Da mettere alla fine degli esperimenti

Alessandro Stefani[1], Caterina Buranelli[2], and Cristi Gutu[3]

[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) e' quello di reperire documenti rilevanti relativi a una certa esigenza informativa[4]; dunque i documenti sono il primo ingresso del sistema, mentre il secondo e' costitutito dalle interrogazioni; i documenti devono essere indicizzati e nell'indice creato, si andra' a effettuare la ricerca per reperire documenti rilevanti[5]. Questa seconda parte e' detta reperimento e non si occupa solo di ricercare tra i documenti, ma anche di riordinare secondo un certo ordine di rilevanza. Cio' che descrive i documenti e cio' che descrive le interrogazioni deve essere confrontabile, infatti nei programmi di indicizzaizone e di reperimento si usa uno schema, che deve essere uguale in entrtambi i casi. L'indicizzazione e' un trade-off tra il miglioramento della rapprerentazione del contenuto informativo dei documenti (efficacia) e la gestione degli indici (efficienza). Benche' ci sia sempre tensione tra queste due caratteristiche, ad oggi la questione piu' studiata e' quella dell'efficienza.

[4] insieme delle circostanze in cui unapersona ha un problema da risolvere o un compito da svolgeree richiede informazioni importanti, utili o necessarie per larisoluzione del problema o lo svolgimento del compito

[5] la rilevanza e' la proprieta' 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 pesatura[6] BM25F [1].

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 OHSUMED.

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 ciè 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 BibTEX se si scrive con LATEX.

2 Base di partenza

La base di partenza è formata dai metodi documentati nei libri di testo. Si eviti di trascrivere pari pari, si cerchi piuttosto di rielaborare i contenuti in modo da renderli *coerenti* col resto della relazione; in particolare, si descrivano tutti e solo i metodi usati negli esperimenti e si eviti di parlare di quei metodi che poi non sono stati usati; ad esempio, se si conducono degli esperimenti con BM25F, si deve descrivere questo schema di pesatura in questa sezione.

3 Esperimenti

Per gli esperimenti si è utilizzato la collezione sperimentale chiamata OHSUMED[7], la collezione contiene circa 54711 documenti. Gli esperimenti effettivi sono iniziati solamente dopo aver stabilito quale e' la baseline dalla quale partiamo, in altre parole ci siamo chiesti quale sia la base di parteza dalla quale dobbiamo migliorare il nostro sistema di ricerca.

[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

Affinchè Whoosh possa indicizzare una collezione di documenti, necessita la specificazione di uno schema, che include per ogni campo, nome del campo e tipo di quel campo.

```
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: Schema necessario all'indicizzazione dei documenti.

Definito lo schema abbiamo sfruttato la configurazione strettamente necessaria senza alcun tuning per avere un sistema di reperimento "minimale".

La configurazione baseline ideale, è stata scelta in base al Mean Average Precision (M.A.P.)[3] variando il parametro che indica quale schema di pesatura usare nel processo di reperimento tra cui TF_IDF e BM25F .

Il processo e codice di indicizzazione sono facilmente comprensibili visionando il file *indicizzazione_batch_baseline.py*.

Per eseguire l'indicizzazione baseline è sufficiente lanciare il seguente script python con il comando:

```
python indicizzazione_batch_baseline.py \
cartella_indice file_documenti.xml
```

Per eseguire il reperimento che poi produce il file treceval baseline basato sullo schema TF_IDF è sufficiente lanciare lo script python con il comando:

```
python reperimento_batch_baseline.py cartella_indice \
file_query.xml 1 > reperimento_baseline.treceval
```

Per eseguire il reperimento che poi produce il file treceval baseline basato sullo schema BM25F è sufficiente lanciare lo script python con il comando:

```
python reperimento_batch_baseline.py cartella_indice \
file_query.xml 2 > reperimento_baseline.treceval
```

3.1 Risultati baseline con schema di pesatura TF IDF

Descrizione: "1 Campo" significa che il reperimento e' stato eseguito soltanto valutando il campo title, "2 Campi" titolo e abstract, "3 Campi" titolo e abstract e terms.

3.2 Risultati baseline con schema di pesatura BM25F

Descrizione: "1 Campo" significa che il reperimento e' stato eseguito soltanto valutando il campo title, "2 Campi" titolo e abstract, "3 Campi" titolo e abstract e terms.

1 Campo	**2 Campi**	**3 Campi**
num*q all 63*	num*q all 63*	num*q all 63*
num*ret all 37454*	num*ret all 57356*	num*ret all 58456*
num*rel all 670*	num*rel all 670*	num*rel all 670*
num*rel*ret all 305	num*rel*ret all 380	num*rel*ret all 382
map all 0.0833	map all 0.0591	map all 0.0829

Figura: Risultati treceval, nessuna manipolazione del testo, numero risultati restituiti per ogni query = 1000, pesatura TF IDF.

1 Campo	**2 Campi**	**3 Campi**
num*q all 63*	num*q all 63*	num*q all 63*
num*ret all 37454*	num*ret all 57356*	num*ret all 58456*
num*rel all 670*	num*rel all 670*	num*rel all 670*
num*rel*ret all 307	num*rel*ret all 387	num*rel*ret all 383
map all 0.1073	map all **0.1289**	map all 0.1227

Figura: Risultati treceval, nessuna manipolazione del testo, numero risultati restituiti per ogni query = 1000, pesatura BM25F.

Alla luce dei risultati si e' scelto come valori per i parametri baseline: *Documenti rilevanti reperiti*: 387; $M.A.P$: 0.1289.

3.3 Primo tentativo: uso delle stopword

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

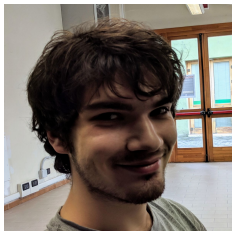
Inizialmente si è pensato di utilizzare le stopword generali della lingua inglese per eliminare le parole che non portano informazione significativa. I risultati sono stati accettabili, ma siamo riusciti a migliorarli togliendo anche stopword cliniche, cioè strettamente inerenti al contesto medico e abbiamo ottenuto i seguenti risultati:

1 Campo	**2 Campi**	**3 Campi**
num*q all 63*	num*q all 63*	num*q all 63*
num*ret all 54873*	num*ret all 612669*	num*ret all 61908*
num*rel all 670*	num*rel all 670*	num*rel all 670*
num*rel*ret all 477	num*rel*ret all 570	num*rel*ret all 538
map all 0.2045	map all **0.2752**	map all 0.1665

Figura: Risultati treceval, rimozione delle stopword cliniche, numero risultati restituiti per ogni query = 1000, pesatura BM25F.

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
4. Discovering Related Clinical Concepts Using Large Amounts of Clinical Notes. Ganesan, Kavita and Lloyd, Shane and Sarkar, Vikren, (2016), pp. 27-33



(a) Alessandro Stefani

(b) Caterina Buranelli

(c) Cristi Gutu