Motori di ricerca

Andrea Marin

Università Ca' Foscari Venezia
SVILUPPO INTERCULTURALE DEI SISTEMI TURISTICI
SISTEMI INFORMATIVI PER IL TURISMO

a.a. 2013/2014

Section 1

Information Retrieval e Motori di Ricerca

Un mare di pagine

- Google stima che esistano circa 1000 miliardi di pagine web
- Circa 9 miliardi sono indicizzate da Google
- Non tutte le pagine sono aggiornate/attuali
- Non tutte le pagine sono accessibili da tutti
- Problema: come trovare l'informazione di cui abbiamo bisogno?

Problemi da risolvere

- Due principali problemi da risolvere

1. Inserire e gestire l'informazione estratta dal web in una base di dati
2. Rispondere alle ricerche degli utenti interrogando la base di dati

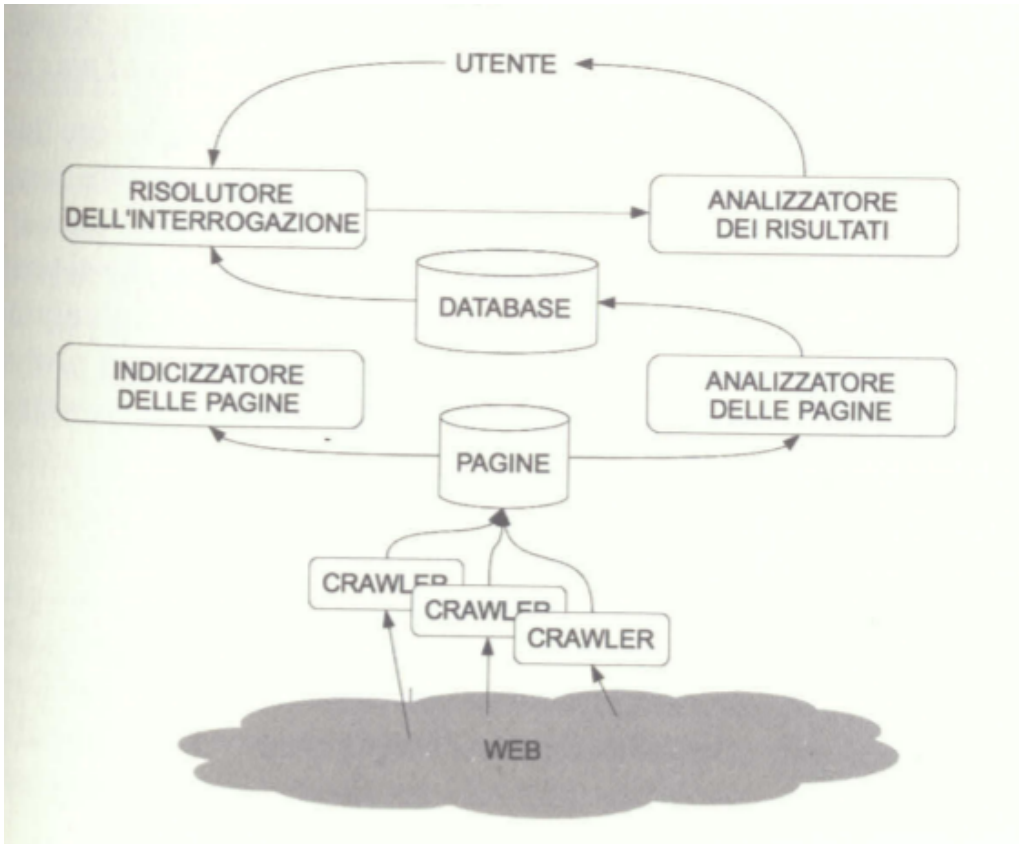
- Perchè è importante capire come funzionano i motori di ricerca?

Soluzioni

1. Il primo problema si risolve con agenti automatici che sondano il web esplorando le pagine ad aggiungere al database
2. Il secondo problema si risolve offrendo agli utenti un linguaggio di interrogazione per esprimere le loro ricerche

- Un problema aggiuntivo consiste nel presentare i risultati della ricerca dando maggior risalto ai risultati più rilevanti...
  - Cosa si intende per rilevante?

Architettura dei motori di ricerca



I crawler

- Il crawler è un agente automatico che esplora il web
  - Googlebot è il crawler di Google!
- Il web viene esplorato a partire da pagine conosciute inseguento i link ipertestuali trovati
- Una lista di pagine visitate è tenuta per evitare comportamenti ciclici
  - Si torna su una pagina solo se è stata aggiornata dall'ultima visita
- Le informazioni raccolte dai crawler verranno indicizzate successivamente

Indicizzazione delle pagine

- Indicizzare significa creare un struttura dati che consente l'accesso veloce a delle informazioni
  - Esempio: indice analitico nei libri
- L'indice consente di calcolare i risultati della ricerca in modo efficiente

Di cosa parla una pagina web?

- Per scoprire il contenuto principale di una pagina web si possono usare le frequenze di occorrenza
  - Se in un testo di m parole, una particolare parola compare n volte allora la sua frequenza è n/m
- Legge di Zipf: ordinando le parole di un testo per frequenza, la seconda ha frequenza dimezzata rispetto alla prima, la terza pari ad un terzo e così via
- L'80% di un testo è composto per il solo 20% di parole diverse
- La frequenza di occorrenza serve al motore per decidere la pertinenza di una pagina con la ricerca

Linguaggi per la ricerca nel web (Google) /1

query:

- termine
- query AND query
- query OR query
- -query
- +termine
- site:x
- link:x
- related:x
- allintitle:query
- ...

Linguaggi per la ricerca nel wel (Google) /2

termine:

- parola
- “frase”
- termine * termine

Esempio

- Interrogare Google per trovare dei file *pdf* che si trovino sotto il dominio *it*. I file devolo parlare di *world wide web* oppure avere soltanto il *web* ma usando il termine nel titolo ed escludendo i file che di *www*

Soluzione:

```
filetype:pdf ((‘‘-www intitle:web) OR (’’world wide
                    web‘‘)) site:it
```

Analisi dei risultati

- Risultato: Un insieme di pagine che rispettano i criteri dell'interrogazione
- Possibilità di imprecisioni
  - Pagine pertinenti non elencate
  - Pagine non pertinenti comunque elencate
- Come valutare l'accuratezza e la rilevanza dei risultati?

Accuratezza

- Insieme delle pagine volute: P
- Insieme delle pagine trovate: T
- Pertinenza:
$$\frac{|T \cap P|}{|T|}$$
- Precisione:
$$\frac{|T \cap P|}{|P|}$$
- Silenzio:
$$\frac{|P \setminus T|}{|P|}$$
- Rumore:
$$\frac{|T \setminus P|}{|T|}$$

Rilevanza

- I contenuti di un risultato possono essere ordinati per rilevanza
  - Alcuni motori prediligono i siti paganti
- Google stabilisce la rilevanza di un item del risultato mediante un algoritmo chiamato PageRank
  - Più del contenuto (per la rilevanza) conta il numero di riferimenti alla pagina

Riferimenti

Libro di testo:

- Ch. 5: tutto tranne 5.5.2

Dal web:

- http://www.googleguide.com/
- http://www.googleguide.com/advanced_operators.html
- Voce "PageRank" di Wikipedia (versione inglese)