IV. Algoritmi, che passione!

La crescita vertiginosa di Google non ha intaccato la sua fama di motore rapido ed efficiente, affidabile e completo: tutti abbiamo sentito dire che "se non c'è su Google, non esiste!" e che "Google è più veloce". Alla base di questo successo, oltre agli elementi che abbiamo analizzato finora, si trova l'algoritmo di PageRank, già citato in apertura, che guida lo spider di Google alla scoperta delle Reti. Vediamo in dettaglio di cosa si tratta e come funziona.

Algoritmi e vita reale

Un algoritmo[64] è un metodo risolutivo applicato a un problema, un procedimento che si compone di passi semplici da eseguire in sequenza per ottenere un dato risultato. Un algoritmo che perviene alla soluzione del problema è detto corretto, e se la soluzione viene ottenuta in tempi brevi è detto efficiente. Esistono molti diversi tipi di algoritmi, impiegati nei campi più disparati delle scienze; non si tratta però di astruse procedure che riguardano un'esigua minoranza di studiosi, bensì di pratiche che influenzano le nostre vite quotidiane molto più di quanto non sembri di primo acchito.

Ad esempio, le tecniche per registrare un programma televisivo utilizzano algoritmi, ma anche i metodi per ordinare un mazzo di carte o per pianificare le soste di un viaggio particolarmente lungo. In un tempo relativamente prevedibile, realizzando una serie di passi semplici e replicabili in maniera identica, scegliamo più o meno implicitamente gli algoritmi adeguati alla soluzione che stiamo cercando. Semplici significa soprattutto specificati in modo non ambiguo, immediatamente evidenti per chi applicherà l'algoritmo, cioè per il suo esecutore. In questo senso, una ricetta è un algoritmo: "fate bollire tre litri d'acqua in una pentola, salate e gettate cinquecento grammi di riso, scolate dopo dodici minuti, aggiungete spezie a volontà" è una descrizione di passi semplici e non ambigui, se il destinatario della ricetta è in grado di disambiguare passaggi come "salare", oppure "aggiungete spezie a volontà".

Gli algoritmi non sono necessariamente metodi per raggiungere una soluzione nel minor tempo possibile. Infatti ne esistono alcuni che si occupano di ottenere soluzioni accettabili senza preoccuparsi del fattore tempo; altri ancora permettono di raggiungere un risultato nel minor numero di passaggi, oppure hanno come priorità il risparmio di risorse[65].

Importa qui sottolineare, al di là di qualsiasi approfondimento specialistico, la natura pratica, applicativa degli algoritmi. Gli algoritmi riguardano tutti noi perché sono pratiche concrete per raggiungere un dato obiettivo. In campo informatico vengono utilizzati per risolvere problemi ricorrenti nella programmazione dei software, nella progettazione delle reti e nella costruzione di

[64] "Algoritmo: insieme di regole o direttive atte a fornire una risposta specifica a uno o più dati in input". Per una prima introduzione, http://it.wikipedia.org/wiki/Algoritmo. Il termine algoritmo deriva dal nome di "al-Khwarizmi", importante matematico arabo del nono secolo. Muhammad ibn Musa al-Khwarizmi ha introdotto l'uso dei numeri arabi nella matematica: la sua opera "(Libro) di al-Khwarizmi sui numeri indiani" fu tradotta in latino come "Algorismi de numero Indorum". Su questo testo l'Europa intera imparò ad usare il sistema di notazione decimale posizionale ancora oggi in vigore; la numerazione romana venne progressivamente abbandonata. Le procedure che permettevano di effettuare calcoli in notazione decimale divennero così note come "Algorismi" o "Algoritmi" e più tardi lo stesso termine fu applicato in generale alle procedure di calcolo necessarie per ottenere un determinato risultato.

[65] Il metodo migliore per raggiungere Parigi può essere quello di partire con un volo diretto dall'aeroporto più vicino alla propria città, oppure quello di prendere il primo volo in partenza dallo stesso aeroporto, scendere alla stazione d'arrivo, risalire sul primo aereo, scendere e nuovamente ripartire, continuando finché Parigi non sarà raggiunta. È abbastanza certo che entrambi i metodi ci permetteranno di raggiungere la destinazione: con il primo approccio arriveremo alla meta nel minor tempo possibile e probabilmente con il minor spreco di denaro; il secondo, invece, ci permetterà di apprezzare i principali aeroporti delle capitali europee nei diversi periodi dell'anno. Algoritmi differenti descrivono le diverse possibilità.

apparecchiature hardware. Negli ultimi anni, soprattutto a causa della crescente importanza dei modelli reticolari di analisi e interpretazione della realtà, molti ricercatori hanno focalizzato i loro studi sulle metodologie di costruzione e di percorrenza delle reti e dei dati che ne costituiscono la materia viva. L'economia della ricerca di cui parla John Battelle[66] è resa possibile dal perfezionamento di algoritmi per la ricerca di informazioni, studiati per accrescere le possibilità di reperimento e condivisione dei dati in maniera sempre più efficiente, veloce, affidabile e sicura. Il caso più noto al grande pubblico è il fenomeno del peer-to-peer: invece di creare enormi banche dati a cui è necessario accedere per trovare video, audio, testi, software e ogni genere di informazioni, vengono sviluppati in continuazione algoritmi sempre più ottimizzati per facilitare la creazione di reti altamente decentrate, nelle quali ogni utente si può mettere in contatto direttamente con altri utenti e attuare scambi proficui[67].

La strategia dell'oggettività

L'aumento vertiginoso della qualità e della quantità di banda dei nostri computer, insieme alla costante diminuzione dei costi, ci ha permesso di navigare in internet meglio, per più tempo e più velocemente. Solo vent'anni fa i modem a pochi *baud* (numero di simboli trasmesso al secondo) erano un lusso per pochi, mentre ora anche in Italia la fibra ottica, attraverso cui viaggiano milioni di *byte* al secondo, è una tecnologia accessibile.

Dieci anni fa erano necessarie elevate competenze informatiche per creare contenuti adatti alle reti digitali; ora invece la maggiore facilità di pubblicazione di contenuti sul web, l'onnipresenza della posta elettronica, il miglioramento dei sistemi di scrittura collettiva online, come blog, wiki, portali, mailing list e parallelamente l'abbassamento dei costi di registrazione e manutenzione dei domini e degli spazi Internet favoriscono la trasformazione degli utenti: da semplici fruitori di informazioni messe a disposizione da specialisti dell'IT, essi divengono sempre più creatori di informazioni.

Il miglioramento della connettività procede dunque di pari passo con una crescita esponenziale dei dati immessi in rete e quindi, come già abbiamo avuto modo di notare, implica la pressante necessità di strumenti di ricerca sempre migliori. L'urgenza diffusa a ogni livello di servizi di ricerca attira forzatamente l'interesse di sociologi, informatici, ergonomisti, designer, studiosi della comunicazione in genere. D'altra parte, il diluvio informazionale delle reti globali non è una banale "messa in rete" delle società così come le conosciamo, ma un fenomeno estremamente complesso, che esige interpretazioni non banali. Crediamo pertanto che tale impegno teorico e pratico, non possa essere delegato agli specialisti, ma debba essere frutto di un'elaborazione collettiva.

Infatti se da un lato la costruzione di reti autogestite può essere un'occasione per ampliare e collegare fra loro zone autonome, dall'altro il controllo sociale trova nelle tecnologie dell'informazione uno strumento di repressione formidabile.

La realizzazione di questo secondo scenario, di cui il caso Echelon[68] è solo la manifestazione più

[66] Si veda il capitolo II.

[67] Il peer-to-peer. Generalmente con peer-to-peer (o P2P) si intende una rete di computer o qualsiasi rete che non possieda client o server fissi, ma un numero di nodi equivalenti ("peer" significa, appunto, "pari") che fungono sia da client che da server verso altri nodi della rete. Questo modello di rete è l'antitesi dell'architettura client-server, dove esiste un rapporto gerarchico tra il fornitore di servizio (server) e il ricevente (client). Mediante questa configurazione distribuita, qualsiasi nodo è in grado di avviare o completare una transazione. I nodi equivalenti possono differire nella configurazione locale, nella velocità di elaborazione, nell'ampiezza di banda e nella quantità di dati memorizzati. L'esempio classico di P2P è la rete per la condivisione di file (fonte: http://www2.autistici.org/inventa/doku.php?id=glossario_no-copyright_ecc).

[68] Echelon è il nome di un sistema di sorveglianza e spionaggio elettronico globale messo in opera dagli USA. La giustificazione ufficiale è oggi la lotta al terrorismo: comunicazioni telefoniche, mail e ogni comunicazione digitale viene intercettata e analizzata per fini politici ed economici. Per un'analisi approfondita, si veda: Duncan Campbell, Il mondo sotto sorveglianza - Echelon e lo spionaggio elettronico globale, Elèuthera, Milano, 2003; http://home.hiwaay.net/~pspoole/echres.html raccoglie moltissime risorse disponibili.

clamorosa, appare certamente più probabile alla luce del costante aumento del numero di individui che producono informazioni, contrapposto alla diminuzione continua dei fornitori di strumenti di ricerca. L'accesso alle informazioni prodotte da un numero sempre più imponente di individui è gestito da un pugno di monopolisti che riducono una delicata questione sociale e politica a una gara di marketing senza esclusione di colpi, nella quale l'adozione di un algoritmo migliore risulta essere l'elemento vincente.

Infatti un algoritmo di ricerca è uno strumento tecnico che attiva un meccanismo di marketing estremamente sottile: l'utente si fida del fatto che i risultati non siano filtrati e corrispondano alle preferenze di navigazione che la comunità di utenti genera. In sostanza, si propaga un meccanismo di fiducia nell'oggettività della tecnica (nello specifico, la procedura algoritmica che genera il risultato dell'interrogazione) che viene ritenuta "buona" in quanto non influenzata dalle idiosincrasie e dalle preferenze di individui umani. Le macchine "buone", figlie di una scienza "oggettiva" e di una ricerca "disinteressata", non manipoleranno i risultati, non ci diranno bugie perché non possono mentire e comunque non avrebbero alcun interesse a farlo. La realtà è ben diversa e questa credenza si rivela un'ipotesi demagogica, dietro alla quale le macchine del marketing e del controllo accumulano profitti favolosi.

Il caso di Google è l'esempio lampante di questa "strategia dell'oggettività" legata alla tecnica: infatti il motore di ricerca "buono per motto" sfrutta e traccia interamente e in maniera continuativa i comportamenti degli utenti che utilizzano i suoi servizi, al fine di profilare le loro abitudini e inserire nelle loro attività (navigazione, posta, gestione file, ecc.) pubblicità personalizzate, contestuali, leggere, onnipresenti, e possibilmente in grado di generare *feedback*, in modo che gli utenti siano in grado di fornire nel modo più semplici informazioni utili per i venditori e anzi giungano a migliorare essi stessi i "suggerimenti pubblicitari", esprimendo le proprie preferenze. La richiesta continua dell'opinione degli utenti, oltre a lusingare le persone che si sentono partecipi di una vasta "democrazia elettronica", è in effetti il modo più semplice ed efficace per ottenere informazioni preziose dal punto di vista commerciale sui gusti dei consumatori. Sono le preferenze (e l'inconsapevolezza) degli utenti a far vincere un motore di ricerca sugli altri, poiché un sito molto visitato può modificare i sui contenuti in base a "suggerimenti" commerciali e attivare di conseguenza virtuosi movimenti economici .

Da un punto di vista squisitamente informatico, ai motori di ricerca compete la gestione di quattro elementi: la ricerca di dati nella rete (spider), la memorizzazione delle informazioni in appositi archivi (basi di dati), un valido algoritmo per ordinare i dati secondo le ricerche formulate (interrogazione), e infine lo sviluppo di un'interfaccia capace di soddisfare l'utente; ciascuno dei primi tre aspetti viene curato da un apposito tipo di algoritmo: ricerca, memorizzazione/archiviazione, interrogazione.

La potenza di Google, come di Yahoo! e altri giganti della ricerca in rete, è dunque basata su:

8. "spider", ovvero un software per prelevare contenuti dalle reti;
9. hard-disk di enorme capienza per memorizzare i dati su supporti affidabili e ridondanti, onde evitare qualsiasi perdita accidentale;
10. un sistema rapido per trovare (e ordinare) i risultati di un' interrogazione in base al valore di ranking delle pagine;
11. infine un'interfaccia utente via web (ma non solo: Google Desktop e Google Earth, ad esempio, sono programmi da installare sulla macchina dell'utente) per rispondere alle richieste riguardanti queste informazioni.

Spider, basi di dati e ricerche

Lo spider è un applicativo che, nella maggior parte dei casi, viene sviluppato nei laboratori di

ricerca degli stessi motori di ricerca. Il suo scopo è quello di navigare saltando tra un link e l'altro sulle pagine del web raccogliendo informazioni: formati dei documenti, parole chiave, autori delle pagine, ulteriori links, ecc. Al termine delle sue esplorazioni il software-spider consegna il tutto alla base di dati che archivierà le informazioni. Inoltre lo spider deve preoccuparsi di captare le variazioni di ogni sito e quindi programmare una successiva visita per immagazzinare nuovi dati. In particolare lo spider di Google gestisce due tipologie di scansioni dei siti, una mensile approfondita, *Deep-crawl*, e una giornaliera di aggiornamento, *Fresh-crawl*. In questo modo la base dati di Google viene costantemente aggiornata dallo spider sulle evoluzioni delle reti. Dopo una scansione approfondita Google impiega qualche giorno per aggiornare le varie indicizzazioni e propagare i nuovi risultati in tutti i *datacenter*. Questo lasso di tempo è noto come *Google dance*, (danza di Google): i risultati delle ricerche differiscono anche sensibilmente, poiché fanno riferimento a indici diversi. A partire dal 2003 Google ha modificato le sue metodologie di catalogazione e aggiornamento, limitando drasticamente gli effetti della "danza" e spalmandoli nel tempo; in effetti, ora i risultati delle ricerche variano in modo dinamico e continuativo senza nessuno stravolgimento periodico. In realtà i risultati delle ricerche differiscono anche in base alle precedenti navigazioni degli utenti, che vengono archiviate e utilizzate per "migliorare", nel senso di "semplificare", il reperimento delle informazioni[69].

La sequenza di scelte che l'applicativo compie per indicizzare un sito è la vera potenza dell'algoritmo di Google. Mentre l'algoritmo di base PageRank è depositato sotto brevetto da Stanford, e quindi pubblico, questi ulteriori passaggi algoritmici non sono rilasciati pubblicamente né da Google, né da nessuno dei motori di ricerca attualmente in uso; allo stesso modo non sono pubblici i processi di salvataggio nella base di dati.

In ambito informatico, una base di dati (*database*) è, in sostanza, un archivio digitale; nella sua forma più semplice – e attualmente più diffusa – è rappresentabile sotto forma di una o più tabelle in relazione fra loro che presentano valori in entrata e valori in uscita: si parla allora di database relazionale. Come ogni archivio, una base di dati è organizzata secondo precise regole di stoccaggio, estrazione e continuo miglioramento dei dati stessi (recupero di dati corrotti, correzione di voci duplicate, costante reingegnerizzazione dei processi di acquisizione dei dati, ecc.).

Gli informatici studiano da decenni le metodologie di ricerca, immissione e miglioramento dei dati in database, sperimentando linguaggi di programmazione e approcci differenti (gerarachico, reticolare, relazionale, a oggetti, ecc.). La progettazione di una base di dati è una componente cruciale del processo di sviluppo di un sistema informativo complesso come Google, poiché da essa dipende essenzialmente la sua funzionalità. Per ottenere una rapida estrazione dei dati e, in generale, una gestione efficiente, è quindi fondamentale la corretta individuazione degli scopi del database e, nel caso dei database relazionali, delle tabelle, da definire attraverso i loro campi e le relazioni che le legano. Naturalmente è necessario adottare approssimazioni inevitabili nei passaggi fra le lingue naturali, analogiche, e i dati immessi, digitali, che sono evidentemente discreti: zero o uno, l'informazione è presente oppure no, non esistono vie di mezzo. Il punto dolente è la segretezza di queste metodologie: come avviene in tutti i progetti di sviluppo proprietari, a differenza di quelli liberi, è molto difficile sapere quali strumenti e quali algoritmi siano stati utilizzati.

Attraverso i testi redatti dai centri di ricerca e dalle università è possibile reperire le scarse informazioni rese pubbliche a proposito dei progetti proprietari. Su questi testi si trovano informazioni utili per comprendere la struttura dei computer e la gestione dei dati da parte dei motori di ricerca. Per dare un'idea della potenza di calcolo attualmente disponibile, vengono descritti computer capaci di convertire indirizzi Internet in sequenze univoche di byte utili come indici per i database in 0.5 microsecondi e capaci di eseguire 9000 spider in contemporanea; scendendo nel concreto, si tratta di sistemi in grado di analizzare e immagazzinare circa 50 milioni

[69] Si veda il cap. V.

di nuove pagine al giorno[70].

L'ultimo elemento algoritmico che si cela dietro alla "semplice" facciata di Google è il dispositivo di ricerca, ovvero quel sistema che, data una interrogazione utente, è capace di trovare i risultati più congrui, ordinarli per importanza e ranking, infine inviarli all'interfaccia.

Alcune università e laboratori hanno deciso di rendere pubbliche le loro ricerche in tal senso, in particolare le soluzioni raggiunte e i differenti approcci utilizzati per ottimizzare la velocità di accesso alle informazioni, la complessità dell'ordinamento e la selezione dei parametri di input più interessanti.

I motori di ricerca, infatti, devono essere in grado di fornire risultati ottimali quasi istantaneamente, offrendo nel contempo un ventaglio di possibilità di scelta il più ampio possibile. Google rappresenta senz'altro lo stato dell'arte dei motori di ricerca: simili straordinari risultati si possono ottenere solo grazie all'implementazione di opportuni filtri, come vedremo approfonditamente nel prossimo capitolo.

Per ora è importante sapere che l'esito migliore viene assicurato attraverso il giusto bilanciamento tra potenza di calcolo e qualità dell'algoritmo di ricerca. Ricercare un'informazione tra i *terabyte* (1 TB = 1000 GigaByte) o *petabyte* (1 PB = 1000 TB = 1 milione di GigaByte) necessita l'impiego di straordinari supporti di archiviazione e formidabili sistemi di indicizzazione, con il compito di individuare sia in quale punto dell'enorme archivio si trova l'informazione che calcolare il tempo necessario per prelevarla.

La Rete trabocca di leggende non sempre verificate né verificabili a proposito della capacità computazionale di Google, anche perché l'azienda rivela pochi particolari della propria infrastruttura tecnologica. Alcune fonti parlano di centinaia di migliaia di computer collegati fra loro in migliaia di giganteschi *cluster* che montano apposite distribuzioni GNU/Linux; altre di supercomputer, dispositivi la cui estetica rimanda a scenari fantascientifici: enormi silos super refrigerati nei quali uno o più bracci meccanici spostano alla massima velocità migliaia di dischi rigidi. Entrambe le soluzioni sono plausibili, insieme ad altre ancora, e non sono necessariamente in contraddizione. Di certo, l'estrema scalabilità delle macchine di Google consente prestazioni eccezionali, dal momento che il sistema è "aperto" a continui miglioramenti.

Dalla brand-identity all'interfaccia partecipativa

Ricerca, archiviazione e reperimento dei dati sono procedure estremamente complesse e necessitano, per essere comprese a fondo, conoscenze e approfondimenti che esulano dagli intenti di questo testo. Vedremo più avanti alcuni dettagli del loro funzionamento. Un'attenzione particolare va dedicata all'interfaccia perché mentre le performance dell'algoritmo e l'architettura della base di dati sono elementi strutturali del motore di ricerca che rimangono invisibili all'utente, l'interfaccia è progettata e gestita come immagine di Google stesso.

Per interfaccia intendiamo innanzitutto il "blank box"[71], quello spazio vuoto nel quale si immettono le proprie domande o "intenzioni di ricerca" nel quadro della pagina universale di Google, studiata per risultare accogliente, confortevole, famigliare.

Si tratta di un'impostazione detta universale perché viene declinata in numerose lingue (al momento, oltre 104 fra lingue e dialetti personalizzabili per oltre 113 paesi) e in ognuna di queste presenta un modello di interazione che rimane invariato e che unifica i comportamenti di ricerca in

[70] Si veda ad esempio la documentazione resa pubblica da IBM Almaden Research Center: http://www.almaden.ibm.com/webfountain/publications/

[71] Nel gergo informatico, "black box" si riferisce a una "scatola nera" che riceve input, li elabora in maniera non trasparente per l'utente e restituisce un output. Il concetto di "blank box" ricalca questo metodo, ma in maniera implicita, e perciò ambigua, perché pur trattandosi di uno "spazio pulito, vuoto" (*blank*, appunto) è carico di significati e funzioni di ricerca altamente differenziate.

uno schema unico e omogeneo.

Sulla pagina di Google ci troviamo di fronte un'interfaccia lineare composta da elementi essenziali, ciascuno con una funzione ben precisa e universalmente riconosciuta. Essa è in grado di accettare indicazioni di ricerca di diversa natura e complessità, dall'introduzione di semplici parole chiave (es. "ippolita") a parole composte, che vanno poste tra virgolette (es. "comunità scrivente"), fino a ricerche mirate: ad esempio, le ricerche possono essere limitate a un sito particolare, oppure a una lingua specifica, a pagine provenienti solo da un determinato dominio, o ancora a documenti di un certo formato, e così via, a seconda del grado di raffinatezza che si vuole ottenere. Si tratta cioè di un esempio riuscito d'interfaccia che raggiunge il non semplice obiettivo di associare un significato positivo allo spazio bianco della pagina. L'interfaccia si presenta senza orpelli, quasi vuota, o meglio riempita da un unico elemento "vuoto": il *blank box*, che rassicura l'utente e tende a indurre comportamenti attivi, invece di provocare lo smarrimento dovuto all'assenza di punti di riferimento, o viceversa dalla presenza di input visivi sovrabbondanti. Si evita così la confusione generata dalle pagine troppo piene, quasi fossero affette da una sorta di *horror vacui*, da un'ansia comunicativa che, nel tentativo di attirare l'utente con mille banner, effetti grafici, giochini, ottiene spesso l'effetto contrario.

Non esiste una navigazione vera e propria sulla pagina di Google: le diverse componenti della pagina hanno un significato funzionale, servono per accedere a servizi, non per condurre l'utente in un percorso; il loro utilizzo innesca comportamenti che diventano parte molto rapidamente di una routine di ricerca, al punto da apparire istintivi dopo poco tempo. L'interfaccia del motore di ricerca è studiata in modo che l'utilizzo, la dinamica di funzionamento e le aspettative dell'utente, un utente generico, si ripetano; anzi, anche dopo aver immagazzinato e digerito le "personalizzazioni" dell'utente stesso, le pratiche di ricerca rimangono sostanzialmente identiche, tanto che possiamo parlare di uno strumento "universale".

La disposizione di testi e immagini è lineare e si avvale dell'utilizzo di elementi grafici ricorrenti, ad esempio l'impiego dei colori elementari; le immagini usate sono qualitativamente omogenee. Lo stile di progettazione dell'interfaccia è sobrio, quasi scarno e, a dispetto del design di tendenza delle *brand-identity* (e della *corporate-identity*)[72] orientato alla ricerca di una specificità estetica, fa riferimento a qualità percettive elementari ma molto efficaci nella loro semplicità.

Da questa identificazione visiva immediata deriva una facilità d'uso nettamente superiore rispetto ai motori di ricerca concorrenti. Il livello di ergonomia raggiunto è stupefacente: Google non ha la necessità di mostrarsi come un accentratore di servizi attraverso la propria interfaccia; in altre parole, la sua architettura visiva è quella tipica dei portali multiservizio. Le interfacce dei diversi servizi sono autonome e sostanzialmente indipendenti, caratterizzate tutte dalla presenza della "blank box" e non linkate le une con le altre in maniera diretta. Ad esempio, sono necessari molti passaggi non intuitivi per raggiungere il servizio di code.google.com, pensato per tecnici di vario livello, partendo dal servizio base di ricerca delle immagini, ovvero images.google.com, indirizzato a un pubblico più generico: è necessario scendere "in profondità" nel sito google.com e sapere cosa cercare. Nonostante questa frammentazione, siamo tutti in grado di riconoscere la rete di servizi offerta da Google; inoltre i fruitori sono in grado di utilizzare in maniera combinata e integrata le risorse informative messe a disposizione, sia per coloro che si limitano al semplice uso del browser, sia per Google-dipendenti, i *Google-totally-addicted*[73] che si precipitano entusiasti su ogni nuovo servizio.

---

[72] L'ideazione di una nuova immagine per un prodotto o un servizio è nota come *brand identity*; quando riguarda una società specifica si parla di *corporate identity*. Ormai il concetto di "brand" ha ampiamente superato l'idea di "marchio distintivo", giungendo a configurarsi piuttosto come una "marca" che ha necessità di espansione psichica, territoriale, commerciale. Per una prima introduzione, si veda http://it.wikipedia.org/wiki/Marca

[73] La Google-mania dilaga e genera nuovi linguaggi; per una panoramica, si veda: http://en.wikipedia.org/wiki/Google_(search_engine); un elenco di servizi e strumenti correlati a Google, http://en.wikipedia.org/wiki/List_of_Google_services_and_tools

Questa deterritorializzazione dei servizi genera un peculiare meccanismo relazionale: gli utenti non vengono a conoscenza delle nuove sezioni direttamente da Google, ma dalla rete informale degli utilizzatori, da altri siti sui quali i visitatori espongono i loro gusti e discutono delle loro abitudini. La vasta gamma dei servizi offerta da Google viene automaticamente localizzata dal fruitore stesso nel momento in cui si interessa a un nuovo servizio: ad esempio, per quanto riguarda la zona geografica, viene presentata immediatamente l'interfaccia linguistica appropriata all'utente. D'altra parte, è semplice inquadrare la tipologia di utenti a cui un servizio è indirizzato, e valutare il grado di preparazione tecnica richiesto, o il grado di affinità con gli altri utilizzatori. Il meccanismo di passaparola diventa dunque simile a un "PageRank relazionale".

In prima approssimazione, esistono una dimensione relazionale locale, nella quale il passaparola avviene fra amici e conoscenti, e una dimensione relazionale tipologica, nella quale un certo tipo di utenti, identificabili in base a parametri statistici (età, sesso, impiego, ecc.) utilizza un particolare servizio e mette in moto l'economia relazionale.

I dieci problemi relativi all'usabilità dei siti web, discussi da Jakob Nielsen[74], fra i più noti studiosi di interfacce utente, sembrano non intaccare minimamente il sito di Google che, nonostante sia scritto in linguaggio HTML totalmente fuori standard[75], riesce ad assicurare la piena visibilità su tutti i browser, grafici o testuali che siano.

La pulizia grafica delle pagine viene esaltata da un'ottima gestione visiva degli aspetti commerciali. Nessun link pubblicitario in homepage o nelle pagine di documentazione e informazione: la pubblicità in Google si trova solo tra i risultati delle ricerche, appositamente separata dai risultati proposti ma non estranea agli argomenti ricercati. Si può dire quindi che Google è capace di esprimere, quantomeno circa la disposizione scenica delle sue interfacce, il giusto compromesso tra rispetto degli utenti e necessità di ritorno economico. La pubblicità, principale fonte di introiti di Google, viene progettata e realizzata in modo da non diventare invasiva e non distrarre gli utenti dal loro utilizzo dei servizi.

I link pubblicitari sono sponsorizzati in modo dinamico per seguire il percorso compiuto da un utente all'interno del motore di ricerca e quindi, in seconda istanza, sui siti Internet.

I collegamenti commerciali dunque non sono statici, ma si modificano e accompagnano le ricerche degli utenti; questo è possibile anche grazie ai feed RSS (acronimo di *RDF Site Summary*, o di *Really Simple Syndication*), uno dei formati più utilizzati per la distribuzione di contenuti Web, e in virtù delle diverse sorgenti informative digitali (quotidiani, riviste, agenzie di stampa, ecc.) in grado di modificare dinamicamente l'homepage di Google. Infatti Google mette la sua homepage a disposizione degli utenti registrati, rendendola totalmente configurabile grazie all'aggiunta di feed RSS: è così possibile impostare le previsioni del tempo automatiche per le città che si desidera monitorare, oppure scandagliare l'archivio storico delle ricerche effettuate. Si possono organizzare i segnalibri e gli ultimi messaggi di posta ricevuti, ma anche tenere sotto controllo i file del proprio computer senza soluzione di continuità rispetto ai contenuti web, grazie all'applicativo Google desktop.

Il meccanismo di promozione pubblicitaria, i servizi e i sofisticati meccanismi di profilazione dell'utente sembrano costituire un tutt'uno a livello estetico e contenutistico; dal canto loro, i link

[74] Jakob Nielsen, cinquantenne informatico danese, è una delle voci più autorevoli nel campo dell'usabilità del web. Nielsen è noto, fra l'altro, per le sue critiche all'eccesso di grafica e animazioni (ad esempio Flash) che affliggono molti siti popolari a spese dell'usabilità, pratica dannosa soprattutto per i disabili. Il suo ultimo volume pubblicato è Jakob Nielsen, Marie Tahie, Homepage Usability, Apogeo, Milano, 2002. I dieci problemi dell'usabilità: http://www.shinynews.it/usability/1005-errori.shtml

[75] HTML (acronimo per Hyper Text Mark-Up Language) è un linguaggio usato per descrivere i documenti ipertestuali disponibili nel Web. Non è un linguaggio di programmazione, ma un linguaggio di markup, ossia descrive il contenuto, testuale e non, di una pagina web. Si veda: http://it.wikipedia.org/wiki/HTML, ma soprattutto il sito della W3 sullo standard HTML: http://www.w3.org/MarkUp/

sponsorizzati sarebbero in questo senso nient'altro che semplici suggerimenti, graficamente compatibili e concettualmente allineati con l'operazione di ricerca che si sta compiendo. L'economia di Google è altamente integrata con l'interfaccia, al punto da poter essere esclusa a livello visivo da chi non ne è interessato e sfruttata da chi invece trova interessante i link e i percorsi commerciali proposti.

Anche Yahoo![76] e molti altri motori di ricerca e portali mettono a disposizione strumenti analoghi per la personalizzazione della propria homepage; tuttavia la quantità e la qualità delle offerte di Google, al momento, rimane insuperata. Si tratta di configurazioni piuttosto semplici, ma che richiedono in ogni caso una certa dimestichezza con le interfacce web e un po' di tempo per essere realizzate. In ambito web la soglia di attenzione è notoriamente bassissima, le pagine vengono visualizzate e abbandonate in tempi molto rapidi, dell'ordine di pochi secondi; perciò un utente che investe parecchi minuti, o decine di minuti, opera delle scelte che rivelano molto di sé e delle proprie abitudini di consumatore. Queste informazioni, accuratamente archiviate dalla compagnia di turno (Google o Yahoo! che sia) costituiscono la vera ricchezza prodotta dall'utente stesso, e sono fondamentali per proporre beni e servizi mirati da parte delle aziende sponsor.

La personalizzazione delle pagine rende un sito più amichevole, il sito stesso diventa come uno strumento personale in cui l'utente investe tempo, scegliendo colori, aspetto, contenuti. Un visitatore abituale in grado di configurare la propria pagina iniziale viene cooptato e reso partecipe nella costruzione dell'interfaccia web. Concedere il potere e il controllo su alcune pagine all'utente significa promuoverlo da semplice bersaglio di campagne pubblicitarie a consumatore "intelligente" ed è senz'altro il modo migliore e più sottile per creare fidelizzazione promuovendo l'interazione. SI profilano ambienti dotati di interfacce partecipative e non esclusive per ricevere pubblicità sempre più personalizzate, per entrare tutti insieme nel dorato mondo di Google.

PageRank, o l'autorità assoluta di un mondo chiuso

L'algoritmo che permette a Google di assegnare un valore alle pagine indicizzate dallo spider è noto come PageRank.

Sappiamo già che il funzionamento del PageRank si basa sulla popolarità di una pagina web, calcolata in base al numero di siti che hanno almeno un link puntato a essa. A parità di numero link, due pagine web avranno PageRank diversi in base all'importanza di chi li ha linkati: con questo meccanismo si valuta l'aspetto "qualitativo" dei siti. I link presenti nelle pagine web più linkate otterranno importanza superiore rispetto a quelli presenti nelle pagine meno linkate.

Facciamo un esempio concreto: spesso, controllando le statistiche di accesso relative a un sito, si riscontrano un numero enorme di contatti provenienti da siti pornografici. Questo avviene perché Google attribuisce un ranking dipendente dagli accessi, che a loro volta vengono visualizzati nelle pagine di statistiche pubbliche. Esistono perciò programmi che sfruttano la pervasività di questa logica di connessione e valutazione dei nodi di una rete per innalzare il proprio rank; come spesso accade, i primi sperimentatori sono i siti pornografici (come è stato per le gallerie di immagini su web, o per il commercio online).

In pratica vengono utilizzati alcuni programmi che si occupano di cercare i siti con statistiche di accesso pubbliche; viene quindi effettuato un numero molto elevato di richieste, simulando visite provenienti da un finto link presente in un altro sito, che nella maggior parte dei casi è appunto un sito pornografico. Questo meccanismo di bombardamento fa letteralmente impennare il numero di accessi al sito in questione e di conseguenza le statistiche mostrano incrementi evidenti; in questo modo aumenterà sensibilmente il Google-Ranking del sito e in ultima analisi anche quello del sito pornografico da cui i link sono arrivati: insomma, un guadagno per tutti, almeno a livello di visibilità.

[76] Si veda ad esempio MyYahooo!, my.yahoo.com

Questo tipo di operazione non è illegale: nessuno vieta di fare richieste a un sito Internet; grazie a questa pratica i siti a statistica pubblica ottengono un ranking più elevato. Inoltre, tale meccanismo ci consente di illustrare come anche la magia tecnologica del ranking di Google, ritenuto oggettivo e veritiero, sia legata ai "bassifondi" della rete niente affatto autorevoli e a pratiche di linking piuttosto equivoche.

Altre pratiche non illegali che sfruttano l'approccio all'indicizzazione di Google sono note come SEO (*Search Engine Optimization*); si tratta di un insieme di attività svolte per migliorare il posizionamento di un sito web nei risultati della ricerca. L'offerta di un posto di primo piano avviene spesso attraverso e-mail di spam provenienti da indirizzi improbabili, evidentemente tradotte con programmi automatici, che promettono strabilianti risultati:

"Noi registriamo il Suo sito internet in 910 motori di ricerca, registro e catalogo web. Noi portiamo il Suo sito internet sui primi posti di Google e Yahoo! Provateci! Non si corre nessun rischio. Al posto di 349€ soltanto 299€ (costo unico, senza abbonamento)." Ovviamente Google continua a rivendicare la propria trasparenza: "nessuno può garantire che il vostro sito compaia al primo posto nei risultati di Google"[77].

Dal punto di vista matematico, una conseguenza del PageRank basato sull'analisi dei link è l'integrità della base di dati; ovvero, la determinazione di uno spazio circoscritto, per quanto ampio, nel quale compiere ricerche. Infatti, se le pagine sono valutate e scoperte solo attraverso link ciò significa che non esistono pagine non linkate o isole di documenti slegati dal resto del web: in sostanza, nel mondo di Google esiste sempre un percorso che porta da una pagina a una qualsiasi altra presente nella base di dati, cioè nelle reti indicizzate.

Le ricerche quindi saranno tendenzialmente funzionali, evitando al massimo la possibilità di link rotti (*broken link*) o di informazioni diverse da quelle precedentemente archiviate, presenti nella memoria nascosta (*cache memory*). Il problema è che in questo modo gli utenti sono indotti a credere erroneamente che Internet sia un mondo chiuso, connesso, completo, privo di strade poco illuminate o di percorsi preferenziali, poiché sembrerebbe che, data un'interrogazione, si giunga sempre al risultato "giusto".

Ciò dipende dal fatto che la visione googoliana di Internet scaturisce interamente dai percorsi che lo *spider* compie nel suo rimbalzare da un collegamento all'altro. Se una pagina non è citata da nessun altro sito, allora essa non comparirà mai in nessuna interrogazione compiuta da un utente, perché lo spider non ha mai avuto modo di trovarla, pesarla e valutarla. Tuttavia questo non significa affatto che siano assenti isole di dati, tutt'altro.

Ne sono un esempio i siti dinamici, nei quali le funzionalità offerte si basano totalmente sulle scelte dell'utente. Uno di questi siti è trenitalia.com: compilando l'apposita scheda (*form*), il sito è capace di fornire in tempo reale gli orari dei treni, le coincidenze, i percorsi più veloci per raggiungere una destinazione. Google non è in grado di comprendere le richieste di questo *form* e quindi non indicizza gli orari e i percorsi generati dinamicamente da trenitalia.com. Solo l'intervento umano può superare questo scoglio. L'unica soluzione proposta da Google è di inglobare nella sua interfaccia strumenti di redirezionamento sui siti di compagnie aeree o ferroviarie nel momento in cui l'utente ricerca un percorso, destinazione e arrivo.

L'integrità referenziale proposta dalla base dati di Google deve essere rivista, perché sottintende l'idea di un mondo unico per tutti, chiuso e finito. Al contrario, tracciare un percorso in una rete complessa significa compiere un'esplorazione che determina sempre dei risultati relativi e parziali.

Il sogno di un Google contenitore di tutta Internet è un'idea demagogica particolarmente comoda,

[77] Testo di una email ricevuta su info@ippolita.net nel maggio 2005. La posizione di Google sui SEO: http://www.google.it/intl/it/webmaster/seo.html. Per approfondimenti tecnici, rimandiamo alla guida strategica al posizionamento su Google, http://www.googlerank.com/, di cui è disponibile anche una versione italiana.

utile per sostenere la completezza e l'affidabilità delle informazioni disponibili, insomma tutte le caratteristiche che rendono Google un "servizio unico", un dispensatore di verità. Nelle ricerche della vita quotidiana tale chiusura assoluta è molto utile, perché conduce rapidamente a un risultato; in realtà però ci illude che la libertà consista nell'ottenere una qualità totale. Sappiamo invece che in un sistema reticolare complesso non esistono verità assolute, ma solo autorità distribuite a seconda del percorso che si desidera affrontare, o anche solamente in funzione del tempo che si è disposti a investire nella ricerca. La qualità dipende interamente dalla nostra soggettiva percezione dell'accettabilità del risultato. Le reti che siamo in grado di analizzare, apprezzare e vivere, sono oggetti complessi i cui nodi e collegamenti sono in costante mutamento. Poiché il compito di accettare un elaborato di navigazione relativo a una ricerca dipende in ultima analisi dall'utente, risulta essenziale l'esercizio della capacità critica, la consapevolezza della soggettività del proprio punto di vista. Per generare il percorso che davvero ci interessa analizzare è necessario ipotizzare l'esistenza di una rete finita e limitata, un mondo chiuso solo dalle nostre esigenze personali, sapendo tuttavia che si tratta di una localizzazione soggettiva, non assoluta né costante nel tempo. Esplorare una rete implica la capacità di dividere le reti in sottoreti di analisi e corrisponde alla creazione di piccoli mondi localizzati e temporanei[78].

In pratica nella navigazione quotidiana i collegamenti casuali sono di primaria importanza: la creazione di collegamenti nuovi e inaspettati non può in alcun modo essere prevista dall'analisi degli elementi della rete suggerita dal ranking Google. Questi collegamenti hanno la funzione di "porte dimensionali" e consentono la diminuzione o addirittura l'annullamento delle distanze fra due nodi della rete.

PageRank, o la moneta della scienza

Inoltre, l'algoritmo del PageRank, a differenza di quanto riporta la vulgata, non è una invenzione originale di Google, ma si fonda sulle scoperte matematico-statistiche di Andrej Andreevic Markov, che nei primi anni del XX secolo analizzò i fenomeni statistici su sistemi chiusi, cioè quei sistemi in cui ogni elemento è causa o effetto solo di altri elementi del sistema stesso[79].

Sergey Brin e Larry Page sono sicuramente partiti da questa base teorica, ma i miglioramenti che sono stati apportati non sono stati del tutto resi pubblici, al di là del brevetto depositato da Stanford.

L'esempio migliore per chiarire la morfologia di questo algoritmo è il passa parola fra amici e conoscenti. Nelle relazioni interpersonali più si parla di un dato evento, più questo assume importanza e contemporaneamente diventa parte di un patrimonio comune. Se si limita la diffusione di quel dato evento a una sfera ristretta la sua popolarità sarà minore. Lo stesso vale per gli uomini di spettacolo: più riescono a far parlare di sé maggiore sarà il loro ranking, più saranno conosciuti e più saranno famosi (è per questo che esistono trasmissioni autoreferenziali come "L'Isola dei Famosi"...). Questa stessa logica viene applicata da Google ai dati online.

Google propaganda questo suo metodo in maniera molto convincente, diffondendo l'immagine di Internet come una grande democrazia, poiché l'algoritmo agisce come se i link fossero voti relativi ai siti: poco importa se si linka per dire male o dire bene: l'importante è che se ne parli. La forzatura di questa "democrazia globale" ottenuta attraverso un algoritmo è evidente a chiunque: come se la democrazia dipendesse dalla tecnologia e non dalle pratiche degli individui.

L'origine culturale di questa pratica, come già accennato[80], è derivata dal sistema, estremamente elitario, della revisione dei pari (*peer-rieview*) da parte dei *referees* delle pubblicazioni scientifiche: in questo modo il contributo individuale di ogni ricercatore si inserisce in una rete di rapporti,

[78] Per approfondimenti sul concetto di mondi chiusi localizzati (LCW, *Localized Closed World*), si veda la dispensa sulle reti a cura di Andrea Marchesini: http://www2.autistici.org/bakunin//doc/reti/index.xml
[79] Per un'introduzione sulle catene di Markov si veda: http://en.wikipedia.org/wiki/Markov_chain.
[80] Si veda il cap. III.

verifiche e valutazioni che consente la trasmissione e il controllo dei risultati della ricerca. La democrazia globale di Google si configura insomma come l'applicazione del "metodo scientifico" delle pubblicazioni alla Rete, grazie all'algoritmo di PageRank, una sorta di "*referee* tecnologico" in grado di valutare in maniera oggettiva le informazioni del web, tenendo conto delle preferenze espresse dal "popolo dei navigatori" attraverso i link, e proporle nell'ordine giusto.

Il parallelo è stringente: da un lato, le pubblicazioni scientifiche acquistano peso e autorevolezza in base al loro collocamento nel quadro del loro specifico campo di ricerca; tale collocamento viene ottenuto tramite le citazioni, ovvero i riferimenti alla letteratura. In questo modo la ricerca scientifica garantisce la propria continuità, poiché ogni nuovo articolo non nasce nel vuoto, ma si pone come il "presente" del lungo percorso della tradizione scientifica. Dall'altro lato, i link delle pagine web vengono interpretati dallo spider di Google come "citazioni", che aumentano appunto l'autorevolezza, cioè il ranking, di quella pagina.

L'elitarismo scientifico, base del sentimento di timorato rispetto che incute la "scienza" si basa curiosamente sulla pratica della pubblicazione: del resto, rendere "pubblico" non implica rendere "accessibile" e "comprensibile"[81]. Infatti "le scoperte degli scienziati, teoriche o sperimentali che siano, non sono e non possono essere considerate conoscenza scientifica finché non siano state registrate in modo permanente", come sosteneva negli anni Settanta il sociologo Robert Merton[82]. L'affermazione è forse eccessivamente perentoria (la scienza antica si tramandava in modo tutt'altro che pubblico: si pensi alla scuola pitagorica in Grecia, alla distinzione fra scritti esoterici ed essoterici, ecc.), ma evidenzia correttamente il carattere eminentemente pubblico della conoscenza scientifica moderna. La comunicazione non è quindi un sottoprodotto della ricerca, bensì parte integrante di una forma di sapere i cui presupposti sono il carattere cumulativo e quello cooperativo. La scienza, almeno a partire dal XVI secolo, da una parte è orientata al conseguimento di risultati nuovi, che possano rappresentare un aumento del patrimonio conoscitivo, dall'altra assume come punto di partenza i frutti delle ricerche precedenti. Possiamo abbozzare una storia della comunicazione scientifica che si evolve insieme ai media destinati a supportarla: dalla fitta corrispondenza epistolare fra gli scienziati alla stampa periodica su riviste erudite, fino alla comunicazione digitale. Non a caso i primi nodi di Internet furono centri di ricerca accademica, che avevano la necessità di comunicare e condividere le proprie informazioni.

Tuttavia la mutazione del supporto non ha prodotto un sostanziale cambiamento nel metodo di connessione tipico di questa forma comunicativa, che rimane quello delle citazioni. Descritte come "moneta della scienza", le citazioni sarebbero una sorta di tributo degli scienziati ai loro maestri e ispiratori. Più concretamente, collegano la ricerca presentata con quelle già svolte dallo stesso autore o da altri. Tuttavia è ragionevole assumere che il numero di citazioni ricevute da un determinato lavoro possa rappresentare un'indicazione della sua importanza o almeno del suo impatto sulla comunità scientifica. Negli anni questo sistema è diventato materia di studio specifica: l'analisi bibliometrica è una disciplina che utilizza tecniche matematiche e statistiche per analizzare i modelli di distribuzione dell'informazione, e in particolare delle pubblicazioni. Attualmente la bibliometria, e in particolare il suo più noto indicatore, l'*impact factor*[83], viene comunemente usata come criterio "oggettivo" per valutare la qualità del lavoro scientifico svolto da un singolo ricercatore o da un'istituzione. Un grande archivio per l'analisi bibliometrica è stato messo online

[81] L'impressione che la scienza sia troppo difficile da capire per chiunque non sia uno specialista è socialmente radicata in tutti coloro che, a partire dalla loro vita quotidiana, se ne sentono alieni. Le mura del lavoro tecnico sembrano inviolabili. La comune espressione inglese "non è scienza per razzi" (*it's not rocket science*), di solito sarcastica osservazione fatta a qualcuno che ha insoliti problemi nello svolgimento di compiti facili, è solo un esempio della manifestazione di pubblica riverenza verso l'intensità intellettuale della scienza e la sua separazione dalle comuni attività di ogni giorno. Si veda a tal proposito l'attività di CAE, Critical Art Ensemble, www.critical-art.net

[82] Robert K. Merton, Scienza, tecnologia e società nell'Inghilterra del XVII secolo, Franco Angeli, Milano, 1975.

[83] Eugene Garfield, The Impact Factor, in "Current Contents", n. 37 (25) 1994, pp. 3-8; http://www.isinet.com/isi/hot/essays/journalcitationreports/7.html

nel 1993 proprio a Stanford, la culla di Google. Il progetto SPIRES (Stanford Public Information REtrieval System)[84] nacque nel 1974 dalla raccolta di note bibliografiche sugli articoli di fisica delle alte energie curata dalla biblioteca universitaria di Stanford. Limitatamente al ristretto campo d'analisi (la fisica delle alte energie), SPIRES è un database completo e ad accesso gratuito, che consente ricerche complesse anche sulle citazioni, una palestra che Brin e Page hanno saputo sfruttare al meglio per mettere a punto l'algoritmo di PageRank. Accanto all'algoritmo vi sono poi alcuni accorgimenti che contribuiscono a rendere Google un vero e proprio strumento di mediazione globale del web.

[84] Si veda l'articolo "L'evoluzione delle abitudini di citazione nella comunicazione scientifica primaria. Il caso della fisica delle alte energie", Marco Fabbrichesi, Barbara Montolli; http://jekyll.comm.sissa.it/notizie/lettere02_01.htm