Pescare pagine nell'oceano del Web

Ad oggi, si stima che siano circa 50 miliardi le pagine web disponibili online in Internet, e il numero cresce costantemente (anche se alcune pagine vecchie scompaiono…)

Come pescare la o le pagine più pertinenti quando facciamo una ricerca online usando alcune parole chiave?

Con un **motore di ricerca.** Il più famoso oggi è sicuramente Google, ma non è stato assolutamente il primo.

Pescare pagine nell'oceano del Web

La rete Internet si sviluppa a partire dalla fine degli anni '60, ma per molto tempo rimane una infrastruttura di collegamento tra computer riservata principalmente agli specialisti informatici.

Il Word Wide Web come lo conosciamo noi, che sfrutta lo strato di comunicazione hardware e software offerto da internet, nasce all'inizio degli anni '90

Pescare pagine nell'oceano del Web

A partire da quel momento incominciano a venire sviluppati sempre più siti web, e nasce quindi l'esigenza di potersi orientare all'interno di un mondo virtuale in cui il numero di siti web e delle pagine di cui sono composti cresce vertiginosamente:

Nasce l'esigenza di un motore di ricerca all'interno del World Wide Web.

Prima dell'era di Google, il più famoso dei motori di ricerca era sicuramente **Altavista** (a partire più o meno dal 1995)

Pescare pagine nell'oceano del Web

A partire dal 1996 Larry Page e Sergey Brin mettono a punto l'idea alla base di Google e la sua implementazione, e nel giro di 10 anni Google si afferma come il miglior motore di ricerca disponibile.

Cosa vuol dire *migliore*: che in poco tempo (frazioni di secondo) riesce a restituire, con estrema accuratezza, le pagine più pertinenti rispetto a un qualsiasi insieme di parole chiavi di ricerca

Pescare pagine nell'oceano del Web

Cosa fa un motore di ricerca:

Fase di matching: cerca tutte le pagine web che hanno a che fare con le parole chiave (**keywords**) specificate dall'utente

Fase di ranking: ordina le pagine per grado di pertinenza, in modo da sfrondare le pagine meno significative.

L’indicizzazione delle pagine

L’indice analitico di un qualsiasi libro riporta un elenco di parole che occorrono nel testo, con a fianco il numero delle pagine in cui quella parola occorre.

Ad esempio: "ghepardo, 124, 156" significa che, nel testo, troviamo la parola ghepardo alla pagina 124 e alla pagina 156.

L’indice analitico facilita enormemente la capacità di trovare velocemente le pagine di un testo in cui si parla di un certo argomento (ad esempio: il ghepardo).

L'indicizzazione delle pagine

L'idea di base di un motore di ricerca è simile, ma con due complicazioni in più:

Le pagine Web attualmente disponibili sono infinitamente di più di quelle di un libro
(e oltre tutto il loro numero cresce costantemente)

Quando facciamo una ricerca in rete usiamo spesso un insieme di parole chiave, e vogliamo trovare tutte le pagine che hanno a che fare con tutte le parole usate (ad esempio "*eventi Torino*")

L’indicizzazione delle pagine: il matching

Già, ma le pagine Web sono miliardi! Come si fa a rispondere ad una ricerca in meno di un secondo? Dobbiamo prepararci in anticipo.

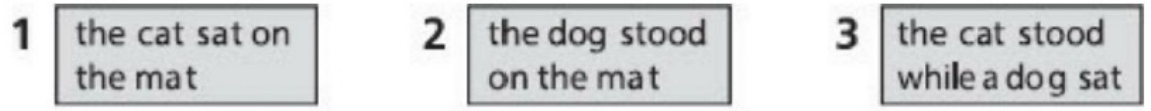
Un programma detto "**Web crawler**" esplora periodicamente tutto il Web e fa una copia delle pagine, che vengono numerate.

Vengono poi estratte e ordinate le parole che occorrono nelle varie pagine trovate dal crawler

E a fianco di ogni parola viene scritto il numero della pagina in cui quella parola occorre

L’indicizzazione delle pagine: il matching

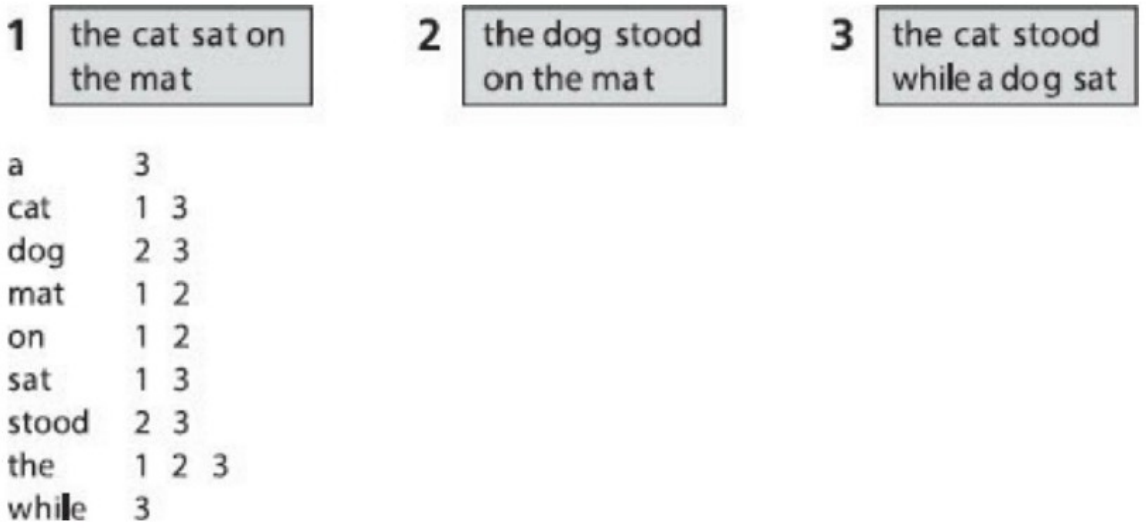
Ad esempio, se tutto il web fosse fatto solo dalle tre pagine indicate qui sotto, con i rispettivi testi, l’indice costruito a partire dalle tre pagine sarebbe formato dall’elenco di parole riportato sotto:
la parola *sat* si trova nella pagine 1 e 3.



a	3
cat	1 3
dog	2 3
mat	1 2
on	1 2
sat	1 3
stood	2 3
the	1 2 3
while	3

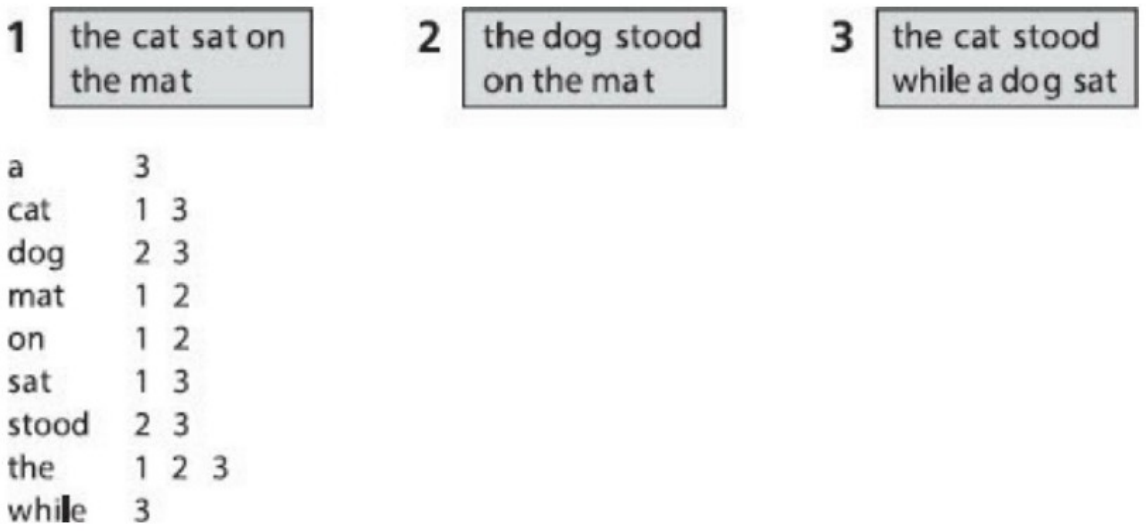
L'indicizzazione delle pagine: il matching

Dunque, per rispondere ad una ricerca che usasse la parola chiave *sat*, basterebbe individuarla nell'elenco e restituire i numeri 1 e 3 (o meglio, l'indirizzo web delle corrispondenti pagine).



L’indicizzazione delle pagine: il matching

Se invece la richiesta fosse relativa a più parole chiave, ad esempio "*cat dog*" allora il motore di ricerca restituirebbe l’insieme delle pagine Web associate ad entrambe le parole chiave, ossia la pagina 3 (o meglio, il suo indirizzo Web)



L'indicizzazione delle pagine: il matching

Nel caso reale, le parole dell'elenco sono milioni (visto che possiamo fare ricerche in qualsiasi lingua) e le pagine web sono miliardi, ma il principio è lo stesso.

In tutto ciò torna ovviamente utile l'enorme capacità di memorizzazione dei nostri computer e la velocità a cui sanno organizzare (ad esempio ordinare) e poi cercare l'informazione rilevante (ad esempio, calcolare l'intersezione di tutte le pagine in cui occorrono le diverse parole chiave)

L’indicizzazione delle pagine: il ranking

Ma c’è un problema: le pagine Web sono miliardi, una qualsiasi ricerca di un insieme di parole chiave potrebbe restituire centinaia o migliaia di pagine potenzialmente significative. Quali lo sono veramente?

Dopo la fase di matching, abbiamo bisogno di una fase di ranking che restituisca solo le pagine effettivamente rilevanti, e a partire da quelle più rilevanti in assoluto

Dobbiamo quindi stabilire un criterio di rilevanza

L’indicizzazione delle pagine: il ranking

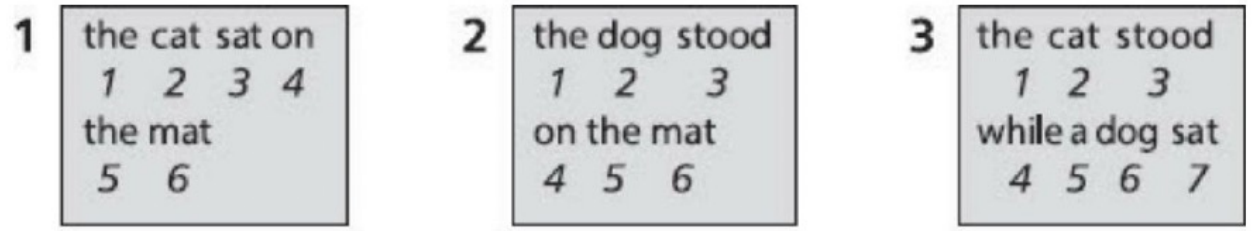
Ad esempio, in una certa pagina Web, una parola chiave è particolarmente rilevante se compare nel titolo della pagina, o almeno nelle prime righe.

Due parole chiave sono reciprocamente molto rilevanti se compaiono l’una vicina all’altra (e nello stesso ordine) piuttosto che se una compare solo all’inizio e una solo alla fine della pagina.

Per tenere conto di ciò, nella fase di indicizzazione si riporta non solo in quale pagina occorre una certa parola, ma anche in quale posizione della pagina.

L'indicizzazione delle pagine: il ranking

Nello stesso esempio di prima, ora vediamo che a ogni parola è associato non solo il numero delle pagine in cui compare, ma anche in quale posizione: *dog* compare in posizione 2 nella pagina 2, e in posizione 6 nella pagina 3.



a	3-5
cat	1-2 3-2
dog	2-2 3-6
mat	1-6 2-6
on	1-4 2-4
sat	1-3 3-7
stood	2-3 3-3
the	1-1 1-5 2-1 2-5 3-1
while	3-4

L’indicizzazione delle pagine: il ranking

Il linguaggio di base con cui sono scritte le pagine Web, l’HTML (Hypertext Markup Language) facilita questo lavoro (che ovviamente è fatto da opportuni programmi), perché permette di individuare facilmente il titolo di una pagina, il suo corpo, e così via.

L’indicizzazione delle pagine: il ranking

Infine, ovviamente le pagine Web non hanno tutte la stessa rilevanza rispetto a un qualsiasi insieme di parole chiave di ricerca: anche se sulla mia pagina Web ho scritto da qualche parte *Elenco degli eventi a Torino,* difficilmente la mia pagina verrà considerata più rilevante della pagina degli eventi a Torino del quotidiano cittadino se cerco "*eventi Torino"*.

La possibilità di sponsorizzare alcune pagine Web rispetto a specifiche parole chiave modifica poi ulteriormente i risultati del processo di ranking.

Quali sono le pagine più importanti: PageRank

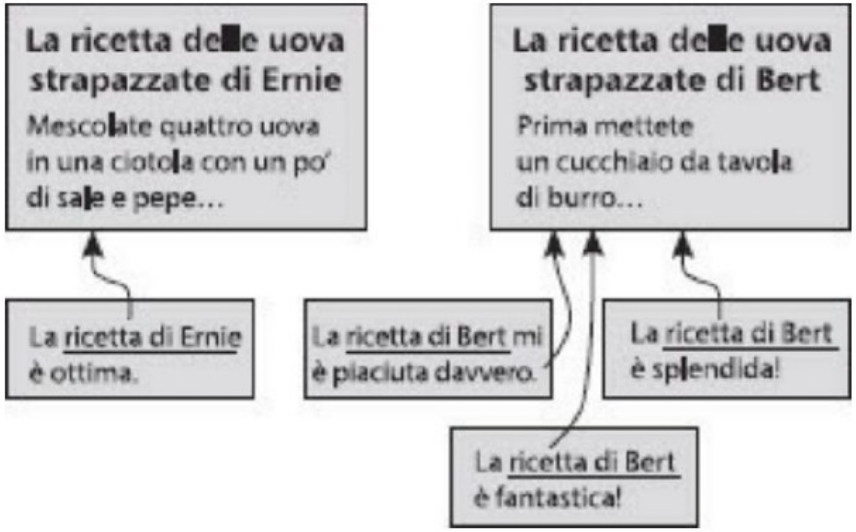
Il suggesso di Google si deve in buona parte all'algoritmo **PageRank** inventato da Page (Omen nomen!) e Brin.

PageRank è particolarmente capace di filtrare le pagine Web trovate nella fase di matching per restituire solo quelle effettivamente rilevanti.

Oltre a utilizzare tutte le tecniche già viste per stabilire la rilevanza di una pagina rispetto a una certa ricerca, l'idea di fondo di PageRank è di cercare di capire quanto una pagina è *popolare*, cioè nota all'interno del Web

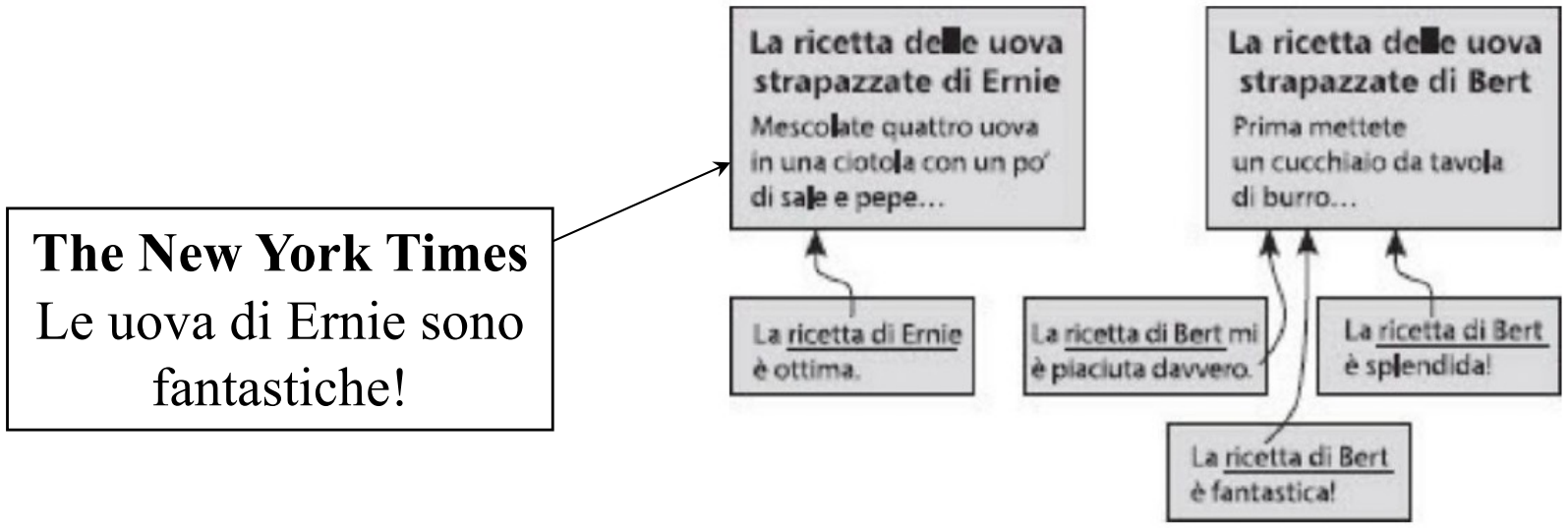
Quali sono le pagine più importanti: PageRank

Nel Web, una pagina (e il sito che la contiene) è tanto più popolare quanto più viene citata da altri siti, ossia da quanti siti contengono un riferimento (un link) a quella pagina.



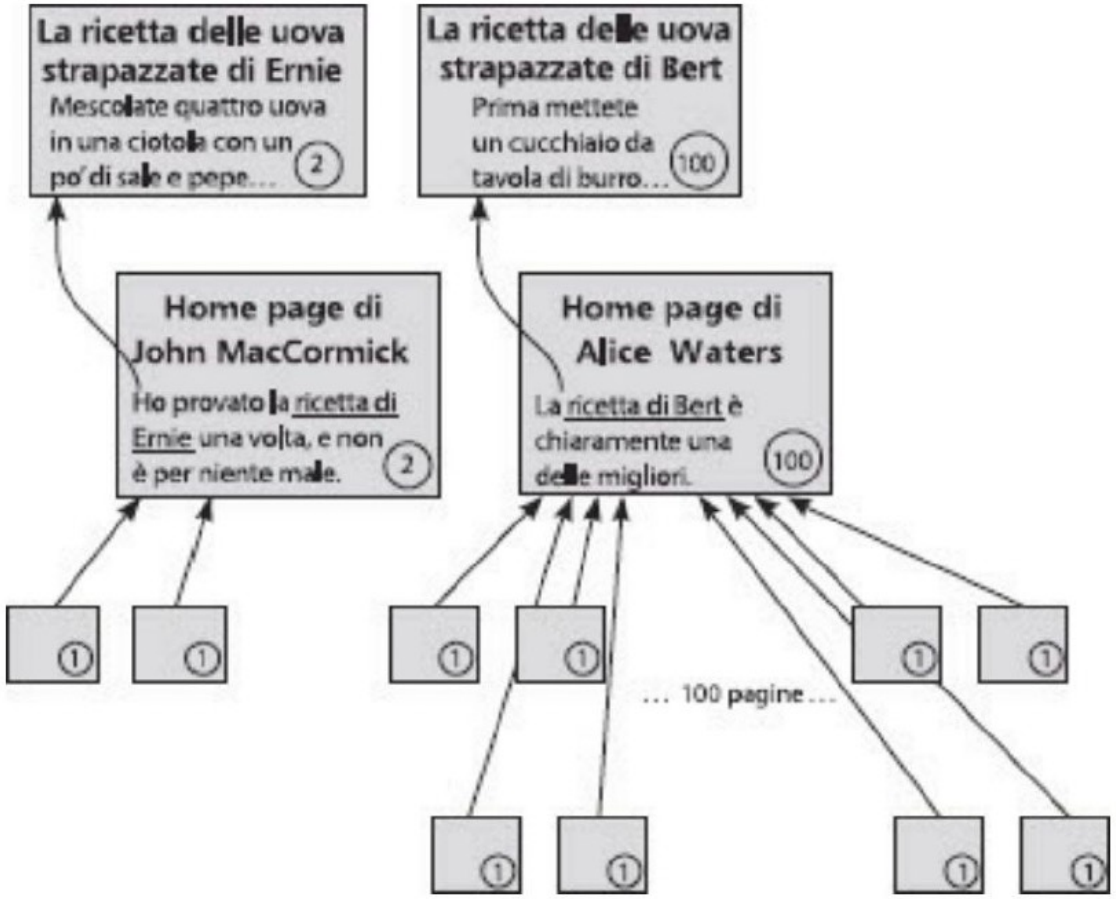
Quali sono le pagine più importanti: PageRank

Anche in questo caso, non tutti i siti sono uguali:
è più importante essere citati (linkati…) da un sito autorevole, prestigioso e noto in tutto il mondo piuttosto che da tanti siti insignificanti e sconosciuti



Quali sono le pagine più importanti: PageRank

Nasce quindi il problema di valutare l'autorevolezza di un sito Web. Il solo contare quanti altri siti citano (hanno un collegamento verso) un certo sito A non è sufficiente, e può produrre errori grossolani:

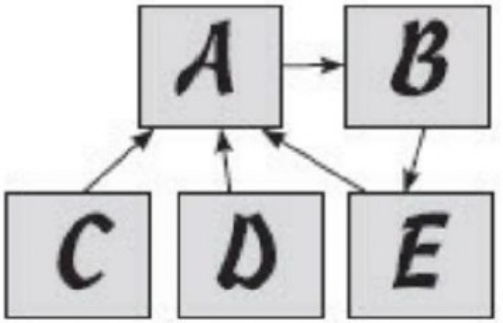


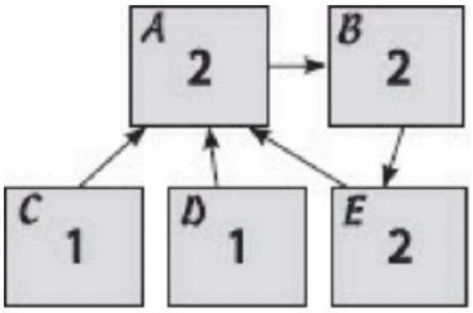
Quali sono le pagine più importanti: PageRank

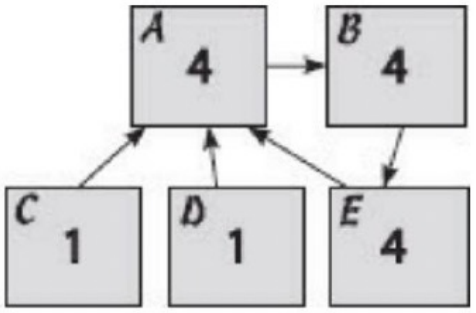
Infatti, nel Web è del tutto normale che un insieme di citazioni (link verso un certo sito) formino un ciclo:

Sito A cita il sito B che cita il sito C che cita il sito A

Un Web crawler vede che B è citato da A:
B = 1. E è citato da B che è
citato da A: dunque E=1+1=2.
Ma A è citato da E che è citato da B,
dunque A=1+2=3. Ma
allora B=3, allora E=3,
allora A=4, allora…





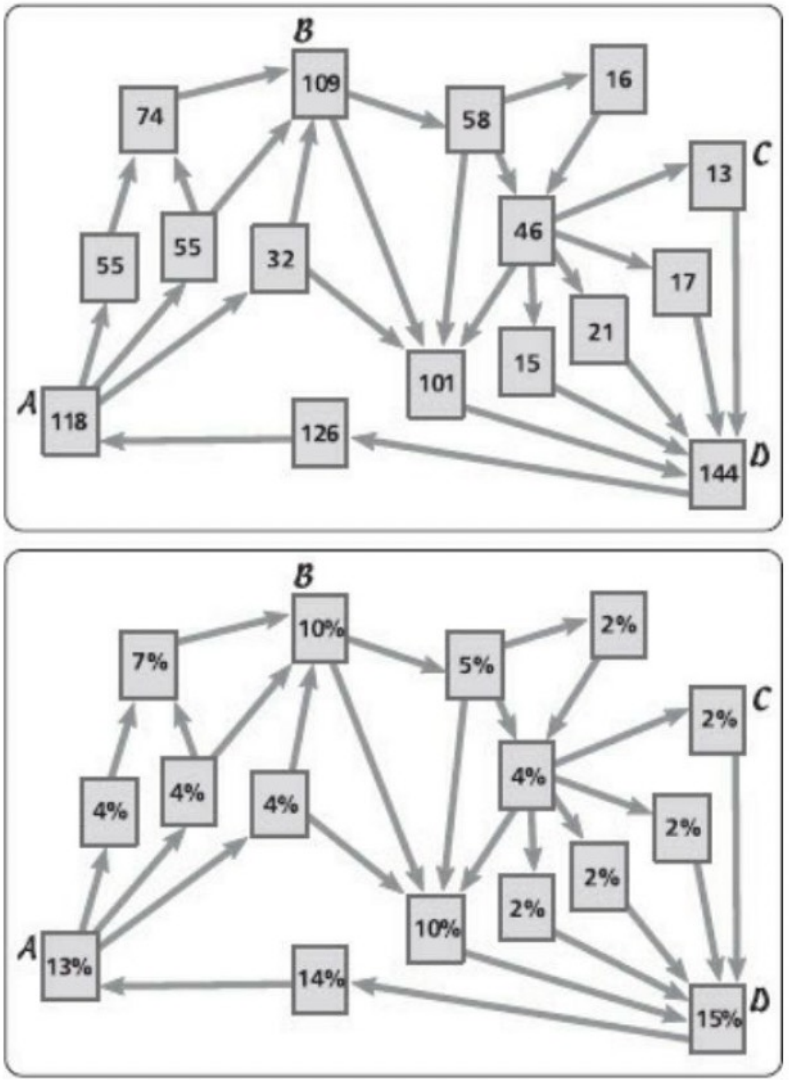


PageRank: il trucco del navigatore casuale

Scegli una pagina a caso, e al suo interno segui un link a caso verso un altro sito. Ripeti milioni di volte.

I siti su cui sei ritornato più volte sono i siti più autorevoli.

Infatti, se ritorni più volte sullo stesso sito, vuol dire che ci sono molti collegamenti verso quel sito, e questo è un indice di popolarità e autorevolezza.



PageRank: il caso reale

Nel caso reale, PageRank non si basa solo sulle tecniche che abbiamo visto, ma usa in totale circa 200 attributi diversi per stabilire il ranking delle pagine restituite in una ricerca

L'algoritmo PageRank viene continuamente aggiornato e modificato, e molte delle sue parti sono coperte da segreto industriale.