

Lezione 8

Motori di Ricerca

Basi di dati

- Un campo prevalente dell'applicazione informatica è quello costituito dall'archiviazione e dalla gestione dei dati (*basi di dati*).
- *Sistema Informativo*. Un sistema informativo è un insieme di strumenti automatici, uomini, procedure, informazioni e norme, orientato alla gestione delle informazioni di interesse di un ente per i propri fini.
- Di un Sistema Informativo viene detto *Sistema Informatico* la parte automatizzata che comprende sia l'hardware che l'interfaccia per l'utente e le procedure sui dati.
- Tecnologia delle *basi di dati*: la gestione ed il trattamento dei dati vengono centralizzati in un unico sistema logico di archivi, che costituisce una rappresentazione non ridondante ed integrata dei dati, ed a cui tutti gli utenti possono accedere per la porzione di loro competenza. Si evitano in tal modo i problemi di
 - *ridondanza* dei dati
 - possibilità di *inconsistenza* nella rappresentazione dei dati
 - impossibilità di una *gestione in tempo reale*
 - *rigidità* del sistema informativo risultante

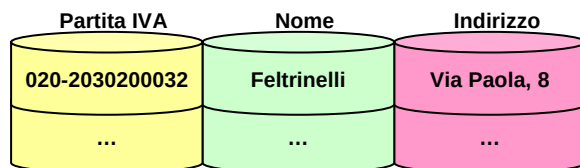
Data Base Management System

- Le applicazioni informatiche che gestiscono il flusso di informazioni ed i dati, ponendosi fra il database e l'utente, si chiamano *Data Base Management System* (DBMS).
- Le funzioni che svolge un DBMS sono essenzialmente quattro:
 - *Controllo centralizzato dei dati*
 - *Separazione degli aspetti logici dagli aspetti fisici*
 - *Controllo dell'accesso contemporaneo ai dati*
 - *Controllo sulla sicurezza dei dati*
- Per realizzare questi obiettivi, il DBMS si struttura con un organizzazione basata su tre livelli:
 - **livello interno**: è relativo al modo con cui i dati sono organizzati nelle strutture fisiche di memorizzazione
 - **livello logico**: è relativo alla rappresentazione logica delle basi di dati, ossia al modo con cui i dati sono descritti e alle loro relazioni indipendentemente dalla loro memorizzazione fisica
 - **livello esterno**: è relativo al modo con cui gli utenti percepiscono i dati. Uno stesso schema logico può essere rappresentato in maniera diversa in dipendenza dalle necessità dell'utente.
- Riferendoci al secondo livello, ovvero a quello logico, è senz'altro utile conoscere alcuni metodi di organizzazione logica dei dati. I principali metodi sono
 - il modello gerarchico;
 - il modello reticolare;
 - **il modello relazionale**

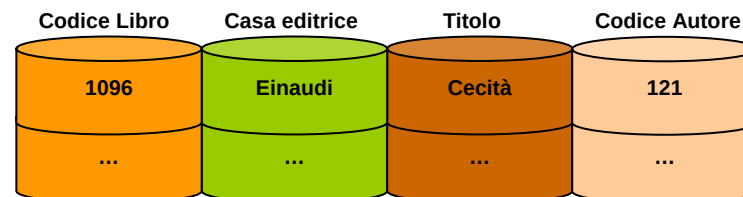
Il database relazionale

- Nel database relazionale i dati sono organizzati in tabelle chiamate **relazioni**.
- Ogni tabella si compone di un certo numero di colonne, dette gli **attributi** della relazione
- Ogni riga invece contiene gli insiemi dei **dati** relativi ad una informazione.

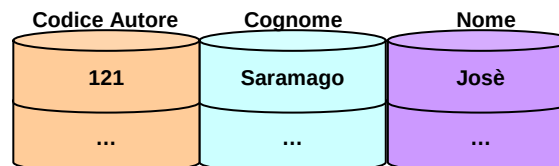
Negozio		
attributo 1	attributo 2	attributo 3
Partita IVA	Nome	Indirizzo
020-2030200032	Feltrinelli	Via Paola, 8
...	...	...



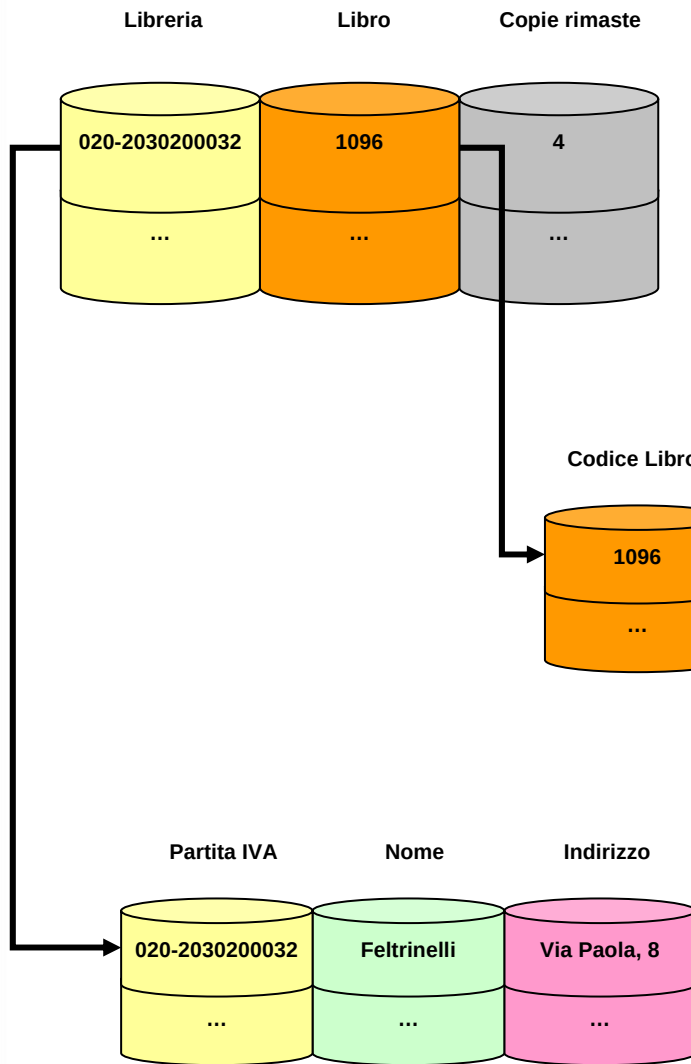
Libro			
attributo 1	attributo 2	attributo 3	attributo 4
Codice Libro	Casa Editrice	Titolo	Codice Autore
1096	Einaudi	Cecità	121
...	...	...	...



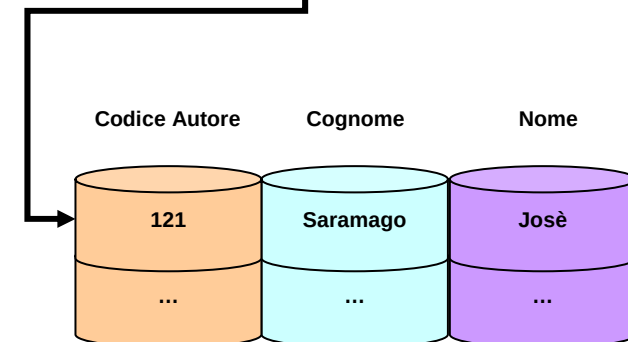
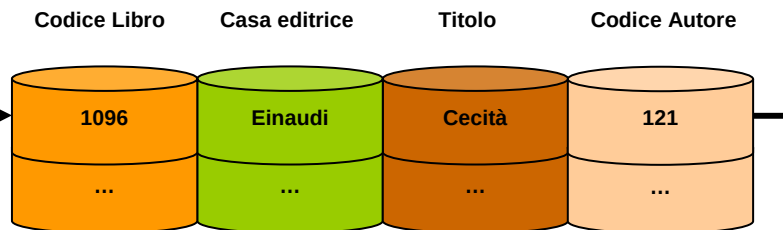
Autore		
attributo 1	attributo 2	attributo 3
Codice Autore	Cognome	Nome
121	Saramago	Josè
...	...	...



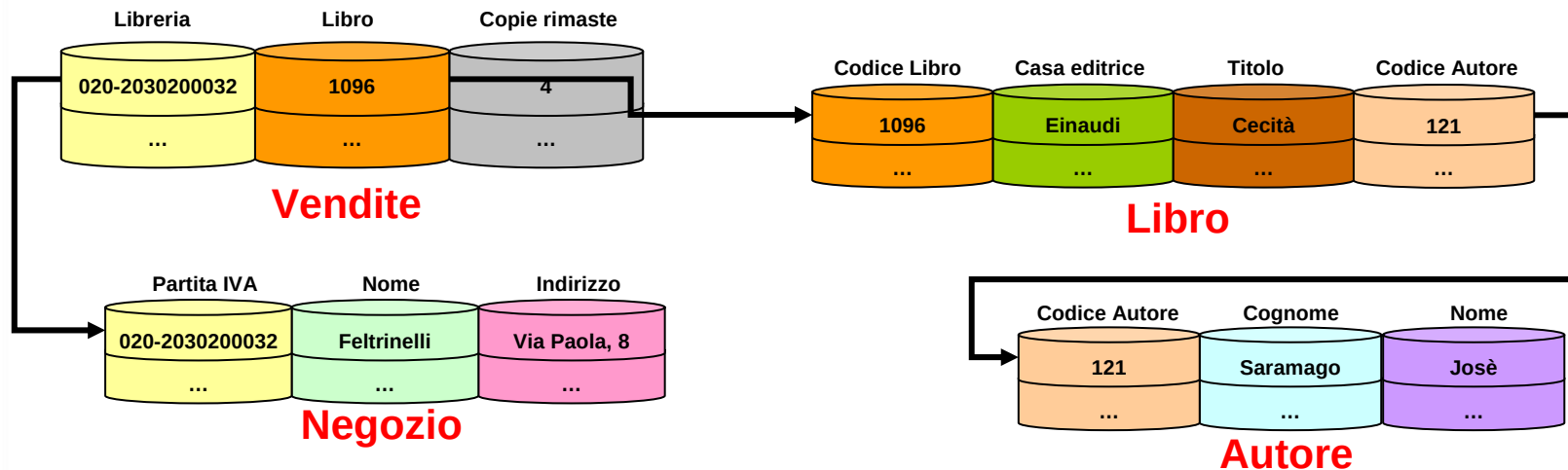
Esempio



Vendite		
attributo 1	attributo 2	attributo 3
Libreria	Libro	Copie rimaste
020-2030200032	1096	4
...	...	...



Osservazioni



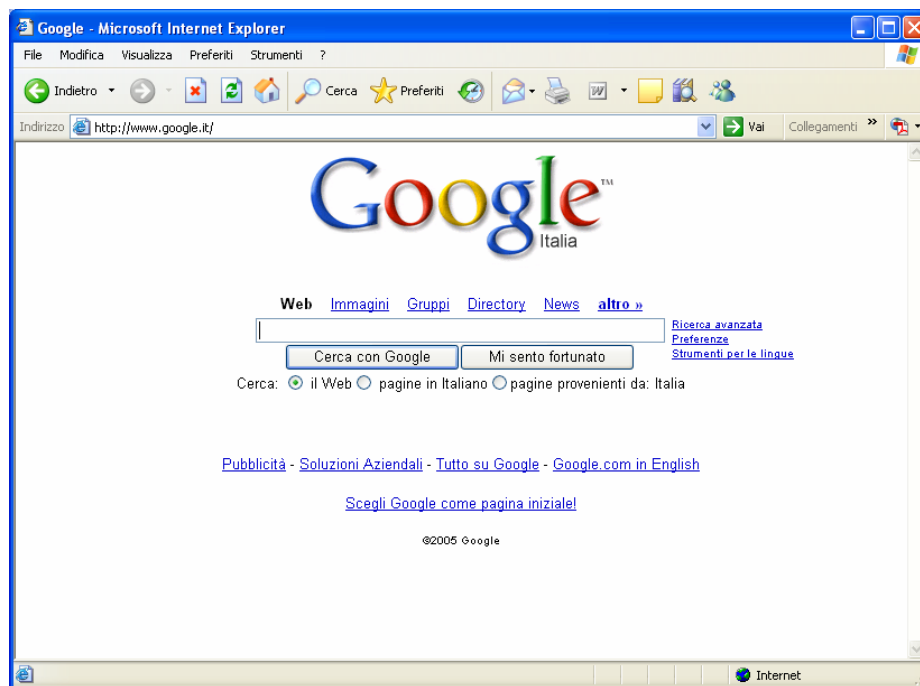
- Dai valori contenuti nella tabella **Vendite** si può risalire a quelli contenuti nelle altre.
- La colonna **Libreria** della tabella **Vendite** e la colonna **Partita-IVA** della tabella **Negozi** sono definite sullo stesso dominio di valori. Con questo meccanismo, una riga della relazione **Negozi** viene messa in corrispondenza con tutte le righe della tabella **Vendite** che presentano lo stesso valore dell'attributo **Libreria**.
- Il valore dell'attributo **Partita-IVA** della tabella **Negozi** identifica una ed una sola libreria. Quando occorre questa evenienza in un attributo, questo si chiama **chiave della relazione**, e la sua importanza risiede nel fatto che la chiave identifica univocamente una ed una sola riga in una tabella.
- Questo tipo di organizzazione dei dati è estremamente flessibile. Eventuali modifiche, aggiunte di informazioni alle tabelle, consultazioni o nuove aggregazioni di dati, sono semplicemente ottenibili gestendo le relazioni fra le varie tabelle.

Ricerche in Internet

- Gli strumenti a disposizione per le ricerche in Internet, sono essenzialmente quattro:
 - **Motori di ricerca:** applicazioni che catalogano il materiale presente in rete automaticamente con l'utilizzo di programmi chiamati *spider* o *crawler*, creando così dei database consultabili tramite l'inserimento di parole chiave e restituendo l'URL delle pagine che meglio esaudiscono i criteri della ricerca
 - **Metamotori:** applicazioni in grado di estendere la ricerca interrogando i principali motori di ricerca presenti in rete
 - **Directory:** collezioni di indirizzi web (URL) reperiti manualmente e catalogati secondo aree di interesse
 - **Database specializzati:** applicazioni di interesse specifico contenenti informazioni specializzate e dettagliate consultabili tramite rete

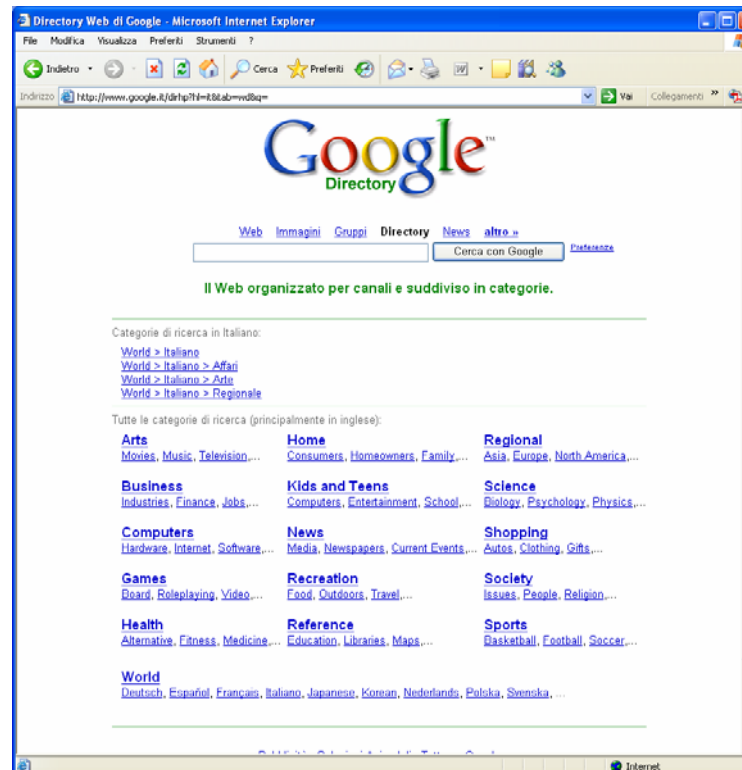
Motori di ricerca e metamotori

- Disponibilità di tutto il testo della pagina web selezionata
- Ricerca per parole chiave che devono essere presenti esattamente nel testo
- Nessuna catalogazione per argomenti
- Database generato da *spider* in maniera automatica con un minimo intervento degli operatori
- Grandi dimensioni del materiale censito
- I meta-motori cercano su più motori di ricerca in maniera superficiale, tipicamente ritengono il 10% del risultato della ricerca su ciascun motore



Directory

- Siti selezionati ed inseriti manualmente a volte da esperti della materia
- A volte mantenuti aggiornati
- Tipicamente organizzati in maniera gerarchica
- Possibilità di navigare attraverso le categorie tramite l'indicazione di termini generici
- *Non è necessario ricercare i documenti conoscendo a priori parte del contenuto*



Spider e webcrawler

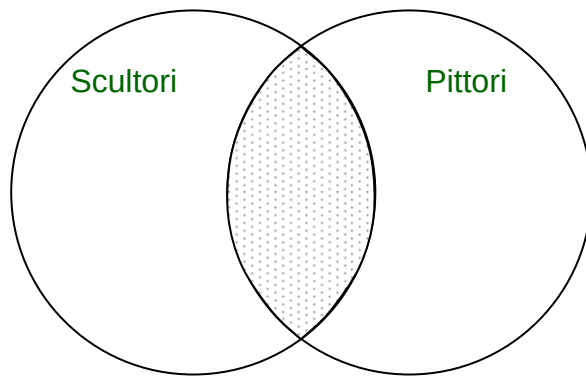
- I motori di ricerca, per trovare le informazioni richieste fanno uso di programmi denominati *spider*. Sebbene sia difficile conoscere l'algoritmo degli spider, in quanto dall'efficienza di quest'ultimo dipende la bontà dell'archivio di un motore di ricerca, si possono delineare i suoi principi di funzionamento.
- Per prima cosa occorre dire che l'esplorazione della rete parte sempre da un link (pagine non referenziate non saranno mai indicizzate).
- Una volta raggiunta una pagina, parte l'esplorazione del sito da parte del programma automatico, il quale esamina il testo della pagina e lo riporta alla applicazione che genera il *database*.
- A volte lo spider parte da una pagina e seguendo i link esplora più pagine di un sito: questo metodo è detto *deep crawl*. Lo spider esamina il testo escludendo parole di uso talmente comune da essere eliminate da quelle in ricerca, le cosiddette *stop words*.
- Quando un motore di ricerca presenta l'elenco delle pagine che soddisfano i criteri di ricerca, li elenca in una classifica che accoppia sia l'attinenza della pagina alla ricerca che il cosiddetto *linkage weighting*, ovvero una metodologia di classifica basata sul numero di link verso una pagina: più una pagina è raggiunta da link maggiore sarà la sua importanza e quindi verrà proposta prima come risultato di ricerca.

La rete invisibile

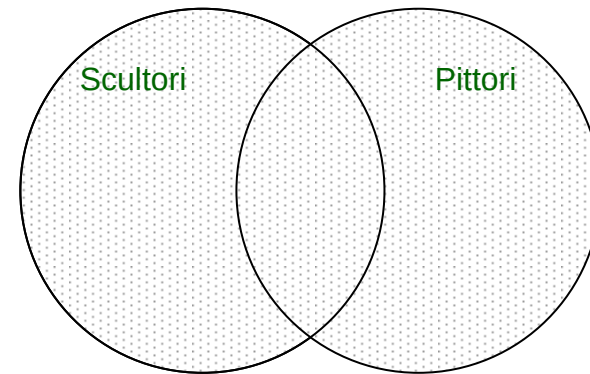
- Esiste una categoria di pagine web che non sono rintracciabili dagli spider: è la cosiddetta *rete invisibile*.
- I motivi tecnici possono essere dovuti a:
 - *Pagine dinamiche.*
 - *Pagine con richiesta di login*
 - *Pagine non in formato HTML* (es. documenti PDF ed i formati grafici)
 - Pagine generate da script (Java o Javascript), contenenti nell'URL il carattere ? che viene interpretato dallo *spider* come una stop word.
- I motivi non tecnici per la non rintracciabilità sono invece essenzialmente dovuti a:
 - *Pagine escluse per scelta*, ad esempio pagine contenenti argomenti considerati offensivi o provenienti da siti considerati non legali o con contenuti offensivi.

Il linguaggio di ricerca

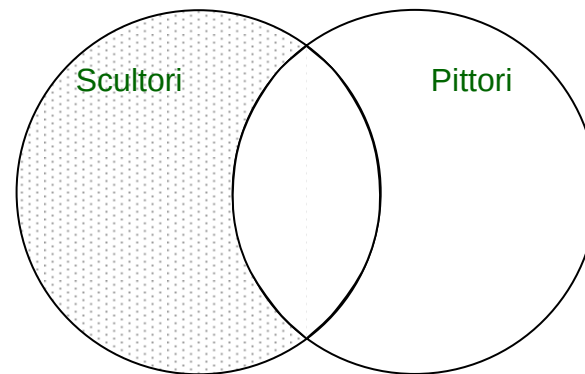
Si basa sugli operatori della logica booleana



scultori **AND** pittori



scultori **OR** pittori



scultori **AND NOT** pittori

Altri operatori

- **NEAR** ed altri operatori di prossimità
 - Richiede la presenza di entrambi i termini, come AND, ma richiede inoltre che essi siano vicini nel testo, tipicamente ad una distanza massima di dieci vocaboli
 - Es: scultori NEAR pittori
- Parentesi **()** *nesting*
 - Obbliga a ricercare ed eseguire per primi i termini e le operazioni logiche incluse fra le parentesi
 - Le parentesi *devono essere usate* per concatenare termini con OR o NEAR se altri operatori logici sono presenti nella ricerca
- Carattere jolly *****
 - Usato per evitare di specificare un carattere
 - Es: cavall* sostituisce cavallo, cavalla, cavalli cavallerizzo... e tutte le parole che cominciano per cavall
- **“ ”** virgolette
 - Richiede che il testo contenuto nelle virgolette appaia esattamente così come viene scritto
 - E' più restrittivo di AND e NEAR
 - Vincola anche l'ordine di ricerca delle parole: “Giuseppe Verdi” è diverso da “Verdi Giuseppe”