CORSO
Sistemi per il recupero delle informazioni

Nicola Drago
nicola.drago@univr.it
I motori di ricerca

Le origini

- Già nel 1990 una grossa mole di dati è disponibile in "INTERNET" il WEB ancora non esiste, ma centinaia di siti FTP mettevano a disposizione miglia di programmi e documenti.

  Si chiama **Archie** ed è stato creato da uno studente canadese della McGill University of Montreal, Alan Emtage, in grado di "indicizzare il contenuto degli archivi FTP"

  Funzionalità ridotte a quelle che siamo abituati a usare oggi come ad esempio era necessario conoscere il nome esatto del file da "ricercare"

Le origini

- **1993**

  Aliweb (Archie Like Indexing on teh Web) viene creato da Martij Koster sulla scia di Archie, e ha come obiettivo la raccolta di tutti i siti presenti sul web. La grande innovazione che porta con sé è la possibilità, per ogni singolo utente, di inviare l'indirizzo del proprio sito, così da includerlo nella lista indicizzata.

Le origini

- **1994**

  Lycos (http://www.lycos.it/), un motore di ricerca che dimostra in poco tempo la sua grande potenza nell'ispezionare il web e il proprio innovativo sistema di ricerca basato sull'importanza attribuita alle ricerche già effettuate in precedenza, nonché sulla possibilità di approssimare le parole usate per l'indagine. Il suo archivio, composto da 394.000 documenti a solo un mese dalla sua nascita, si espanderà sino a contenerne più di 60 milioni nel novembre del 1996. (Lycos si evolverà poi in portale web d'intrattenimento, con fornitura di servizi e-mail e social network.)

Le origini

- **1994**

  Yahoo! (https://it.yahoo.com/), grazie all'impegno di David Filo e Jerry Yang. Ideato **inizialmente come una raccolta delle pagine web preferite dai due informatici** statunitensi, Yahoo! si impone sulla concorrenza grazie ad un'importante novità: ogni sito web che viene indicizzato è dotato di una descrizione. Yahoo! negli anni non solo cresce notevolmente, ma inizia ad inserire al proprio interno siti commerciali, facendo pagare un canone annuo per far parte del suo archivio.

  Attualmente "**un portale**" contenete notizie mail e altro.

  Ora parte di Oath (https://www.oath.com/it/my-data/#meetoath)

Le origini

- **1995**

  AltaVista, dotato di una larghezza di banda quasi illimitata per i tempi e primo a offrire le query in un linguaggio naturale. Diviene uno dei motori di ricerca più popolari tra gli utenti per la sua velocità, poi surclassato da Google e acquistato, nel 2003, da Yahoo!

Le origini

- **1996 - 98**

  SEO Per la prima volta, i webmaster iniziano a “ottimizzare” i siti, ovvero a svolgere tutte quelle attività finalizzate a ottenere la migliore rilevazione, analisi e lettura del sito web da parte dei motori di ricerca.

  Si pubblicano sul web i primi documenti che parlano di analisi e di estrazione dei dati dalle pagine web. Per la prima volta, John Audette e Bruce Clay utilizzano il termine SEO (Search Engine Optimization), presentato ufficialmente nel 1997, riferendosi all’opera di tecnici qualificati che “ottimizzano” le pagine online (struttura, testi, immagini etc) così da essere più facilmente trovati dai motori di ricerca.

Le origini

- **1998**

  Google (https://www.google.it/)nato da Sergey Brin e Larry Page, due studenti di Stanford appassionati di matematica, porta alla creazione del **PageRank**, **un algoritmo di analisi atto alla catalogazione dei siti attribuendovi importanza o meno sulla base del peso assegnato a ogni elemento di un collegamento ipertestuale**. Con il PageRank nasce Google, destinato a diventare il sito e motore di ricerca più visitato.

Le origini

- **2001**

  Una massiccia migrazione degli utenti verso Google che, di fatto, diviene il motore di ricerca più usato dagli internauti. Nello stesso anno, l’introduzione dei primi Social Network segnano una nuova rivoluzione e pongono la necessità di integrare, negli algoritmi di ricerca, i segnali e le tendenze provenienti da essi.

Le origini

- **2009**

  Microsoft concretizza il bagaglio di conoscenze acquisite da precedenti esperienze (MSN Search, Live Search, Windows Live Search) in un nuovo motore di ricerca: Bing. L'anno successivo anche Yahoo! abbraccia la tecnologia di Bing, cominciando ad utilizzare il motore di ricerca di Microsoft nel suo portale.

Le origini

- **2010**

  Google lancia **Google Instant**, ovvero la casella di ricerca in tempo reale, che mostra i risultati considerati più pertinenti dal sistema nel momento stesso in l'utente digita la ricerca

Le origini

- **2011-14**

  In questi ultimi anni gli aggiornamenti dell'algoritmo di ricerca di Google si fanno sempre più intensi, e sono volti da un lato a migliorarne le performance semantiche, dall'altro penalizzare comportamenti scorretti nella rete.

  Ecco i principali aggiornamenti, dai nomi più strani:

  – 2011 – Google Panda: filtro di ricerca che penalizza i siti con contenuti di bassa qualità.

  – 2012 – Google Penguin: intercetta i siti che, utilizzando tecniche di spam, hanno generato link in entrata non in maniera naturale, ma a pagamento.

Le origini

- **2011-14**
  – 2013 – Google Hummingbird: algoritmo studiato per comprendere il significato semantico di una ricerca piuttosto che basarsi sulle singole parole.
  – 2014 – PigeonUpdate: i risultati di ricerca vengono geolocalizzati in base alla zona in cui la ricerca stessa viene eseguita.

Riassumendo

In generale per "motori di ricerca" si intendono tutte le risorse di catalogazione di siti internet disponibili sul web. Usato in questo modo pero' il termine e' improprio in quanto in realta' il mondo dei motori di ricerca e' contraddistinto dalla presenza di due generi diversi fra loro: i "motori di ricerca" veri e propri ed le "directories".

- Quali sono le differenze?
- Motore di ricerca = spider
- Indicizzazione con mezzi automatizzati (robot)

Riassumendo

Directory = persona fisica

- Compilazione di un form che deve essere giudicato da persona preposta

Riassumendo

Analizziamo quanto detto:

Motori di ricerca: i motori di ricerca funzionano attraverso un programma che e' diverso fra motore e motore e che tecnicamente si chiama "spider" o "robot" o "crawler". Non c'e' dunque una persona fisica che guarda il sito per giudicarlo ai fini del posizionamento; c'e' invece un programma (software) che dietro richiesta del webmaster che richiede l'indicizzazione del sito, visita le pagine del sito scandagliando tutti i contenuti ed osservando la tecnica delle pagine web. Lo spider controlla il testo contenuto nelle pagine, il codice html, parole, frasi, immagini, tag di ogni tipo ecc. e posiziona il sito sulla base del grado di corrispondenza tra le pagine web e l'algoritmo con cui e' programmato.

Riassumendo

Analizziamo quanto detto:

Ogni directory invece funziona attraverso l'operato di persone fisiche preposte al compito di controllare i siti web di cui viene richiesta l'indicizzazione, al fine di darne un giudizio. Ai fini del posizionamento nella directory e' determinante cio' che viene scritto nei moduli per la registrazione in quanto la directory non registra tutti i contenuti del sito ma solo una sua pagina (solitamente la home page) ed i dati immessi nel modulo di registrazione.

Riassumendo

Il Page Rank?

L'algoritmo reale del Page Rank non è conosciuto, e peraltro viene continuamente aggiornato e modificato, ma per capire il senso del suo funzionamento possiamo prendere come riferimento la versione dell'algoritmo iniziale e semplificato, scritto dei fondatori stessi di Google, Larry Page e Sergey Brin.

Ipotizziamo di avere una pagina A che riceve link da alcune pagine (P1, P2, P3) e di cui vogliamo calcolare il Page Rank.

Vediamo qui l'equazione, dove

Riassumendo

Il Page Rank?

$$PR(A) = (1-d) + d \left(\frac{PR(P1)}{C(P1)} + \frac{PR(P2)}{C(P2)} + \ldots + \frac{PR(Pn)}{C(Pn)} \right)$$

Pr[A] sta per il Page Rank della pagina A che vogliamo calcolare;

d è il cosiddetto Damping factor: è un fattore definito da Google che indica la probabilità che un visitatore decida di passare ad un'altra pagina (originariamente questo elemento aveva il valore di 0,85, ma Google lo ha più volte modificato, per modificare a sua volta la percentuale di Page Rank che deve passare da una pagina all'altra; alzando il valore d si abbasserà il valore del Page Rank totale della pagina);

Riassumendo

Il Page Rank?

$$PR(A) = (1-d) + d \left(\frac{PR(P1)}{C(P1)} + \frac{PR(P2)}{C(P2)} + \ldots + \frac{PR(Pn)}{C(Pn)} \right)$$

C[P1], C[P2] il numero complessivo di link contenuti nella pagina che offre il link.

Da questo algoritmo, seppur assolutamente semplificato, è possibile capire la filosofia di Google riguardo il Page Rank.

Google, con questo algoritmo, vuole definire quanto una pagina sia popolare nel Web, e la popolarità è determinata essenzialmente dal numero e dall'importanza dei link che puntano ad una pagina.