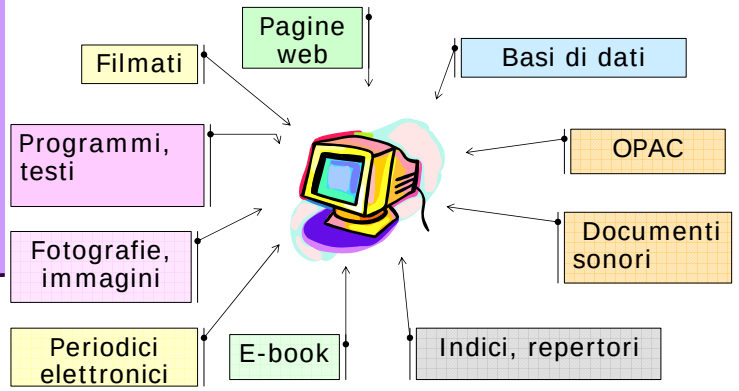
# Informazioni sul Web

## Tipi di risorse su Web




2

## Problemi per IR su Web

- Dati distribuiti: I documenti sono sparsi su milioni di server differenti
- Dati Volatili: Molti documenti appaiono e spariscono (così detti *dead links*).
- Enormi volumi di dati: Miliardi di documenti diversi
- Dati non strutturati e ridondanti: Non esiste una struttura uniforme, ci sono errori html, circa 30% di documenti duplicati.
- Qualità dei dati: Non ci sono controlli editoriali, le informazioni possono essere false, possono esserci errori, testi mal scritti..
- Dati eterogenei: multimediali (immagini, video, suoni..) diversi linguaggi, diversi formati (pdf, ps..)

## Caratteristiche dell'informazione su Web

- Fluidità:
  nascita, morte, migrazione e/o cambiamenti di siti o pagine web
- Fossilizzazione:
  link morti, documenti superati da versioni più recenti, ecc.
- Disintermediazione
  passaggio diretto del testo, dall'autore al lettore:vantaggi e pericoli
- Basso grado di strutturazione e di caratterizzazione semantica
  complica la ricerca di documenti e la valutazione della loro rilevanza

Il Web offre un'immensa quantità di informazioni, ma non le integra e organizza

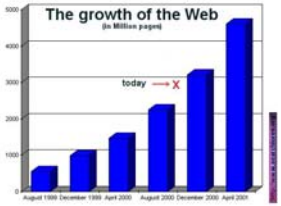
4

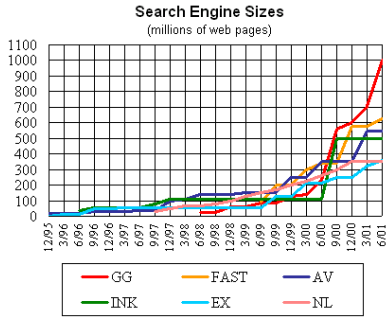
# Informazioni su Web

- Il web si espande a ritmi esponenziali
  - 5 o 6 miliardi di pagine web, ma il dato è in costante crescita
- **Information overload**: eccesso di informazione che non si riesce più a padroneggiare.

5

# Crescita delle pagine indicizzate




*SearchEngineWatch, Aug. 15, 2001*

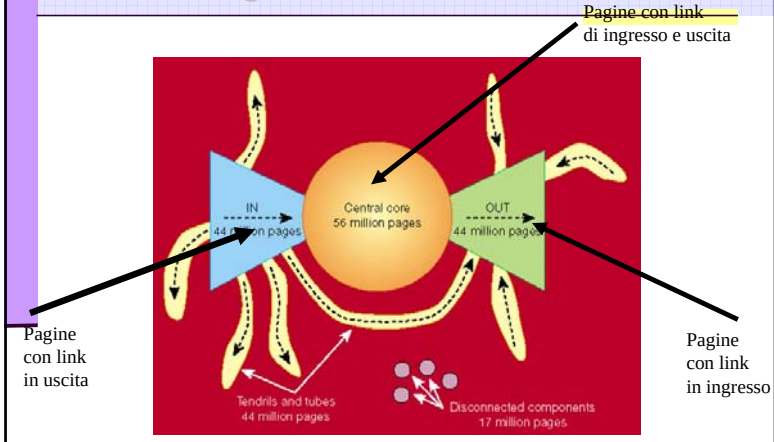
Google elenca il numero di pagine correntemente indicizzate.

# Indicizzazione delle informazioni su Web

- Alti tempi per l'aggiornamento dei link
- Non tutto è interessante
  - Secondo Inktomi (uno dei maggiori produttori di sistemi per motori di ricerca): il 25% dei documenti presenti nel Www è irrilevante un altro 25% ha interesse solo per un numero molto ristretto di persone
- L'obiettivo è selezionare solo le risorse più interessanti per gli utenti
  - cercando di dare maggior peso alle informazioni "locali"

7

# Struttura a grafo del Web




http://www9.org/w9cdrom/160/160.html

## Il web invisibile

- Termine tradizionale indicante risorse impenetrabili ai motori di ricerca:
  - risorse interne a database, o file difficilmente identificabili e classificabili quali file di immagine, audio e video
- Tutte queste risorse stanno diventando meno invisibili grazie all'evoluzione dei motori di ricerca
  - Tuttavia buona parte del Web continua ad essere di fatto irraggiungibile e quindi invisibile

9

## Ricerca di informazioni su Web

- Due grandi categorie:
  - Sistemi informativi (es., basi di dati, biblioteche digitali, …) con interfaccia web
  - Sistemi di recupero di informazione distribuita su web: Indici
- Alcune categorie di strumenti di IR su web
  - Directory
  - Motori di ricerca
  - Portali generali/tematici
  - Virtual reference desk
  - Recommendation systems

10

## Directory

- Indici consultabili tramite browsing
  - Riferiscono siti o, comunque, unità informative compiute
  - a differenza dei motori di ricerca che trovano singole pagine (che solo talvolta costituiscono la home page di un sito)
- Automatiche:
  - Costruiti con tecniche di clustering/classificazione per l'analisi automatica del contenuto dei documenti
- Manuali:
  - Sono costruiti o supervisionati da personale specializzato che indicizza pagine Web in base al loro argomento
    - talvolta aggiungendo un breve riassunto e una valutazione
  - ciò migliora la qualità, fa diminuire il rumore
  - costi maggiori: gli archivi sono più piccoli

11

## Esempi di siti con directory

- Yahoo: http://www.yahoo.com
  - basato sul lavoro di un centinaio di redattori professionisti
- Open directory: http://dmoz.org
  - basato sulle segnalazioni di decine di migliaia di volontari
- Looksmart: http://www.looksmart.com
- About: http://www.about.com
  - diviso in circa 700 aree tematiche diverse, ognuna affidata ad un esperto volontario
- Supereva: http://www.supereva.it

12

## Motori di ricerca (search engine)

- Uso tipico:
  - Query: lista di parole chiave immesse in una finestra di ricerca (form)
  - Risposta: una serie di URL di documenti ordinati per rilevanza, e con informazioni sintetiche sul contenuto
- Principio di funzionamento:
  - individuano e indicizzano automaticamente le occorrenze di uno o più termini presenti in una buona parte delle pagine web mondiali o di un singolo Stato
  - il problema maggiore consiste nell'individuare automaticamente (senza l'intervento di catalogatori umani) il contenuto semantico delle pagine web

## Alcuni motori di ricerca

- Google, http://www.google.com
  - il più usato al mondo: 200 milioni di richieste al giorno
- Altavista, http://www.altavista.com
  - dotato di numerose opzioni per affinare la ricerca
- Alltheweb, http://www.alltheweb.com
- Hotbot, http://www.hotbot.com
- MSN, http://www.msn.com/
- Lycos, http://www.lycos.com
- Northern light, http://www.northernlight.com
- Excite, http://www.excite.com
- Go.com, http://www.go.com
- Overture, http://www.overture.com

14

## ... alcuni motori di ricerca

- Ci sono poi motori studiati per rispondere a domande poste in linguaggio naturale (in inglese)
  - es. Ask Jeeves http://www.ask.com
  - In questo caso a ogni interrogazione viene fornita sempre una sola risposta e non una lista di indirizzi.
- Nel webspace italiano da segnalare:
  - Arianna, http://arianna.libero.it
  - Il Trovatore, http://www.iltrovatore.it
  - Janas, http://www.tiscali.it
  - Supereva, http://www.supereva.it

15

## Esempio: Altavista

- Ricerca normale
  - Si digitano una serie di parole
  - Se racchiuse tra " e " si ricercano parole consecutive
  - Uso di + e – (Es. +Clinton –Levinski)
- Ricerca avanzata
  - Si possono utilizzare gli operatori booleani (AND, OR, NOT – anche &, |, !)
  - Si possono ordinare i risultati
- Page ranking:
  - frequenza di occorrenza delle parole e se compaiono nel titolo piuttosto che nel testo …

## Esempio: Google

- Ricerca normale
  - Come sopra
- Ricerca avanzata
  - with **all** of the words
  - with the **exact phrase**
  - with **at least one** of the words
  - **without** the words
  - Ristrette dalla lingua, data, parole nel titolo …
- Page ranking
  - si considera il numero e l'autorevolezza dei link entranti

## Motori di ricerca specifici

- Nati di recente, sono degli ibridi fra le due categorie precedenti, delle quali cercano di unire i pregi:
  - applicano la potenza «cieca» dei motori di ricerca solo a siti dedicati a una particolare disciplina o argomento e indicizzati da personale specializzato
- Esempi:
  - Argos, http://argos.evansville.edu, per argomenti che riguardano la storia classica e medievale
  - Hippias, http://hippias.evansville.edu,per ricerche di filosofia



## Pay for placement

- Le aziende pagano per garantirsi una buona posizione nelle liste ottenuta da ricerche con determinate parole chiave:
  - ciò favorisce i siti delle aziende commerciali a discapito di siti non-profit o comunque privi di finanziamento
  - meccanismo sempre più diffuso
- Non sempre i motori di ricerca dichiarano la propria politica rispetto a queste soluzioni
  - Google vende solo banner pubblicitari che compaiono in associazione a certe parole di ricerca, dichiarando di non piegarsi alla logica del pay for placement



## Meta-indici

- Permettono l'accesso a un certo numero di repertori primari come quelli finora elencati.
  - Alcuni sistemi permettono di immettere una sola volta i termini di ricerca, lasciando al software il compito di ripetere l'interrogazione su tutti i motori selezionati e di produrre una risposta cumulativa
  - risultato «sporco»: cieco rispetto alle peculiarità dei vari archivi e delle relative tecniche di interrogazione
- Possono essere suddivisi in tre sottocategorie:
  - indici di indici
  - multi indici
  - meta-indici in senso stretto

## Indici di indici

- Sono in realtà dei repertori di indici
  - semplici liste di link a indici
  - qualche volta ampiamente commentati
- Esempi:
  - http://riceinfo.rice.edu/Internet
  - http://www.searchenginewatch.com
  - http://www.motoridiricerca.it
  - http://www.notess.com/search

21

## Multi-indici

- Si tratta di pagine con diversi form per la ricerca su vari indici, interrogabili solo uno alla volta.
- Alcuni esempi:
  - http://www.webtaxi.com
  - http://www.humnet.unipi.it/motoridiricerca.html

22

## Meta-indici in senso stretto

- In questo caso un'unica maschera di ricerca permette l'interrogazione cumulativa di vari indici contemporaneamente.
  - Talora i risultati sono «schiacciati» eliminando le ripetizioni e sono ordinati in base alla supposta rilevanza rispetto alla richiesta o ad altri criteri.
- Esempi:
  - http://www.metacrawler.com
  - http://vivisimo.com, raggruppa per voci i risultati suddividendoli in cartelle etichettate con nomi che ne indicano il contenuto: "clusterizzazione".

23

## Meta indici: alcuni esempi

- **Fagan Finder**: http://www.faganfinder.com
  (comprende un'ampia selezione di strumenti di ricerca, di metaricerca e di reference; offre anche un buon "metatraduttore automatico")
- **Ithaki**: http://www.ithaki.net
  (dà la possibilità di effettuare metaricerche limitate ad una specifica area geografica)
- **ProFusion**: http://www.profusion.com
- **Fazzle**: http://www.searchonline.info
- **Ixquick**: http://ixquick.com
- **Kartoo**: http://www.kartoo.com

24

## Portali

- I portali si candidano a guida per i navigatori
  - non solo per la ricerca di informazioni, ma anche per altrr attività (acquisti in linea, prenotazioni di servizi, ecc.).
  - molto usati da utenti meno esperti
- Includono quasi sempre
  - un indice per argomento orientato alle necessità della vita quotidiana,
  - un motore di ricerca sviluppato in proprio o mutuato
  - un insieme di svariati servizi
- Si suddividono in:
  - Portali generali (“portali orizzontali”)
  - Portali tematici (“portali verticali” o “vortali”)



## Portali

- Esempi italiani:
  - Ciaoweb, http://www.ciaoweb.it
  - Jumpy, http://www.jumpy.it
  - Kataweb, http://www.kataweb.it
  - Supereva, http://www.supereva.it
  - Virgilio, http://www.virgilio.it



## Virtual reference desk (Vrd)

- Raccolgono, ordinano e talvolta valutano e commentano le principali fonti informative e i più utili strumenti di ricerca disponibili in rete …
  - … relativamente a Internet in generale (Vrd generali),
  - … o su una determinata disciplina o argomento (Vrd specializzati)
- Esempi
  - per bibliotecari:
    - http://www.burioni.it/forum/ridi/home.htm
    - http://www.cultura.regione.toscana.it/bibl/ref/index.htm
  - per umanisti:
    - http://lettere1.lett.unitn.it/lavori/carl.htm
    - http://www.rassegna.unibo.it/index.html



## I weblog

- Siti weblog (o blog):
  - siti prevalentemente (ma non necessariamente) personali
  - costruiti a partire da ‘articoli’ (post) organizzati cronologicamente, con in testa i più recenti
- Il mondo dei weblog ha creato uno spazio condiviso: la blogosfera
  - popolato da utenti che si scambiano informazioni, le approfondiscono, le discutono collaborativamente
- I weblog come strumento di “public opinion”:
  - la caratteristica di inserire link a siti o risorse di interesse, e quella di gestire commenti al proprio articolo, rendono la blogosfera una vera e propria ragnatela di riferimenti incrociati.
  - Esistono ormai numerosi weblog che costituiscono una fonte informativa diretta e strumenti di comunicazione insostituibili per movimenti dalla natura spesso transnazionale
  - Esempi: guerra in Iraq, mondo no-global, America latina, ecc.

## Gli indici della blogosfera

- Esiste una gran quantità di indici e directory che mappano la blogosfera (ma solo i weblog):
  - Eatonweb, http://portal.eatonweb.com
  - Blogwise, http://www.blogwise.com
- Tuttavia si è ancora molto indietro nella catalogazione semantica:
  - esempio di “aggregatore semantico” è BlogAggregator, http://www.bookcafe.net/blog/aggregator/
- Popularity Index
  - individuano le notizie di volta in volta più discusse nella blogosfera
  - classifiche dei weblog più popolari
  - costellazione di appartenenza di un weblog (cioè l’insieme dei weblog che lo citano e ne sono citati), ecc.

29

## Sistemi personalizzabili

- Mirano a rintracciare autonomamente tutte le risorse di interesse per l’utilizzatore, sulla base della preventiva definizione di un accurato «profilo di ricerca».
- Esempi piuttosto semplici:
  - MyYahoo! (http://www.my.yahoo.com)
  - My Excite (http://www.my.excite.com)

30

## Sistemi personalizzabili – agenti di ricerca

- Si tratta di programmi che svolgono, a intervalli prefissati, ricerche anche molto complesse, e che hanno la capacità di “reagire” autonomamente ai risultati ottenuti
  - ad esempio filtrandoli attraverso criteri pre-impostati e difficilmente eseguibili direttamente sul motore di ricerca
- Ne esistono di molti tipi:
  - Copernic Agent (http://www.copernic.com)
    - è in grado di interrogare oltre 1000 strumenti di ricerca
  - BullsEye (http://www.intelliseek.com)
  - BotSpot, http://www.botspot.com
  - BotKnowledge, http://www.botknowledge.com
  - Agentland, http://www.agentland.com

31

## Web semantico

- Obiettivi:
  - utilizzare sistemi di deduzione logica automatica o euristica per elaborare l’informazione in base alla sua semantica
- Per ogni singola occorre che sia possibile
  - identificarla in modo univoco nel web (URI: Universal Resource Identifier)
  - associarle una descrizione formale del suo significato, espressa in un linguaggio comprensibile anche alle macchine

32

# Web semantico (2)

- Strumenti utilizzabili:
  - Linguaggi per la descrizione delle risorse:
    - RDF (Resource Description Framework), metalinguaggio dichiarativo basato su XML
  - Ontologie formali: sistemi per specificare le relazioni concettuali soggiacenti a tali descrizioni
- Considerazioni:
  - Evidenti difficoltà pratiche
  - Si fonda sulla collaborazione dei vari creatori e utenti delle risorse informative su web

33