

Sistema IR per collezione di articoli medici utilizzando diversi insiemi di stopword

Alessandro Stefani¹, Caterina Buranelli², and Cristi Gutu³

¹ Corso di laurea in Statistica per le tecnologie e le scienze, matricola 1148387
`alessandro.stefani.6@studenti.unipd.it`

² Corso di laurea in Statistica per le tecnologie e le scienze, matricola 1234567
`caterina.buranelli@studenti.unipd.it`

³ Corso di laurea in Statistica per le tecnologie e le scienze, matricola 1147351
`gheorghecristi.gutu@studenti.unipd.it`

Sommario DA FINIRE QUANDO ABBIAMO MESSO A POSTO LA SEZIONE ESPERIMENTI. Questo progetto tratta la realizzazione attraverso il pacchetto *Whoosh*, di un motore di ricerca volto al reperimento di documenti della collezione sperimentale *OHSUMED* indicizzata opportunamente. Il progetto è anche corredato di un webserver che permette all'utente di interrogare il motore di ricerca in forma interattiva attraverso un browser a scelta.

Keywords: *Information Retrieval · IR · statistica · reperimento · indicizzazione · Whoosh · Python · webserver · web · interrogazione web*

1 Introduzione

Lo scopo finale di un Sistema di Information Retrieval (IRS) è quello di reperire documenti rilevanti relativi a una certa esigenza informativa⁴; dunque i documenti sono il primo ingresso del sistema, mentre il secondo è costituito dalle interrogazioni; i documenti devono essere indicizzati e nell'indice creato, si andrà a effettuare la ricerca per reperire documenti rilevanti⁵. Questa seconda parte è detta reperimento e non si occupa solo di ricercare tra i documenti, ma anche di riordinare secondo un certo ordine di rilevanza. Ciò che descrive i documenti e ciò che descrive le interrogazioni deve essere confrontabile, infatti nei programmi di indicizzazione e di reperimento si usa uno schema, che deve essere uguale in entrambi i casi. L'indicizzazione è un trade-off tra il miglioramento della rappresentazione del contenuto informativo dei documenti (efficacia) e la gestione degli indici (efficienza). Benché ci sia sempre tensione tra queste due caratteristiche, ad oggi la questione più studiata è quella dell'efficienza.

⁴ insieme delle circostanze in cui una persona ha un problema da risolvere o un compito da svolgere richiede informazioni importanti, utili o necessarie per la soluzione del problema o lo svolgimento del compito

⁵ la rilevanza è la proprietà che rende l'informazione importante, utile o necessaria a soddisfare l'esigenza informativa dell'utente

Esistono diversi modi per risolvere questo problema, nel nostro caso abbiamo ricercato la configurazione migliore tra un sistema di reperimento con o senza uso di stopword e il tuning dei parametri dello schema di pesatura⁶ BM25F .

L'obiettivo principale della relazione è da una parte la documentazione del progetto di un servizio di IR e dall'altra una misura del grado in cui si sia riusciti a mettere in pratica i contenuti della disciplina illustrati durante le lezioni.

A tal scopo la relazione dovrà illustrare nelle sezioni successive:

- i metodi di indicizzazione,
- i modelli di reperimento,
- l'interfaccia basata su un *browser* per il WWW
- i risultati della *valutazione* condotta con la collezione sperimentale OHSU-MED.

Il lettore della relazione è lo studente medio di un corso di laurea in statistica al quale la relazione deve dare tutti gli strumenti per comprendere il contenuto. Ci si metta nei suoi panni e si scriva tutto ciò e solo ciò che serve. Chiedersi qual è il messaggio che lo studente deve “portarsi a casa”, esplicitarlo in questo paragrafo e concentrarsi su quello nel resto della relazione.

L'introduzione della relazione deve servire al lettore a capire se vale la pena continuare a leggere il resto. Si possono riassumere i contenuti delle sezioni successive e metterne in evidenza i punti principali. La relazione consiste di tre paragrafi principali dopo questa introduzione e prima della bibliografia, per la quale si suggerisce Bib_TEX se si scrive con L^AT_EX.

2 Base di partenza

Un metodo utilizzato per tutti gli esperimenti è il così detto “Best Match 25 Model with Extension to Multiple Weighted Fields” o in breve BM25F

Le stopwords[2] sono parole che non portano informazione significativa al contenuto informativo come congiunzioni, articoli, avverbi..

3 Interfaccia web

4 Esperimenti e Risultati

Per gli esperimenti si è utilizzata parte della già citata collezione sperimentale chiamata OHSUMED⁷.

Si tratta di una collezione di oltre 300'000 articoli provenienti dal database bibliografico MEDLINE pubblicati tra il 1987 ed il 1991, di questi documenti ne sono stati utilizzati 54'711.

⁶ funzione che assegna per ogni documento diversi livelli d'importanza dei termini mediante dei pesi, che possono variare con l'interrogazione

⁷ <https://bit.ly/2wpOynZ>

I programmi utilizzati per effettuare indicizzazione e reperimento sono stati scritti in python(versione 2.7), sfruttando oggetti e funzioni del modulo whoosh⁸.

Per la valutazione dei risultati ottenuti nei vari esperimenti si è invece utilizzato lo strumento standard trec_eval⁹. Questo programma permette di ottenere alcune misure della qualità di un sistema di information retrieval se si conoscono i documenti rilevanti per le query sperimentali utilizzate.

Come misura da usare nel confronto tra risultati di diverse run si è scelta la Mean Average Precision(MAP)[3] , quantità calcolabile sia a livello di singola query che a livello complessivo su un insieme di query. Inoltre, per verificare se i valori di MAP ottenuti in run diverse si discostano in modo statisticamente significativo gli uni dagli altri si effettuano test di Wilcoxon[] su coppie di risultati.

4.1 Scelte iniziali

Prima di iniziare gli esperimenti sono state fatte delle scelte riguardo certi aspetti dei metodi di reperimento che rimanessero costanti durante tutti gli esperimenti. Queste scelte sono state fatte in seguito ad alcune run preliminari.

Per la funzione di reperimento e ranking come schema di pesatura si è deciso di utilizzare il BM25F, descritto nella sezione precedente, inquanto "state-of-the-art" dal punto di vista dei modelli per information retrieval su documenti strutturati, e come operatore logico per raggruppare le parole delle query si è deciso di utilizzare l'operatore OR, ovvero si cercano documenti in cui è presente almeno una parola della query.

Sempre a seguito delle run preliminari si è deciso di utilizzare il campo "desc" delle query anzichè il campo "title" inquanto dà risultati migliori, inoltre, si è notato che in alcune query sperimentali sono presenti parole scritte in modo errato e che quindi non risultano presenti nell'indice. Per questo dopo aver verificato, leggendo i documenti rilevanti per quelle query, quali fossero gli errori si è ritenuto opportuno effettuare una correzione al momento del reperimento. A grandi linee, la correzione viene fatta su parole che non risultano presenti nell'indice cercando in quest'ultimo delle parole alternative che si discostano di al più una lettera dalla parola originale; queste parole vengono quindi aggiunte alla query di partenza e si procede con la ricerca.

Un'ultima scelta è stata fatta riguardo il numero massimo di documenti reperiti. Si è optato per 100 documenti poiche un numero maggiore non porta grandi miglioramenti dal punto di vista del MAP.

⁸ <https://whoosh.readthedocs.io/en/latest/index.html>

⁹ https://trec.nist.gov/trec_eval/

Affinchè Whoosh possa indicizzare una collezione di documenti, necessita la specificazione di uno schema che include, per ogni possibile campo dei documenti della collezione, il nome del campo ed il tipo del campo.

Lo schema di base per la collezione qui utilizzata è il seguente:

```
schema = Schema(docid      = ID(stored=True),
                title      = TEXT(stored=True),
                identifier  = ID(stored=True),
                terms      = NGRAM(stored=True),
                authors     = NGRAM(stored=True),
                abstract    = TEXT(stored=True),
                publication  = TEXT(stored=True),
                source      = TEXT(stored=True))
```

Figura 1. Schema necessario all'indicizzazione dei documenti. "stored=True" indica che il campo viene salvato ed è quindi successivamente accessibile.

4.2 Baseline

I valori di MAP scelti come baseline sono quelli ottenuti effettuando ricerche su un indice con lo schema di base, riportato sopra, senza l'utilizzo di una lista di stopwords e cercando in due campi dei documenti: "title" e "abstract". Cercando su un campo ("title") o tre campi ("title", "abstract" e "terms") risulta solo in un peggioramento dei risultati.

Si possono vedere i risultati ottenuti usando la configurazione "baseline" riassunti in Tabella 1.

| un campo | due campi | tre campi |
|-------------------|-----------------------|-------------------|
| numq all 63 | numq all 63 | numq all 63 |
| numret all 6228 | numret all 6244 | numret all 6271 |
| numrel all 670 | numrel all 670 | numrel all 670 |
| numrelret all 349 | numrelret all 419 | numrelret all 367 |
| map all 0.2130 | map all 0.2744 | map all 0.2155 |

Tabella 1. Risultati treceval complessivi per tutte le query, nessuna manipolazione del testo, numero risultati restituiti per ogni query = 100, pesatura BM25F.

Come si nota dalla tabella il valore baseline di *M.A.P* è 0.2744.

4.3 Esperimenti

Gli esperimenti sono costituiti in tre prove in cui si cambia l'insieme di stopwords usato nell'indicizzazione. In particolare ad ogni prova si utilizza un insieme di stopwords con parole in più rispetto a quello delle prove precedenti.

Come con la baseline, per ciascuna prova si effettua la ricerca per uno, due e tre campi.

Prima prova: Per la prima prova sono state utilizzate le stopwords generali della lingua inglese. Queste comprendono termini come "the", "that", "is"; sono poi state aggiunte anche le lettere dell'alfabeto ed i numeri.

Come si vede in Tabella 2, non sembra si sia ottenuto un miglioramento significativo rispetto alla baseline. Il MAP ottenuto utilizzando due campi "migliora" solo dello 0.0001.

| un campo | due campi | tre campi |
|--------------------------|--------------------------|--------------------------|
| <i>numq all 63</i> | <i>numq all 63</i> | <i>numq all 63</i> |
| <i>numret all 6228</i> | <i>numret all 6244</i> | <i>numret all 6271</i> |
| <i>numrel all 670</i> | <i>numrel all 670</i> | <i>numrel all 670</i> |
| <i>numrelret all 350</i> | <i>numrelret all 418</i> | <i>numrelret all 367</i> |
| map all 0.2137 | map all 0.2745 | map all 0.2152 |

Tabella 2. Risultati treceval, rimozione delle stopwords generali, numero risultati restituiti per ogni query = 100, pesatura BM25F.

Seconda prova: Nella seconda prova si è allora cercato di migliorare i risultati aggiungendo alle stopwords generali, alcune parole che non sono normalmente considerate stopwords ma che nel contesto medico, come nel caso della collezione OHSUMED, diventano tali.

Per questo abbiamo utilizzato un insieme di stopwords denominate *stopword_cliniche* che oltre ad avere le stopwords generali ha anche parole come "medical", "condition" o "family"¹⁰.

Con questa prova sono stati ottenuti risultati accettabili, infatti come si vede in Tabella 3, l'aumento del MAP rispetto alla baseline è stato quasi dello 0.01:

| un campo | due campi | tre campi |
|--------------------------|--------------------------|--------------------------|
| <i>numq all 63</i> | <i>numq all 63</i> | <i>numq all 63</i> |
| <i>numret all 6228</i> | <i>numret all 6244</i> | <i>numret all 6271</i> |
| <i>numrel all 670</i> | <i>numrel all 670</i> | <i>numrel all 670</i> |
| <i>numrelret all 350</i> | <i>numrelret all 419</i> | <i>numrelret all 376</i> |
| map all 0.2156 | map all 0.2837 | map all 0.2166 |

Tabella 3. Risultati treceval, rimozione delle stopwords cliniche, numero risultati restituiti per ogni query = 100, pesatura BM25F.

¹⁰ <https://github.com/kavgan/clinical-concepts/blob/master/clinical-stopwords.txt>

Terza prova: Nell'ultima prova sono state aggiunte altre parole alla lista delle stopwords. La scelta delle parole da aggiungere è stata fatta in due fasi principali.

Nella prima è stata ricavata la frequenza con cui le parole delle query appaiono nei rispettivi documenti rilevanti. Dopo di ciò le parole con frequenza prossima allo 0 sono state prese in considerazione come possibili stopwords.

Nella seconda fase sono state prese le query che, basandosi sui risultati della seconda prova, hanno avuto un MAP relativamente basso (inferiore a 0.2). Sono state quindi prese alcune parole di queste query come altre possibili stopwords anche basandosi sulla loro frequenza nei documenti rilevanti.

Per alcune parole la scelta se includerle nell'insieme di stopwords è stata resa difficile dal fatto che, essendo presenti in più di una query la loro rimozione poteva significare il miglioramento di un MAP ed il peggioramento di un altro.

Per questo, prima di arrivare all'insieme finale sono stati fatti alcuni tentativi.

Come si può notare dalla Tabella 4 si ha un aumento del MAP rispetto alla baseline decisamente superiore alle prove precedenti, con circa lo 0.075 in più.

| un campo | due campi | tre campi |
|--------------------------|------------------------------|--------------------------|
| <i>numq all 63</i> | <i>numq all 63</i> | <i>numq all 63</i> |
| <i>numret all 5282</i> | <i>numret all 5690</i> | <i>numret all 5788</i> |
| <i>numrel all 670</i> | <i>numrel all 670</i> | <i>numrel all 670</i> |
| <i>numrelret all 386</i> | <i>numrelret all 463</i> | <i>numrelret all 435</i> |
| <i>map all 0.2720</i> | <i>map all 0.3508</i> | <i>map all 0.2813</i> |

Tabella 4. Risultati treceval, rimozione delle stopwords cliniche migliorate, numero risultati restituiti per ogni query = 100, pesatura BM25F.

La Figura 2 riassume i risultati delle tre prove e si può notare un netto miglioramento utilizzando le stopwords migliorate.

4.4 Test di significatività :

Per avere conferma che le differenze tra i MAP ottenuti nelle varie prove e nella baseline siano significativamente diversi si è ritenuto opportuno effettuare dei test. Tra i test che si utilizzano comunemente a questo scopo si è scelto di utilizzare il test di Wilcoxon che a differenza dell'alternativo t-test risulta avere più potenza statistica in caso di non normalità dei dati.

I test sono stati fatti tra i MAP delle prove utilizzando due campi in quanto hanno dato generalmente risultati migliori.

L'ipotesi alternativa è unilaterale ed equivale a dire che, tra le due prove confrontate, la seconda prova ha MAP maggiore della prima.

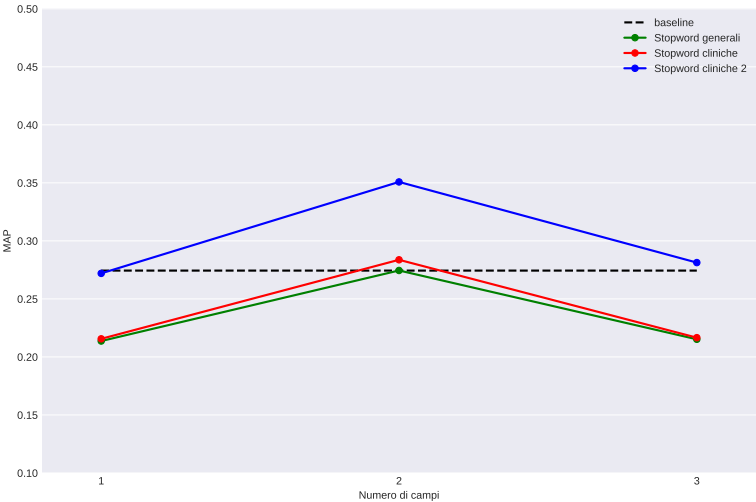


Figura 2. Variazione del MAP cambiando il numero di campi per i tre insiemi di stopwords, la linea orizzontale indica la baseline di 0.2744.

| Prove a confronto | statistica test | p-value |
|-----------------------|-----------------|------------|
| BASELINE contro STOP1 | 582 | 0.0393 |
| BASELINE contro STOP2 | 551.5 | 0.0092 |
| STOP1 contro STOP2 | 549.5 | 0.0056 |
| BASELINE contro STOP3 | 276.0 | 1.2749e-06 |
| STOP1 contro STOP3 | 291.0 | 1.2936e-06 |
| STOP2 contro STOP3 | 290.0 | 3.5437e-06 |

Tabella 5. Statistiche test e livelli di significatività osservati per i test sulla differenza dei MAP. STOP1 indica le stopwords generali, STOP2 indica le stopwords cliniche e STOP3 indica le stopwords cliniche migliorate.

5 Conclusione

Riferimenti bibliografici

1. W. Bruce Croft and Donald Metzler and Trevor Strohman. Search Engines: Information Retrieval in Practice. Addison Wesley, (2009), pp. 250-252
2. W. Bruce Croft and Donald Metzler and Trevor Strohman. Search Engines: Information Retrieval in Practice. Addison Wesley, (2009), pp. 90
3. W. Bruce Croft and Donald Metzler and Trevor Strohman. Search Engines: Information Retrieval in Practice. Addison Wesley, (2009), pp. 313



(a) Alessandro Stefani



(b) Caterina Buranelli



(c) Cristi Gutu