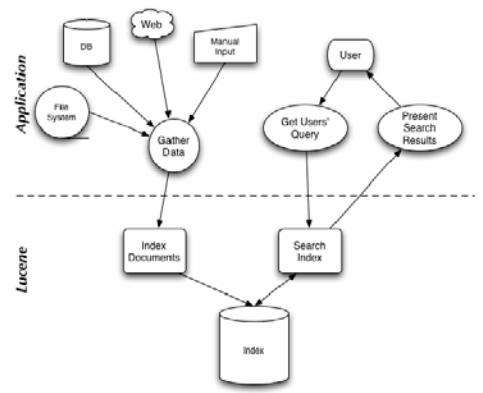
Lucene

Lucene

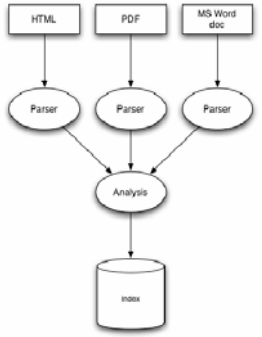


Lucene

- E' una API (Application Programming Interface) sviluppata in Java
  - Estremamente efficiente e semplice da usare
- Mette a disposizione i blocchi fondamentali per costruire un indicizzatore e un motore di ricerca
- Fa parte del progetto Apache
  - Disponibile online: http://lucene.apache.org

Indicizzazione in Lucene

- Tre fasi:
  - Conversione in testo
  - Analisi
  - Salvataggio nell'indice



Preliminari

- Impostare la variabile di ambiente CLASSPATH
  - **c:\lucene\lucene-core.jar; c:\lucene\lucene-demos.jar**
- Lanciare il promt dei comandi
  - “start” -> “programmi” -> “accessori” -> “prompt dei comandi”
- Spostarsi nella cartella “c:\lucene”
  - cd c:\lucene

Indicizzazione di documenti

- Eseguire il programma di indicizzazione
  - **java org.apache.lucene.demo.IndexFiles <cartella_docs>**
  - es: java org.apache.lucene.demo.IndexFiles c:\lucene\docs
- Compare una cartella “index”
  - contiene l’indice creato da Lucene
  - per evitare che venga sovrascritta nelle prove successive, rinominarla opportunamente e spostarla nella cartella c:\lucene\indexes
- Il contenuto dell’indice può essere consultato con il programma “luke”
  - Clickare sul file “lukeall-0.7.jar” nella cartella “c:\lucene” (richiede la JVM 1.5)
  - o accedere al sito http://www.getopt.org/luke/webstart.html

Interrogazione dell’indice

- Eseguire il query parser
  - **java org.apache.lucene.demo.SearchFiles <cartella_index>**
  - <cartella_index> contiene un indice creato precedentemente
- Digitare le query a linea di comando
  - E’ possibile specificare un insieme di keywords
  - Esempi:
    - vector
    - query parser

Sintassi del query parser

- Una query è costituita da termini e operatori
- Due tipi di termini:
  - Termine Singolo: costituito da una sola parola
    - Es: test
  - Frase: un gruppo di parole racchiuse fra doppi apici
    - Es: "hello dolly"
- Più termini possono essere combinati con operatori Boolean per costituire query più complesse

Campi

- Lucene permette di indicizzare vari campi di un documento:
  - Es: nome, path, contenuto (o differenti sezioni del testo)
- Per ogni termine si può indicare il campo in cui cercarlo, altrimenti si usa il campo di default
  - Sintassi **<nome_campo>:<termine>**
- Esempio:
  - Assumiamo che l'indice abbia i campi *modified* e *contents* (default)
  - Per cercare un documento modificato il 20/02/2006 e contenente il testo "vector", si possono usare le seguente query:
    modified:20060220* AND contents:vector
    modified:20060220* AND vector
- Attenzione: l'indicatore di campo vale solo per il termine immediatamente successivo
  - La query *modified:20060220* 2007** cerca solo "*20060220**" nel campo modified, mentre "*2007**" è cercato nel campo di default

Operatori booleani

- OR (default)
  - Es: documenti contenenti "jakarta apache" o solo "jakarta":
    - "jakarta apache" jakarta
    - "jakarta apache" OR jakarta
- AND
  - "jakarta apache" AND "Apache Lucene"
- **+** ***(required operator)***
  - Il termine dopo "+" symbol deve comparire
  - Es.: **+jakarta apache**
    documenti che devono contenere "jakarta" e possono contenere "lucene"

Operatori booleani

- **NOT**
  - Esclude i documenti che contengono il termine
  - Es.: **"jakarta apache" NOT "Apache Lucene"**
    documenti che contengono "jakarta apache" ma non "Apache Lucene"
  - L'operatore NOT non può essere usato da solo (NOT "jakarta apache")
- ***- (prohibit operator)***
  - Differenza fra insiemi: equivalente al NOT
  - Es: "jakarta apache" -"Apache Lucene"
    documenti che contengono "jakarta apache" ma non "Apache Lucene"

Espressioni complesse

- Si possono usare le parentesi tonde per raffruppare clausole e formare sotto-query
- Es: (jakarta OR apache) AND website
  Il termine *website* deve esistere mentre uno dei termini *jakarta* e *apache* possono esistere
- Si può usare anche per raggruppare i termini ce si riferiscono allo stesso campo
  - title:(+return +"pink panther")

Wildcard Searches

- Lucene supporta ricerche con caratteri jolly
- "?": single-character wildcard
  - Sostituisce un singolo carattere
  - Es: “te?t” per cercare sia "text" sia "test"
- "*": multiple character wildcard
  - sostituisce 0 o più caratteri
  - Es: “test*” per cercare sia “tests” sia “tester”
- E’ possibile usarli in qualunque parte del termine tranne che nel primo carattere

Fuzzy Searches

- Si basano sulla distanza di Levenshtein fra stringhe (o Edit Distance)
  - # operazioni di edit (inserimento, cancellazione, sostituzione di un carattere) necessarie per trasformare una stringa nell’altra
    - Esempi: *dist(casa,casta)=1*, *dist(cassa,casta)=2*
  - I valori di distanza sono normalizzati e convertiti in similarità
- Sintassi: usare il simbolo "~“ alla fine di un Termine Singolo
  - Se "~“ non è sulla tastiera, digitare **1 2 6** tenendo premuto il tasto *Alt*
- Un parametro addizionale permette di specificare una soglia minima di similarità, compreso fra 0 e 1
  - La similarità è 1 (massima) quando i due termini coincidono
  - Similarità di default = 0.5
- Esempio:
  - operat~0.6
  - Si troveranno termini come “operator” e “operation”

Proximity Searches

- E’ possibile chiedere che le parole appaiano vicine nel testo
- Sintassi: **<Frase> ~ distanza**
- Esempio:

  "jakarta apache"~10

  richiede che il documento contenga "apache“ e "jakarta" ad una distanza massima di 10 parole

Range Searches

- Permettono di cercare documenti i cui campi sono compresi in un certo intervallo
  - Si applicano a campi non testuali
  - Viene considerato l’ordinamento lessicografico
- Due modi:
  - **<nome_campo>:[<valore1> TO <valore2>]**

    Il campo deve assumere un valore compreso tra i due estremi *valore1* e *valore2*, inclusi
  - **<nome_campo>:{<valore1> TO <valore2>}**

    Il campo deve assumere un valore nell’intervallo, estremi esclusi
- Esempio
  - **modified:[20060302 TO 20060303]**

    Trova documenti modificati tra il 20060302 ed il 20060303, estremi inclusi

Pesatura dei Termini

- Per dare maggiore importanza ad un termine, si usa il simbolo "^" seguito da un fattore di boost
- Sintassi **<termine>^<fattore_boost>**
- Esempio: **jakarta^4 apache**

  Sono preferiti i documenti in cui *jakarta* ha peso maggiore
- Fattore di boost
  - Deve essere positivo
  - Per default, è 1.
  - Può essere minore di 1 (es. 0.2): il termine viene "declassato"

Sintassi delle query – sommario (1)

Query	Example	Notes
single term	document	Searches for documents that contain "document" term in the default field.
phrase	"important document"	Searches for documents that contain the phrase "important document" in the default fiels.
searching fields	title:document	Searches for documents that contain "document" term in the "title" field.
wildcard search	doc?ment	Single-character wildcard search. It will match "document" and "dociment" but not "docooment".
	document*	Multi-character wildcard search. It will match "document" and "documentation".
fuzzy search	document~	Search based on similar spelling.
	document~0.9	Search based on similar spelling. 0.9 is the required similarity (default: 0.5)
proximity search	"important document"~5	Find words of a phrase that are not next to each other. Maximum distance in this example is 5 words.
range search	author:{Einstein TO Newton}	Searches for document with "author" field value between specified values.
	date:{20050101 TO 20050201}	Searches for document with "date" field (DateTime type) value between specified dates.

Sintassi delle query – sommario (2)

Query	Example	Notes
relevance	important^4 document	Set boost factor of the term "important" to 4. Default boost factor is 1.
	"important document"^4 "search engine"	You can set boost factor for phrases too.
OR operator	important document	"OR" is the default operator.
	important OR document	The default field must contain either "important" or "document".
AND operator	important AND document	The default field must contain both word.
+ operator	important +document	The default field must contain "document" and may contain "important".
NOT/- operator	-important document	The default field must contain "document" but not "important".
grouping	(important OR office) AND document	Use parentheses for expression grouping.
	author:(Einstein OR Newton)	Parentheses work with fields as well.
relevance	important^4 document	Set boost factor of the term "important" to 4. Default boost factor is 1.

Sintassi delle query – Query non consentite

Query	Examples
wildcard at the beginning of a term	?ocument, *ocument
stop words	a, the, and
special characters: +- && || ! ( ) { } [] ^ " ~ * ? :\	\+, \: