stato svolto, finora, per le cardinalità da 1 a 12 (l'alto numero di sottoinsiemi rende il problema computazionalmente pesante, come discusso in [3]). Per ognuno dei 10 migliori sottoinsiemi ottenuti su TREC96 e TREC87 e per ogni cardinalità è stata calcolata la MAP parziale sui sistemi della collezione Terrier; questo valore è poi stato correlato, mediante sia la correlazione sia la Tau di Kendall, con la MAP totale sui sistemi della collezione Terrier. In questo modo si sono ottenuti 10 valori di correlazione per ognuna delle 12 cardinalità, riferiti ai migliori sottoinsiemi calcolati su TREC96/87 e generalizzati su Terrier.

Le figure 4 e 5 riportano i risultati, rispettivamente in termini di correlazione e Tau. Nelle figure, le tre linee rappresentano i valori di correlazione massimi, il 95esimo percentile e medi: sono analoghe alle tre linee più in alto di figura 2 (sono differenti perché qui è stata usata la metrica MAP anziché logit(AP)). I punti rappresentano i valori di correlazione per i 10 best set ottenuti su una popolazione di sistemi differente.

Si può notare come la maggior parte dei punti in figura 4 stia al di sopra della linea media; per TREC87 molti sono anche al di sopra del 95esimo percentile. Questo risultato è più positivo di quello ottenuto in [7]: se si considerano i 10



(a)



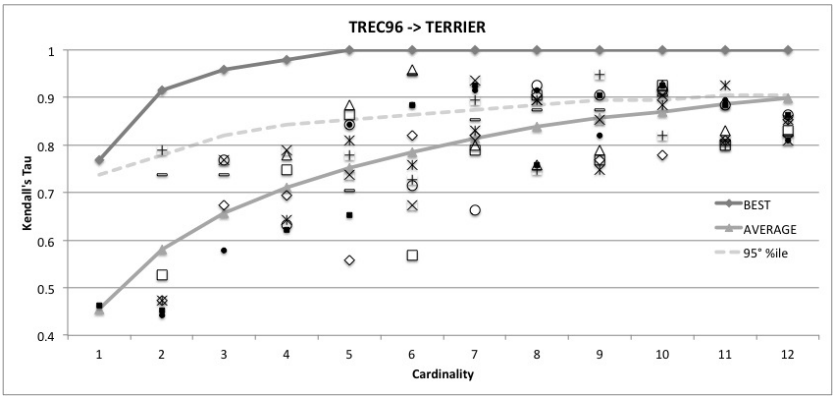
(b)

Figura 4. Generalizzazione: correlazioni dei 10 best set secondo TREC96 (a) e TREC87 (b) su Terrier.

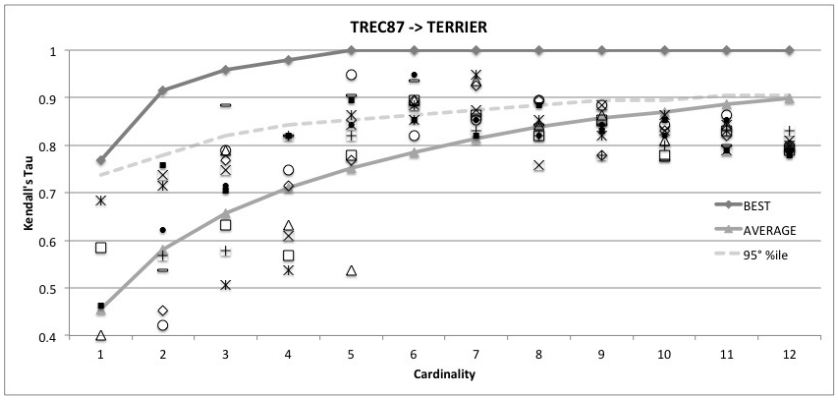
migliori sottoinsiemi di topic ottenuti sulla base di una certa popolazione di sistemi, fra di essi molti sono adeguati a misurare le prestazioni anche di altre popolazioni di sistemi. Il fatto che TREC87 si comporti sistematicamente meglio di TREC96 inoltre è positivo, in quanto lascia intravedere un modo di scegliere la popolazione di sistemi in cui cercare i sottoinsiemi di topic generali (è meglio se è omogenea).

Tau di Kendall presenta risultati un po' più negativi della correlazione lineare: in figura 5 molti punti sono al di sotto non solo del 95esimo percentile ma anche della linea mediana. Questo significa che i best set sono più efficaci nel predire il valore di MAP che nell'ordinare i sistemi allo stesso modo dell'insieme di tutti e 50 i topic.

Viene spontaneo a questo punto porsi una terza domanda:



(a)



(b)

Figura 5. Generalizzazione: Tau di Kendall dei 10 best set secondo TREC96 (a) e TREC87 (b) su Terrier.

D3. L'ordine dei "best set" si ripercuote sulla capacità di generalizzazione dei sottoinsiemi? Ossia: il primo best set tende ad essere migliore (quando usato su una popolazione di sistemi differente) del secondo, e questo a suo volta tende ad essere migliore del terzo e così via?

Una prima risposta negativa viene già dal risultato di [7], ma si può essere più sistematici ed analizzare tutti i migliori 10 best set. Le figure 4 e 5 non consentono di rispondere, e quindi nelle figure 6 e 7 vengono riportati gli stessi risultati (i valori di correlazione e Tau al variare delle cardinalità per i 10 best set) in una forma grafica più appropriata. Dall'andamento ondulato (più evidente per la Kendall di Tau, in figura 7) è chiaro che la risposta a **D3** è negativa. Quindi per trovare il sottoinsieme di topic che generalizza meglio non ci si può basare solo sulla bontà di tale sottoinsieme sulla popolazione di partenza, ma bisogna

considerare vari sottoinsiemi.

5 Conclusioni e sviluppi futuri

In questo lavoro abbiamo rivisto ed esteso alcuni risultati ottenuti in [3, 7]. Sulla base degli esperimenti effettuati, e ancora in corso, sembra che:

- se si cerca di predire le prestazioni di una popolazione di sistemi usando un sottoinsieme di topic di cardinalità ridotta rispetto agli usuali 50 topic di TREC, esistono *molti* sottoinsiemi di topic "buoni";
- se si selezionano i sottoinsiemi di topic "buoni" su una popolazione di sistemi, anche se il migliore di tali sottoinsiemi per ogni cardinalità sembra non essere generale (ossia, sembra non adeguato a valutare le prestazioni su un'altra popolazione di sistemi [7]), in realtà la situazione migliora se si considerano i successivi "buoni" sottoinsiemi: molti fra questi sono invece adeguati.

Gli esperimenti di generalizzazione presentati in questo lavoro riguardano soltanto le cardinalità da 1 a 12 e prendono in considerazione solamente la metrica MAP. Questa limitazione è dovuta alla complessità computazionale nel calcolo di tutti i possibili sottoinsiemi a cardinalità maggiore di 12 (e, specularmente, minore di 38), soprattutto per quanto riguarda la Tau di Kendall. Per poter confrontare in maniera più diretta i risultati ottenuti con i risultati presentati in [7], è in corso di elaborazione un esperimento che utilizza come metrica logit(AP), la stessa di [7], invece di MAP.

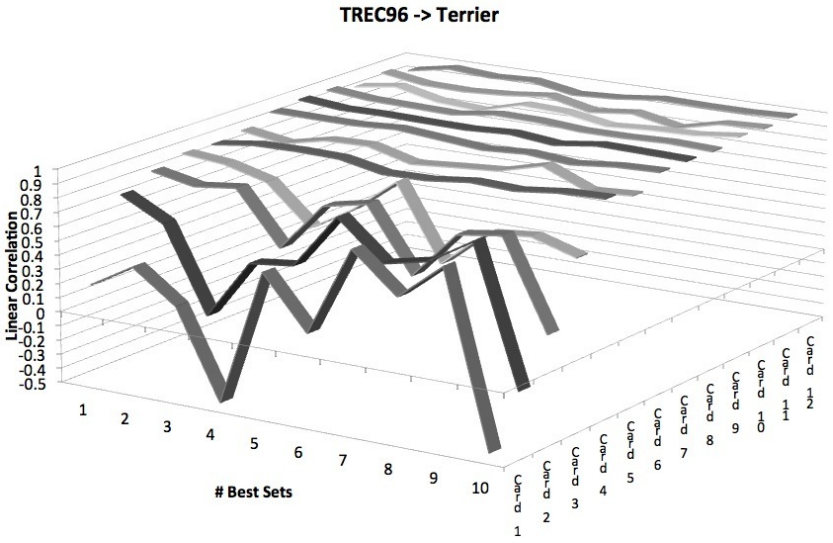
Un'altra possibile estensione del lavoro riguarda lo studio della generalizzazione per le cardinalità da 38 a 50 (i cui dati sono calcolabili in tempi accettabili). Tuttavia, le cardinalità di maggior interesse per lo scopo che si prefigge lo studio (la sensibile riduzione del numero di topics), sono probabilmente quelle comprese tra circa 5 e circa 20, ragionevolmente coperte dal lavoro presentato. Inoltre, come fatto già in [3], sarà importante verificare i risultati, oltre che sui dati di TREC, anche sui dati delle altre iniziative di valutazione.

Ringraziamenti

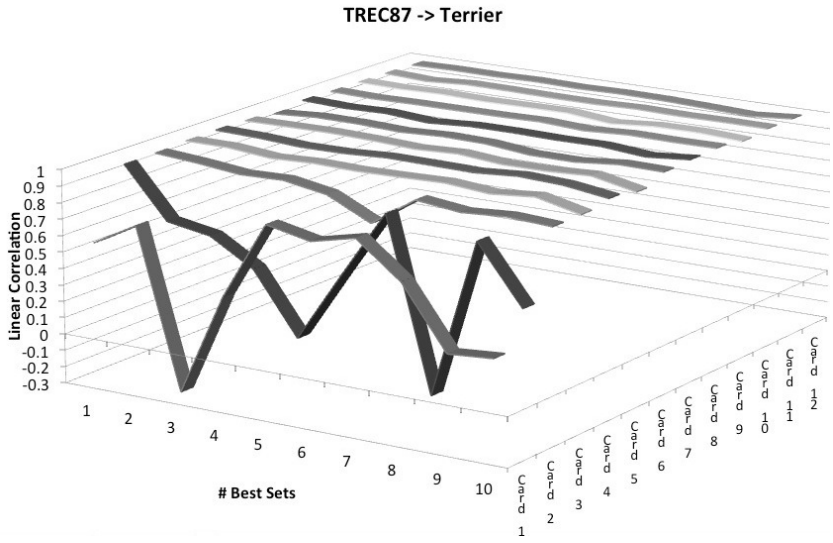
Ringraziamo Steve Robertson per aver fornito alcuni dati per gli esperimenti e per alcuni utili suggerimenti.

Riferimenti bibliografici

1. C. Buckley and E. Voorhees. Evaluating evaluation measure stability. In N. J. Belkin, P. Ingwersen, and M.-K. Leong, editors, *SIGIR 2000: Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 33–40, New York, 2000. ACM Press.

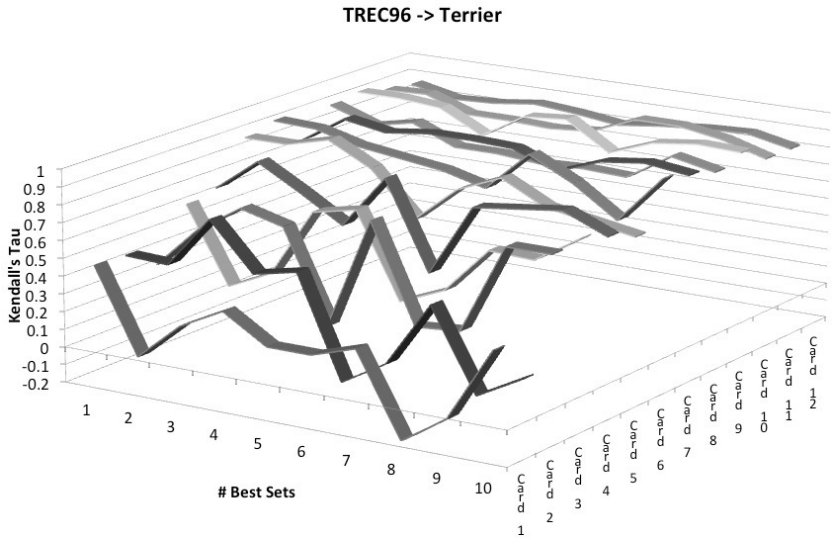


(a)

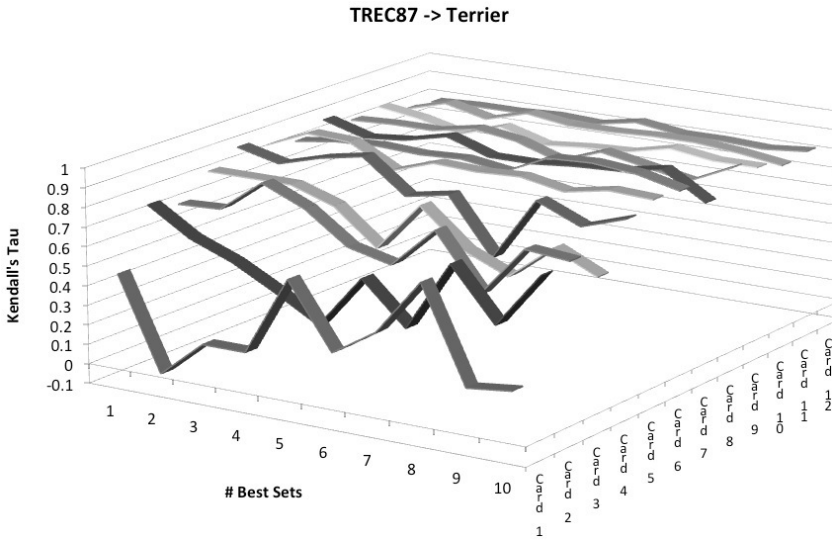


(b)

Figura 6. Andamento della correlazione dei 10 best set secondo TREC96 (a) e TREC87 (b) su Terrier.



(a)



(b)

Figura 7. Andamento della Tau di Kendall dei 10 best set secondo TREC96 (a) e TREC87 (b) su Terrier.

2. B. Carterette, J. Allan, and R. Sitaraman. Minimal test collections for retrieval evaluation. In E. N. Efthimiadis, S. T. Dumais, D. Hawking, and K. Järvelin, editors, *SIGIR 2006: Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 268–275, New York, 2006. ACM Press.
3. J. Guiver, S. Mizzaro, and S. Robertson. A few good topics: Experiments in topic set reduction for retrieval evaluation. *ACM Transactions on Information Systems (TOIS)*, 27(4), November 2009.
4. S. Mizzaro and S. Robertson. HITS hits TREC — exploring IR evaluation results with network analysis. In C. L. A. Clarke, N. Fuhr, N. Kando, W. Kraaij, and A. P. de Vries, editors, *SIGIR 2007: Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 479–486, New York, 2007. ACM Press.
5. I. Ounis, G. Amati, P. V., B. He, C. Macdonald, and Johnson. Terrier Information Retrieval Platform. In *Proceedings of the 27th European Conference on IR Research (ECIR 2005)*, volume 3408 of *LNCS*, pages 517–519. Springer, 2005.
6. I. Ounis, C. Lioma, C. Macdonald, and V. Plachouras. Research directions in terrier. *Novatica/UPGRADE Special Issue on Web Information Access, Ricardo Baeza-Yates et al. (Eds), Invited Paper*, 2007.
7. S. Robertson. On the Contributions of Topics to System Evaluation. In *Advances in Information Retrieval*, volume 6611 of *Lecture Notes in Computer Science*, pages 129–140. Springer Berlin / Heidelberg, 2011.
8. M. Sanderson and J. Zobel. Information retrieval system evaluation: effort, sensitivity, and reliability. In G. Marchionini, A. Moffat, J. Tait, R. Baeza-Yates, and N. Ziviani, editors, *SIGIR 2005: Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 162–169, New York, 2005. ACM Press.
9. E. Voorhees and C. Buckley. The effect of topic set size on retrieval experiment error. In M. Beaulieu, R. Baeza-Yates, S. H. Myaeng, and K. Jarvelin, editors, *SIGIR 2002: Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 316–323, New York, 2002. ACM Press.
10. E. M. Voorhees and D. Harman. Overview of the Eighth Text REtrieval Conference (TREC-8). In *TREC*, 1999.
11. W. Webber, A. Moffat, and J. Zobel. Statistical power in retrieval experimentation. In *CIKM '08: Proceeding of the 17th ACM conference on Information and knowledge management*, pages 571–580, New York, NY, USA, 2008b. ACM.
12. E. Yilmaz and J. A. Aslam. Estimating average precision with incomplete and imperfect judgements. In P. S. Yu, V. J. Tsotras, E. A. Fox, and B. Liu, editors, *CIKM 2006: Proceedings of the 13th ACM Conference on Information and Knowledge Management*, pages 102–111, New York, 2006. ACM Press.
13. J. Zobel. How reliable are the results of large-scale information retrieval experiments? In W. B. Croft, A. Moffat, C. J. van Rijsbergen, R. Wilkinson, and J. Zobel, editors, *SIGIR'98: Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 307–314, New York, 1998. ACM Press.