Sistemi informativi (corso progredito)
a.a. 2014/2015

Laboratorio n. 7

Hyperlinked Induced Topic Search

Massimo Melucci

Obiettivi

- Comprensione di Hyperlinked Induced Topic Search (HITS).
- Un'implementazione di HITS.
- Una funzione di reperimento che integra HITS.

Come funziona

Dopo aver implementato l'algoritmo di HITS per una data interrogazione, ci sono diversi modi per integrarlo, ad esempio:

1. si produce il primo *ranking* (vedi laboratorio n. 2) con o senza RF, a scelta (vedi laboratorio n. 4)
2. si fissa un numero $N \geq 1$ e si usano i primi N documenti reperiti
3. si produce una matrice di adiacenza di dimensione $N \times N$ sulla base delle citazioni (si veda il laboratorio n. 5)
4. si calcola HITS
5. si combinano i punteggi di autorevolezza, centralità e rilevanza secondo una funzione definita appositamente
6. si riordinano i primi N documenti; gli altri documenti sono accodati ai primi N riordinati oppure si combinano i punteggi anche per essi (autorevolezza e centralità saranno nulli)

Base di partenza

- Si vedano i pacchetti dei laboratori precedenti.
- I dati della lezione precedente.

Procedimento

Il procedimento è in un due passi, uno per ogni consegna:

- Alla prima consegna si deve produrre un documento di testo che illustra la propria funzione di reperimento integrata con HITS. Il nome del documento deve essere nel seguente formato: `lab-7-gruppo-`*n*`.txt`, dove *n* è il numero del gruppo; il formato può anche essere Microsoft Word (si usi l'estensione `.doc` o `.docx`) o Adobe PDF (si usi l'estensione `.pdf`).
- continua...

Procedimento

- ...
- Alla seconda consegna, si devono produrre i seguenti risultati:
  - La versione definitiva del testo fornito alla prima consegna.
  - I risultati dell'esecuzione della funzione di reperimento nel seguente formato testuale:

    Id.Int. `Q0` Id.Doc. Rango Punteggio EtichettaRun

    dove EtichettaRun è l'etichetta che identifica la *run* nel formato `G`*n*`R`*m*`HITS` dove *n* è il numero del proprio gruppo, *m* è il numero della run da 0 a 9 e `HITS` sta per Latent Semantic Analysis; ad esempio `G17R3HITS` è la *run* n. 3 del gruppo n. 17 con HITS
- continua...

Procedimento

- ...
- Ad esempio:

```
1   Q0   1234     1   1.234   G17R3HITS
1   Q0   345      2   1.056   G17R3HITS
1   Q0   2909     3   1.056   G17R3HITS
...
1   Q0   12     114   0.056   G17R3HITS
1   Q0   879    115   0.034   G17R3HITS
1   Q0   3204   116   0.023   G17R3HITS
2   Q0   12       1   3.467   G17R3HITS
2   Q0   879      2   3.123   G17R3HITS
...
```

- **Attenzione**: non si aggiunga l'intestazione appena descritta.
- Si metta la *run* in un file di testo omonimo, ad esempio, `G17R3HITS.txt`
- Si carichi un archivio compresso con all'interno tutti i file.