I segreti di Google & C.

I motori di ricerca? Superficiali, poco efficaci e ignoranti. Ma in futuro saranno come quelli di "Star Trek". Parola del padre di Google.

Immaginate un bibliotecario che non sa una parola di italiano. Che riesce solo a riconoscere 2 parole identiche in una pagina, senza capirne il significato. E si limita a guardare una stanza su 4 della biblioteca, senza scendere negli enormi scantinati... Gli affidereste la ricerca di un libro che vi interessa? Probabilmente no. Eppure, è quello che fate quando digitate una ricerca su Google e sugli altri motori di ricerca.

Per mettere a nudo i limiti del Web di oggi, e capire come sarà fra 20 anni, non c'è bisogno di andare negli Usa: basta entrare all'Università di Padova e fare una chiacchierata con Massimo Marchiori, docente associato di informatica. Ha solo 40 anni, ma è grazie a lui che i motori di ricerca si sono evoluti: nel 1997 ha inventato l'algoritmo che ha permesso la nascita di Google e dei motori di 2ª generazione.

E dire che, l'anno prima, era disperato: non riusciva a vincere i concorsi universitari, spesso chiusi a chi non ha raccomandazioni. Così ha inviato il curriculum negli Usa, al Mit di Boston, dove l'inventore del Web, Tim Berners-Lee, gli ha detto: «La nostra università è al completo, ma per te faremo un'eccezione: benvenuto». Ma Marchiori non è rimasto un "cervello in fuga": dopo aver inventato l'algoritmo di Google (senza guadagnare un €), nel 2000 ha rifiutato offerte da 50 mila dollari al mese, per tornare a insegnare in Italia, a 970 € al mese: «I soldi non sono tutto. Preferisco formare i giovani nel nostro Paese e fare ricerca in libertà». Gli studenti hanno ricambiato la sua dedizione, intitolandogli un "fan club" 1 su Internet.

Come funzionavano i motori di ricerca prima di Google?

Cercavano in ogni pagina web la parola inserita nella ricerca, e premiavano le ricorrenze: se una pagina citava più volte quella parola, finiva in cima alla lista dei risultati. Ma era un modo stupido di procedere: i motori – Altavista, Lycos ed Excite – si limitavano a cercare nelle singole pagine, senza indagare le relazioni con altre pagine. E i risultati delle ricerche erano inadeguati: se digitavi *rose* (rosa) ottenevi pagine su monumenti, città, società ma non sul fiore.

Allora che cosa ha fatto?

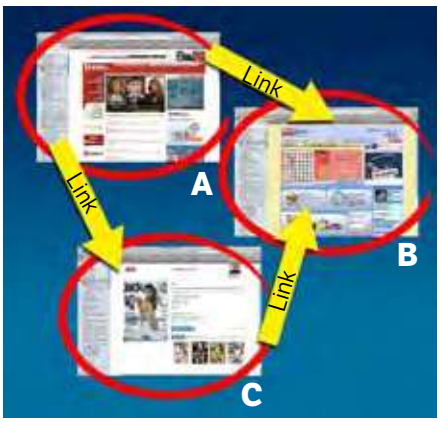
Ho allargato la prospettiva. Se voglio capire una persona puntando il binocolo sul suo viso, ottengo informazioni parziali: ma se, allargando il campo, vedo che stringe la mano a un mafioso, ottengo un'informazione più significativa... Sono le relazioni con le altre pagine Web, i link, a far capire se una pagina è utile. Così ho elaborato un modello matematico che assegnava punteggi decrescenti (da 1 a 0) a ogni sito linkato, e l'ho testato. Dopo aver messo a punto l'algoritmo, battezzato "hypersearch" 2, i risultati delle ricerche miglioravano del 60%. Nel 1997 ho presentato la ricerca alla Conferenza mondiale del Web a Santa Clara (Usa): lì Lawrence Page (uno dei 2 fondatori di Google, *ndr*) mi si attaccò come una »

Massimo Marchiori

Docente di informatica a Padova e membro del W3C, l'ente che definisce gli standard del Web. Nel 2004 ha vinto il premio TR100 della *Technology Review*, assegnato ai migliori 100 ricercatori del mondo.

1 Fan club

http://mmfanclub.altervista.org/
Il fan club di Marchiori, con articoli e video.



↑ **Punti.** Tre siti linkati. Con l'algoritmo di Marchiori il sito A (più utile) ottiene più punti; con l'algoritmo di Google ottiene più punti B (il più popolare).

↙ **Osservati.** Una foto di Google Street View: mostra le strade e chi le frequenta. Solo dopo varie proteste i volti dei passanti sono stati resi anonimi.

I motori riescono a leggere solo il 25% del Web superficiale, che è 1/500 del totale

» cozza per chiedere informazioni, e nel 1998 lanciò Google [3], che si basava proprio su quell'algoritmo. Ma con una correzione: mentre il mio algoritmo premia le pagine più utili, cioè più ricche di link, quello di Google, "pagerank", premia i siti più popolari (*v. sopra*).

Da allora, cosa è cambiato?

L'avvento di Google ha rivoluzionato il Web: per avere più visibilità, i siti hanno inserito molti link, anche in automatico. E questo ha impoverito il valore dei link: mentre 10 anni fa erano una scelta consapevole, oggi si sono inflazionati. Per questo, i motori di ricerca continuano ad aggiustare il tiro con i "tweak", le correzioni all'algoritmo: sono centinaia all'anno. Con queste pezze i motori se la cavano, ma, bene o male, si equivalgono tutti. I motori di oggi sono come la tv generalista: non scontenta nessuno ma non approfondisce nulla.

Quali sono i limiti di Google?

Diversi. Molti pensano che Google scandagli tutti i siti del mondo. Ma non è così: secondo mie stime, oggi Google fa ricerche al massimo sul 25% delle pagine (ma solo sul Web superficiale, 1/500 del Web totale: *v. disegno alla prossima pag.*). Deve operare una scelta, per non allungare i tempi di risposta, e per motivi tecnici: non esiste un computer capace di fare, in un tempo accettabile, una ricerca sull'intero Web.

2 Hypersearch
www.w3.org/People/Massimo/papers/WWW6/
La presentazione di hypersearch.

3 Google
http://infolab.stanford.edu/~backrub/google.html
La presentazione di Google, che cita Marchiori.

66,8
per cento
Le ricerche fatte con Google rispetto ad altri motori.

121
milioni
I domini Web nel mondo: uno ogni 56 persone.

1/4
di secondo
Tempo di risposta a una ricerca su Google.

Con quale criterio Google seleziona dove cercare?

Nessuno lo sa: è un segreto industriale. Se fosse un criterio imparziale e automatico, sarebbe accettabile. Ma c'è anche un intervento umano: tutte le voci di Wikipedia sono considerate "di serie A" perché danno informazioni generali, e quindi finiscono sempre nella top list. Ma le "raccomandazioni" possono essere meno imparziali. Google possiede Doubleclick, la più grande concessionaria pubblicitaria di Internet: chi ci assicura che non premi i propri clienti con una maggior visibilità nelle ricerche? Nel 2003 Google aveva pubblicato il pagerank: per alcuni siti, come Google, era superiore a 10, il che è impossibile a meno di un intervento umano. E nulla vieta di pensare che si possano creare corsie preferenziali in economia o in politica: le censure di Google in Cina ne sono un esempio... Il fatto che Google non sveli i propri criteri, è come le elezioni senza controllori allo spoglio delle schede: è inevitabile che qualcuno possa approfittarsene. Ma questi problemi valgono per tutti i motori di ricerca attuali, non solo per Google.

Come rimediare alla situazione?

L'ideale sarebbe un motore trasparente, "open source", che sveli i criteri adottati per selezionare e premiare le pagine. Ma anche questa soluzione presenta controindicazioni: se si sa come funziona un motore di ricerca, qualcuno ne può approfittare per costruire siti che si adattino a esso, ottenendo così una corsia preferenziale...

E i problemi di privacy?

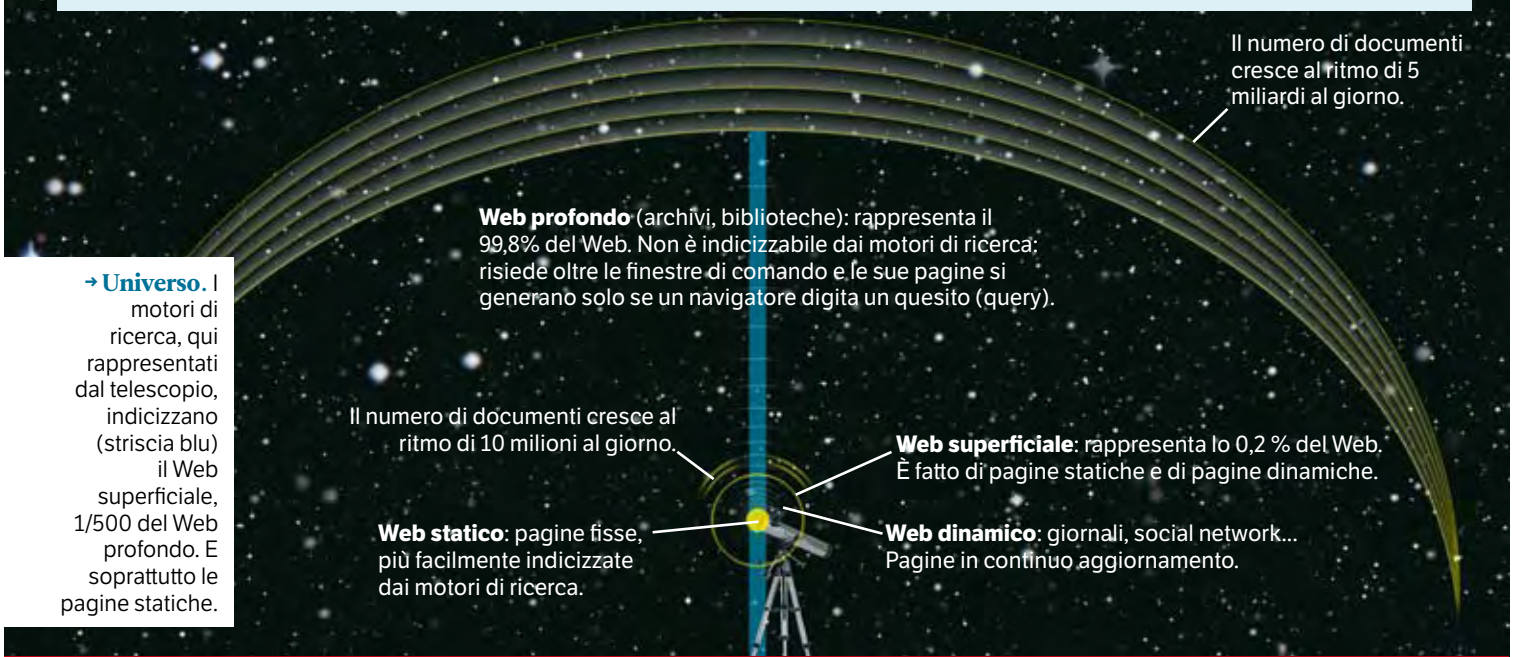
Sono l'altra faccia oscura. Google conservava per sempre i dati di ogni utente sui siti visitati e le ricerche effettuate. Solo dal 2007, dopo le proteste dell'Ue, ha deciso di renderli anonimi entro 2 anni dalla loro registrazione: ma chi controlla che lo faccia davvero? E che cosa fanno gli altri motori come Yahoo e Bing? Tutti hanno un patrimonio pazzesco: possono memorizzare gli acquisti, le preferenze sessuali, culturali, politiche, religiose, di ciascuno di noi. Google legge, seppur in automatico, le nostre mail: se digitiamo "Parigi", ci appariranno pubblicità di viaggi a Parigi. E quest'anno, in Germania, hanno scoperto che le auto di StreetView, che fotografano le strade per Google Maps, hanno registrato "per errore" 600 gigabytes di dati sul traffico delle reti wireless agganciate durante il percorso. E probabilmente questo è successo anche altrove. I motori di ricer- »

Il Web visibile? Solo lo 0,0034%...

Quanta parte di Internet riesce a scandagliare Google? Difficile dirlo: il Web è come il cosmo, in continua espansione. E, come il cosmo, è composto per lo più di "materia oscura", cioè invisibile ai motori di ricerca: il Web profondo (v. disegno sotto).

Silenzio. Solo nel 2000 Michael Bergman, del Nec Research Institute, è riuscito a ricavarne una misura: è ampio almeno 500 volte il Web superficiale, quello indagato dai motori di ricerca. E non interamente: al massimo ne coprono il 16%, hanno accertato nel 1999 Steve Lawrence e Lee Giles del Nec Research Institute. A quell'epoca Google copriva il 7,8%: il risultato migliore era di Northern Light (www.nlsearch.com).

Nel 2005, l'amministratore delegato di Google, Eric Schmidt, disse che il Web contiene 5 milioni di terabytes, di cui sono indicizzabili solo 170: lo 0,0034% del Web. Da allora, Google non rivela più le proprie capacità di calcolo.



→ Universo. I motori di ricerca, qui rappresentati dal telescopio, indicizzano (striscia blu) il Web superficiale, 1/500 del Web profondo. E soprattutto le pagine statiche.

Fonti: Michael Bergman (2000), Michael Lesk (2000), Steve Lawrence-Lee Giles (1999), www.google.com/press/podium/ana.html

I motori sanno tutto di noi: gusti, sesso, politica... Ma nessuno li controlla

» ca hanno un mostruoso database di informazioni: che uso ne fanno? La polizia dovrebbe fare controlli ciclici, per evitare abusi.

Un potere senza limiti...

E potrebbe aumentare. Oggi il Web è liberale: chi digita una ricerca non può scegliere quale tragitto faranno i propri dati. Il pacchetto di bytes corre lungo i cavi telefonici e arriva a un server, che li smista senza distinzioni. Ma nel 2008 Google ha chiesto alle società di telecomunicazioni di avere corsie preferenziali per i propri dati: se ciò avvenisse, le pagine di Google (e dei siti più potenti) si caricherebbero subito, e gli altri più lentamente. E finirebbe la democrazia del Web.

Quale sarà la prossima evoluzione dei motori di ricerca?

Il Web semantico. Oggi i computer non capiscono ciò che fanno: sono come un uomo che non sa il cinese e sa solo notare che lo stesso ideogramma si ripete in varie pagine. Nel Web semantico, invece, i computer saranno capaci di capire che cosa vuole l'utente. Già oggi se inserisco *drink*, i motori fanno la ricerca anche su siti che contengono sinonimi (*cocktail*, *aperitivo*, *whisky*...). Entro il 2020, in modo graduale e invisibile, i motori saranno capaci di capire il senso di frasi complesse come: "Dov'è il ristorante più vicino dove il vino non costa troppo?". L'ideale sarebbe riprodurre nei computer le associazioni del cervello; ma per gestire la mole di dati occorrerebbero supercomputer che comunque sarebbero troppo lenti. È più realistico arrivare al livello di un bambino di 10 anni, molto veloce e con una memoria straordinaria. Un bimbo che sa tutto e risponde a qualsiasi domanda in meno di 1 secondo.

50
petaflops
ovvero 50 miliardi di miliardi di operazioni al secondo: è la capacità di calcolo stimata di Google con circa 500 mila server on line (con processori da 2,5 a 3 GhZ). Jaguar, il più potente supercomputer al mondo, ha una potenza 50 volte più bassa (1,75 petaflops). *Fonte: http://blogs.broughturner.com*

Come immagina il Web del futuro?

Un motore di ricerca stile *Star Trek*, che esegue gli ordini vocali. Capace di adattarsi alle nostre capacità di linguaggio, senza costringerci a studiare un manuale di istruzioni. E con interfacce 3D: manipoleremo le immagini. Il Web, grazie al Gps, ci seguirà ovunque: saprà dove siamo e ci farà interagire con lo spazio. L'informazione e la persona saranno la stessa cosa: potremo chiedere se nei paraggi c'è qualcuno che vuol fare due chiacchiere. Il Web non sarà solo una grande biblioteca, ma una piazza di incontri: i social network lo stanno già dimostrando. Non so se avverrà fra 20 anni, ma tra 50 anni di sicuro. ■

Vito Tartamella