Indicizzazione nei motori di ricerca: trovare aghi nel pagliaio più grande del mondo

Huck, stando qui potresti toccare quella fessura, da cui sono uscito, con una canna da pesca. Prova a vedere se ci riesci.
~ Mark Twain, **Le avventure di Tom Sawyer**

I motori di ricerca hanno cambiato profondamente la nostra vita. La maggior parte di noi effettua interrogazioni molte volte al giorno, ma di rado ci fermiamo a pensare come possano funzionare questi meravigliosi strumenti. La grande quantità di informazioni disponibili e la velocità e la qualità dei risultati ci sembrano ormai così normali che proviamo un senso di frustrazione se una domanda non trova risposta nel giro di qualche secondo. Tendiamo a dimenticare che ogni ricerca estrae un ago dal più grande pagliaio del mondo: il World Wide Web.
Il servizio superbo che ci forniscono i motori di ricerca non è solo il risultato dell'impiego di una grande quantità di tecnologia speciale per affrontare il problema. Certo, ognuna delle grandi aziende che gestisce uno dei motori di ricerca principali utilizza una rete internazionale di enormi centri dati, con migliaia di server e apparecchiature di rete all'avanguardia, ma tutto questo hardware sarebbe inutile senza i brillanti algoritmi necessari per organizzare e recuperare le informazioni che chiediamo. In questo capitolo e nel successivo, quindi, andremo a scoprire alcune delle gemme algoritmiche che lavorano per noi ogni volta che effettuiamo una ricerca sul Web. Come vedremo, due dei compiti principali di un capitolo sono l'identificazione di corrispondenze (matching) e l'ordinamento (ranking). Questo capitolo si occupa di una bella tecnica di identificazione di corrispondenze, il trucco delle metaparole. Nel prossimo capitolo, ci dedicheremo al problema dell'ordinamento ed esamineremo il famoso algoritmo PageRank di Google.

Corrispondenza e ordinamento

Sarà utile iniziare con uno sguardo generale su quel che succede quando inviamo una richiesta di ricerca sul Web, una interrogazione. Ci saranno, come abbiamo già detto, due fasi principali: l'identificazione di corrispondenze e l'ordinamento. Nella pratica, i motori di ricerca combinano le due attività in un processo unico, per ragioni di efficienza, ma le due fasi sono concettualmente distinte, perciò ipotizzeremo che l'identificazione delle corrispondenze sia completata prima che inizi l'ordinamento. La Figura 2.1 illustra un esempio, dove l'interrogazione è "London bus timetable" o "Orari autobus Londra". La fase di identificazione delle corrispondenze risponde alla domanda "quali pagine web soddisfano l'interrogazione?" - in questo caso, tutte le pagine che citano gli orari degli autobus londinesi.

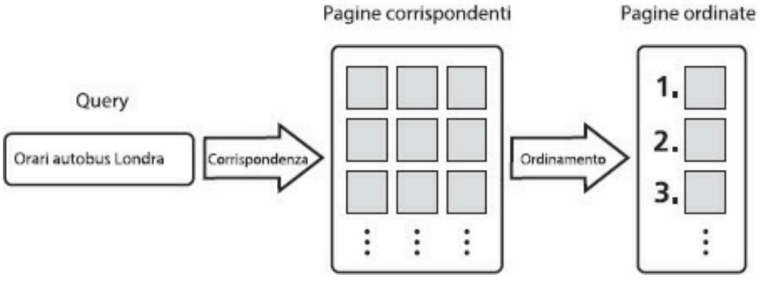

Figura 2.7 *Le due fasi della ricerca nel Web: identificazione di corrispondenze e ordinamento. È possibile che ci siano migliaia o milioni di corrispondenze dopo la prima fase, che poi vanno ordinate per pertinenza decrescente nella seconda fase.*

Ma per molte interrogazioni dei motori di ricerca, nella realtà, possono esistere centinaia, migliaia o addirittura milioni di hit, di pagine che corrispondono al criterio di ricerca, e gli utenti dei motori in genere preferiscono esaminare solo una manciata di risultati, magari cinque, dieci al massimo. Perciò un motore di ricerca deve essere in grado di scegliere, fra un numero molto grande di risultati, i migliori. Un buon motore di ricerca non solo estrarrà i risultati migliori, ma li presenterà anche nell'ordine più utile, cioè con la pagina più adatta in prima posizione, seguita dalla successiva in ordine di pertinenza e così via.
Il compito di selezionare i pochi risultati migliori nell'ordine giusto è il compito dell'ordinamento o ranking, seconda e cruciale fase che segue quella iniziale di ricerca di corrispondenze. In un campo in cui la concorrenza è micidiale, i motori di ricerca vivono o muoiono a seconda della qualità dei loro sistemi di ordinamento. Nel 2002, la quota di mercato dei tre maggiori motori di ricerca negli USA era all'incirca uguale: Google, Yahoo e MSN ricevevano ciascuno poco meno del 30% di tutte le interrogazioni di ricerca del paese {MSN poi ha cambiato nome ed è diventato Live Search e infine Bing). Nell'arco di pochi anni, Google è riuscito ad aumentare drasticamente la sua quota di mercato, retrocedendo Yahoo e MSN a quote inferiori al 20%. Quasi tutti pensano che l'ascesa di Google alla cima del settore sia stata merito dei suoi algoritmi di ordinamento, perciò non si esagera dicendo che i motori di ricerca vivono o muoiono a seconda della qualità dei loro sistemi di ordinamento. Però, lo abbiamo già detto, degli algoritmi di ordinamento ci occuperemo nel prossimo capitolo; per ora, ci concentriamo sulla fase dell'identificazione di corrispondenze.

AltaVista: il primo algoritmo di matching alla scalata del Web

Dove inizia la nostra storia? Verrebbe spontaneo dire con Google, il maggiore successo tecnologico degli inizi del ventunesimo secolo, ma sarebbe sbagliato. La storia degli inizi di Google, come progetto di ricerca per il Ph. D. di due studenti della Stanford University è tanto appassionante quanto sorprendente. Larry Page e Sergey Brin sono riusciti nel 1998 a mettere insieme un assortimento eterogeneo di macchine per realizzare un nuovo tipo di motore di ricerca: meno di dieci anni dopo la loro azienda era diventata il più grande gigante digitale dell'era di Internet.
Ma l'idea delle ricerche sul Web circolava già da anni. Fra le prime offerte commerciali ci furono lnfoseek e Lycos (lanciati entrambi nel 1994) e AltaVista, lanciato nel 1995. Per qualche anno, alla metà degli anni Novanta, AltaVista fu il re dei motori di ricerca. A quel tempo ero specializzando in informatica e ricordo bene di essere rimasto colpito dall'ampiezza dei risultati di AltaVista: per la prima volta un motore di ricerca aveva indicizzato completamente tutto il testo di tutte le pagine del Web, non solo, ma restituiva i risultati in un batter d'occhi. Il nostro viaggio alla comprensione di questo sensazionale risultato tecnologico inizia con un concetto (letteralmente) antico: l'indicizzazione.

Buona, vecchia indicizzazione

Il concetto di *indice* è l'idea fondamentale alla base di ogni motore di ricerca, ma i motori di ricerca non hanno inventato gli indici: si tratta, in effetti, di un'idea vecchia quanto la scrittura stessa. Gli archeologi hanno scoperto una biblioteca di un tempio babilonese di 5000 anni fa che catalogava per argomento le sue tavolette in cuneiforme, perciò l'idea dell'indicizzazione può ambire a buon diritto al titolo di più vecchia fra le idee utili per l'informatica.
Oggi la parola "indice" di solito indica l'indice analitico, una sezione alla fine di un testo di saggistica o di consultazione: tutti i concetti che si potrebbe voler trovare in quel libro sono elencati in un ordine ben

preciso (solitamente in ordine alfabetico) e a ciascun concetto segue un elenco delle posizioni (di solito numeri di pagina) in cui si parla di quel concetto. Un libro di zoologia può avere una voce di indice del tipo "ghepardo 124, 156", che significa che la parola "ghepardo" si trova alle pagine 124 e 156. (Come piccola curiosità, potreste cerca la parola "indice" nell'indice di questo libro. Se non ci sono errori, vi rimanderà a questa pagina.)
L'indice di un motore di ricerca funziona nello stesso modo dell'indice di un libro: le "pagine" sono pagine del World Wide Web e i motori assegnano un numero di pagina diverso a ogni singola pagina web. (Sì, ci sono tantissime pagine, parecchi miliardi, ma i computer sono bravissimi nel trattare numeri molto grandi.) La Figura 2.2 presenta un esempio che renderà tutto più concreto. Supponiamo che il World Wide Web sia fatto solo di tre pagine web, a cui sono assegnati i numeri 1, 2, 3.

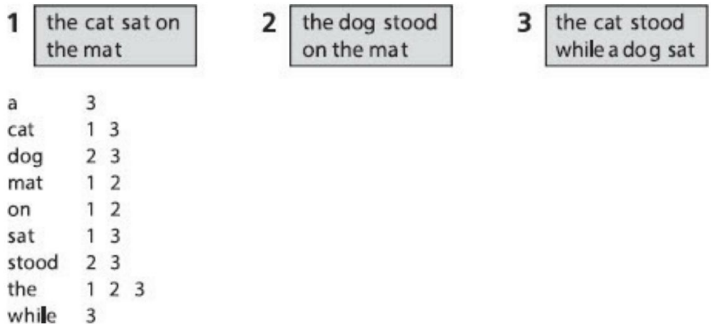

Figura 2.2 *In alto: un World Wide Web immaginario costituito da tre pagine solamente, identificate dai numeri 1, 2, 3. In basso: un indice molto semplice, con relativi numeri di pagina.*

Un computer può costruire un indice di queste tre pagine innanzitutto stendendo un elenco di tutte le parole che compaiono nelle pagine e poi ordinando alfabeticamente l'elenco. L'**elenco delle parole** (*word list*) in questo caso particolare sarà "a, cat, dog, mat, on, sat, stood, the, while". Poi il computer esplorerà le pagine parola per parola e per ogni parola prenderà nota del numero della pagina in cui compare. Dal risultato finale, in figura, si può vedere subito che la parola "cat" è presente nelle pagine 1 e 3, ma non nella 2 e che la parola "while" appare solo nella pagina 3.
Con questo metodo molto semplice, un motore di ricerca può già dare una risposta a molte domande semplici. Per esempio, supponiamo che inseriate l'interrogazione cat. Il motore di ricerca può rapidamente saltare al lemma cat nel suo elenco di parole. (Dato che l'elenco delle parole è in ordine alfabetico, un computer può trovare rapidamente un lemma, come un essere umano può trovare rapidamente un lemma in un dizionario.) Appena trova il lemma cat, il motore di ricerca può subito fornirvi l'elenco delle pagine in cui compare, in questo caso 1 e 3. I motori di ricerca moderni arricchiscono la presentazione dei risultati con un frammento di ciascuna pagina, ma in genere trascureremo questi particolari e ci concentreremo su come fa il motore a sapere quali siano i numeri di pagina che sono hit per l'interrogazione avanzata.
Un altro esempio semplicissimo: verifichiamo la procedura per l'interrogazione dog. In questo caso, il motore di ricerca trova rapidamente la voce dog e restituisce i risultati 2 e 3. Ma che cosa succede se l'interrogazione contiene più parole, per esempio cat dog? Questo significa che volete trovare pagine che contengano sia la parola "cat" sia la parola "dog". Ancora una volta, è una cosa facile per il motore di ricerca, grazie all'indice esistente. Prima cerca le due parole singolarmente, per sapere in quali pagine siano presenti e ottiene le risposte 1, 3 per "cat" e 2, 3 per "dog". Poi può rapidamente esaminare i due elenchi di risultati, per vedere se ci siano numeri di pagina uguali. In questo caso le pagine 1 e 2 non soddisfano questo requisito, ma la pagina 3 è presente in entrambi gli elenchi, perciò la risposta finale è costituita da un solo risultato, la pagina 3. Una strategia molto simile si applica per interrogazioni con

più di due parole. Per esempio, l'interrogazione cat the sat restituisce come risultati le pagine 1 e 3, che sono gli elementi comuni agli elenchi per "cat" (1, 3), "the" (1, 2, 3) e "sat" (1, 3).
Fin qui, sembrerebbe che costruire un motore di ricerca sia molto facile: la più semplice possibile delle tecnologie di indicizzazione sembra funzionare bene anche per interrogazioni con più parole. Purtroppo, però, questa impostazione semplice è del tutto inadeguata per i motori di ricerca moderni: le ragioni sono diverse, ma per ora ci concentreremo su un solo problema, quello delle *interrogazioni a sintagma*. Sono le interrogazioni che cercano un sintagma preciso, un'espressione complessa e non semplicemente le occorrenze di varie parole in qualsiasi punto di una pagina. Nella maggior parte dei motori di ricerca, queste interrogazioni si indicano con l'uso delle virgolette: così, per esempio, l'interrogazione "cat sat" ha un significato molto diverso dall'interrogazione cat sat. L'interrogazione cat sat cerca pagine che contengano le parole "cat" e "sat" in qualsiasi punto e in qualsiasi ordine, mentre "cat sat" cerca pagine che contengano la parola "cat" seguita immediatamente dalla parola "sat". Nel nostro banale esempio con tre sole pagine, cat sat ha come risultati le pagine 1 e 3, ma "cat sat" restituisce un solo risultato, la pagina 1.
In che modo un motore di ricerca può eseguire efficacemente una interrogazione a sintagma? Restiamo all'esempio di "cat sat". Sembrerebbe che il primo passo sia fare la stessa cosa che si fa per una normale interrogazione a più parole cat sat: recuperare dall'elenco delle parole l'elenco delle pagine in cui ciascuna parola occorre, in questo caso 1, 3 per "cat" e 1, 3 anche per
"sat". Ma a questo punto il motore è bloccato. Sa per certo che entrambe le parole sono presenti nelle pagine 1 e 3, ma non ha modo di dire se vi compaiano una subito dopo l'altra e nell'ordine giusto. Si potrebbe pensare che il motore possa tornare a consultare le pagine web originali per stabilire se vi compaia o meno quell'espressione precisa. In effetti sarebbe una soluzione, ma molto, molto inefficiente, perché comporta che si vadano a rileggere tutti i contenuti di tutte le pagine che potrebbero contenere quell'espressione, e il numero di quelle pagine potrebbe essere enorme. Non dimenticate che stiamo ragionando su un esempio estremamente piccolo, con tre sole pagine, mentre un motore di ricerca reale deve dare risultati corretti per decine di miliardi di pagine web.

Il trucco della posizione della parola

La soluzione a questo problema è la prima idea davvero ingegnosa che ha contribuito al buon funzionamento dei motori di ricerca moderni: l'indice non deve conservare solo i numeri di pagina, ma anche le posizioni all'interno delle pagine. Queste posizioni non sono nulla di misterioso: indicano semplicemente la posizione di una parola all'interno della pagina, per cui la terza parola ha la posizione 3, la ventinovesima la posizione 29 e via elencando. Il nostro intero insieme di dati costituito dalle tre pagine è mostrato nella Figura 2.3, dove sono state aggiunte le posizioni delle parole. Sotto si vede l'indice costruito tenendo conto anche della posizione delle parole. Vediamo un paio
di esempi, per essere sicuri di aver capito bene. La prima riga dell'indice è "a 3-5" e significa che la parola "a" si presenta una sola volta nell'insieme dei dati, cioè nella pagina 3, e che è la quinta parola di quella pagina. La riga più lunga dell'indice è "the 1-1 1-5 2-1 2-5 3-1 ", che dà le posizioni esatte di tutte le occorrenze della parola "the" nell'insieme dei dati: due volte nella pagina 1 (alle posizioni 1 e 5), due volte nella pagina 2 (alle posizioni 1 e 5) e una volta nella pagina 3 (alla posizione 1).
Ricordiamo perché abbiamo introdotto le posizioni delle parole nelle pagine: per risolvere il problema di come rispondere in modo efficiente a una interrogazione a sintagma. Vediamo allora dove ci porta il nuovo indice. Riprendiamo la stessa interrogazione di prima, "cat sat". I primi passi sono gli stessi del vecchio indice: estraiamo le posizioni delle singole parole dall'indice, cosicché per "cat" otteniamo 1-2, 3-2 e per "sat" otteniamo 1-3, 3-7. Fin qui tutto bene: sappiamo che gli unici risultati possibili per l'interrogazione "cat sat" sono le pagine 1 e 3. Ma, come prima, ancora non siamo sicuri che in quelle pagine si presenti proprio quella espressione precisa - è possibile che le due parole siano nella pagina, ma

non una vicina all'altra oppure non nell'ordine giusto. Per fortuna, è facile verificarlo adesso con le informazioni sulla posizione. Concentriamoci sulla pagina 1, per cominciare. Dalle informazioni dell'indice, sappiamo che "cat" compare nella posizione 2 della pagina 1 (è questo il significato di 1-2), e sappiamo che "sat" compare nella posizione 3 della pagina 1 (è questo il significato di 1-3). Ma se "cat" è nella posizione 2 e "sat" è nella posizione 3, allora sappiamo che "sat" compare subito dopo "cat" (perché 3 è il successore immediato di 2) e perciò l'esatta espressione che stiamo cercando, "cat sat", deve essere presente in questa pagina, a partire dalla posizione 2!

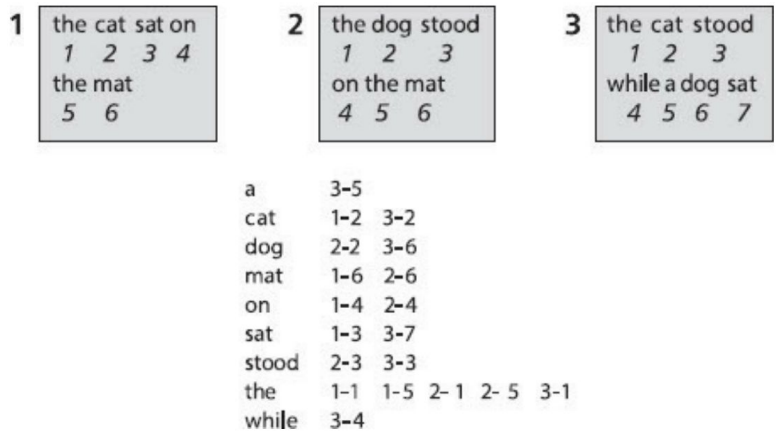

Figura 2.3 *In alto: le nostre tre pagine web a cui è stata aggiunta l'indicazione delle posizioni delle parole. In basso: un nuovo indice che include sia i numeri delle pagine, sia le posizioni delle parole nella pagina.*

So che sono un po' noioso, ma il motivo per esaminare questo esempio in tutti i suoi particolari è capire esattamente quali informazioni si usano per arrivare alla risposta. Notate che abbiamo trovato un risultato per il sintagma "cat sat" esaminando solo le informazioni dell'indice (1-2, 3-2 per "cat" e
1-3, 3-7 per "sat"} e non le pagine web originali. Questo è un aspetto fondamentale, perché abbiamo dovuto esaminare solo due voci dell'indice, invece di rileggere tutte le pagine che potevano essere risultati validi, e potrebbero esserci letteralmente milioni di pagine simili nel caso di un motore di ricerca reale che cerca di rispondere a una interrogazione a sintagma reale. Per riassumere: l'inclusione nell'indice della posizione delle parole nelle pagine ci ha permesso di rispondere a un'interrogazione
a sintagma esaminando solo un paio di righe nell'indice, invece di leggere per esteso un gran numero di pagine web. Questo semplice trucco della posizione delle parole è uno degli elementi chiave per il funzionamento dei motori di ricerca!
In realtà, non abbiamo ancora finito di elaborare l'esempio
"cat sat". Abbiamo elaborato le informazioni per la pagina 1, ma non quelle per la pagina 3. Il ragionamento però è simile: vediamo che "cat" compare nella posizione 2, e "sat" nella posizione 7, perciò non possono essere vicine, perché 7 non è il successore immediato di 2. Perciò sappiamo che la pagina 3 non è un risultato valido per l'interrogazione a sintagma "cat sat", anche se è un risultato valido per la ricerca a più parole cat sat.
Incidentalmente, il trucco della posizione delle parole è importante anche per altri motivi. Per esempio, consideriamo il problema di trovare parole che siano vicine l'una all'altra. In certi motori di ricerca, lo si può fare inserendo la parola chiave NEAR nell'interrogazione; AltaVista offriva questo strumento sin dagli inizi e lo fa ancora nel momento in cui scrivo. Come esempio, supponiamo che in un certo specifico motore di ricerca l'interrogazione cat NEAR dog trovi le pagine in cui la parola "cat" occorre a distanza di non più di cinque parole dalla parola "dog". Come possiamo rispondere in modo efficiente a questa interrogazione per il nostro insieme di dati? Utilizzando le posizioni delle parole è facile. La voce di

indice per "cat" è 1-2, 3-2 e la voce di indice per "dog" è 2-2, 3-6, perciò vediamo subito che la pagina 3 è l'unico risultato possibile. Sulla pagina 3, poi, "cat" occorre nella posizione 2, mentre "dog" è nella posizione 6, perciò la distanza fra le due parole è 6 - 2, cioè 4. Quindi "cat" occorre
a una distanza non superiore alle cinque parole da "dog" e la pagina 3 è un risultato valido per l'interrogazione cat NEAR dog. Anche in questo caso, notate l'efficienza con cui è stato possibile rispondere: non c'è stato bisogno di leggere i contenuti effettivi di alcuna pagina web, ma sono stati consultate solo due voci dell'indice.
In realtà, le interrogazioni NEAR non sono molto importanti per gli utenti dei motori di ricerca, nella pratica: quasi nessuno le usa e la maggior parte dei motori di ricerca non le consente nemmeno. Nonostante questo, la possibilità di eseguire interrogazioni NEAR è davvero fondamentale per i motori di ricerca reali, perché i motori eseguono costantemente, dietro le quinte, interrogazioni di questo tipo. Per capire il perché, dobbiamo prima dare uno sguardo a uno degli altri grandi problemi dei motori di ricerca moderni: il problema dell'ordinamento.

Ordinamento e prossimità

Fin qui ci siamo concentrati sulla fase della corrispondenza: il problema di trovare in modo efficiente tutti i risultati validi, gli hit, per una interrogazione data. Ma, come abbiamo detto, la seconda fase, l'ordinamento, è essenziale per un motore di ricerca di qualità: è la fase in cui si selezionano pochi risultati ottimali per presentarli all'utente,
Analizziamo un po' meglio il concetto di ordinamento o ranking. Da che cosa dipende realmente la posizione di una pagina in questa classifica? La domanda qui non è "Questa pagina soddisfa l'interrogazione?" bensì "Questa pagina è rilevante per l'interrogazione?". Gli informatici usano la parola "rilevanza" o "pertinenza" per parlare di quanto una data pagina sia adeguata utile come risposta a una interrogazione particolare.
Concretamente, supponiamo siate interessati a sapere che cosa causa la malaria, e che inseriate in un motore di ricerca l'interrogazione `malaria causa`. Per semplificare le cose, supponiamo che ci siano due soli risultati per l'interrogazione, le due pagine mostrate nella figura 2.4. Date un'occhiata a queste pagine. Dovrebbe esservi subito chiaro, in quanto esseri umani, che la pagina 1 parla effettivamente della causa della malaria, mentre la pagina 2 sembra la descrizione di qualche campagna militare in cui semplicemente capita, per puro caso, che vengano usate entrambe le parole "causa" e "malaria". Perciò la pagina 1 è senza alcun dubbio più "pertinente" per la ricerca malaria causa rispetto alla pagina 2. Ma i computer non sono esseri umani e non esiste un modo facile per far "capire" a un computer gli argomenti delle due pagine, perciò potrebbe sembrare impossibile che un motore di ricerca riesca a ordinare correttamente i due risultati.

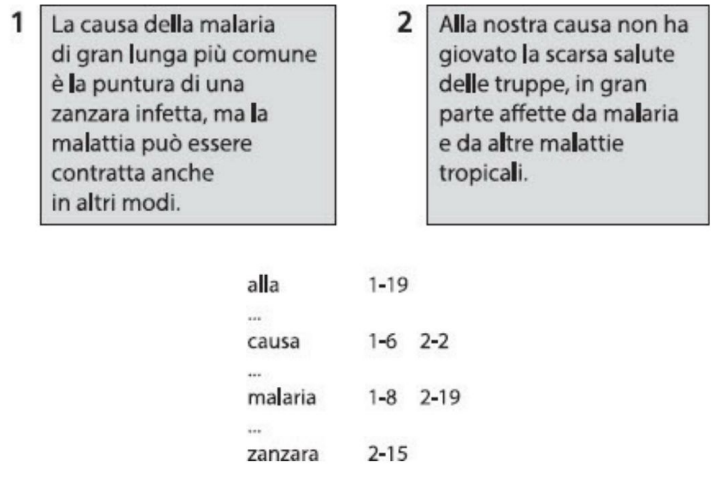

Figura 2.4 *In alto: due esempi di pagine web che citano la malaria. In basso: parte dell'indice costruito a partire dalle due pagine in alto.*

In questo caso, però, esiste un modo semplice per ottenere l'ordinamento giusto. Statisticamente, le pagine in cui le parole dell'interrogazione occorrono vicino l'una all'altra è più probabile siano pertinenti, rispetto alle pagine in cui le parole dell'interrogazione sono molto distanti. Nell'esempio della malaria, vediamo che le parole "malaria" e "causa" sono a distanza di due sole parole nella pagina 1, ma sono separate da 14 parole nella pagina 2. {Ricordate che il motore di ricerca può stabilirlo rapidamente esaminando solo le voci dell'indice, senza dover tornare a leggere le pagine web stesse.) Perciò, anche se il computer non "capisce" realmente l'argomento dell'interrogazione, può congetturare che la pagina 1 sia più pertinente della pagina 2, perché le parole dell'interrogazione sono molto più vicine fra loro nella pagina 1 che non nella pagina 2.
Per riassumere: gli essere umani non usano molto le interrogazioni NEAR, ma i motori di ricerca usano costantemente le informazioni sulla prossimità per migliorare le loro classifiche, e il motivo per cui possono farlo in modo efficiente è l'uso del trucco della posizione delle parole.
Sappiamo che i Babilonesi utilizzavano l'indicizzazione 5000 anni prima che nascessero i motori di ricerca. Questi ultimi non hanno inventato neanche il trucco della posizione delle parole: si tratta di una tecnica ben nota, utilizzata in altri tipi di recupero delle informazioni prima che Internet facesse il suo ingresso sulla scena.
Nel prossimo paragrafo invece parleremo di un nuovo trucco che a quanto pare è stato inventato proprio dai progettisti di motori di ricerca, il **trucco della metaparola**. Un uso molto abile di questo trucco e di varie altre idee collegate ha contribuito a catapultare AltaVista sulla vetta del mondo dei motori di ricerca verso la fine degli anni Novanta.

Il trucco della metaparola

Fin qui abbiamo usato esempi di pagine web molto semplici ma, come probabilmente saprete, le pagine web in genere sono molto strutturate, con titoli, intestazioni, collegamenti e immagini, mentre fin qui le abbiamo trattate come semplici elenchi di parole. Ora andiamo a scoprire come i motori di ricerca tengono conto anche della struttura delle pagine web. Per mantenere
il massimo livello di semplicità possibile, introdurremo solo un elemento di strutturazione: consentiremo alle nostre pagine di avere un **titolo** (*title*) in testa, seguito dal **corpo** (*body*) della pagina. La Figura 2.5 mostra il nostro esempio di tre pagine che ora hanno anche un titolo.

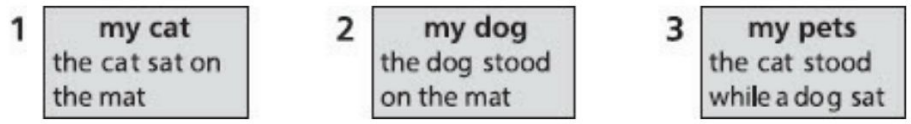

Figura 2.5 *Un insieme di pagine web che hanno un titolo e un corpo.*

In realtà, per analizzare la struttura delle pagine nello stesso modo dei motori di ricerca, dobbiamo sapere qualcosa di più su come le pagine sono scritte. Le pagine web sono composte in un linguaggio particolare, che consente ai browser di visualizzarle con una presentazione elegante. (Il linguaggio più usato si chiama HTML, ma come sia fatto nei particolari HTML non è importante per noi qui.) Le istruzioni di formattazione per intestazioni, titoli, collegamenti, immagini e così via vengono scritte utilizzando parole speciali definite metaparole (metaword). Per esempio, la metaparola utilizzata per indicare l'inizio del titolo di una pagina web può essere <titleStart> e la metaparola per indicare la fine

del titolo può essere <titleEnd>. Analogamente, si può indicare l'inizio del corpo della pagina con <bodyStart> e concluderlo con <bodyEnd>. Non fatevi confondere dai simboli "<" e ">": sono presenti su quasi tutte le tastiere e spesso vengono indicati con il loro significato matematico di "minore" e "maggiore". Qui però non hanno proprio nulla a che vedere con la matematica, sono semplicemente simboli che servono a contrassegnare le metaparole come qualcosa di diverso dalle normali parole di una pagina web.

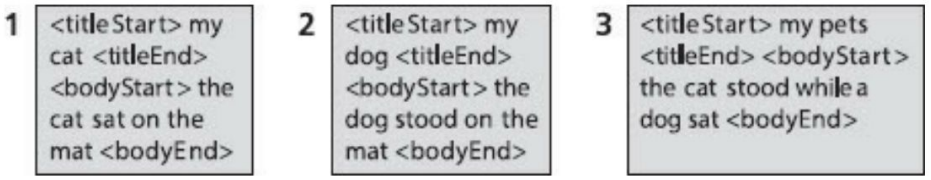

Figura 2. 6 *Lo stesso insieme di pagine web della figura precedente, ora presentate come potrebbero essere scritte con l'uso di metaparole, anziché come verrebbero visualizzate in un browser.*

Date un'occhiata alla Figura 2.6, che mostra esattamente lo stesso contenuto della figura precedente, ma ora come sarebbe effettivamente scritto nelle pagine web, anziché come verrebbe visualizzato in un browser. La maggior parte dei browser consente di esaminare i contenuti grezzi di una pagina web con un comando da menu come "Visualizza sorgente" o "Origine": vi consiglio di fare una prova alla prima occasione. (Notate che le metaparole usate qui, come <titleStart> e <titleEnd> sono di mia invenzione, esempi facilmente riconoscibili per facilitare la comprensione. Nel vero HTML, le metaparole sono denominate *tag* o **marcatori**. I tag che indicano inizio e fine di un titoli in HTML sono <title> e </title>: cercateli dopo aver usato il comando "Visualizza sorgente".)
Quando si costruisce un indice, è semplice includere tutte le metaparole. Non servono nuovi trucchi: si memorizzano le posizioni delle metaparole esattamente come si fa per le parole normali. Le Figure 2.7 e 2.8 mostrano l'indice costruito a partire dalle tre pagine web con le loro metaparole. Date un'occhiata alle figure e sincerateVi che non sta succedendo niente di misterioso. Per esempio, la voce "mat" indica 1-11, 2-11, il che significa che "mat" è l'undicesima parola della pagina 1 e anche l'undicesima parola della pagina 2. Le metaparole sono trattate nello stesso modo, perciò la voce di indice per "<titleEnd>", con 1-4, 2-4, 3-4, significa che "<titleEnd>" è la quarta parola di pagina 1, pagina 2 e pagina 3.
Questo trucco, indicizzare le metaparole come le parole normali, è il "trucco della metaparola". Può sembrare di una semplicità disarmante, ma questo trucco è fondamentale per consentire ai motori di ricerca di eseguire ricerche accurate ed effettuare ordinamenti di buona qualità. Vediamo un esempio semplice. Supponiamo che un motore di ricerca ammetta un tipo particolare di interrogazione con la parola chiave IN, così che un'interrogazione boat IN TITLE restituisca come risultati solo pagine che hanno la parola "boat" nel titolo e giraffe IN BODY trovi solo pagine che contengono "giraffe" nel loro corpo. Badate che la maggior parte dei motori reali non consente interrogazioni IN esattamente in questo modo, ma alcuni permettono di avere l'equivalente se si fa clic su una opzione "ricerca avanzata", dove si può specificare che le parole cercate debbono essere nel titolo o in qualche altra parte specifica di un documento. Facciamo finta che esista la parola chiave IN solo per semplificare la spiegazione. In realtà, nel momento in cui scrivo, Google permette di effettuare una ricerca nei titoli con la parola chiave intitle:, perciò l'interrogazione intitle:boat restituisce pagine con "boat" nel titolo. Provate!

a	3-10
cat	1-3 1-7 3-7
dog	2-3 2-7 3-11
mat	1-11 2-11
my	1-2 2-2 3-2
on	1-9 2-9
pets	3-3
sat	1-8 3-12
stood	2-8 3-8
the	1-6 1-10 2-6 2-10 3-6
while	3-9
<bodyEnd>	1-12 2-12 3-13
<bodyStart>	1-5 2-5 3-5
<titleEnd>	1-4 2-4 3-4
<titleStart>	1-1 2-1 3-1

Figura 2. 7 *L'indice per le pagine web mostrate nella figura precedente, metaparole comprese.*

Vediamo come fa un motore di ricerca a rispondere in modo efficiente alla interrogazione dog IN TITLE per le tre pagine dell'esempio nelle ultime figure. Innanzitutto, estrae la voce di indice per "dog", che è 2-3, 2-7, 3-11. Poi (e forse non ve l'aspettavate, ma seguitemi un attimo) estrae le voci di indice sia per <titleStart> sia per <titleEnd>. Le informazioni estratte fin qui sono riepilogate nella Figura 2.8 (per il momento potete ignorare circonferenze e rettangoli).

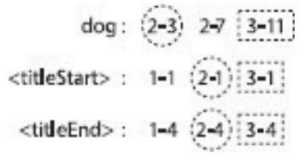

Figura 2.8 *Ed ecco come un motore effettua la ricerca di dog IN TITLE.*

Il motore di ricerca poi comincia a esplorare la voce "dog", esamina ciascuna delle sue occorrenze e verifica se sia o meno all'interno di un titolo. Il primo risultato per "dog" è l'elemento 2-3, racchiuso in una circonferenza, che corrisponde alla terza parola della pagina numero 2. Esaminando gli elementi di <titleStart>, il motore può scoprire dove inizia il titolo della pagina 2 (deve essere il primo numero che inizia con "2-"). In questo caso arriva all'elemento circolettato 2-1, il che significa che il titolo della pagina 2 inizia con la parola numero 1. Nello stesso modo, può scoprire dove termina il titolo della pagina 2. Non fa altro che esaminare gli elementi della voce <titleEnd>, alla ricerca di un numero che inizi con "2-", e si ferma quando incontra l'elemento circolettato 2-4. Dunque, il titolo della pagina 2 finisce con la parola 4.
Tutto quello che sappiamo fin qui è riassunto dagli elementi circolettati nella figura, che ci dicono che il titolo della pagina 2 inizia con la parola 1 e finisce con la parola 4, e che la parola "dog" è la parola 3. Il passo finale è facile: poiché 3 è maggiore di 1 e minore di 4, siamo sicuri che questa occorrenza della parola "dog" è effettivamente all'interno di un titolo; di conseguenza la pagina 2 sarà un risultato valido per l'interrogazione dog IN TITLE.
Il motore di ricerca può passare adesso all'occorrenza successiva di "dog", che è 2-7 (la settima parola della pagina 2), ma, poiché sappiamo già che la pagina 2 è un risultato valido, possiamo lasciar perdere questo caso e passare all'occorrenza successiva, 3-11, che è evidenziata da un rettangolo. Questo ci dice che "dog" è la parola 11 della pagina 3. Perciò cominciamo a esaminare gli elementi successivi alle posizioni attualmente circolettate nelle righe di <title-Start> e <titleEnd>, alla ricerca di elementi che inizino con "3-". (È importante notare che non dobbiamo tornare all'inizio di ogni riga: possiamo riprendere da dove eravamo arrivati nell'esplorazione per l'occorrenza precedente.) In questo semplice

esempio, "3-" è casualmente proprio il numero successivo in entrambi i casi: 3-1 per <titleStart> e 3-4 per <titleEnd>, evidenziati dai rettangoli per comodità. Ancora una volta dobbiamo stabilire se l'occorrenza di "dog" che stiamo esaminando, a 3-11, si trova all'interno di un titolo o no. Le informazioni nei rettangoli ci dicono che nella pagina 3 "dog" è la parola 11, mentre il titolo inizia con la parola 1 e finisce con la 4. Poiché 11 è maggiore di 4, sappiamo che questa occorrenza di "dog" è dopo la fine del titolo e perciò non è nel titolo, e di conseguenza la pagina 3 non è un risultato valido per l'interrogazione dog IN TITLE.

Quindi il trucco delle metaparole consente a un motore di ricerca di fornire risposta a interrogazioni sulla struttura di un documento in modo estremamente efficiente. L'esempio riguardava solo la ricerca all'interno dei titoli delle pagine, ma tecniche molto simili consentono di cercare parole nei collegamenti ipertestuali, nelle descrizioni delle immagini e in varie altre parti utili delle pagine web. E tutte queste interrogazioni possono trovare risposta con la stessa efficienza dell'esempio precedente: proprio come per le interrogazioni viste prima, il motore di ricerca non deve tornare a esaminare le pagine web originali, ma può rispondere consultando solo un piccolo numero di voci d'indice. E, cosa altrettanto importante, deve esaminare ciascun elemento di ogni voce una volta sola. Ricordate quello che è successo quando abbiamo completato l'elaborazione per la prima occorrenza
della pagina 2 e siamo passati al possibile risultato di pagina 3: anziché tornare all'inizio delle voci <titleStart> e <titleEnd> il motore può continuare la sua esplorazione dal punto in cui era arrivato. Questo è un aspetto fondamentale per l'efficienza delle interrogazioni IN.

Le interrogazioni sui titoli e altre "interrogazioni strutturali" che dipendono dalla struttura di una pagina web sono simili alle interrogazioni NEAR di cui abbiamo parlato prima, nel senso che gli esseri umani le utilizzano raramente, mentre i motori di ricerca le applicano internamente di continuo. Il motivo è lo stesso: i motori di ricerca vivono o muovono in base alla qualità delle loro capacità di ordinamento, e queste possono essere migliorate significativamente sfruttando la struttura delle pagine web. Per esempio, le pagine che hanno la parola "dog" nel titolo è molto più probabile che contengano informazioni sui cani delle pagine che contengono la parola "dog" solo nel corpo. Perciò, quando un utente invia la semplice interrogazione dog, un motore di ricerca può effettuare internamente una ricerca dog IN TITLE (anche se l'utente non l'ha richiesto esplicitamente) per trovare pagine che è più probabile parlino di cani, anziché semplicemente citare la parola incidentalmente.

I trucchi dell'indicizzazione e della ricerca di corrispondenza non sono tutto

Costruire un motore di ricerca non è impresa facile. Il prodotto finale è come una macchina di enorme complessità con molti ingranaggi e molte leve, che devono essere tutti montati correttamente perché il sistema sia utile. Perciò è importante rendersi conto che i due trucchi presentati in questo capitolo da soli non risolvono il problema di costruire un indice efficace per un motore di ricerca. Il trucco della posizione delle parole e quello delle metaparole però danno sicuramente il gusto di come i veri motori di ricerca costruiscono e usano gli indici.

Il trucco delle metaparole ha contribuito al successo di AltaVista là dove altri avevano fallito, cioè nel trovare corrispondenze in modo efficiente in tutto il Web. Lo sappiamo perché il trucco delle metaparole è descritto in una richiesta di brevetto del 1999 per gli Usa presentata da AltaVista, con il titolo "Constrained Searching of an Index". Nonostante la sua raffinata costruzione, l'algoritmo per le corrispondenze di AltaVista non è stato sufficiente per rimanere a galla nella turbolenza dei primi tempi del settore delle ricerche. Come già sappiamo, l'efficacia nel trovare corrispondenze è solo una faccia della medaglia: l'altra grande sfida è ordinare i risultati. E, come vedremo nel prossimo capitolo, la comparsa di un nuovo tipo di algoritmo di ordinamento è stata sufficiente a eclissare Alta-Vista e a portare Google in vetta.