Documento finale per la prova d'esame di Sistemi Informativi a.a. 2014/2015

Giuseppe Bianchi
giuseppe.bianchi@studenti.unipd.it
Bianca Marroni
bianca.marroni@studenti.unipd.it
Mario Rossi
mario.rossi@studenti.unipd.it

Sommario Questo documento illustra il perscorso effettuato per arrivare ad ottenere una buona conoscenza del ramo dell'Information Retrieval. Partendo dalle conoscenze di base ad ogni esercitazione stato aggiunto un tassello, il quale non altro che un'attributo in pi da considerare per il calcolo della precisione dei file ritenuti da noi importanti. All'interno del documento saranno spiegati i diversi progetti effettuati. Come siamo passati da un'indicizzazione manuale ad una automatica (laboratorio 2), l'algoritmo di reperimento implementato per effettuare il ranking dei documenti presenti nella collezione per ogni richiesta, query (laboratorio 3). L'introduzione dei giudizi di rilevanza e la loro influenza sul risultato elaborato dall'algoritmo di reperimento scelto (laboratorio 4). Il calcolo effettuato del pagerank di ogni documento e come quest'ultimo in fluisca sul ranking (laboratorio 5). L'analisi della relazione tra una collezione di documenti e i termini contenuti in essi tramite l'utilizzo della tecnica **LSA** (laboratorio 6). Concludento con l'introduzione e lo studio di **HITS** (Hyperlink-Induced Topic Search): algoritmo di analisi dei link (laboratorio 7).

1 Introduzione

Quando si scrive l'introduzione, si deve tenere presente che il lettore atteso sarà a conoscenza degli elementi di base. Si dovranno quindi introdurre i concetti non normalmente trattati in un corso di base in Information Retrieval e che sono invece utilizzati nel proprio documento.

Il resto del documento sarà scritto con i criteri seguenti:

- esaustività
- precisione
- chiarezza
- correttezza
- sintesi

2 Metodologia

In questo paragrafo, si illustreranno i metodi sviluppati e sperimentati con le attività di laboratorio. Le notazioni e tutti gli aspetti non banali dovranno essere spiegati. Naturalmente, la notazione di un paragrafo non dovrà essere reintrodotta nei paragrafi successivi, di conseguenza, la notazione non dovrà essere ambigua.

2.1 Approccio

In questo paragrafo si illustra l'approccio generale, ad esempio:

- se c'è un'idea generale
- come ci si è organizzati nel lavoro di gruppo
- quali strumenti software particolare sono stati utilizzati
- altri aspetti ritenuti rilevanti alla metodologia in generale

2.2 Indicizzazione

Qui va il contenuto del laboratorio n. 2.

2.3 Reperimento

Qui va il contenuto del laboratorio n. 3. Esso rappresenta la *baseline*.

2.4 Relevance Feedback

Qui va il contenuto del laboratorio n. 4.

2.5 PageRank

Qui va il contenuto del laboratorio n. 5.

2.6 Latent Semantic Analysis

Qui va il contenuto del laboratorio n. 6.

2.7 Hyper-linked Induced Topic Search

Qui va il contenuto del laboratorio n. 7.

2.8 Ottimizzazione

Qui va il contenuto del laboratorio n. 8 (solo per chi ha in programma questo laboratorio).

2.9 Altri metodi

Se sono stati sviluppati altri metodi, descriverli qui.

3 Risultati sperimentali

Questo paragrafo presenta e discute i risultati sperimentali. Si dovranno scegliere tre *run* al massimo per ciascuno dei metodi illustrati nei paragrafi 2.3, 2.4, 2.5, 2.6 e 2.8.

Si dovranno confrontare le misure di efficacia (ad esempio, *Mean Average Precision*, MAP) mediante illustrazioni anche grafiche. Un'analisi della significatività statistica delle differenze tra i valori di MAP sarebbe opportuna.

Un confronto particolare dovrà essere fatto tra la *baseline* del paragrafo 2.3 e i metodi dei paragrafi successivi.

La parte preziosa di questo paragrafo è la discussione dei risultati. Si dovrà dare un'interpretazione ragionata, chiara ed esaustiva delle ipotesi per cui sono state osservate o meno le differenze tra i valori di MAP.

4 Conclusioni

In questo paragrafo si possono aggiungere delle osservazioni di carattere generale sugli esperimenti; ad esempio, si può concludere se un proprio metodo di reperimento o una variazione dei metodi più avanzati hanno portato a qualche miglioramento rispetto alla *baseline*.