Lezioni Lincee
Dipartimento di Informatica, Sistemistica e Comunicazione
24 febbraio 2017

Tracciare una rotta nell'oceano del Web alla scoperta di informazioni utili: Approcci scientifici e modelli

GABRIELLA PASI

UNIVERSITÀ DEGLI STUDI DI MILANO-BICOCCA / DISCO

Internet e il World Wide Web

Il **World Wide Web** è stato ideato da **Tim Berners-Lee** e **Robert Cailliau**. È stato implementato da Berners-Lee nel **1991** presso il CERN di Ginevra.

Può essere visto come una **collezione di documenti (pagine Web)** collegati tra loro tramite ***link***.

Si basa su **URL**, protocollo **HTTP** e **Web Browser**.

Il WWW è uno dei **servizi** supportati da Internet.

Internet è una **rete di reti**, una interconnessione fisica e logica finalizzata al trasporto di dati su una pluralità di dispositivi terminali.

Tale interconnessione è resa possibile da una *suite* di **protocolli di comunicazione** comunemente chiamata «TCP/IP»

Come scoprire ciò di cui abbiamo bisogno sul Web?

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?

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 …

Rischio: ottenere *disinformazioni* anziché *informazioni*.

Come scoprire ciò di cui abbiamo bisogno sul Web?

Browsing? Inefficace!

Può andare bene quando abbiamo un ragionevole punto di partenza

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

Quali sistemi?

Motori di ricerca

Richiedono la formulazione di una QUERY (tecnologia ***pull***)

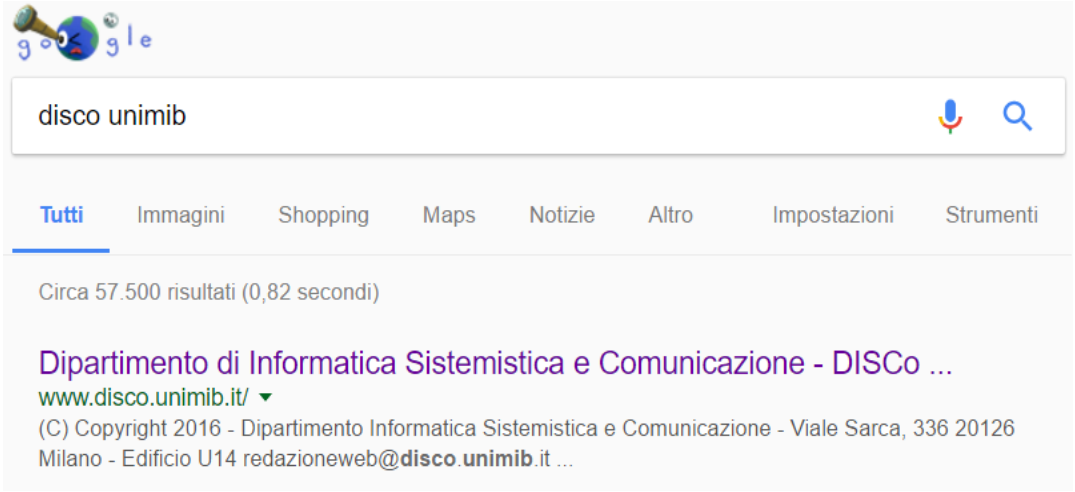
Sistemi per la Raccomandazione di Informazioni

Richiedono PROFILI UTENTE (assenza di QUERY, tecnologia ***push***)

I Motori di Ricerca sul Web

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

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

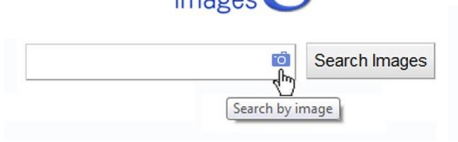


YouTube

Spotify

SHAZAM

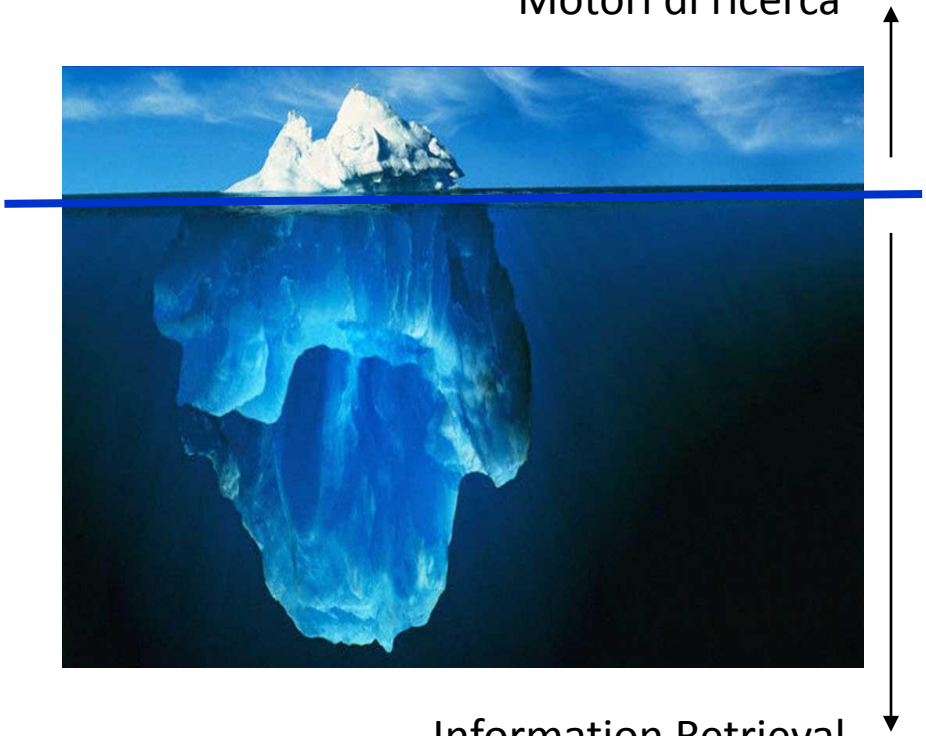
Google images



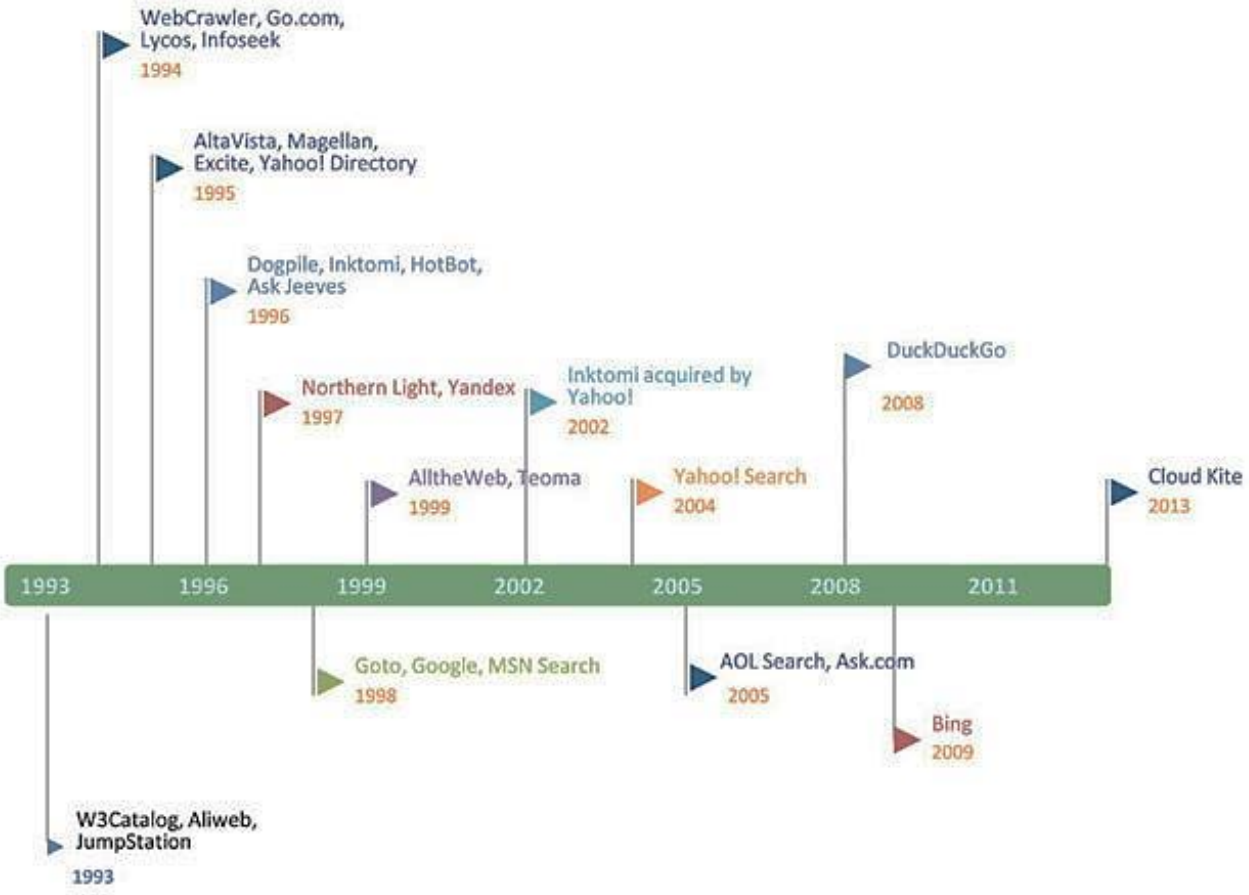
Le «radici» dei Motori di Ricerca

I motori di ricerca sono sistemi software usati su Web in modo intensivo.

Ciò che molti non sanno è che essi rappresentano la punta dell'iceberg dell'**Information Retrieval**, una disciplina fondata alla fine degli anni sessanta.



I Motori di Ricerca sul Web



Cosa deve fare un Motore di Ricerca?

L'obiettivo di un motore di ricerca su Web è reperire tutte le pagine Web utili per l'utente che ha formulato la query, possibilmente non presentando le pagine non utili. Compito **difficile** perché pervaso da soggettività.

Soggettività!!!

Necessità Informative

Query

Documenti reperiti

Informazioni che il sistema **stima utili** per l'utente

Documenti

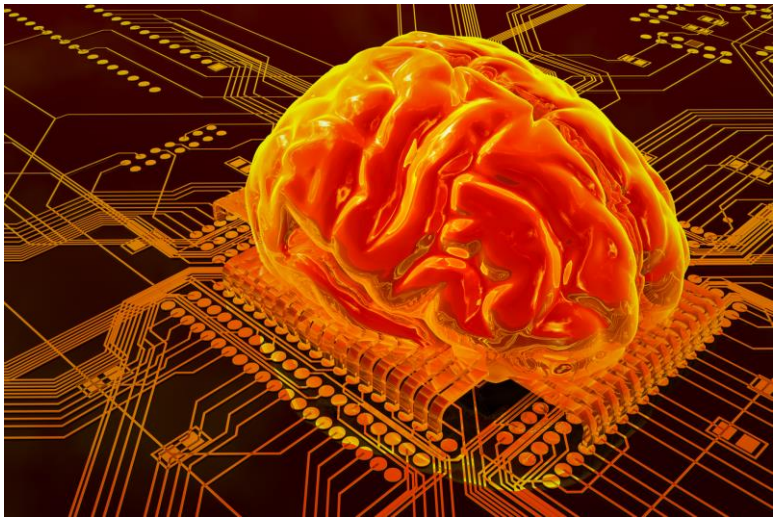
Sistema di Information Retrieval

I Motori di Ricerca sul Web

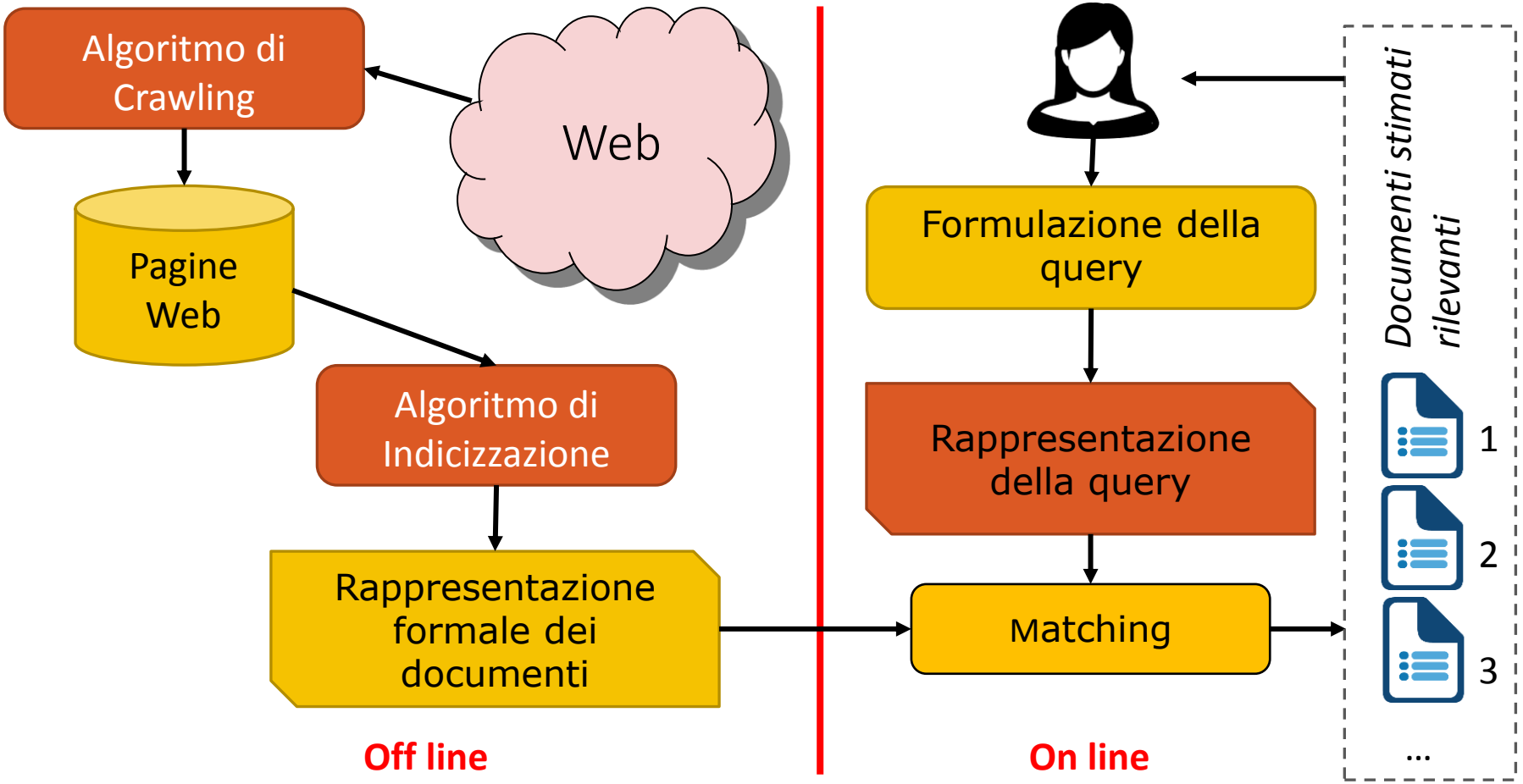
Un motore di ricerca offre una soluzione a un **problema decisionale**: come identificare e stimare **l'utilità** delle pagine Web che soddisfano le preferenze dell'utente? Occorre:

- interpretare il contenuto di: testi, immagini, video, audio
- interpretare le esigenze dell'utente → **oltre la query**!

Un motore di ricerca è un sistema software complesso basato sulla **definizione di modelli**



Architettura di un Motore di Ricerca



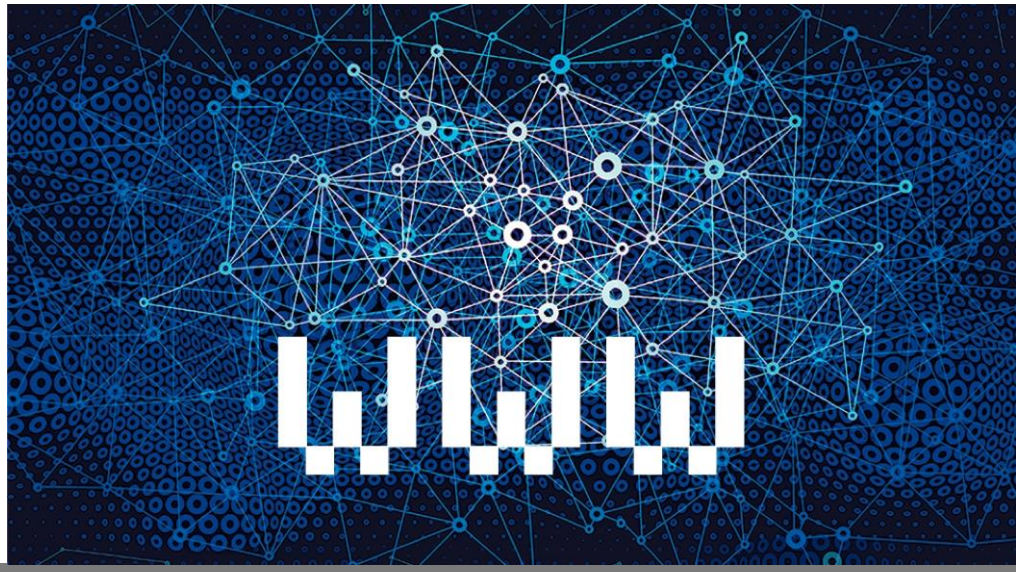
Come viene rappresentato il Web?

Come un **Grafo Orientato**

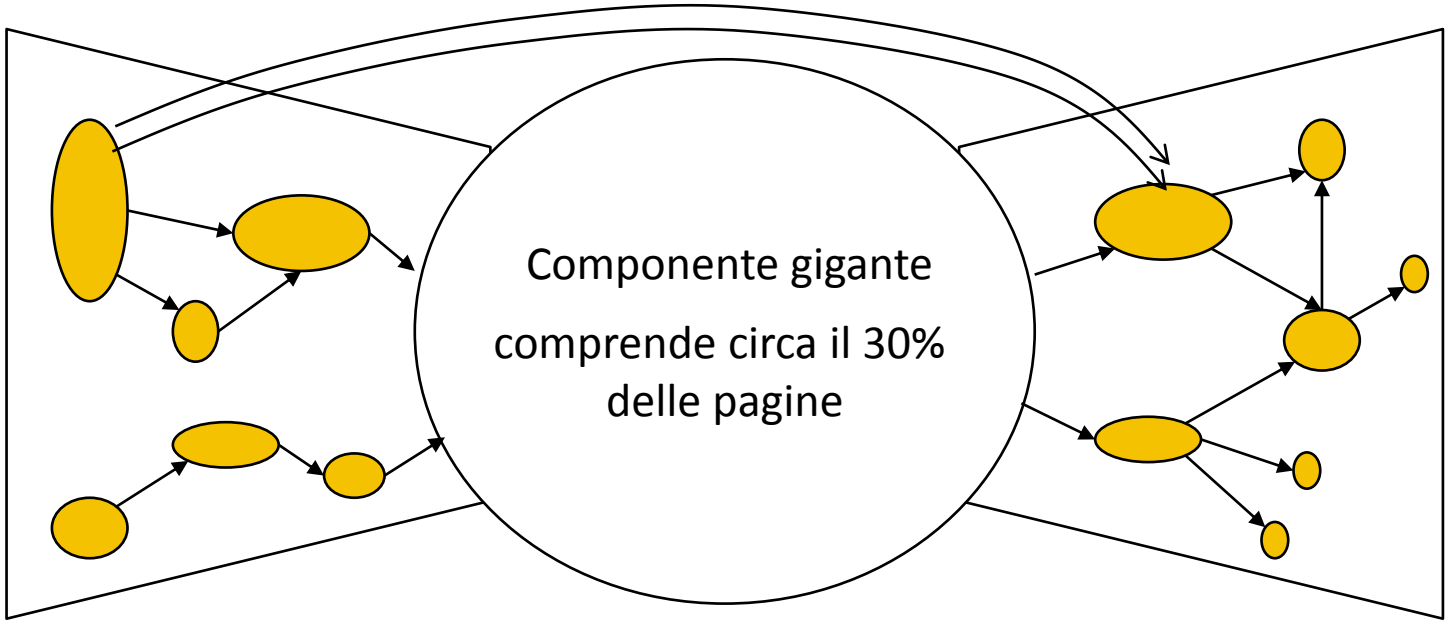
Le dimensioni del Web sono difficili da valutare, il grafo è enorme:

- numero di nodi (= pagine): circa 5 miliardi (escludendo le pagine non accessibili);
- numero di archi: circa 1200 miliardi;
- numero di host: circa 3 miliardi;
- numero di utenti: oltre 3 miliardi.

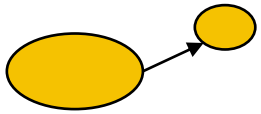
- Fonte: WorldWideWebSize.com



La struttura «a cravattino» del Web



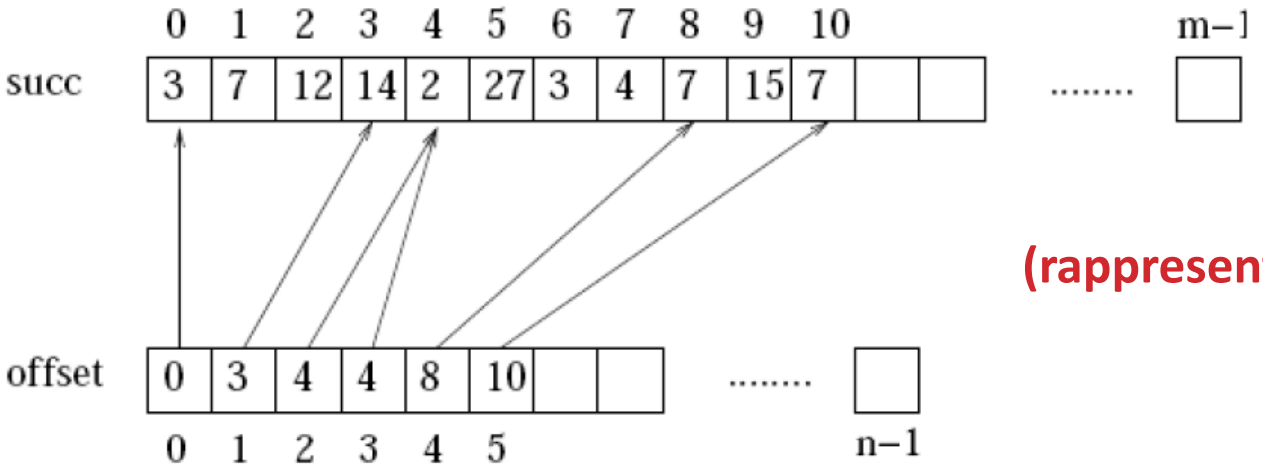
Componenti sorgente (circa 24%)



Componenti «pozzo» circa 24%:

Componenti «isolate» e tentacoli

Come viene rappresentato il Web?



Mediante vettori (rappresentazione semplificata)

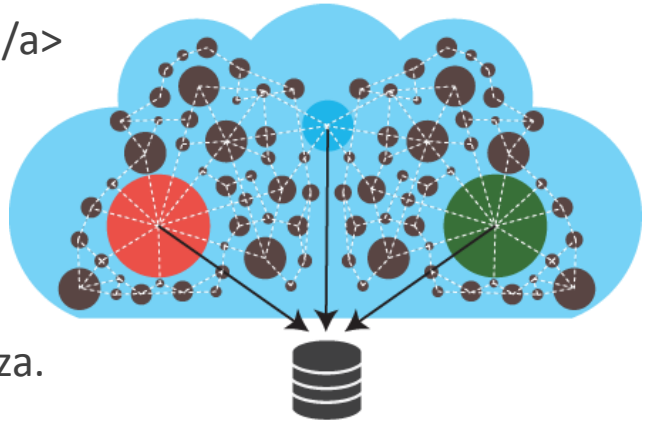
Il vettore degli offset dice da che indice del vettore dei successori partono i successori di un certo nodo.

Contiene implicitamente l'indicazione di quale sia il grado (positivo) del nodo.

OGNI MOTORE DI RICERCA COSTRUISCE LA PROPRIA RAPPRESENTAZIONE DEL WEB!

Il processo di «Crawling»

1. Inizializza una coda di pagine con alcuni URL noti (popolari o inviati da utenti):
   - e.g http://www.unimib.it
2. Seleziona un indirizzo dalla coda.
3. Seleziona la pagina.
4. Cerca nella pagina indirizzi di altri URL
   - Ad esempio publications
5. Scarta gli URL che:
   - non possono essere analizzati es. .exe , …
   - sono già stati visitati.
6. Aggiunge gli URL alla coda:
   - utilizza una strategia di visita in profondità o in ampiezza.
7. se non è scaduto il tempo torna al punto 2.



La valutazione di una query

La valutazione di una query viene effettuata dalla componente di ***Matching***, che per ogni documento calcola un **valore numerico** che rappresenta la **stima di utilità** (rilevanza) del documento rispetto alla query.

Tale valutazione prende in considerazione di versi aspetti (**dimensioni**) che possono rappresentare l'utilità di una pagina per l'utente:

1) il motore di ricerca deve identificare le pagine Web che trattino gli argomenti specificati dalle parole nella query → valutazione della ***pertinenza***

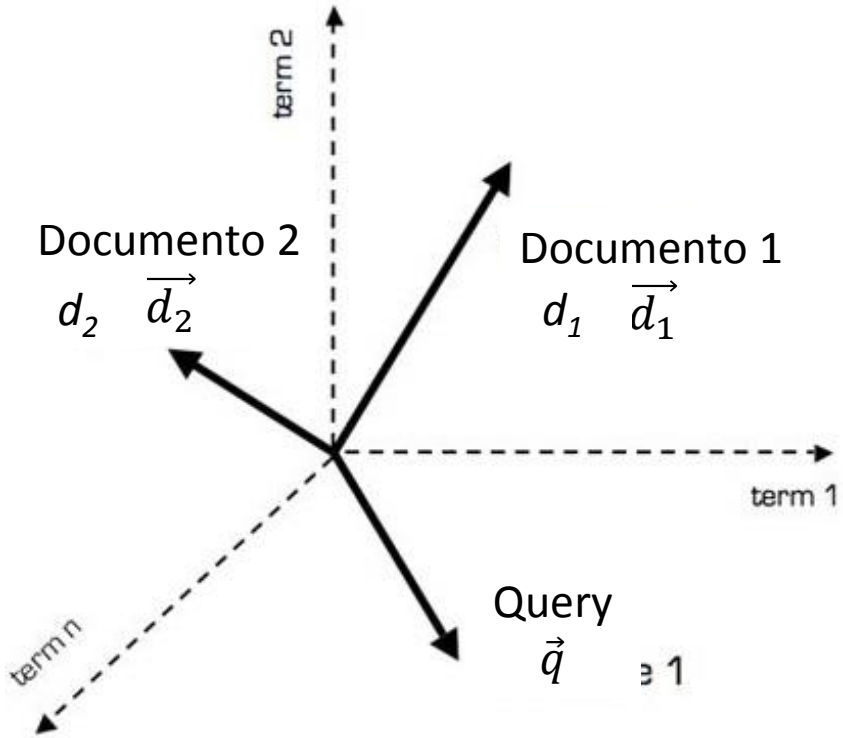
2) Le pagine più ***popolari*** sono più rilevanti → ***analisi dei link*** (introdotta da Google nel 1998).

…

Ad ogni dimensione corrisponde un **modello**

Esempio di valutazione della pertinenza: il modello vettoriale

I metodi con modello vettoriale sono tra i più utilizzati ed efficaci



- Documenti e query rappresentati come vettori
- Angolo tra i vettori come misura Di similarità

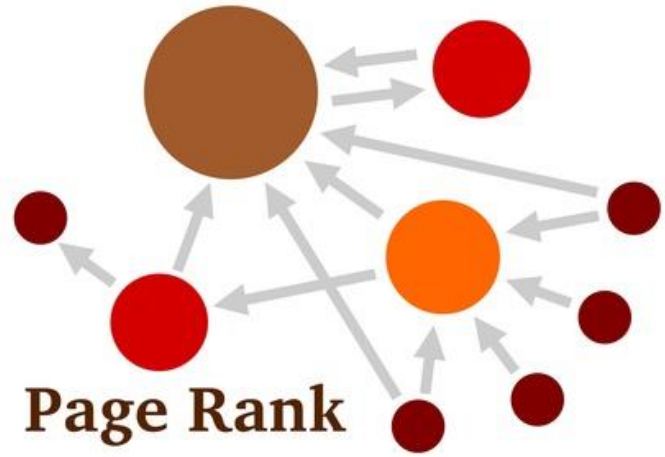
Esempio: l'algoritmo di PageRank

È stato definito da **Sergey Brin** e da **Larry Page**, i fondatori di Google.

Utilizza la struttura del Web per stimare la **popolarità** delle pagine. Le pagine popolari hanno più probabilità di contenere informazioni rilevanti rispetto alle pagine non popolari.

Simula una *navigazione* del Web da pagina in pagina; il PageRank è un *valore numerico* associato a ogni pagina che esprime la *probabilità* di raggiungere quella pagina.

Tale probabilità dipende dal numero di *in-link* della pagina, dalla popolarità delle pagine che hanno un link ad essa e dal di numero di *out-link* di tali pagine



Il futuro: la ricerca personalizzata

In un negozio, le streghe di Hocus Pocus …

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.

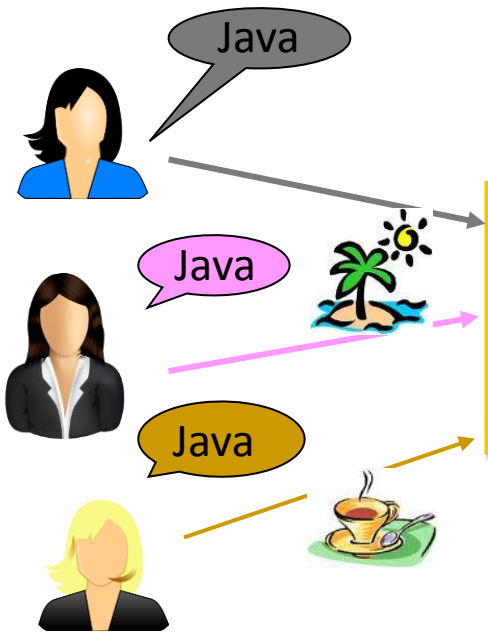
- Evento (obiettivo)
  - matrimonio
  - sport
  - lavoro
  - festa
  - sabba ☺

Contesto:

- contesto dell'utente (preferenze)
  - taglia
  - tessuto
  - gusto
  - prezzo

Approccio *"classico"*

Nell'approccio classico, la stessa query formulata da diversi utenti fornisce sempre la stessa lista di risultati:



Obiettivi della ricerca:

- Trovare un software
- Fare un viaggio
- Comprare un prodotto

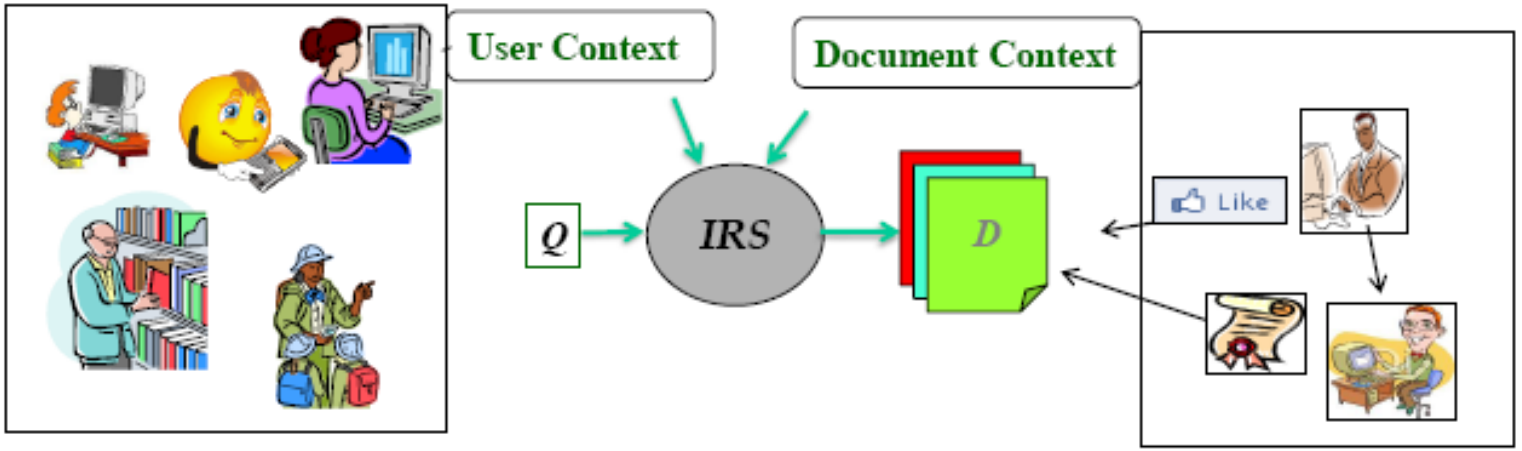
Motore di Ricerca

Java SUN developers
Dev Journal: Top 10 Java guides
Java Coffee
...
Holidays in Java island

One size fits all

Il contesto viene ignorato, eppure è in un contesto ben preciso che si genera una data necessità informativa

Approccio personalizzato



Il contesto viene integrato nel processo di IR

E' in atto una migrazione da *IR query centered* a *IR context-centered*

- *Necessità informativo = query + ?*
- *Rilevanza = utilità basata anche sul contesto o sulla situazione*

I Social Media

Social Media, o media sociali, è un termine che indica tecnologie e pratiche *online* che le persone adottano per condividere contenuti testuali, immagini, video e audio.

Utilizzano le tecnologie del **Web 2.0**, l'insieme di tutte quelle applicazioni online che permettono un elevato livello di interazione tra il sito Web e l'utente come i blog, i forum, le chat, i wiki, le piattaforme di condivisione di media

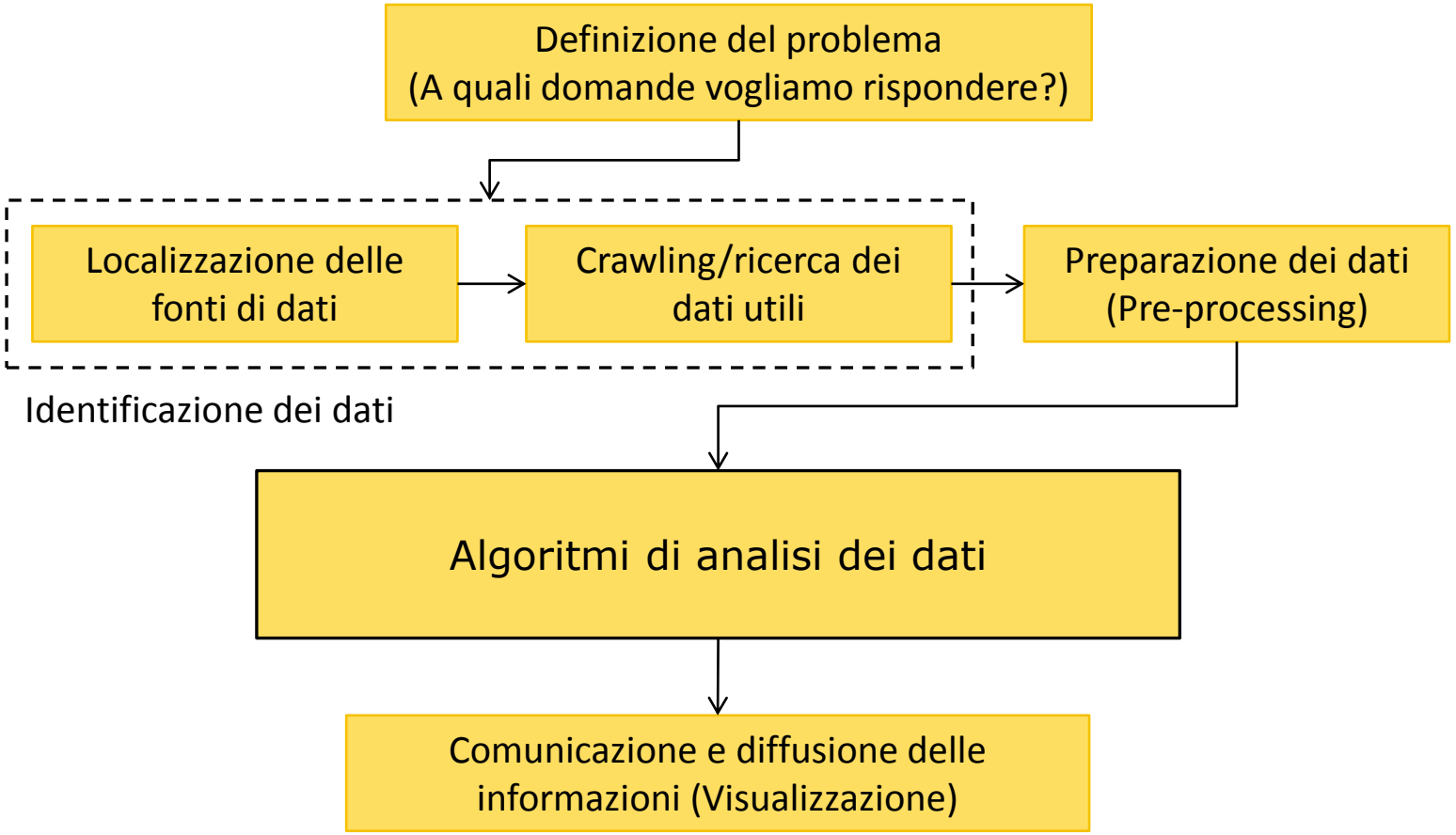
Analisi dei dati sui Social Media

Il processo che analizza i dati sui Social Media prende il nome di **Social Media Analytics** (a volte si usano i temini più generici o più specifici di *Social Media Listening*, *Social Media Monitoring*, o *Social Media Intelligence*)

Riguarda il processo di **raccolta** dei dati sui Social Media e di trasformazione di tali dati in **informazioni** che permettono di prendere ulteriori **decisioni** (in particolare di business)

Lo scopo di tali analisi è di fornire maggiore centralità all'**utente** (o cliente in ambito business)

Schema di analisi dei dati



Analisi sui Social Media

Tipologie popolari di analisi sui Social Media:

- **Analisi delle reti sociali** (Social Network Analysis)

- **Analisi dei sentimenti/opinioni** (Sentiment Analysis / Opinion Mining)
- **Valutazione della credibilità** dell'informazione (Information Credibility Assessment)

Domande?

Grazie per l’attenzione