Sistemi per il recupero delle informazioni

Gabriele Pozzani

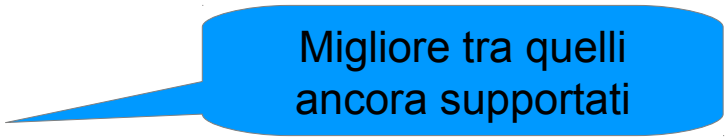
A.A. 2012/2013

Corso di Laurea Magistrale in Editoria e Giornalismo

Software Open-Source per l'Information Retrieval

SW per l'IR testuale

- Esempi di software per l'IR testuale sono:
  - HtDig
  - Indri
  - Lucene
  - MG4J
  - Omega
  - OmniFind
  - SWISH-E
  - SWISH++
  - Terrier
  - Zettair



Apache Lucene

- Lucene è una libreria (insieme di funzioni) per l'indicizzazione di documenti testuali
  - Sviluppato da Apache Software Foundation
- Supporta l'indicizzazione e la ricerca full-text
- Utilizzato anche da:
  - Wikipedia
  - La Repubblica
  - Apple
  - IBM
  - AOL
  - LinkedIn
  - Libreria Nazionale di Firenze

Apache Lucene, indicizzazione

- L'indicizzazione avviene costruendo un indice inverso posizionale
  - L'indice è composto da più file che memorizzano tra l'altro
    - I termini del vocabolario
    - La frequenza dei termini nei documenti
    - La posizione dei termini

Apache Lucene, scoring

- Lucene è basato sul modello booleano esteso
  - Combina insieme il modello booleano e quello vettoriale
    - Le query sono booleane
    - I documenti che rispettano le query booleane sono pesati utilizzando il modello vettoriale
  - Il peso di un termine in un documento è calcolato usando TF-IDF
  - La similarità di un documento ad una query è calcolata usando la misura del coseno

Apache Lucene, querying (I)

- Lucene supporta innanzitutto le query booleane
  - OR

    `cane OR gatto`
  - AND

    `cane AND gatto`
  - NOT

    `cane NOT gatto`
  - +
    - Richiede l'assoluta presenza di un termine

    `+cane gatto` [documenti in cui compare sicuramente cane ed eventualmente anche gatto]

Apache Lucene, querying (II)

- Lucene supporta le query basate su campi
  - Possono essere indicizzati anche metadati dei documenti
    - Titolo, autore, data di pubblicazione, …
  - La ricerca può essere eseguita all'interno di tutti i campi o solo in un campo specifico

    `title:"La sindrome di Anastasia"`

Apache Lucene, querying (III)

- Lucene supporta inoltre i cosiddetti modificatori
  - Ricerca di frasi

    `"oro nero"`
  - Caratteri jolly
    - ? : rappresenta un qualunque carattere

      `m?go` → mugo, mago
    - * : rappresenta una qualunque sequenza (anche nulla) di caratteri

      `tim*` → tim, timo, timido, timore, …
  - Ricerca per somiglianza sintattica (fuzzy search)

    `sweet~` → sweet, sweety, tweet, ween, swept, …

Apache Lucene, querying (IV)

- Ricerca per prossimità

  `"cane gatto"~10`

  [documenti in cui compaiono sia cane che gatto a non più di 10 termini di distanza]
- Ricerca per intervallo

  `[cane TO canide]`

  [documenti con un qualunque termine che nell'ordine alfabetico si trovi tra “cane” e “canide”]
- Boosting dei termini
  - Permette di dire che un termine è più importante di altri e deve avere un peso maggiore nel calcolo della similarità tra query e documenti

  `cane^3 gatto`

Progetti basati su Lucene

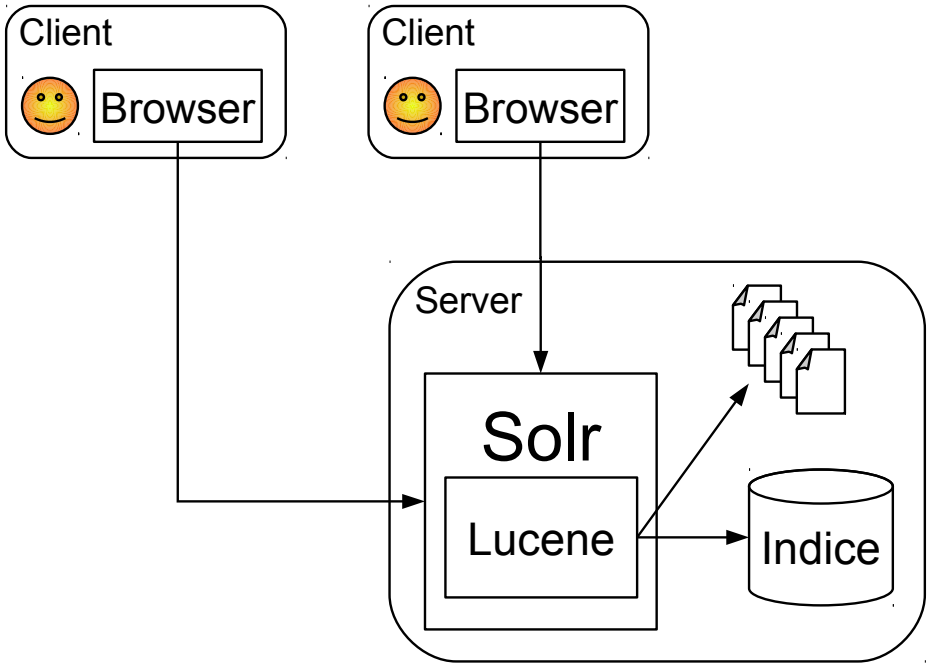
- Lucene non ha un'interfaccia grafica
  - È solo una libreria
- Diversi altri progetti open-source usano o estendono Lucene
  - Luke – Lucene Index Toolbox
    - Fornisce un'interfaccia grafica
  - Apache Solr
    - Server WEB per la ricerca in indici creati con Lucene
    - Usato per creare la propria libreria digitale consultabile via WEB
    - Usato da Repubblica.it
  - Apache Nutch
    - Web crawler che poi usa Lucene per indicizzare pagine WEB
    - Comunica con Solr per rendere fruibili i propri indici
  - DocFetcher
    - Motore di ricerca desktop (simil Spotlight, Nepomuk, Windows Search Tool)

Luke

- È uno software di sviluppo e diagnostico
  - Accede ad indici di Lucene già esistenti
  - Permette tra l'altro di:
    - Scorrere i documenti per ID o per termine
    - Visualizzare i documenti e copiarli nella clipboard
    - Recuperare una lista dei termini più frequenti
    - Eseguire una ricerca e visualizzare i risultati
    - Eliminare selettivamente documenti dall'indice
    - Ottimizzare gli indici

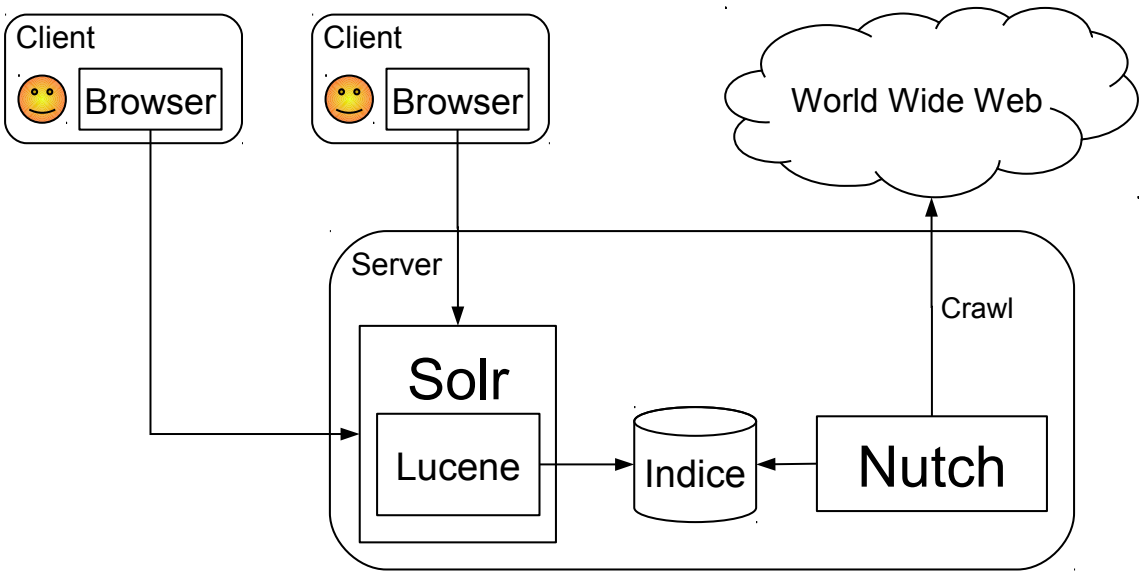
Apache Solr

- Server Web basato su Java



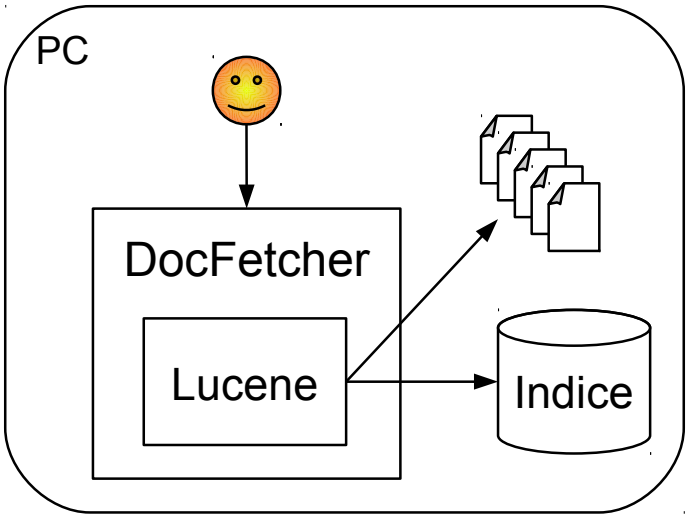
Apache Nutch

- Web crawler
  - Permette di pubblicare l'indice creato dal crawling tramite Solr



DocFetcher

- Indicizzazione e ricerca file desktop
  - Supporta molti formati di file
    - HTML, PDF, MS Office, Open Document Formats
    - Zip, rar, tar, 7z



Riferimenti

- Lucene
  - http://lucene.apache.org/
- Luke
  - http://code.google.com/p/luke/
- Solr
  - http://lucene.apache.org/solr/
- Nutch
  - http://nutch.apache.org/
- DocFetcher
  - http://docfetcher.sourceforge.net/