INTRODUZIONE A INTERNET

per studenti universitari

Dr. Giorgio F. Signorini

▷ signo@chim.unifi.it

Dipartimento di Chimica

▷ http://www.chim.unifi.it/~signo

Università di Firenze

- 2-2: ricerca di informazione in WWW

28 ottobre 2004

MATERIALE CHE SI TROVA SU WWW

WWW può essere visto “come” una grande biblioteca, analizzabile con strumenti di ricerca automatici

- non è organizzato dall’alto, né omogeneo
- insieme a fonti autorevoli, molto materiale di scarso valore o attendibilità; come distinguere?
- contiene materiale spontaneo, non filtrato (è facile pubblicare il vostro romanzo su WWW)
- inglese! *

*Secondo una statistica (fonte non più reperibile), nel 2000 il 68% del WWW era scritto in inglese; solo l’1.6% in Italiano. [http://global-reach.biz/globstats/refs.php3]

INDICI GLOBALI

- Non esiste un indice “ufficiale” del WWW

- Nessun indice può essere completo

  – ci sono ca. 46 milioni di server WWW al mondo (1/2004, cfr http://www.zakon.org/robert/internet/timeline/ `http://www.zakon.org/robert/internet/timeline/`), alcuni con migliaia di pagine

  – ogni giorno moltissimi documenti nascono, muoiono, cambiano, o cambiano indirizzo (URI)

INDICI

- per soggetto

- per parola

Raccolta di documenti da inserire nell'indice:

- scelta fatta dai compilatori (es. Encyclopaedia Britannica, Open Directory)
- su notifica da parte degli autori (es. LookSmart)
- a pagamento (es. Ouverture)
- automatica (es. AltaVista, AllTheWeb.com, Google)

INDICI O CATALOGHI PER SOGGETTO

C'è sempre una qualche scelta di chi compila il catalogo:

- schema di classificazione è arbitrario
- assegnazione a (una o più) classi è arbitraria (può usare indicazioni fornite dall'autore)

Queste scelte possono essere affidate anche ad un software specializzato

Classificazione è generalmente gerarchica.
Es: `computer > software > games`

INDICI PER PAROLA

Indici per parola sarebbero troppo grandi da consultare direttamente.

La pagina di un indice contiene una casella dove inserire le parole da cercare. Avviando la ricerca, un programma sul sito remoto ricerca le parole nell'indice.

Come risultato, si ottiene un elenco di pagine che contengono le parole cercate.

Ricerca “avanzata”:

- operatori logici e di posizione (AND, OR, NOT, NEAR)
- frasi intere
- frammenti di parola
- campi particolari (titolo, URI,...)
- proprietà (data, lingua,...)
- sottoricerche (ricerche più ristrette tra risultati di ricerche ampie)
- ordinare secondo importanza di parole “chiave”

MOTORI DI RICERCA

Motore di ricerca *("search engine"):* un programma che visita periodicamente (quasi) tutto il WWW e registra il contenuto delle pagine (più altre informazioni) in un indice.

Normalmente si intende: motore di ricerca = la sua sede (home-page) su WWW, che permette l'accesso all'indice

- per parola (sempre)
- per soggetto (se c'è)

Se una pagina viene aggiornata, entro un certo tempo il motore di ricerca aggiorna corrispondentemente l'indice

Alcuni motori di ricerca:

google!	`http://www.google.com`
AllTheWeb.com	`http://www.alltheweb.com`
Yahoo!	`http://www.yahoo.com`
...	

Lista di motori di ricerca:

`http://www.searchenginewatch.com/links/major.html`

Motori di ricerca “intelligenti”

Forniscono le risposte in ordine di rilevanza, valutando, ad es:

- importanza delle parole (numero di ripetizioni, posizione)
- vicinanza tra le parole
- importanza della pagina (se altre pagine le citano, se il titolo contiene il termine “Home Page”, etc.)

Altre forme di intelligenza:

- correzione a supposti errori di battitura (Google)
- suggerimento di ricerche simili a quella data (Lycos)
- indicazione di categorie (soggetti di un catalogo) pertinenti alla ricerca data

Google! (`http://www.google.com`) è un motore di ricerca intelligente

1. ricerca base:
   - tutte le parole (ordine è irrilevante)
   - risposte in ordine di importanza
   - si può accedere alla copia della pagina visitata da google
2. ricerca avanzata permette di specificare meglio
3. ha anche un catalogo per soggetto

IL CATALOGO PER SOGGETTO (DIRECTORY) DI GOOGLE

`http://dir.google.com`

- include solo una piccola parte* delle pagine del WWW, recensite e classificate da volontari (Open Directory)

- le pagine di ogni categoria sono ordinate per importanza come i risultati dell'indice per parola (motore di ricerca) di google

- collegamento tra directory e motore di ricerca:
  - sulla directory si può cercare con il motore di ricerca
  - sui risultati del motore di ricerca è riportata la categoria corrispondente della directory (se c'è)

*3.800.000(Open Directory)/1.500.000(base google) [10/2002]

STRATEGIE DI RICERCA

1. Ricerca per soggetto:
   è sconsigliata perché
   - cataloghi sono incompleti e poco aggiornati
   - difficile trovare categoria giusta
2. Ricerca per parola:
   - Scopo ideale:
     - SOLO risposte rilevanti
     - TUTTE le risposte rilevanti (non sempre)
   - Ragionevole approssimazione:
     - 10-100 risposte
     - le più rilevanti tra le prime

OBIETTIVI E STRUMENTI

possibili obiettivi di una ricerca su WWW (in ordine di precisione decrescente)

- pagine che contengano un testo preciso [es. legge, letterario]
- pagine che contengano un oggetto non-testuale preciso [es. programma, riproduzione di un quadro]
- pagine che *si riferiscano* ad un oggetto preciso [es. homepage ditta, orario FS]
- un'informazione precisa nel contenuto ma non nella forma [un dato storico o scientifico, un'informazione tecnica o commerciale]
- una "bibliografia" generale su un argomento

possibili strumenti

- ricerca per parola, diretta:
  - con motore di ricerca intelligente (che ordini le risposte per rilevanza e importanza)
  - per frase, per parole molto inconsuete, per host/URI
  - per parole con operatori logici e di posizione
- ricerca per parola, indiretta:
  - ricerca di pagine di accesso a database off-line [es. dizionari], poi ricerca su quelle
  - ricerca di cataloghi specifici o “nodi di scambio”, poi uso di quelli
  - ricerca di pagine simili a quelle cercate per trovare nuove parole chiave
- ricerca per soggetto:
  - su cataloghi settoriali
  - su cataloghi generali ma selettivi (es. Britannica)
  - facendo anche ricerca per parola all’interno di un soggetto (es. Google Directory)

COME VALUTARE L'ATTENDIBILITÀ DELL'INFORMAZIONE

Se l'informazione non può essere valutata da voi stessi (opinioni, ragionamenti), occorre sapere se ci si può fidare

Non c'è un modo universale.

Attendibilità è basata su reputazione ("Chi è che l'ha detto?") più che sul numero di conferme (a volte, informazioni sbagliate sono ripetute molte volte)

Questo è ad esempio il motivo per cui google considera una pagina "importante" non solo se essa è citata da molte altre pagine, ma anche se queste pagine a loro volta sono molto citate (cioè godono di buona reputazione)

INFORMAZIONE CONTENUTA NELL'URI

Il nome a domini che appare nell'URI a volte dà indicazione sulla fonte.

Es: una pagina su un presunto virus (GoodTimes):

`http://www.hr.doe.gov/goodtime.html`

I domini di primo e secondo livello (`.doe.gov`) dicono che si tratta di

- un sito governativo USA
- precisamente, il Ministero dell'Energia (Department Of Energy, DOE)

1. una fonte autorevole dovrebbe contenere:

   - nome dell'autore
   - qualifica dell'autore
   - appartenenza dell'autore a enti o società
   - data di creazione

2. ci sono anche altri indicatori di qualità

INDICATORI DI QUALITÀ

- Credibilità
- Accuratezza
- Ragionevolezza
- Sostegno esterno

INDICATORI DI QUALITÀ (2)

- Credibilità:
  - qualifica e/o titoli (accademici, professionali,...) dell'autore o dell'organizzazione
  - l'autore è facilmente contattabile, anche per verificare quanto sopra
  - l'organizzazione prevede un controllo di qualità sull'informazione (revisione, tipo riviste scientifiche)
  - l'informazione è pubblicata su libro (!)
- Accuratezza:
  - aggiornamento
  - concretezza (verificabilità)
  - completezza

INDICATORI DI QUALITÀ (3)

- Ragionevolezza:
  - tono equilibrato
  - mancanza di interessi diretti
  - assenza di specifiche richieste o pressioni sul lettore
- Sostegno esterno:
  - recensioni indipendenti
  - documento contiene riferimenti

BIBLIOGRAFIA

Ed Krol, *The Whole Internet,* 2nd Edition

Ed Krol, *Internet,* Milano, Jackson libri, 1994.

D. J. Barrett, *NetResearch: Finding Information Online*, Songline and O'Reilly & Associates, 1997, (ISBN 1-56592-245-X)

Riccardo Ridi, *Internet in Biblioteca,* Bibliografica

F. Metitieri, R. Ridi, *Ricerche Bibliografiche in Internet,* Apogeo (ISBN 88-7303-431-4)

R. Harris, *Evaluating Internet Research Sources*, 17/11/1997. `http://www.virtualsalt.com/evalu8it.htm`