UNIVERSITÀ DEGLI STUDI DI MILANO BICOCCA
DIPARTIMENTO DI INFORMATICA, SISTEMISTICA E COMUNICAZIONE

STRUMENTI E APPLICAZIONI DEL WEB

Il problema dell'accesso alle informazioni (1)

Marco Viviani

Anno Accademico 2025-2026

Introduzione al problema

Come scoprire ciò di cui abbiamo bisogno sul Web? (1/2)

UNIVERSITÀ DEGLI STUDI DI MILANO BICOCCA
DIPARTIMENTO DI INFORMATICA, SISTEMISTICA E COMUNICAZIONE

- Il WWW nasce nel 1991. Da allora i siti proliferano generando una mole crescente e inarrestabile di dati: testi / immagini / audio / …
- Dal 2004 con l'avvento di Facebook nascono e si moltiplicano i Social Media:
  - quantità immani di contenuto generato dagli utenti su una moltitudine di argomenti! Musica, sport, medicina, cinema, …
- Potenzialmente ci si dovrebbe trovare risposta ad ogni nostra domanda ma…
- … a volte è come trovare un ago in un pagliaio!

Come scoprire ciò di cui abbiamo bisogno sul Web? (2/2)

UNIVERSITÀ DEGLI STUDI DI MILANO BICOCCA
DIPARTIMENTO DI INFORMATICA, SISTEMISTICA E COMUNICAZIONE

Il «sovraccarico informativo» (1/2)

UNIVERSITÀ DEGLI STUDI DI MILANO BICOCCA
DIPARTIMENTO DI INFORMATICA, SISTEMISTICA E COMUNICAZIONE

- Il termine "sovraccarico di informazioni" è stato coniato da **Bertram Gross**, professore di scienze politiche all'Hunter College, nella sua opera del 1964 - The Managing of Organizations.

- È stato reso popolare da **Alvin Toffler**, lo scrittore americano e futurista, nel suo libro "Future Shock" nel 1970.

Il «sovraccarico informativo» (2/2)

UNIVERSITÀ DEGLI STUDI DI MILANO BICOCCA
DIPARTIMENTO DI INFORMATICA, SISTEMISTICA E COMUNICAZIONE

"Information overload occurs when the amount of input to a system exceeds its processing capacity. Decision makers have fairly limited cognitive processing capacity. Consequently, when information overload occurs, it is likely that a ***reduction in decision quality*** *will occur."*

Soluzioni? (1/3)

- Browsing?
  - Inefficace!
  - Può andare bene quando abbiamo un ragionevole punto di partenza.

Soluzioni? (2/3)

UNIVERSITÀ DEGLI STUDI DI MILANO BICOCCA
DIPARTIMENTO DI INFORMATICA, SISTEMISTICA E COMUNICAZIONE

- Spesso cerchiamo informazioni sulle piattaforme offerte dai Social Media, anche adottando il paradigma del **passaparola**.

- **Pericolo**: contenuti generati in modo incontrollato da chi si crede esperto e non lo è…
- **Risultato**: ottenere disinformazioni anziché informazioni.

Soluzioni? (3/3)

- L'enorme e crescente quantità di informazioni disponibili (BIG DATA) ha motivato la definizione di **sistemi software** che ci aiutino ad orientarci in questo oceano.

Accesso automatico alle informazioni

UNIVERSITÀ DEGLI STUDI DI MILANO BICOCCA
DIPARTIMENTO DI INFORMATICA, SISTEMISTICA E COMUNICAZIONE

- Sviluppo di sistemi che aiutano l'utente a **identificare le informazioni rilevanti** per le sue necessità (informare, cioè ridurre l'ignoranza).
- La definizione di tali sistemi si basa sulla soluzione di un **problema di decisione**: come identificare e definire l'importanza delle informazioni che soddisfano le preferenze dell'utente? È necessario:
  - Interpretare il contenuto di testi, immagini, video, audio;
  - Interpretare le necessità informative dell'utente.
- Ruolo centrale della nozione di **rilevanza**: la rilevanza è una proprietà soggettiva: difficile da definire e misurare!

Sistemi di accesso alle informazioni

UNIVERSITÀ DEGLI STUDI DI MILANO BICOCCA
DIPARTIMENTO DI INFORMATICA, SISTEMISTICA E COMUNICAZIONE

- *DataBase Management Systems* (DBMS)
  - Richiedono la **formulazione di una *query***.
- *Information Retrieval Systems* (*Search Engines*) – Motori di ricerca
  - Richiedono la **formulazione di una *query***.
- *Information Filtering Systems* (*Recommender Systems*) – Sistemi di raccomandazione
  - Richiedono **profili utente**, cioè descrizioni di necessità informative e preferenze aggiornate dinamicamente, anche in base al comportamento dell'utente (NO *query*).

Il nostro focus

- Motori di ricerca
  - Richiedono la formulazione di una *query* (tecnologia *pull*)

- Sistemi per la raccomandazione di informazioni (*Recommender Systems*)
  - Richiedono profili utente (assenza di *query*, tecnologia *push*)

I motori di ricerca

L'*Information Retrieval* (1/2)

UNIVERSITÀ DEGLI STUDI DI MILANO BICOCCA
DIPARTIMENTO DI INFORMATICA, SISTEMISTICA E COMUNICAZIONE

- L'IR è la disciplina informatica che si occupa dell'**archiviazione e del recupero dei documenti**.

- Il suo obiettivo è la creazione di sistemi software che permettano la memorizzazione di grandi quantità di documenti in un archivio, per permettere un **recupero efficiente ed efficace** dei documenti rilevanti per le necessità informative degli utenti.

L'*Information Retrieval* (2/2)

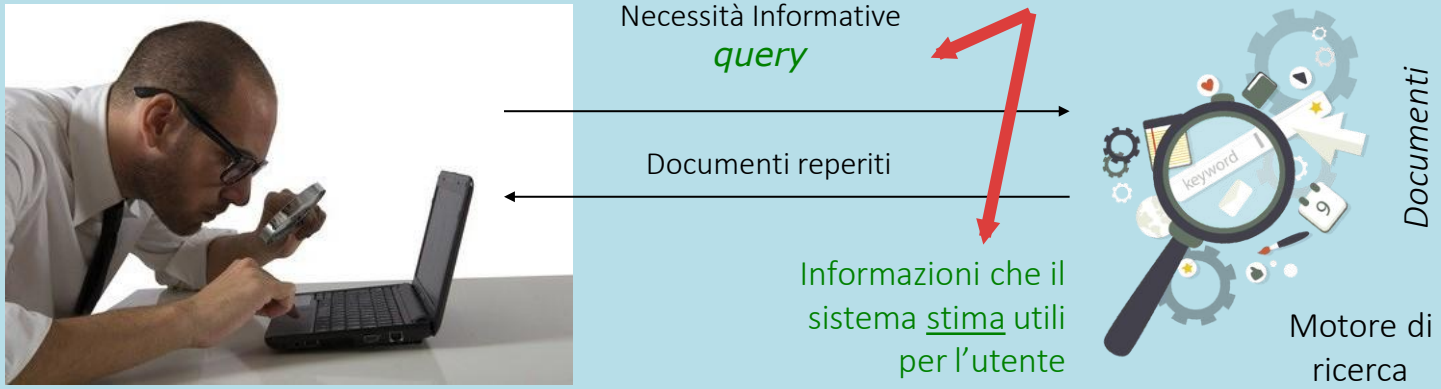
UNIVERSITÀ DEGLI STUDI DI MILANO BICOCCA
DIPARTIMENTO DI INFORMATICA, SISTEMISTICA E COMUNICAZIONE

"IR is the name for the process or method whereby a prospective user of information is able to convert his need for information into an actual list of citations to documents in storage containing information useful to him (...). IR embraces the intellectual aspects of the description of information and its specification for search, and also whatever systems, techniques, and machines that are employed to carry out the operation." [Mooers, 1951]

Cosa deve fare un motore di ricerca?

UNIVERSITÀ DEGLI STUDI DI MILANO BICOCCA
DIPARTIMENTO DI INFORMATICA, SISTEMISTICA E COMUNICAZIONE

- L'obiettivo di un motore di ricerca è reperire tutti i documenti utili per l'utente che ha formulato la *query*, possibilmente non presentando i documenti non utili
  → Compito difficile perché pervaso da **soggettività**.



Glossario (1/2)

UNIVERSITÀ DEGLI STUDI DI MILANO BICOCCA
DIPARTIMENTO DI INFORMATICA, SISTEMISTICA E COMUNICAZIONE

- **Documento** (d): unità di informazione recuperabile, espressa in formato libero (senza l'applicazione di formati o schemi specifici). I documenti hanno un contenuto informativo.
- **Archivio** (D): insieme di documenti accessibili tramite un IRS; può essere centralizzato o distribuito.
- **IR testuale**: articoli scientifici, lettere, articoli di giornale, legende di immagini o grafici, trascrizioni audio, ...
- **IR multimediale**: immagini, grafici, audio (parlato o non parlato), video, ..., memorizzati in formato digitale.

Glossario (2/2)

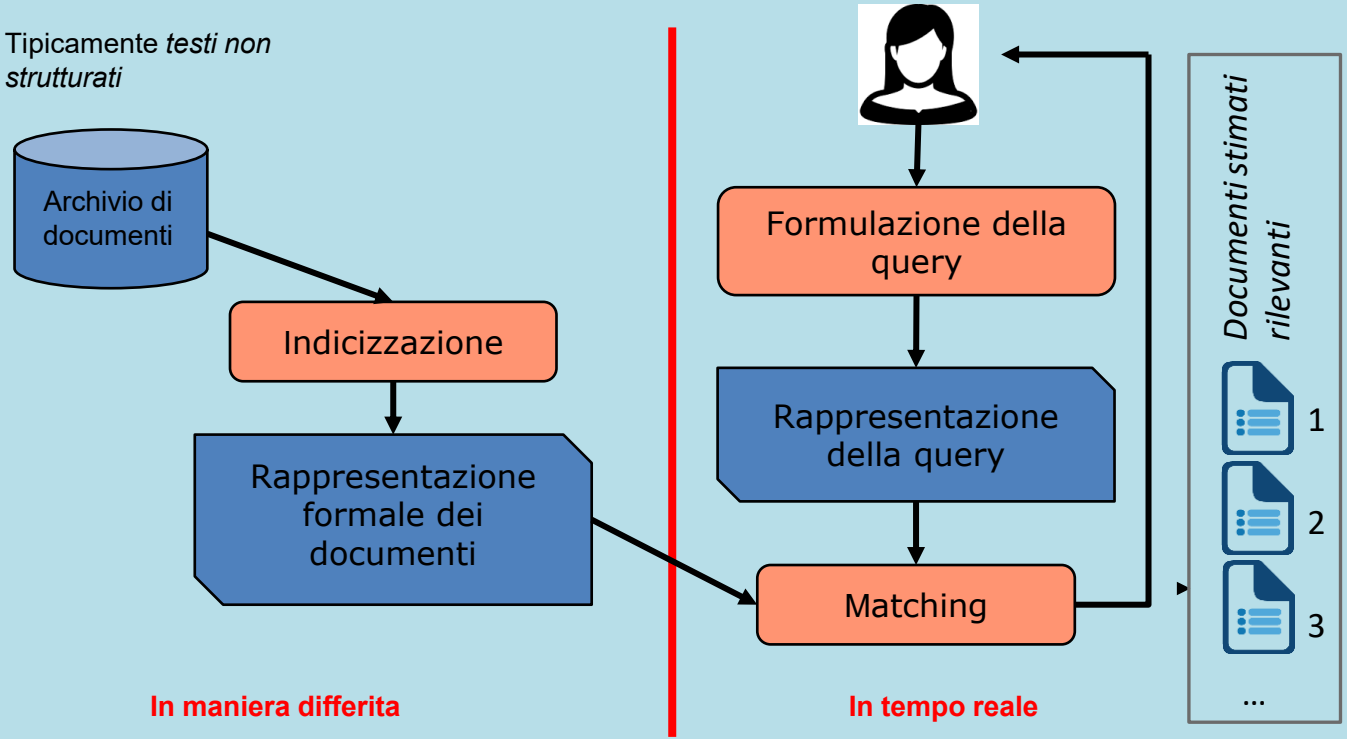
UNIVERSITÀ DEGLI STUDI DI MILANO BICOCCA
DIPARTIMENTO DI INFORMATICA, SISTEMISTICA E COMUNICAZIONE

- **Grandi dimensioni**: i sistemi distribuiti e la diffusione del Web hanno generato basi molto grandi di documenti (archivi) (ad esempio, Google sta indicizzando miliardi di pagine Web).

- **Necessità informative**: necessità di informazioni utili per risolvere un problema; porta alla formulazione di una *query*.

- **Rilevanza**: utilità di un documento per l'utente, stimata dal motore di ricerca rispetto alla *query* che l'utente ha formulato (ma non solo...).

Struttura base di un motore di ricerca (generico)



Componenti di un motore di ricerca

UNIVERSITÀ DEGLI STUDI DI MILANO BICOCCA
DIPARTIMENTO DI INFORMATICA, SISTEMISTICA E COMUNICAZIONE

- **Archivio di documenti**: il documento è l'unità di informazione reperibile. Può essere costituito da un testo in forma narrativa (testuale) o essere composto da parti narrative, pittoriali, codificate, etc. (multimediale).
- **Rappresentazione formale dei documenti**: sintetizza il contenuto informativo dei documenti. È ottenuta mediante il processo di indicizzazione.
- **Linguaggio di interrogazione**: in una *query* sono espresse le condizioni per la selezione dei documenti di possibile interesse.
    - Le query sono solitamente espresse come insiemi di parole chiave (*keywords*).
- **Meccanismo di confronto** (*matching*): confronta la rappresentazione dei documenti archiviati con le condizioni di selezione espresse nella *query*.

Modelli per i motori di ricerca

UNIVERSITÀ DEGLI STUDI DI MILANO BICOCCA
DIPARTIMENTO DI INFORMATICA, SISTEMISTICA E COMUNICAZIONE

- Un motore di ricerca è basato su un **modello matematico** che fornisce una descrizione formale:
  - del documento;
  - della *query*;
  - del modo in cui confrontare rappresentazioni di *query* e documenti per effettuare una **stima della rilevanza** dei documenti e produrre la lista (ordinata) dei documenti stimati rilevanti.
- N.B. Un motore di ricerca semplifica la realizzazione dell'attività di reperimento di risultati rilevanti → i risultati prodotti non sono «perfetti» (stima di rilevanza).

Il processo di indicizzazione

UNIVERSITÀ DEGLI STUDI DI MILANO BICOCCA
DIPARTIMENTO DI INFORMATICA, SISTEMISTICA E COMUNICAZIONE

- Problema:
  - Come descrivere il contenuto «semantico» di un documento in un modo automaticamente gestibile?
- Il processo di **indicizzazione** è basato sull'estrazione di «elementi» (*feature*) che costituiscono la base della descrizione (rappresentazione) del documento
  - Per i testi tali elementi (detti indici) sono generalmente parole;
  - I documenti sono rappresentati come insiemi (pesati) di parole;
  - Idea concepita negli anni 60 e ancora in larga parte utilizzata!

Documenti rilevanti rispetto alla *query*

UNIVERSITÀ DEGLI STUDI DI MILANO BICOCCA
DIPARTIMENTO DI INFORMATICA, SISTEMISTICA E COMUNICAZIONE

- Scopo di un motore di ricerca: **reperire i documenti rilevanti per l'utente** → la rilevanza di un documento è relativa alla *query* formulata.
- Documenti e *query* hanno una **rappresentazione formale**.
- Tali rappresentazioni vengono **confrontate**:
  - **Confronto esatto**: nozione binaria di «rilevanza».
    - Rilevante / Non rilevante.
  - **Confronto parziale**: nozione «graduale» di rilevanza:
    - Idea: confronto «parziale» tra documento e *query*;
    - I documenti «sufficientemente simili» alla *query* vengono reperiti.

Modelli base nei motori di ricerca (1/4)

UNIVERSITÀ DEGLI STUDI DI MILANO BICOCCA
DIPARTIMENTO DI INFORMATICA, SISTEMISTICA E COMUNICAZIONE

- Modello booleano:
  - Documenti e *query* sono rappresentati come **insiemi di parole**.
  - Le ***query* sono di tipo booleano**: due o più termini di ricerca connessi da operatori Booleani:
    - AND, OR, è possibile negare un termine con l'operatore NOT.
  - Il meccanismo di *matching* applica **operazioni insiemistiche** (Es., unione, intersezione).
  - La **rilevanza** à modellata come proprietà binaria dei documenti:
    - **Rilevante/non rilevante** → 1/0

Modelli base nei motori di ricerca (2/4)

UNIVERSITÀ DEGLI STUDI DI MILANO BICOCCA
DIPARTIMENTO DI INFORMATICA, SISTEMISTICA E COMUNICAZIONE

- **Esempio** di rappresentazione formale nel modello booleano:
  - Documento 1 = {Corso, di, Informatica, per, le, Scienze, della, Terra}
  - Documento 2 = {Corso, di, Informatica, per, Editoria, Musicale}
  - Documento 3 = {Corso, di, Informatica, per, le, Scienze, dei, Materiali}
  - *Query* = corso AND informatica AND scienze (TUTTE le parole devono essere contenute)
  - Risultato del *matching* → {Documento1, Documento3}

Modelli base nei motori di ricerca (3/4)

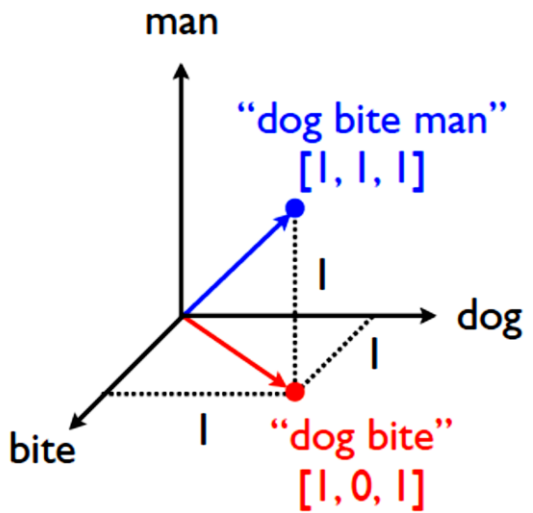
UNIVERSITÀ DEGLI STUDI DI MILANO BICOCCA
DIPARTIMENTO DI INFORMATICA, SISTEMISTICA E COMUNICAZIONE

- Modello vettoriale:
  - È basato sull'**algebra lineare** e rappresenta sia i documenti sia le *query* in uno **spazio vettoriale** ***n*-dimensionale**, dove *n* è il numero totale di termini indice (parole considerate).
  - L'angolo tra i vettori viene usato come misura di **similarità** tra documento e *query*.
  - La rilevanza è modellata come **proprietà graduale** dei documenti rispetto alla query → valori reali.

Modelli base nei motori di ricerca (4/4)

UNIVERSITÀ DEGLI STUDI DI MILANO BICOCCA
DIPARTIMENTO DI INFORMATICA, SISTEMISTICA E COMUNICAZIONE

	dog	*man*	*bite*
documento	1	1	1
query	1	0	1



La rilevanza tematica

UNIVERSITÀ DEGLI STUDI DI MILANO BICOCCA
DIPARTIMENTO DI INFORMATICA, SISTEMISTICA E COMUNICAZIONE

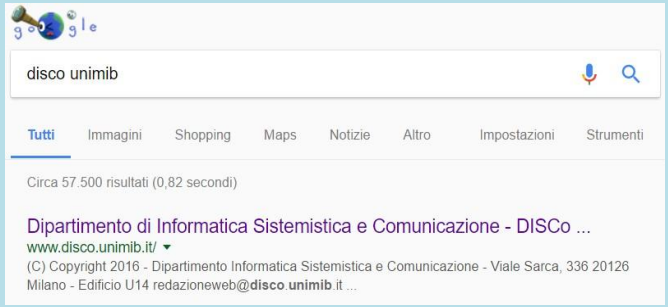
- La valutazione di una *query* viene effettuata dalla componente di *matching*, che per ogni documento calcola un **valore numerico** che rappresenta la **stima della rilevanza** dei documenti rispetto alla *query*.
- Il motore di ricerca deve identificare le pagine Web che trattano gli argomenti specificati dalle parole nella query → **valutazione della pertinenza**.
- In questo caso il valore numerico rappresenta la **rilevanza tematica** del documento rispetto alla query.
- In realtà **la rilevanza è un concetto multi-dimensionale** → Prossime slide.

I motori di ricerca sul Web

UNIVERSITÀ DEGLI STUDI DI MILANO BICOCCA
DIPARTIMENTO DI INFORMATICA, SISTEMISTICA E COMUNICAZIONE

- Nascono per **reperire pagine Web** a fronte di una richiesta esplicita dell'utente (*query*)

- Oggi tanti **motori di ricerca verticali**: immagini, video, ecc.



Una curiosità: la nascita di *Google Images*

UNIVERSITÀ DEGLI STUDI DI MILANO BICOCCA
DIPARTIMENTO DI INFORMATICA, SISTEMISTICA E COMUNICAZIONE

- Jennifer Lopez è arrivata ai Grammy del 2000 in quell'ormai famosissimo Versace verde.
- In un saggio pubblicato a gennaio su Project Syndicate, Eric Schmidt, presidente esecutivo di Google, scrisse: "*At the time, it was the most popular search query we had ever seen. But we had no surefire way of getting users exactly what they wanted: J.Lo wearing that dress.*"

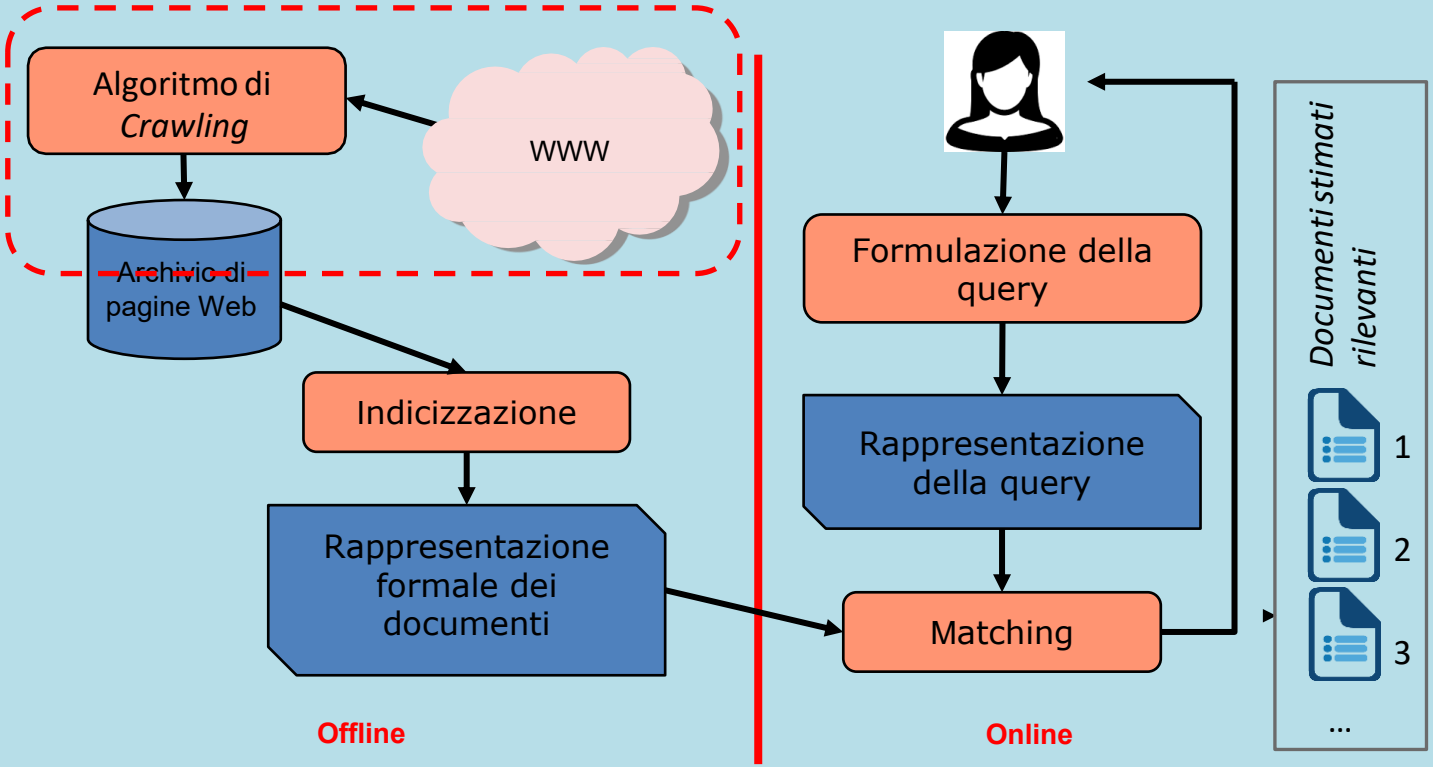
https://www.businessinsider.com/jennifer-lopezs-grammys-dress-inspired-google-image-search-2015-5?IR=T

Nascita e scomparsa di motori di ricerca sul Web

UNIVERSITÀ DEGLI STUDI DI MILANO BICOCCA
DIPARTIMENTO DI INFORMATICA, SISTEMISTICA E COMUNICAZIONE

Struttura di un motore di ricerca sul Web

UNIVERSITÀ DEGLI STUDI DI MILANO BICOCCA
DIPARTIMENTO DI INFORMATICA, SISTEMISTICA E COMUNICAZIONE



Raccolta (*gathering*) delle pagine Web

UNIVERSITÀ DEGLI STUDI DI MILANO BICOCCA
DIPARTIMENTO DI INFORMATICA, SISTEMISTICA E COMUNICAZIONE

Esistono **due modalità**:

- Le pagine Web vengono **fornite (spedite) direttamente** al motore di ricerca dai proprietari.
- Il motore di ricerca è dotato di un **agente software** (*information agent*) detto *crawler*, (*spider*, *worm*, *robot*, *Web Search Agent*) che attraversa il Web per spedire pagine nuove o aggiornate a un server che le indicizza.
    - Il *crawler* naviga su Web usando come punti di partenza URL noti per essere punti di accesso interessanti e successivamente visita altre pagine Web percorrendo i *link* che vanno da una pagina all'altra.

Il *crawler* (1/2)

UNIVERSITÀ DEGLI STUDI DI MILANO BICOCCA
DIPARTIMENTO DI INFORMATICA, SISTEMISTICA E COMUNICAZIONE

- Un *crawler* deve raccogliere le pagine, essenzialmente effettuando una **visita del grafo del Web**.
- I *crawler* vengono eseguiti su server locali e spediscono richieste a server remoti.
  - Infatti i crawler possono essere distribuiti.
- La raccolta inizia a partire da una (o più) pagine, detta(e) **seme/i di raccolta** (*seed set*).

Problemi nella scelta dei semi

- La scelta del seme
  - Influisce sull'insieme delle pagine visitate (*coverage*).
  - Se si parte da una pagina situata nella «componente gigante» si riesce a visitare tutto il Web, tranne il lato sinistro del cravattino.
  - Esiste comunque una parte di Web non raggiungibile usando i *link*: è il cosiddetto «Web nascosto» o «Web profondo» (*hidden Web* / *deep Web*).
  - Secondo molte stime la porzione di Web visibile è minima e copre circa il 16-20% del totale (stima variabile).

Il *Deep Web*

UNIVERSITÀ DEGLI STUDI DI MILANO BICOCCA
DIPARTIMENTO DI INFORMATICA, SISTEMISTICA E COMUNICAZIONE

- ***Deep Web*** (o *hidden Web*) indica le parti del World Wide Web il cui contenuto non è indicizzato dai motori di ricerca Web standard.

- Qualche esempio:

  - **Pagine dinamiche**, generate dinamicamente da applicazioni specifiche, quindi scartare. Per esempio, pagine di prenotazione di voli;

  - **Pagine con contenuti ad accesso limitato**, in cui l'accesso è limitato in modo tecnico (ad esempio, utilizzando CAPTCHA);

  - **Pagine Web private**, che richiedono registrazione e accesso (risorse protette da password);

  - **Contenuto scollegato**, ad esempio pagine che non sono collegate da altre pagine.

E il *Dark Web*?

UNIVERSITÀ DEGLI STUDI DI MILANO BICOCCA
DIPARTIMENTO DI INFORMATICA, SISTEMISTICA E COMUNICAZIONE

- Il cosiddetto «Web oscuro» o ***Dark Web*** è costituito dai contenuti del World Wide Web nelle *darknet* (reti oscure) che si raggiungono via Internet attraverso specifici software, configurazioni e accessi autorizzativi.
- Il *Dark Web* è una piccola parte del *Deep Web*, anche se a volte i due termini vengono confusi.
- Le *darknet* che costituiscono il *Dark Web* includono sia piccole reti, sia reti più grandi (Es., Tor, Freenet, e I2P), in cui operano organizzazioni e singoli individui.
- Il *Dark Web* è basato su informazione crittografata.

Il *crawler* (2/2)

UNIVERSITÀ DEGLI STUDI DI MILANO BICOCCA
DIPARTIMENTO DI INFORMATICA, SISTEMISTICA E COMUNICAZIONE

- Quali *link* seguire?
- A volte non si vuole scaricare l'intero Web, ma solo una parte; ad esempio, le pagine del «Web italiano».
- Occorre stabilire dei criteri (Cos'è il «Web italiano»? .it? qualunque pagina in lingua italiana? E come si valuta se una pagina è in italiano?)
    - Tali criteri indicano quali URL seguire e quali non seguire.

Il processo di *crawling*

UNIVERSITÀ DEGLI STUDI DI MILANO BICOCCA
DIPARTIMENTO DI INFORMATICA, SISTEMISTICA E COMUNICAZIONE

- Inizializza l'insieme dei semi con alcuni URL noti (popolari o inviati da utenti):
  - Ad esempio: http://www.unimib.it, http://www.comune.milano.it, …
- Seleziona un indirizzo URL dall'insieme.
- Seleziona la pagina.
- Cerca nella pagina indirizzi di altri URL.
  - ad esempio alumni
- Scarta gli URL che:
  - non possono essere analizzati, es., .exe .jpg .ps, …
  - sono già stati visitati.
- Aggiunge gli URL all'insieme delle pagine da visitare:
- Se non è scaduto il tempo torna al punto 2.

Frequenza di aggiornamento

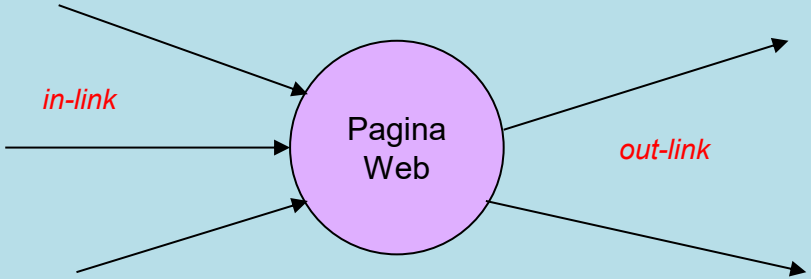
UNIVERSITÀ DEGLI STUDI DI MILANO BICOCCA
DIPARTIMENTO DI INFORMATICA, SISTEMISTICA E COMUNICAZIONE

- Le pagine Web già indicizzate sono state esplorate in tempi diversi e possono non esistere più.
- Possono essere scadute da 1 giorno a 2 mesi. Per questa ragione i motori di ricerca visualizzano la data di indicizzazione delle pagine reperite.
- Ci sono motori che imparano la frequenza di cambiamento delle pagine e le visitano di conseguenza.
- I *crawler* sono in grado di visitare circa milioni di pagine Web al giorno.

Ricerche su Web e *Web link*

UNIVERSITÀ DEGLI STUDI DI MILANO BICOCCA
DIPARTIMENTO DI INFORMATICA, SISTEMISTICA E COMUNICAZIONE

- La differenza principale tra motori di ricerca (tradizionali) e motori di ricerca sul Web riguarda la **presenza di *Web link*** (i collegamenti ipertestuali).
- *Web link*: rappresentano una relazione tra pagine connesse.
  - Documento sorgente (pagina Web) che contiene il link;
  - Documento *target* (pagina Web) che è riferito dal link.
- *In-link* e *out-link*
  - in-link della pagina *p* è un link da una pagina Web alla pagina *p*;
  - out-link della pagina *p* è un link dalla pagina *p* a una pagina Web.



Ordinamento (*ranking*) dei risultati Web

UNIVERSITÀ DEGLI STUDI DI MILANO BICOCCA
DIPARTIMENTO DI INFORMATICA, SISTEMISTICA E COMUNICAZIONE

- Il concetto di **rilevanza tematica** resta fondamentale anche nei motori di ricerca per il Web.
    - I motori di ricerca ordinano le pagine reperite in base alla loro rilevanza (tematica) rispetto alla *query*.
- La differenza principale tra motori di ricerca (tradizionali) e motori di ricerca sul Web è il fatto che i *link* che puntano a una pagina forniscono una misura della sua **popolarità**.

La rilevanza come concetto multi-dimensionale

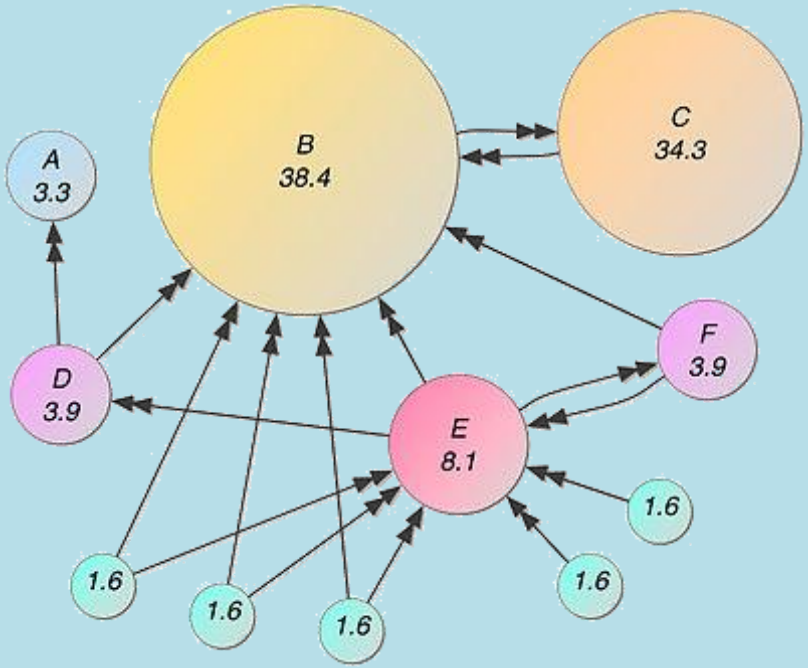
UNIVERSITÀ DEGLI STUDI DI MILANO BICOCCA
DIPARTIMENTO DI INFORMATICA, SISTEMISTICA E COMUNICAZIONE

- La **popolarità** di una pagina può essere calcolata tramite l'**analisi dei link sul grafo del Web**.
  - Algoritmo *PageRank*.
- Il **valore globale di rilevanza** è dato quindi da:
  - rilevanza tematica;
  - popolarità della pagina;
  - …
- In questo senso la rilevanza è un **concetto multi-dimensionale**.
  - Negli ultimi anni altri aspetti costituiscono attributi ulteriori, come la *novelty*.

Google e l'algoritmo *PageRank*

UNIVERSITÀ DEGLI STUDI DI MILANO BICOCCA
DIPARTIMENTO DI INFORMATICA, SISTEMISTICA E COMUNICAZIONE

- Ideato da **Sergey Brin** e **Larry Page**.
- Idea di base: *PageRank* simula la **navigazione casuale** dell'utente su Web: una pagina ha un *PageRank* alto se la somma dei *PageRank* dei suoi *in-link* è alta.
- Una pagina con *PageRank* alto ha:
  - o molti *in-link*;
  - oppure pochi *in-link* con *PageRank* alto.



La ricerca personalizzata

UNIVERSITÀ DEGLI STUDI DI MILANO BICOCCA
DIPARTIMENTO DI INFORMATICA, SISTEMISTICA E COMUNICAZIONE

- In un negozio, a Salem… ☺

La commessa ascolta la richiesta senza vedere le clienti (molto diverse!) → stesso abito: stessa taglia, stesso modello, stesso tipo independentemente dalla cliente e dall'evento.

Ignorato il <u>contesto</u>:

- Contesto dell'utente (preferenze):
  - taglia
  - tessuto
  - gusto
  - prezzo
  - …
- Contesto sociale
  - matrimonio
  - sport
  - lavoro
  - festa
  - sabba ☺

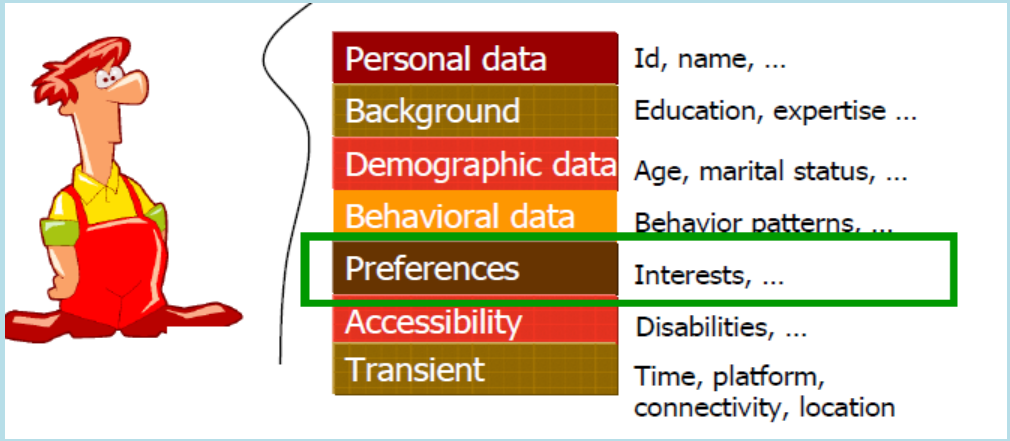
Approccio classico e approccio personalizzato

UNIVERSITÀ DEGLI STUDI DI MILANO BICOCCA
DIPARTIMENTO DI INFORMATICA, SISTEMISTICA E COMUNICAZIONE

- Approccio classico (*query-centered*)
  - La stessa query formulata da diversi utenti fornisce sempre la stessa lista di risultati.
- Approccio personalizzato (*context-centered*)
  - **Necessità informativa** = *query* + ?;
  - **Rilevanza** = utilità basata anche sul contesto dell'utente o sulla situazione;
  - Il contesto dell'utente è rappresentato attraverso un modello di utente (**profilo utente**).

Contesto dell'utente: cosa va nel profilo?

UNIVERSITÀ DEGLI STUDI DI MILANO BICOCCA
DIPARTIMENTO DI INFORMATICA, SISTEMISTICA E COMUNICAZIONE



Tratto da: tutorial "Personalised Systems" tenuto da Yannis Ioannidis e Georgia Koutrika, VLDB 2005, Trondheim, Agosto 2005

- Interessi e comportamento dell'utente:
  - Lista di parole chiave (o concetti), documenti, informazioni sul desktop;
  - Interazioni: dati sui clic, monitoraggio del movimento degli occhi, tempo passato su una pagina, …

Il processo di personalizzazione dei risultati della ricerca

UNIVERSITÀ DEGLI STUDI DI MILANO BICOCCA
DIPARTIMENTO DI INFORMATICA, SISTEMISTICA E COMUNICAZIONE

- Approcci di *pre-processing*: modifica della *query* (*query expansion*).
- Approcci di *post-processing*: riordinamento dei risultati (*re-ranking*)
- Approcci nel processo: personalizzazione integrata nel processo di calcolo del *ranking*.