Sistemi per il recupero delle informazioni

Laurea Magistrale in Editoria e Giornalismo
Prova scritta del 4 settembre 2012

Cognome e nome:_____ **Matricola:**_____

Domanda 1	Domanda 2	Domanda 3	Domanda 4	Domanda 5	Domanda 6	Totale

Istruzioni:

- È vietato portare all'esame libri, eserciziari, appunti e dispense. Chiunque venga trovato in possesso di documentazione relativa al corso, in formato analogico e/o digitale, – anche se non strettamente attinente alle domande proposte – vedrà annullata la propria prova.
- Scrivere solo sui fogli distribuiti, cancellando le parti di brutta con un tratto di penna. Non separare questi fogli. Non utilizzare la penna rossa. Scrivere nome e cognome su tutti i fogli utilizzati.
- Tempo a disposizione: 1 ora e 45 minuti.

1) Descrivere cosa è, come è composto, e quali obiettivi ha un sistema informativo. Elencare (solo elencare!) i diversi tipi di sistemi informativi.
2) Descrivere, fornendo anche un esempio, cosa sono e i principi su cui si basano le interrogazioni Booleane e le interrogazioni Booleane estese.
3) Si calcoli la lunghezza di ricerca attesa supponendo che l'utente voglia 6 documenti rilevanti e che l'insieme dei documenti recuperati venga suddiviso nei seguenti 2 sottoinsiemi:
    - S1 contiene 5 documenti di cui 3 rilevanti e 2 non rilevanti
    - S2 contiene 5 documenti di cui 3 rilevanti e 2 non rilevanti
4) Dare lo schema del processo di indicizzazione e descrivere in particolare cosa sono e come funzionano lo stemming, la stoplist e la legge di Zipf.
5) Dando per assunte le definizioni di precisione, richiamo, fallout e generalità per la misura dell'efficacia dei sistemi per il recupero delle informazioni, si descriva l'equazione che lega queste quattro misure e il significato delle sue diverse parti.
6) Descrivere come avviene il processo di matching nei sistemi di IR vettoriali.

Soluzione dell'esercizio 3)

L'utente legge i documenti in S1 e trova solo 3 documenti rilevanti sui 6 desiderati, ma per farlo ha comunque dovuto esaminare anche gli altri 2 documenti nell'insieme, quindi: 3 documenti rilevanti trovati e 5 documenti letti finora.
Non avendo ancora trovato il numero di documenti rilevanti desiderati, deve leggere anche S2. A questo punto deve trovare tutti tre i documenti rilevanti dei 3 disponibili in S2, quindi il numero di documenti che l'utente deve esaminare in S2 dipende dalla posizione di questi tre documenti rilevanti nella lista dei 5 documenti in S2.
Non sapendo come il sistema di IR ordina i documenti all'interno dei sottoinsiemi essendo ugualmente rilevanti, dobbiamo assumere che l'ordinamento in S2 sia casuale, e qui entra in gioco la teoria delle variabili casuali. Essendo l'ordinamento in S2 casuale, tutte le possibili combinazioni/ordini di tre documenti rilevanti e due non rilevanti hanno la stessa probabilità di essere fornite all'utente. Quindi si calcola il numero medio (o valore atteso) di documenti da leggere per trovare i tre documenti rilevanti sui 5 facendo la media su tutti i possibili ordinamenti dei documenti in S2. Tutti i possibili ordinamenti sono 10:

1.	R	R	R	NR	NR	3
2.	R	R	NR	R	NR	4
3.	R	R	NR	NR	R	5
4.	R	NR	R	R	NR	4
5.	R	NR	R	NR	R	5
6.	R	NR	NR	R	R	5
7.	NR	R	R	R	NR	4
8.	NR	R	R	NR	R	5
9.	NR	R	NR	R	R	5
10.	NR	NR	R	R	R	5

All'utente servono tre documenti rilevanti quindi accanto ad ogni possibile ordinamento è stato riportato il numero di documenti da leggere per arrivare al terzo documento rilevante. Come si vede:

- in un solo caso (il 1°) su 10 l'utente legge 3 soli documenti perché trova subito i documenti rilevanti;
- in 3 casi (il 2°, 4° e 7°) su 10 l'utente legge 4 documenti;
- nei rimanenti 6 casi su 10 l'utente legge 5 documenti.

A questo punto, il valore atteso di documenti che l'utente deve leggere per trovare i primi due documenti rilevanti in S2 è la media del numero di documenti da leggere nei vari casi ma pesato per il numero di combinazioni per ognuno dei casi, cioè:

$$\frac{1}{10}\times 3+\frac{3}{10}\times 4+\frac{6}{10}\times 5=\frac{3}{10}+\frac{12}{10}+\frac{30}{10}=\frac{45}{10}=4{,}5$$

Questo, infine, va sommato al numero di documenti già letti dall'utente in S1, cioè:

$$5+4{,}5=9{,}5$$

La lunghezza di ricerca attesa è quindi 9,5.