PageRank: la tecnologia che ha lanciato Google

Il computer di Star Trek non sembra tanto interessante. Gli chiedono delle cose a caso, lui ci pensa per un po'. Penso che possiamo fare di meglio.
~ Larry Page (co-fondatore di Google)

Dal punto di vista architettonico, il garage è una struttura normalmente modesta, ma nella Silicon Valley i garage hanno un significato imprenditoriale speciale: molte fra le grandi aziende di tecnologia della "valle del silicio" sono nate, o almeno hanno avuto la loro fase di incubazione, in un garage. E non è una tendenza iniziata con il boom delle dot.com negli anni novanta: oltre 50 anni prima, nel 1939, quando l'economia mondiale ancora cercava di riprendersi dalla Grande depressione, la HewlettPackard ha mosso i suoi primi passi nel garage di Dave Hewlett a Palo Alto in California. Parecchi decenni più tardi, nel 1976, Steve Jobs e Steve Wozniak si sono messi a lavorare nel garage di Jobs a Los Altos, sempre in California, dopo aver fondato la loro azienda, quella Apple che oggi è diventata leggendaria.
(Anche se le leggende popolari vorrebbero che la Apple sia stata fondata nel garage, Jobs e Wozniak in realtà hanno cominciato a lavorare in una camera da letto, ma sono rimasti in fretta a corto di spazio e hanno dovuto trasferirsi nel garage.) Forse ancor più degno di nota dei primi passi di HP e Apple è il lancio di un motore di ricerca, chiamato Google, che era in funzione in un garage di Menlo Park in California, quando l'azienda è stata ufficialmente fondata nel 1998.
In quel momento Google aveva già attivato il suo servizio di ricerca per il Web da oltre un anno, inizialmente dai server della Stanford University, dove entrambi i soci fondatori erano studenti di dottorato. Solo quando l'ampiezza di banda necessaria per la crescente popolarità del servizio divenne eccessiva per Stanford i due studenti, Larry Page e Sergey Brin, spostarono la loro sede operativa nel garage, oggi famoso, di Menlo Park. Devono aver fatto proprio le cose per bene, perché solo tre mesi dopo la costituzione legale della società Google è stato indicato da PC Magazine come uno dei 100 migliori siti del 1998.
E qui comincia davvero la nostra storia: secondo le parole di PC Magazine, Google si era meritato quel posto in classifica per la sua "infallibile capacità di restituire risultati estremamente pertinenti". Ricorderete dal capitolo precedente che i primi motori di ricerca commerciali erano stati lanciati quattro anni prima, nel 1994: come ha fatto Google dal suo garage a compensare questo ritardo di quattro anni, saltando a piè pari davanti ai già diffusi Lycos e AltaVista per quel che riguarda la qualità delle ricerche? La domanda non ammette una risposta semplice, ma uno dei fattori più importanti, in particolare in quei primi tempi, è stato l'algoritmo innovativo utilizzato da Google per ordinare i risultati delle ricerche: un algoritmo noto con il nome di **PageRank**.
È un po' un gioco di parole: è un algoritmo che ordina pagine (*page*, in inglese), ma è anche l'algoritmo di ordinamento di Larry Page, il suo principale inventore. Page e Brin hanno reso pubblico l'algoritmo nel 1998, in un saggio di un convegno accademico, "The Anatomy of a Large-scale Hypertextual Web Search Engine". Come fa pensare già il titolo, il saggio fa molto di più che descrivere PageRank: è a tutti gli effetti una descrizione completa del sistema Google, allo stato in cui era nel 1998. Sepolta fra i particolari tecnici del sistema c'è una descrizione di quella che può ben essere considerata la prima gemma algoritmica destinata a emergere nel ventunesimo secolo: l'algoritmo PageRank. In questo capitolo esploreremo come e perché questo algoritmo è in grado di trovare aghi nei pagliai, fornendo coerentemente i risultati più pertinenti come top hit di una interrogazione di ricerca.

Il trucco del collegamento ipertestuale

Probabilmente sapete già che cos'è un collegamento ipertestuale o *hyperlink*: è un'espressione contenuta in una pagina web che porta a un'altra pagina web quando viene attivata (con un clic del mouse o qualche altro metodo) La maggior parte dei browser mostra i collegamenti ipertestuali in caratteri blu sottolineati, in modo che risultino facilmente visibili.

Quella dei collegamenti ipertestuali è un'idea sorprendentemente vecchia. Nel 1945, più o meno contemporaneamente ai primi sviluppi dei computer elettronici stessi, l'ingegnere americano Vannevar Bush pubblicava un saggio visionario, "As We May Think". In quel saggio di ampia portata, Bush descriveva una serie di potenziali nuove tecnologie, fra cui una macchina che chiamava memex. Un memex avrebbe memorizzato documenti e li avrebbe automaticamente indicizzati, ma avrebbe fatto anche molto di più. Avrebbe consentito "l'indicizzazione associativa ... per cui qualsiasi elemento si può fare in modo che a comando ne selezioni immediatamente e automaticamente un altro", in altre parole, una forma rudimentale di collegamento ipertestuale!

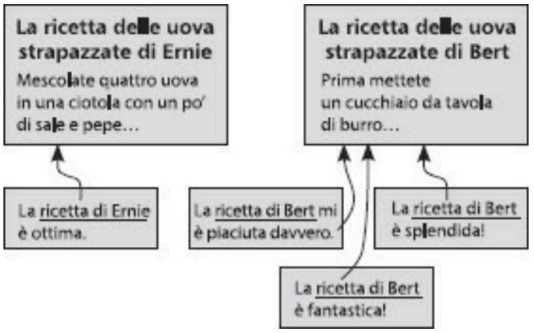

Figura 3.1 *Le basi del trucco del collegamento ipertestuale. Sono mostrate sei pagine web, ciascuna rappresentata da un rettangolo. Due pagine contengono ricette per le uova strapazzate, mentre le altre quattro sono pagine che hanno collegamenti ipertestuali a quelle ricette. Il trucco del collegamento ipertestuale classifica la pagina di Bert al di sopra di quella di Ernie, perché Bert ha tre collegamenti in ingresso mentre Ernie ne ha solo uno.*

I collegamenti ipertestuali hanno fatto molta strada dal 1945. Sono uno degli strumenti più importanti utilizzati dai motori di ricerca per effettuare l'ordinamento e sono fondamentali per la tecnologia PageRank di Google, che ora cominceremo a esplorare seriamente.

Il primo passo per capire PageRank è un'idea semplice, che chiameremo trucco del collegamento ipertestuale. Il modo migliore di spiegarlo è con un esempio. Supponiamo siate interessati a scoprire come si preparano le uova strapazzate e facciate una ricerca nel Web su questo argomento. Qualsiasi ricerca nel Web reale produrrebbe milioni di risultati, ma, per semplicità, supponiamo che escano solo due pagine, una che si chiama "La ricetta delle uova strapazzate di Ernie", mentre l'altra è "La ricetta delle uova strapazzate di Bert". Sono rappresentate nella Figura 3.1, insieme con alcune altre pagine web che hanno collegamenti ipertestuali o alla ricetta di Bert o a quella di Ernie. Per semplicità (ancora), immaginiamo che le quattro pagine mostrate siano le uniche pagine in tutto il Web che hanno un collegamento all'una o all'altra ricetta. I collegamenti ipertestuali sono indicati dal testo sottolineato, le frecce mostrano la destinazione a cui porta il collegamento.

La domanda è: quale dei due risultati deve occupare il primo posto in classifica, la ricetta di Bert o la ricetta di Ernie? Da esseri umani, non faremmo molta fatica a leggere le pagine che si collegano alle due ricette e a farci una nostra valutazione. Sembra che entrambe le ricette siano buone, ma le persone mostrano molto più entusiasmo per la ricetta di Bert che per quella di Ernie. In mancanza di ulteriori informazioni, quindi, probabilmente ha più senso che Bert stia in classifica prima di Ernie.

Purtroppo i computer non sono molto bravi a capire il significato effettivo di una pagina web, perciò un motore di ricerca non può esaminare le quattro pagine che si collegano ai due risultati e valutare con quanto entusiasmo ciascuna ricetta venga consigliata. I computer però eccellono nel contare, perciò un metodo semplice è quello di contare quante pagine si collegano a ciascuna delle ricette (in questo caso sono una per Ernie e tre per Bert) e ordinare le ricette in base al numero dei collegamenti in entrata. Ovviamente questo metodo non ha la stessa precisione che potrebbe avere un essere umano che leggesse tutte le pagine e determinasse l'ordinamento manualmente, ma è comunque una tecnica utile. Statisticamente, in assenza di altre informazioni, il numero dei collegamenti in ingresso può essere un buon indicatore di quanto è probabile che la pagina sia utile, ovvero "autorevole". In questo caso il punteggio è Bert 3, Ernie 1, perciò la pagina di Bert finisce più in alto in classifica quando il motore di ricerca presenta i suoi risultati all'utente.

Figura 3.2 *Le basi del trucco dell'autorevolezza. Ci sono quattro pagine web: due sono ricette per le uova strapazzate, due sono pagine che contengono un collegamento ipertestuale alle ricette. Uno dei collegamenti proviene dall'autore di questo libro (che non è uno chef famoso), uno dalla home page della famosa chef Alice Waters. Il trucco dell'autorevolezza fa salire in classifica la pagina di Bert più di quella di Ernie, perché il collegamento in ingresso alla pagina di Bert ha una maggiore "autorevolezza" di quello in ingresso alla pagina di Ernie.*

Probabilmente avrete già visto qualche possibile problema per il "trucco del collegamento ipertestuale". Una questione ovvia è che a volte i collegamenti sono utilizzati per giudizi negativi anziché positivi. Per esempio, immaginatevi una pagina web con un collegamento alla ricetta di Ernie e che dica "Ho provato la ricetta di Ernie, ed è terribile". Collegamenti come questo, che criticano una pagina anziché consigliarla, fanno effettivamente in modo che il trucco del collegamento ipertestuale attribuisca a certe pagine un posto migliore in classifica di quel che meriterebbero. Ma nella pratica si dà il caso che i collegamenti ipertestuali siano più spesso consigli anziché critiche, perciò il trucco rimane utile nonostante quest'ovvia debolezza.

Il trucco dell'autorevolezza

Forse vi sarete già chiesti perché i collegamenti in ingresso a una pagina debbano essere trattati tutti allo stesso modo. Una raccomandazione da parte di un esperto non vale di più di quella di chi è alle prime armi? Torniamo all'esempio precedente delle uova strapazzate, ma ora con un diverso insieme di collegamenti in ingresso. La Figura 3.2 mostra la nuova situazione: le pagine di Bert e Ernie ora hanno lo stesso numero di collegamenti in ingresso (solo uno), ma quello che va a Ernie parte dalla mia pagina, mentre quello di Bert arriva dalla famosa chef Alice Waters.

In assenza di altre informazioni, quale ricetta sceglieresti? Ovviamente, è meglio scegliere quella consigliata da un famoso chef, anziché quella consigliata dall'autore di un libro di informatica. Questo principio di base è quello che chiameremo il "trucco dell'autorevolezza": i collegamenti che arrivano da pagine con elevata "autorevolezza" contribuiranno a un avanzamento in classifica più di quelli che arrivano da pagine con minore autorevolezza.

Il principio è ottimo, ma in questa forma è del tutto inservibile per i motori di ricerca. Come fa un computer a stabilire automaticamente che l'autorevolezza di Alice Waters quando parla di uova strapazzate è molto maggiore della mia? Ecco un'idea che potrebbe aiutarci: combiniamo il trucco del collegamento ipertestuale con quello dell'autorevolezza. Tutte le pagine inizialmente hanno autorevolezza 1, ma se una pagina ha collegamenti ipertestuali in ingresso, la sua autorevolezza si calcola sommando l'autorevolezza di tutte le pagine che hanno un collegamento ad essa. In altre parole, se le pagine X e Y hanno collegamenti alla pagina Z, l'autorevolezza di Z è la somma dell'autorevolezza di X e di quella di V.
La Figura 3.3 mostra un esempio particolareggiato per il calcolo dell'autorevolezza delle due ricette delle uova strapazzate. I punteggi finali sono racchiusi in un cerchietto. Ci sono due pagine che si collegano alla mia *homepage*; queste pagine a loro volta non hanno collegamenti in ingresso, perciò la loro autorevolezza è 1. L'autorevolezza della mia pagina è la loro somma, quindi è 2. La pagina di Alice Waters ha 100 collegamenti in ingresso, ciascuno dei quali da pagine di autorevolezza 1, perciò ottiene un valore 100. La ricetta di Ernie ha un solo collegamento in ingresso, ma è da una pagina che vale 2, perciò sommando tutti i punteggi relativi ai collegamenti in ingresso {in questo caso ce n'è uno solo), Ernie si guadagna un'autorevolezza di 2. Anche la ricetta di Bert ha un solo collegamento in ingresso, ma vale 100 e così il suo punteggio finale è 100. Dato che 100 è maggiore di 2, la pagina di Bert avrà in classifica una posizione migliore di quella di Ernie.

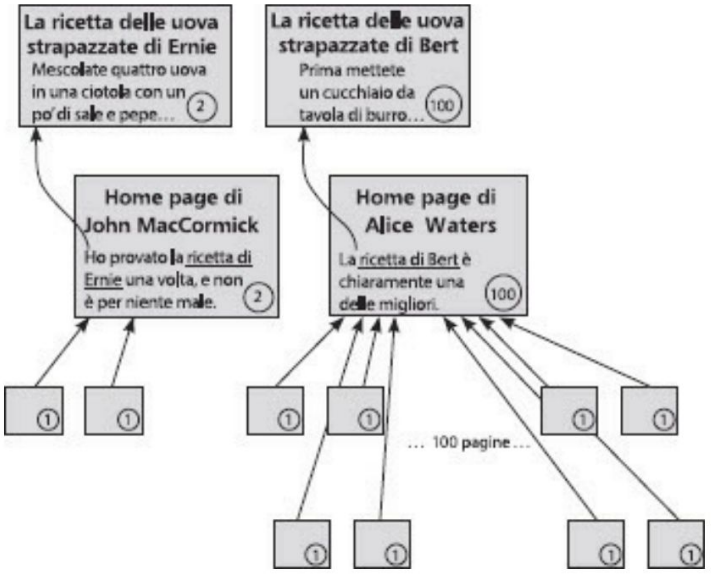

Figura 3.3 *Un semplice calcolo dei "punteggi di autorevolezza" per le due ricette delle uova strapazzate. Il valore dell'autorevolezza è racchiuso in un cerchietto.*

Il trucco del navigatore casuale

Sembrerebbe che abbiamo trovato una strategia per calcolare automaticamente i valori dell'autorevolezza e che funzioni bene, senza che il computer debba realmente comprendere i contenuti di una pagina. Purtroppo, questo metodo incappa in un problema grave: è possibile che i collegamenti ipertestuali formano quello che gli informatici chiamano un "ciclo" o un "anello". Si ha un ciclo se si può tornare al punto di partenza semplicemente seguendo i collegamenti ipertestuali.
La Figura 3.4 ce ne dà un esempio. Ci sono cinque pagine web: A, B, C, D, E. Se partiamo da A, possiamo seguire il collegamento da A a Be poi quello da B a E; da E poi possiamo seguire il collegamento che ci riporta ad A, da dove siamo partiti. Questo significa che A, B ed E formano un ciclo.

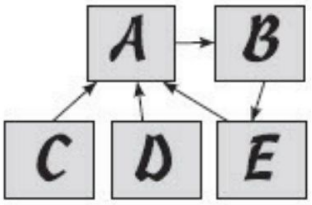

Figura 3.4 *Un esempio di ciclo di collegamenti ipertestuali. Le pagine A, 8, E formano un ciclo perché si può partire da A, seguire il collegamento verso 8, poi da 8 verso E e infine tornare al punto di partenza in A.*

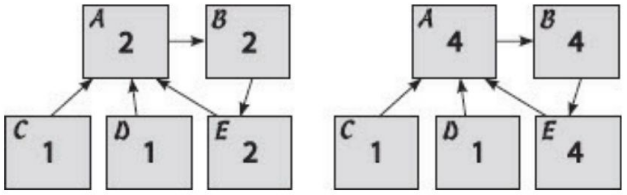

Figura 3.5 *Il problema causato dai cicli. A, 8 ed E non sono mai aggiornati, e il loro punteggio continua a crescere indefinitamente.*

La nostra definizione di "valore (o punteggio) di autorevolezza" (che combina il trucco del collegamento ipertestuale e quello dell'autorevolezza) finisce in guai seri ogni volta che c'è un ciclo. Vediamo che cosa succede in questo particolare esempio. Le pagine Ce D non hanno collegamenti in ingresso, perciò la loro autorevolezza è 1. Ce D hanno entrambi un collegamento ad A, perciò il punteggio di A è la somma di quelli di Ce D, ovvero 1 + 1 = 2. Poi B riceve il punteggio 2 da A, E riceve il punteggio 2 da B. (La situazione fin qui è riassunta nella parte sinistra della figura.) Ma a questo punto A non è più aggiornato: riceve ancora 1 sia da C sia da D, ma riceve anche un 2 da E, per un totale di 4. Ma ora non è più aggiornato B: riceve un 4 da A. Ma poi bisogna aggiornare anche E, che prende un 4 da B. (Ora siamo alla situazione della parte destra della figura.) E così via: ora A è 6, perciò B è 6 ed E è 6, ma allora A è 8, ... Avete sicuramente afferrato l'idea. Continueremmo all'infinito, con i punteggi che continuano ad aumentare a ogni ciclo.

Se si calcola l'autorevolezza in questo modo, si finisce in un problema "uovo o gallina". Se si conosce l'autorevolezza di A, si può calcolare quella di B ed E; e se si conosce l'autorevolezza di B ed E, si può calcolare quella di A. Ma siccome ciascuna dipende dalle altre, il problema sembra insolubile.

Per fortuna il problema si può risolvere con una tecnica che chiameremo trucco del navigatore casuale. Fate attenzione: la descrizione iniziale di questo trucco non ha alcuna somiglianza con i trucchi del collegamento ipertestuale e dell'autorevolezza di cui abbiamo parlato fin qui. Una volta che avremo visto i meccanismi fondamentali del trucco del navigatore casuale, analizzeremo le sue notevoli proprietà: in effetti, combina i tratti positivi degli altri due trucchi, ma funziona anche quando sono presenti cicli di collegamenti ipertestuali.

Immaginiamo una persona che navighi a caso in Internet. Per essere più precisi, il nostro navigatore parte da una pagina scelta a caso da tutto il World Wide Web. Poi esamina tutti i collegamenti ipertestuali che partono da quella pagina, ne sceglie uno a caso e ci fa clic sopra. Poi esamina la nuova pagina e sceglie

a caso uno dei suoi collegamenti ipertestuali. Il procedimento continua, e ogni nuova pagina viene scelta a caso facendo un clic su un collegamento ipertestuale nella pagina precedente. La Figura 3.6 mostra un esempio, in cui immaginiamo che tutto il World Wide Web sia costituito da 16 pagine soltanto. I rettangoli rappresentano pagine web, le frecce rappresentano collegamenti ipertestuali fra le pagine. Ci sono quattro pagine identificate come A, B, C, D, per semplificare i ragionamenti successivi. Le pagine

visitate dal navigatore sono in grigio più scuro, i collegamenti che segue sono in nero e le frecce tratteggiate rappresentano ripartenze casuali, di cui diremo fra poco.
C'è una particolarità nel procedimento: ogni volta che viene visitata una pagina, c'è una probabilità di ripartenza costante (poniamo il 15%) che il navigatore non faccia clic su uno dei collegamenti ipertestuali disponibili ma riavvii il procedimento scegliendo un'altra pagina a caso da tutto il Web. Potete pensare che c'è un 15% di probabilità che il navigatore si annoi, data una qualsiasi pagina, e che decida di seguire una nuova catena di collegamenti. Per qualche esempio, esaminate meglio la figura. Questo particolare navigatore è partito dalla pagina A e ha seguito tre collegamenti ipertestuali a caso prima di arrivare alla pagina 8, annoiarsi e ripartire dalla pagina C. Dopo aver seguito due altri collegamenti a caso, una nuova ripartenza. (Incidentalmente, tutti gli esempi di navigatori casuali in questo capitolo utilizzano una probabilità di ripartenza del 15%, che è la stessa utilizzata dai due fondatori di Google, Page e Brin, nell'articolo originale che descriveva il prototipo del loro motore di ricerca.)

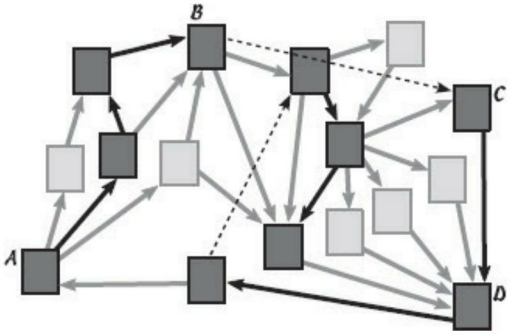

Figura 3.6 *Il modello del navigatore casuale. Le pagine visitate dal navigatore sono in grigio più scuro e le frecce tratteggiate rappresentato ripartenze casuali. La prova inizia dalla pagina A e segue collegamenti ipertestuali scelti a caso, con l'interruzione di due ripartenze casuali.*

È facile simulare questo procedimento con un computer. Ho scritto un programma che fa proprio questo e l'ho fatto girare fino a che il navigatore non aveva visitato 1000 pagine. (Ovviamente, questo non significa 1000 pagine diverse, valgono anche le visite alla stessa pagina e in questo piccolo esempio tutte le pagine sono state visitate molte volte.} I risultati delle 1000 visite simulate sono presentati nella parte superiore della Figura 3.7. Si può vedere che la pagina D è stata quella visitata più spesso, con 144 visite.
Proprio come nei sondaggi dell'opinione pubblica, si può migliorare l'accuratezza della simulazione aumentando il numero dei campioni casuali. Ho fatto girare nuovamente la simulazione, questa volta aspettando che il navigatore visitasse un milione di pagine. (Nel caso ve lo state chiedendo, c'è voluto meno di mezzo secondo, sul mio computer!) Con un numero così grande di visite, è meglio presentare i risultati in forma di percentuali, ed è questo che potete vedere nella parte inferiore della Figura 3.7. Anche questa volta la pagina D era quella visitata più spesso, con il 15% delle visite.

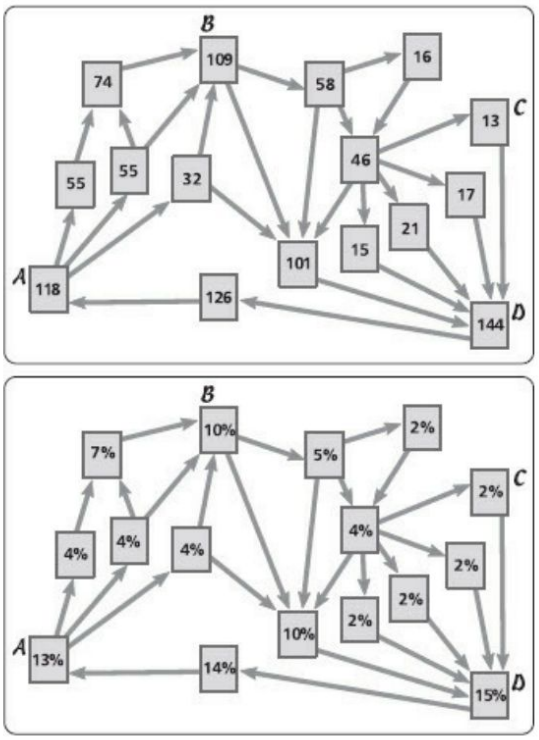

Figura 3.7 *Simulazioni del navigatore casuale. In alto: numero di visite a ciascuna pagina in una simulazione di 1000 visite. Sotto: percentuale di visite a ciascuna pagina in una simulazione di un milione di visite.*

Qual è il legame fra il nostro modello del visitatore casuale e il trucco dell'autorevolezza che vorremmo utilizzare per ordinare le pagine web? Le percentuali calcolate dalle simulazioni del navigatore casuale sono esattamente quello che ci serve per misurare l'autorevolezza di una pagina. Definiamo allora punteggio di autorevolezza del navigatore di una pagina web come la percentuale del tempo che un navigatore casuale dedicherebbe a visitare quella pagina. Cosa notevole, il punteggio di autorevolezza del navigatore incorpora entrambi i nostri trucchi precedenti per ordinare per importanza le pagine web. Li esamineremo uno alla volta.

Innanzitutto, avevamo il trucco del collegamento ipertestuale: l'idea di fondo qui è che una pagina con molti collegamenti in ingresso debba avere una posizione migliore. Questo è vero anche nel modello del navigatore casuale, perché una pagina con molti collegamenti in ingresso ha molte probabilità di essere visitata. La pagina D nella parte ineriore della figura nella pagina seguente è un buon esempio: ha cinque collegamenti in ingresso, più di ogni altra pagina nella simulazione, e finisce per avere il punteggio di autorevolezza del navigatore più alto (15%).

In secondo luogo, avevamo il trucco dell'autorevolezza. L'idea di fondo era che il collegamento proveniente da una pagina molto autorevole debba migliorare il posizionamento di una pagina più di un collegamento in arrivo da una pagina meno autorevole. E il modello del navigatore casuale tiene conto anche di questo. Perché? Perché un collegamento proveniente da una pagina molto popolare avrà maggiori possibilità di essere seguito rispetto a un collegamento da una pagina poco popolare. Per vedere un esempio nella nostra simulazione, confrontate le pagine A e C nel riquadro inferiore della figura: ciascuna ha esattamente un collegamento in ingresso, ma la pagina A ha un punteggio di autorevolezza del navigatore più alto (13% rispetto a 2%) grazie alla qualità del suo collegamento in ingresso.

Notate che il modello del navigatore casuale incorpora simultaneamente sia il trucco del collegamento ipertestuale sia quello dell'autorevolezza. In altre parole, tiene conto sia della qualità sia della quantità dei collegamenti in ingresso. Lo dimostra la pagina B: riceve il suo punteggio relativamente elevato (10%) grazie a tre collegamenti in arrivo da pagine con punteggi moderati, che vanno dal 4 al 7%.La bellezza del trucco del navigatore casuale è che, a differenza di quello dell'autorevolezza, funziona perfettamente

indipendentemente dall'esistenza di cicli di collegamenti ipertestuali. Tornando al nostro esempio precedente delle uova strapazzate (pagina 32), possiamo facilmente eseguire una simulazione del navigatore casuale. Dopo vari milioni di visite, la mia simulazione ha dato i punteggi di autorevolezza del navigatore indicati nella Figura 3.8. Notate che, come nel calcolo precedente con il trucco dell'autorevolezza, la pagina di Bert riceve un punteggio molto superiore a quello della pagina di Ernie (28% contro 1 %), nonostante il fatto che entrambe abbiano esattamente un collegamento in ingresso. Perciò Bert sarebbe in una posizione migliore, fra i risultati di una ricerca sul Web per "uova strapazzate".

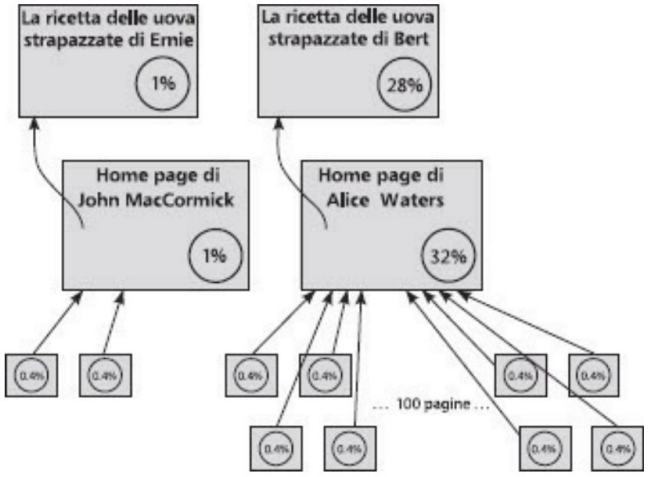

Figura 3.8 *I punteggi di autorevolezza del navigatore per l'esempio delle uova strapazzate di pagina 29. Sia Bert sia Ernie hanno esattamente un collegamento in ingresso che contribuisce a determinare l'autorevolezza delle loro pagine, ma quella di Bert avrà un posizionamento migliore nei risultati di una ricerca nel Web per "uova strapazzate".*

Passiamo ora all'esempio più difficile, quello delle Figure 3.4 e 3.5, che metteva in una difficoltà insormontabile il nostro trucco dell'autorevolezza a causa del ciclo di collegamenti. Anche in questo caso è facile eseguire una simulazione al computer dei navigatori casuali, e si ottengono i punteggi di autorevolezza del navigatore della Figura 3.9. I punteggi determinati con questa simulazione ci dicono quale sia l'ordinamento finale che utilizzerebbe un motore di ricerca al momento di presentare i suoi risultati: la pagina A per prima, seguita da B, poi da E, con C e D alla pari in ultima posizione.

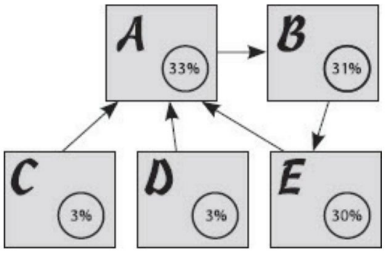

Figura 3.9 *Punteggi di autorevolezza del navigatore casuale per l'esempio precedente con un ciclo di collegamenti ipertestuali (Figure 3.4 e 3.5). Il trucco del navigatore casuale non ha difficoltà a calcolare punteggi appropriati, nonostante la presenza di un ciclo (A ➔ B ➔ E ➔ A)*

PageRank in pratica

Il trucco del navigatore casuale è stato descritto dai due fondatori di Google nel loro famoso saggio del 1998, "The Anatomy of a Large-Scale Hypertextual Web Search Engine" e, insieme con molte altre tecniche, varianti di questo trucco vengono utilizzate tuttora dai grandi motori di ricerca. Esistono però parecchi fattori che complicano le cose, il che significa che le tecniche effettivamente utilizzate dai motori di ricerca moderni sono un po' diverse da quella del navigatore casuale appena descritta.

Uno di questi fattori di complicazione va diritto al cuore di PageRank: l'assunto che i collegamenti ipertestuali conferiscano legittimamente autorevolezza è discutibile. Abbiamo già visto che, anche se i collegamenti possono rappresentare critiche anziché apprezzamenti positivi non si tratta in pratica di un problema significativo. Un problema molto più grave è che è possibile abusare del trucco del collegamento ipertestuale per migliorare artificialmente la posizione delle proprie pagine web. Supponiamo che gestiate un sito Librilibrilibri.com che vende (ma pensa un po') libri. Con una tecnologia automatizzata, è relativamente facile creare un gran numero, diciamo 10.000, di pagine web diverse, ma tutte con un collegamento a Librilibrilibri.com. Quindi, se i motori di ricerca calcolassero l'autorevolezza di PageRank esattamente come abbiamo descritto qui, Librilibrilibri.com potrebbe ottenere senza meritarselo un punteggio migliaia di volte superiore a quello di altre librerie online, con una posizione migliore e quindi con maggiori vendite.

I motori di ricerca definiscono web spam questo tipo di abuso. (Il termine deriva da un'analogia con lo spam nella posta elettronica: i messaggi indesiderati nella vostra casella di posta elettronica sono simili a pagine web indesiderate che ingombrano i risultati di una ricerca nel Web.) Identificare ed eliminare i vari tipi di web spam è importante per tutti i motori di ricerca. Nel 2004, per esempio, alcuni ricercatori alla Microsoft hanno trovato oltre 300.000 siti che ricevevano collegamenti da esattamente 1001 pagine - cosa assai sospetta. Andando a esaminare manualmente questi siti, i ricercatori hanno trovato che la stragrande maggioranza dei collegamenti ipertestuali in ingresso erano web spam.

I motori di ricerca sono impegnati in una corsa agli armamenti contro gli spammer e cercano continuamente di migliorare i loro algoritmi per classificare in modo sensato i risultati. Questa spinta costante a migliorare PageRank ha favorito molta ricerca accademica e industriale su altri algoritmi che usano la struttura a collegamenti ipertestuali del Web per ordinare le pagine. Algoritmi di questo tipo spesso sono indicati come algoritmi di ordinamento basati sui collegamenti.

Un altro fattore di complicazione è legato all'efficienza dei calcoli di PageRank. I nostri punteggi di autorevolezza del navigatore sono stati calcolati eseguendo simulazioni casuali, ma simulazioni simili su tutto il Web richiederebbero troppo tempo per essere utili. I motori di ricerca non calcolano i valori di PageRank simulando navigatori casuali; usano invece tecniche matematiche che danno le stesse risposte ma con un costo computazionale minore. Abbiamo studiato la tecnica della simulazione del navigatore per il suo fascino intuitivo, e perché descrive che cosa calcolano i motori di ricerca, non come lo calcolano.

Val la pena anche di notare che i motori di ricerca determinano i loro ordinamenti utilizzando molto di più di un semplice algoritmo di ordinamento basato sui collegamenti come PageRank. Già nella loro descrizione di Google nel 1998, i suoi due fondatori citavano parecchie altre caratteristiche che contribuivano all'ordinamento dei risultati delle ricerche. Come potete immaginare, la tecnologia è andata avanti: nel momento in cui scrivo, il sito web di Google dichiara che, per valutare l'importanza di una pagina, vengono utilizzati "oltre 200 segnali".

Nonostante le molte complessità dei motori di ricerca moderni, la bella idea al cuore di PageRank, cioè che le pagine autorevoli possano conferire autorevolezza ad altre pagine attraverso i collegamenti ipertestuali, rimane valida. Questa idea ha aiutato Google a detronizzare AltaVista, trasformandola da piccola startup a regina delle ricerche in pochi. Senza l'idea centrale di PageRank, la maggior parte delle interrogazioni annegherebbero in un mare di pagine web irrilevanti. PageRank è in effetti una gemma algoritmica grazie alla quale un ago può salire senza fatica in cima al suo pagliaio.