

# La navigazione di profondità

**Ovvero: ottenere risultati  
soddisfacenti  
nelle ricerche online**

Marco Gualmini  
[www.marcogualmini.it](http://www.marcogualmini.it)  
8 Febbraio 2009



# Un oceano sotto di noi

- Molte pagine dinamiche restano nascoste ai motori di ricerca
- Costituiscono un oceano di informazioni “sommersa”
- Si stima che il “web sommerso” (deep web) sia 500 volte più ampio del “web superficiale”!



M.K. Bergman *“The Deep Web: Surfacing Hidden Value”* University of Michigan, 2001

# L'informazione...

- Esiste?
- É disponibile?
- É “trovabile”?
- Sappiamo cercarla?

# Sommario

1. Introduzione teorica

2. I motori di ricerca

3. Strategie di ricerca

4. Scendere in profondità

# Introduzione storica

- 1969: Nasce Arpanet
- 1971: E-mail
- Inizio anni '90: Primi sistemi di ricerca: FINGER, ARCHIE, WAIS, VERONICA
- 1990: World Wide Web (Tim Berners Lee)
- 1992: NCSA Mosaic
- 1993: primo motore di ricerca (Aliweb)
- Fine anni '90: i portali e la new economy

# Iper testi

- *“L'ipertesto è una struttura informativa costituita di un insieme di testi o pagine leggibili con l'ausilio di un'interfaccia elettronica, in maniera non sequenziale, per tramite di particolari parole chiamate collegamenti ipertestuali (hyperlink o rimandi)...”* (Wikipedia)
- *Sono alla base del World Wide Web*
- Il concetto è antecedente al web, ma in questo trova la sua massima espressione

# URL

- Universal Resource Locator (1994)
- Identifica univocamente una risorsa
- servizio://host:porta/path?querystring
- Esempi:
  - <http://www.marcogualmini.it>
  - [http://www.mensa.it/index.php?option=com\\_wrapper&Itemid=119](http://www.mensa.it/index.php?option=com_wrapper&Itemid=119)
  - <ftp://ftp.kernel.org/pub/linux/kernel/v2.6/linux-2.6.21.6.tar.gz>
  - <telnet://192.168.0.23:8080>

# Anchor HTML

- Concretizzano il concetto teorico di link ipertestuale nel web
- I contenuti informativi sono:
  1. Il nome del link
  2. L'URL di destinazione
- Esempio di codice:
  - `<a href="http://www.linux.it">Comunità Linux italiana</a>`
- Risultato: [Comunità Linux italiana](http://www.linux.it)

# Topologia del web

- Web  $\neq$  Internet
- Configurazione a rete: hub e link
- I link sono unidirezionali
- Link entranti e link uscenti
- Ogni pagina web è indipendente e può essere puntata direttamente da un URL
- Poche pagine sono molto linkate
- Moltissime sono poco linkate

# Continenti e atolli

*Atollo di pagine*

*con soli link*

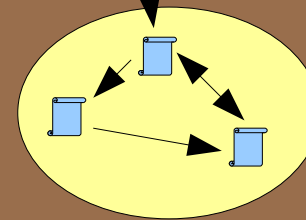
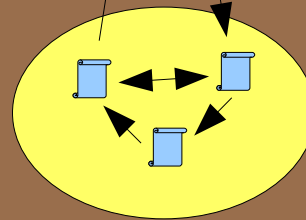
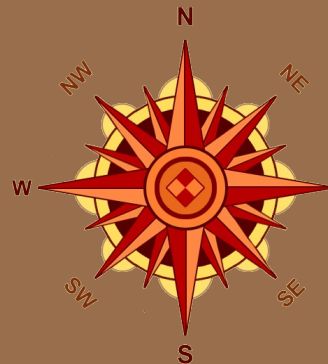
*entranti*

*Terra di mezzo*

*Atollo di pagine*

*con soli link*

*uscanti*



*Penisole*



# Sommario

1. Introduzione teorica

2. I motori di ricerca

3. Strategie di ricerca

4. Scendere in profondità

# Motori di ricerca 1

- *“Un motore di ricerca è un sistema automatico che analizza un insieme di dati spesso da lui stesso raccolti e restituisce un indice dei contenuti disponibili classificandoli in base a formule matematiche che ne indichino il grado di rilevanza data una determinata chiave di ricerca.”* (Wikipedia)
- Raccolta dati dal web
- Analisi (ranking)
- Risposta alle ricerche (query)

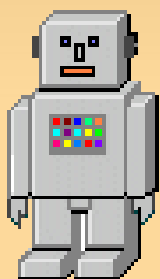
# Motori di ricerca 2

- Indicizzazione manuale/automatica
- Generici/specializzati
- Metamotori
- Multimedia
- Filtri sui contenuti
- Rapidità di aggiornamento
- Link sponsorizzati

Google  
Yahoo!  
MS Live  
Ask  
Altavista  
Excite  
Mooter  
Clusty  
Baidu  
Answers  
Midomi  
Picsearch  
...

# Motori di ricerca 3

## Crawler (o robot)



È il front-end di acquisizione dati. È un software che si muove nel web ininterrottamente seguendo i link ipertestuali.

## Database



È l'archivio di tutte le pagine esplorate dal crawler. Calcola un valore di importanza per ciascuna pagina web in funzione di ciascuna parola contenuta .

## Pagina di ricerca



È l'interfaccia con cui l'utilizzatore finale accede al motore di ricerca, dalla pagina web di inserimento della query alla pagina dei risultati. (SERP)

# Crawler 1

- Costituiscono il front-end di acquisizione dati del motore di ricerca
- Si “muovono” per il web seguendo i link ipertestuali
- Pagine non linkate non vengono indicizzate
- I singoli siti possono definire regole per l'accesso dei crawler (robots.txt)

# Crawler 2

- Cosa leggono:
  - Parole chiave
  - Posizione delle stesse nell'html
  - Link ipertestuali
  - Metainformazioni
  - Documenti non html
- Cosa NON leggono
  - Semantica di immagini e multimedia
  - Animazioni Flash

# Google e Pagerank

- Creato nel 1998 da Larry Page e Sergey Brin, studenti all'Università di Stanford
- Usa un algoritmo ricorsivo chiamato Pagerank che stima la popolarità delle pagine
- Il parametro pagerank è tanto più alto quanto più siti puntano su una pagina e tanto più le pagine referenti hanno PR alto

$$PR = (1-d) + d( PR_1 / N_1 + \dots + PR_i / N_i )$$

# Sommario

1. Introduzione teorica

2. I motori di ricerca

3. Strategie di ricerca

4. Scendere in profondità

# Tattiche di ricerca

- Pazienza
- QI
- Capire la macchina
- Scelta dello strumento adatto
- Ridurre le ambiguità semantiche
- Approssimazione successiva
- Riconoscere i vicoli ciechi
- Serendipity

# Ambiguità

Polo sport cavalli

>

Automobile versioni sport con potenza del motore in cavalli

Luogo geografico

Concetto fisico

Capo di abbigliamento

Caramella di menta

Sport

Politica

Marco

.....

# Conoscere il nemico

- L'informazione esiste? È reperibile?
- Quali siti possono offrire l'informazione?
- Amatoriali o commerciali? Blog? Forum?
- In che forma? (html, img, pdf, testo ecc)
- In che lingua?
- Immedesimarsi nell'offerente
- Proprietà intellettuale

# Ricerca con operatori

- “Ricerca per frasi intere”
- AND OR operatori booleani
- +A -B forza presenza (A) o esclusione (B)
- A\*B richiede una parola tra A e B
- ~x cerca anche i sinonimi di x
- m..n intervallo numerico tra m ed n
- Restrizione ricerca per: data, dominio, tipo di documento, posizione nell'html

# Ricerche multimediali

- Immagini e video sono associati a TAG
  - Tag ricavate da testo vicino o alternativo
  - Tagging manuale
  - Intelligenza artificiale
- Ricerca immagini per pattern recognition
- Ricerca immagini per tipo: volti, clipart, foto
- Ricerca musicale: [midomi.com](http://midomi.com)

# Sommario

1. Introduzione teorica

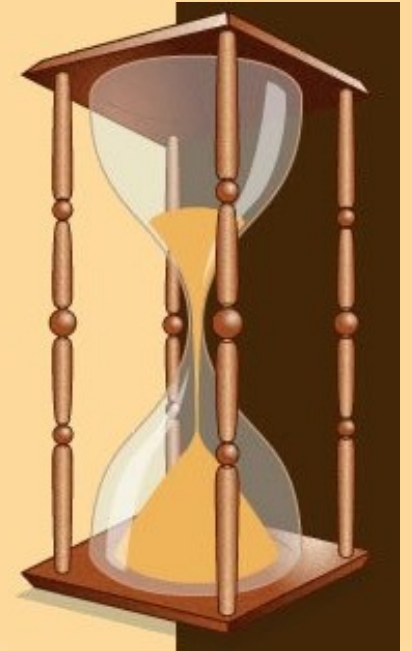
2. I motori di ricerca

3. Strategie di ricerca

4. Scendere in profondità

# Macchine del tempo

- Cercare dati non più presenti online
- Copia cache dei motori di ricerca
- Archivio Usenet (ex dejanews)
- Archive.org



# Google's tip and tricks

- Calcolatrice evoluta
- Conversione di misure e valute
- Costanti matematiche e fisiche
- Ricerche “semantiche”
- Easter eggs !
- Risposte alle Grandi Domande ???

# Specchio delle mie brame...

- Hai controllato la tua immagine online?



# Egosurfing e Self-Googling

- PEW internet & american life style 2007:
  - 47% ha verificato la propria identità
  - 3% lo fa su base regolare
  - 10% deve apparire professionalmente
  - 87% ritiene le informazioni accurate
  - 4% ha avuto esperienze “negative”
  - 54% afferma di controllare attivamente i dati
  - 18% preoccupato ma non limita attivamente
  - Oltre metà ha cercato riferimenti altrui

# Privacy online

- La fine della privacy
- Orme attive ed orme passive
- Ciò che era pubblico prima lo è anche ora...
- ...ma è molto più “disponibile”
- Data crossing
- L'intelligence può essere automatizzata
- Facebook (Mark Zuckerberg, 2004)

# FINE

- 
- [www.marcogualmini.it](http://www.marcogualmini.it)
  - [www.mozillaitalia.org](http://www.mozillaitalia.org)

