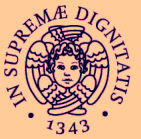
I sistemi di recupero dell'informazione

… grazie a Maria Simi

Dip. Di Informatica – Univ. Pisa



UNIVERSITÀ DI PISA

Sistemi di recupero dell'informazione

- **Sistemi specializzati nella gestione di documenti di testo e nel recupero in base al loro contenuto**
- **Grossa collezione di documenti**
  - Collezioni *full-text*
  - *Digital libraries*
  - Pagine Web (motori di ricerca – *search engines*)
- **In inglese: information retrieval systems**

Differenze con le basi di dati

- Documenti con molto testo piuttosto che dati strutturati.
- Le richieste sono espressioni imprecise del bisogno informativo
- Le risposte sono riferimenti a documenti “che potrebbero contenere le risposte” piuttosto che direttamente le risposte

Domanda tipica a un DBMS

```
SELECT Nome, Ufficio
FROM Impiegati
WHERE AnnoAssunzione > 1970
        AND Stipendio > 2000000
```

Nome	Ufficio
Mario Paoletti	Amministrazione
Guido Carlesi	Amministrazione
Sandra Merlini	Vendite

Domanda tipica a un SRI

FIND architett*

AND (cad OR (progetto AND calcolatore))

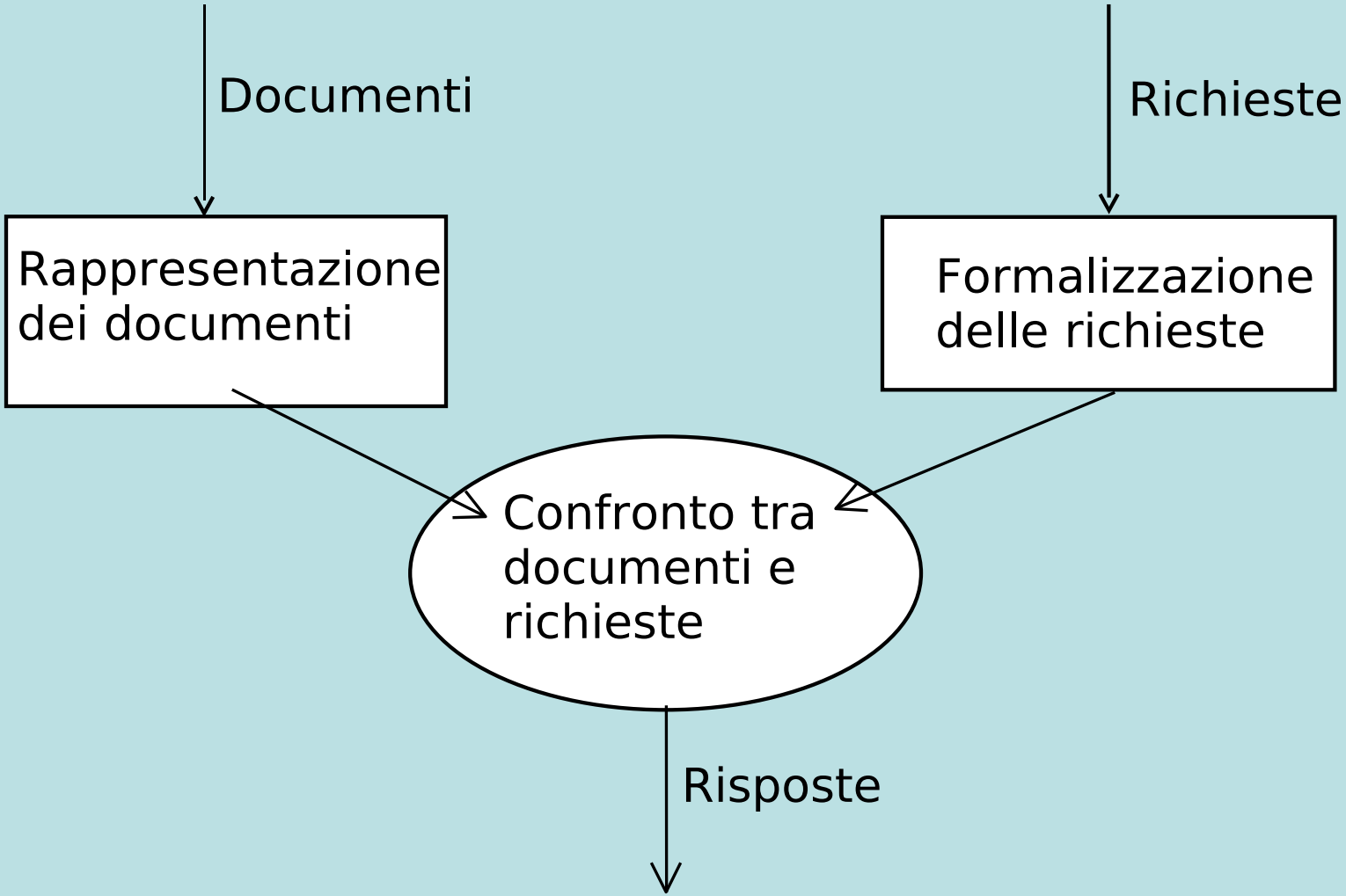
"… l'impiego del calcolatore per lo sviluppo di progetti **architett**onici riguarda il campo di applicazioni dell'informatica conosciuto con il nome di **CAD**, ovvero progetto assistito da calcolatore…"

"… nell'affrontare il **progetto** dell'**architett**ura di un **calcolatore** bisogna tener conto del settore di applicazione in cui verrà utilizzato …"

Sintesi delle differenze

	DBMS	SRI
Tipologia dei dati	Strutturati	Testo
Richiesta	Completa e precisa	Incompleta e vaga
Criterio di scelta	Corrispondenza esatta	Corrispondenza parziale
Risultato	Dati richiesti	Documenti probabilmente rilevanti

Il nucleo di un SRI



Il processo

- **Rappresentazione dei documenti in forma sintetica:**
  - indicizzazione
- **In fase di recupero:**
  - Formalizzazione delle richieste
  - Contronto tra richieste e rappresentazione di documenti
- **Risultato**
  - Binario (si/no) – corrispondenza esatta
  - Probabilistico – corrispondenza parziale

Valutazione di un SRI

- **Richiamo: il numero di documenti rilevanti recuperati in rapporto ai documenti rilevanti presenti nella collezione**
- **Precisione: il numero di documenti rilevanti recuperati in rapporto ai documenti recuperati**
- **A: documenti rilevanti (nella collezione)**
  B: documenti recuperati

  |A|: il numero di elementi di A

  Richiamo = $|A \cap B| / |A|$ Richiamo ottimo: 1 Tipico<1

  Precisione = $|A \cap B| / |B|$ Precisione ottima: 1 Tipica<1

Modelli di SRI

- **Un modello cerca di astrarre le caratteristiche salienti che stanno alla base di una classe di sistemi.**

- **Nel caso degli SRI:**
  - lo stile di rappresentazione dei documenti;
  - lo stile di rappresentazione delle richieste;
  - la modalità del confronto tra rappresentazioni di documenti e richieste.

Due modelli per gli SRI

- **Modello booleano**
  - un modello a corrispondenza esatta
- **Modello vettoriale**
  - un modello a corrispondenza parziale
- **Ne esistono molti altri intermedi: il modello fuzzy, probabilistico …**
- **L'indicizzazione si occupa di come si ottiene la rappresentazione dei documenti (dopo)**

Il modello booleano

- **Rappresentazione dei documenti**
  - Un insieme di termini che ne rappresentano il contenuto (scelti durante l'indicizzazione)
- **Interrogazioni**
  - combinazioni booleane di termini, cioè termini combinati tra loro mediante AND, OR, NOT
- **Criterio di corrispondenza**
  - AND: i termini sono entrambi presenti
  - OR: almeno uno dei due termini è presente
  - NOT: il termine non è presente

Esempio

(film AND amore)

documenti che contengono “film” e “amore”

(dramma OR drammatico)

documenti che contengono “dramma” o “drammatico”

NOT (dramma OR drammatico)

… che non contengono “dramma” o “drammatico”

((film AND amore) NOT (dramma OR drammatico))

Il modello vettoriale: documenti

- **Rappresentazione dei documenti**
  - una sequenza di numeri lunga quanto il numero di tutti i termini utilizzati per rappresentare i documenti nella collezione, un vettore appunto.

    D = (t1, t2, …, tn) n numero di termini
    - ➔ **tk=0 se il termine non è presente**
    - ➔ **altrimenti tk è il peso del termine k-esimo nel documento, una misura di importanza**

Il modello vettoriale: interrogazione

- **Interrogazione: un insieme di termini**
- **Rappresentazione dell'interrogazione:**
  - un vettore, simile ai documenti
  - (con moltissimi 0 e qualche 1 in corrispondenza dei termini specificati dall'utente)
  - **Q(t1, t2, … tn)**

Il modello vettoriale: confronto

- **Una misura di similitudine tra documenti e richiesta.**

- **Esempio**
  - Di(ti1, ti2, ti3, …, tin)
  - Q(q1, q2, q3, …, qn)
  - $S(Q, Di) = q1*ti1 + q2*ti2 + \ldots + qn*tin$

    $= \sum_j qj * tij$ con $0<j \leq n$

Esempio

- **Due documenti che trattano di Papa, Roma e Vaticano …**

  Vettori:

  D1 = [… 0.1, …, 0.1, … , 0.2, …]

  D2 = [… 0.1, …, 0.9, … , 0.9, …]

- **Interrogazione**

  Q=[… 1, …, 1, …, 1, …]

- **Similitudine**

  Sim(D1, Q)=0,1+0,1+0,2=0,4

  Sim(D2,Q)=0,1+0,9+0,9=1,9

Indicizzazione

- **Indicizzazione: processo di rappresentazione dei documenti mediante una descrizione sintetica (*catalogazione per soggetto* in ambito bibliotecario)**
  - Una lista di termini
  - Una lista di termini pesati
- **Serve per costruire indici su collezioni di documenti (vedi organizzazione indicizzata degli archivi)**
- **Linguaggio di indicizzazione: insieme dei termini scelti per indicizzare una collezione di documenti**

Linguaggio di indicizzazione-1

- **Come sono scelte le parole del linguaggio di indicizzazione:**
  - Linguaggio controllato: limitato ad un vocabolario predefinito
  - Linguaggio libero: termini estratti liberamente dal testo del documento e non definiti a priori

Linguaggio di indicizzazione - 2

- **Come sono fatti i termini del linguaggio …**
  - Termini singoli (es. “recupero”, “informazione”, “sistema”…)
  - Termini in contesto: composti da diverse parole (es. “sistemi di recupero dell’informazione”)

Processo di indicizzazione

- **Manuale: è una persona che sceglie quali termini meglio caratterizzano il contenuto di un documento**
  - Più “semantico” e quindi migliore
  - Soggettivo, costoso
- **Automatico: fatto da un programma**
  - Più sintattico, su base statistica e quindi “peggiore”
  - Economico, scalabile

Corrispondenze tipiche

- Indicizzazione manuale
- Linguaggio controllato
- Linguaggio a termini in contesto

- Indicizzazione automatica
- Linguaggio libero
- Linguaggio a termini singoli

Indicizzazione manuale

- **Vantaggi dell'uso di un linguaggio controllato**
  - semplifica il processo
  - lo rende meno soggettivo
  - migliora la comunicazione tra indicizzatori e utenti
- **Svantaggi**
  - Necessità di intermediari

Struttura del linguaggio di indicizzazione

- **Dizionario**: termini ordinati alfabeticamente
- Schema di classificazione - **ontologia**: codici che organizzano i termini gerarchicamente
- **Thesaurus**: termini organizzati in una "rete semantica"

Dizionario

- **Le parole vicine sono “sintatticamente” simili (prefissi simili), ma non nel loro significato**

Esempio:

…	Acido ascorbico
Chimica	…
Chirurgia	…
Chemioterapia	…
…	
Medicina	…
…	Vitamina C

Schema di classificazione

- Es., schema di classificazione decimale di Dewey, usata in ambito bibliografico:

15 psicologia

152 psicofisiologia

1521 percezioni sensoriali

153 processi mentali

154 subconscio

Thesaurus: relazioni semantiche tra termini

- relazioni di preferenza (“US”, usa, e “UF”, usato per)

  Es. elaboratore US calcolatore

  calcolatore UF elaboratore calcolatrice stazione di lavoro

- relazioni di gerarchia (“BT”, termine più generale, e “NT”, termine più specifico)

  Es. felini NT gatti leoni tigri

  gatti BT felini

- relazioni di affinità semantica (“COR”, termine correlato e “SIN”, termine sinonimo)

Un esempio di Thesaurus italiano

Thesaurus delle suppellettili ecclesiastiche dell'ICCD

ampolliera

EAD **ampollina**

ampollina

BT **oggetti liturgici per l'eucarestia**

ACC **ampolliera**

vassoio portampolle

TIP **ampollina con beccuccio**

ampollina con versatoio

EAD - accessorio di
ACC - accessorio
TOP - tipologia particolare
TIP - tipologia generale

Uso di un thesaurus: esempi

- **Recupero per tematismi:**
  - Posate per forchette, coltelli, cucchiai
  - Europa per Italia, Francia, Spagna …
- **Rimando a "forme ufficiali" per aumentare l'oggettività dell'indicizzazione**
- **Uso dei sinonimi per aumentare il richiamo e rendere le ricerche meno "sintattiche"**

Indicizzazione automatica

- **Lo scopo del processo di indicizzazione è duplice:**
  - Rappresentare un documento tramite una lista di termini, che ne caratterizzano il contenuto
  - Assegnare a ciascuno di questi termini un peso, che ne riflette l'importanza

Indicizzazione automatica

- **Assunzione di base: la frequenza di occorrenza di un termine in un documento è indicativo della sua importanza nel caratterizzarne il contenuto.**

"… Uno scrittore normalmente ripete certe parole nell'elaborare aspetti di un certo argomento. L'enfasi è considerata un indicatore di significatività …". [Luhn]

Legge di Zipft

Rango	Termine	Frequenza	(RangoXFrequenza) /1.000.000
1	the	69.971	0,070
2	of	36.411	0,073
3	and	28.852	0,086
4	to	26.149	0,104
5	a	23.237	0,116
6	in	21.341	0,128
7	that	10.595	0,074
8	is	10.099	0,081
9	was	9.816	0,088
10	he	9.543	0.095

Studi statistici

- Principio dello “sforzo minimo”: è più facile ripetere parole di uso comune che usarne di nuove
- Le parole più usate sono quelle di lunghezza breve e di uso comune che hanno un costo di uso molto basso
- Per la lingua inglese il 20% delle parole usate in un testo copre il 70% del testo stesso

Frequenza sì ma …

- **Le parole in assoluto più frequenti sono anche poco significative**
  - avverbi, articoli, preposizioni ecc.
  - le 250 parole più comuni coprono in media il 40-50% di un testo
- **Quello che conta non è la frequenza assoluta ma la frequenza relativa**
  - Es. ‘Computer’ in una biblioteca di informatica

Misura basata sull'inverso della frequenza

$FREQ_{ik}$: frequenza del termine k nel documento i

$DOCFREQ_k$: numero dei documenti in cui compare k

N: il numero dei documenti nella collezione

Inverso della frequenza (IDF)

$$IDF_k = \log_2 \frac{N}{DOCFREQ_k} + 1$$

Pesatura TF-IDF

$$PESO_{ik} = FREQ_{ik} \times IDF_k$$

Processo di indicizzazione automatica – passo 1

1. **Eliminazione di parole di uso comune (uso di *liste di esclusione* – STOP list*)***

Stralcio di una lista di esclusione per la lingua inglese:

A	ALMOST	AMONGST	ANYWHERE
ABOUT	ALONE	AN	ARE
ACROSS	ALONG	AND	AROUND
AFTER	ALREADY	ANOTHER	AS
AFTERWORDS	ALSO	ANY	AT
AGAIN	ALTHOUGH	ANYHOWBE	
AGAINST	ALWAYS	ANYONE	BECAME
ALL	AMONG	ANYTHING	BECAUSE

Processo di indicizzazione automatica – passo 2

1. **Riduzione delle parole alla radice (STEMMING)**
   - Per aumentare il richiamo e ridurre le dimensioni del linguaggio di indicizzazione
   - Si utilizzano liste di suffissi:

Es. anal[isi]	**calcol[are]**
anal[izzare]	**calcol[atore]**
anal[ista]	**calcol[atrice]**
anal[itico]	**calcol[abilità]**

Processo di indicizzazione automatica – passi 3 , 4

3. Pesatura dei termini con una misura tipo TF-IDF: i termini con peso alto vengono assegnati al documento

4. Rappresentazione dei documenti, ad esempio come vettori di pesi.

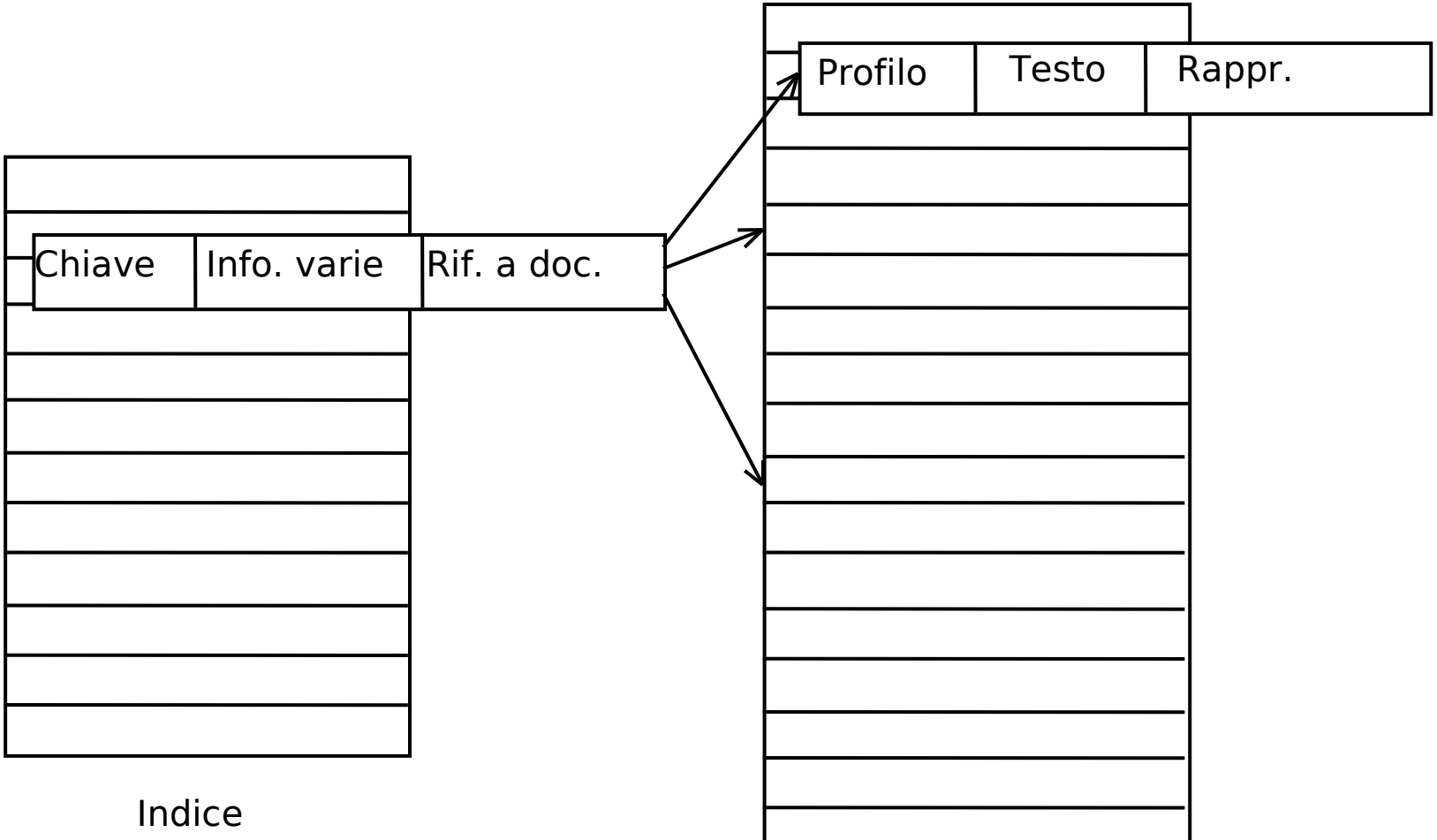
Indicizzazione automatica o manuale?

- In media c'è un accordo del 60% con le due tecniche.
- L'accordo che esiste tra due indicizzatori "umani" d'altra parte non è molto più alto
- L'approccio manuale, anche se qualitativamente superiore, non è scalabile
- In certi domini (es. Web) l'indicizzazione automatica è l'unica possibile

Interrogazione nei sistemi di R.I.

- Si basano su un sistema di archiviazione con organizzazione indicizzata
- Le possibilità offerte dal linguaggio di indicizzazione sono correlate all'informazione che viene mantenuta nell'indice
- AND, OR, NOT sempre possibili

Organizzazione interna dei SRI



Informazione nell'indice

- **Chiave**
  - Tutte le parole del linguaggio di indicizzazione sono i possibili valori per la chiave
- **Riferimenti ai documenti**
  - Riferimenti ai doc che contengono la chiave
  - Un insieme di coppie <doc, pos> per ogni occorrenza del termine nei documenti
- **Informazioni varie**
  - Numero di documenti che contengono la chiave (*postings*)
- **Termini sinonimi**

Operatori booleani

… come operazioni sugli indici

Interrogazione: recupero AND informazione

- S1: insieme dei rif. a doc. associati al termine ‘recupero’ nell’indice

- S2: insieme dei rif. a doc. associati al termine ‘informazione’ nell’indice

- Risultato = S1 ∩ S2 (intersezione di insiemi)

Operatori booleani

Se fosse: recupero OR informazione

1. come sopra
2. come sopra
3. Risultato = S1 ∪ S2 (unione di insiemi)

Se fosse: recupero NOT informazione

…

3. Risultato = S1 – S2 (differenza di insiemi)

Troncamenti e caratteri jolly

Esempio:

*** : una sequenza di caratteri di lunghezza variabile**

? : un qualsiasi carattere

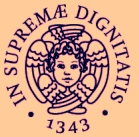
psic* sta per psichiatra, psichiatria, psichiatrico, psicologo, psicologia ecc.

calcolator? i documenti contenenti le parole 'calcolatore' o 'calcolatori'

document??? i documenti contenenti 'documentale', 'documentato', ma non 'documentazione'

Motori di ricerca

...



UNIVERSITÀ DI PISA

Motori di ricerca

- Sono strumenti per localizzare informazioni
- Corrispettivo Web dei Sistemi per il Recupero dell'Informazione
- Uso tipico: digitare una lista di parole chiave in una finestra di ricerca
- Risposta: una serie di URL (indirizzi web) di documenti (pagine web) ordinati per rilevanza, suddivisi in pagine successive

Motori di ricerca più utilizzati

- Arianna, **http://www.arianna.it, MdR italiano con possibilità di collegarsi ai più importanti motori di ricerca internazionali quali AltaVista, Lycos, etc.;**
- Virgilio, **http://www.virgilio.it, MdR italiano;**
- Altavista, **http://www.altavista.digital.com , uno dei motori internazionali più conosciuti.**
- Google, **http://www.google.com, uno dei migliori**
- HotBot, **http://www.hotbot.com**
- Infoseek, **http://www.infoseek.com**
- Lycos **http://www.lycos.com**
- Inktomi: **http://www.inktomi.com**

Altavista: un esempio

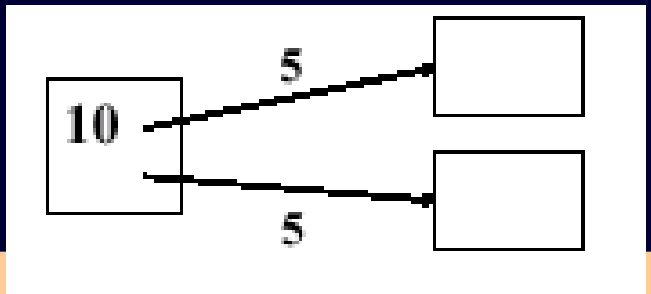
- **Ricerca normale**
  - Si digitano una serie di parole
  - Se racchiuse tra “ e ” si ricercano parole consecutive
  - Uso di + e – (Es. +Clinton –Levinski)
- **Ricerca avanzata**
  - Si possono utilizzare gli operatori booleani (AND, OR, NOT – anche &, |, !)
  - Si possono ordinare i risultati
- **Page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

Google: un altro esempio

- **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: per autorevolezza (si considera il numero e l'autorevolezza dei link entranti)**

Page rank: the Web is a graph

- Idea (Brin & Page, 98)
  - If a page is linked to by many pages, then the page is likely to be important.
  - If a page is linked to by important pages, then the page is likely to be important even though there aren't too many pages linking to it.
  - The importance of a page is divided evenly and propagated to the pages pointed to by it.



Componenti di un search engine

- **Strutture dati**
  - Pagine
  - Indice
- **Crawler**
- **Indicizzatore**
- **Calcolo del page rank**