Studio di euristiche per il miglioramento di algoritmi di ranking per il World–Wide Web

Riassunto della tesi di Laurea in Informatica di Marco Olivo, matr. n. 592150
Relatore: Dr. Sebastiano Vigna
Correlatori: Dr. Paolo Boldi, Dr. Massimo Santini
Anno Accademico 2002-2003

Uno dei problemi aperti dell'informatica e dell'*information retrieval* più affascinanti degli ultimi anni è quello di rispondere alle interrogazioni poste da esseri umani ai motori di ricerca.

I motori di ricerca si distinguono dalle basi di dati e dai sistemi di *information retrieval* tradizionali per diverse ragioni: il World–Wide Web (web) è immenso[1], raddoppia di dimensione ogni dodici mesi e si aggiorna con una frequenza impredicibile; inoltre i dati non sono strutturati e sono altamente eterogenei nel contenuto e nella qualità.

Il recupero di tali quantità di dati è reso difficile non soltanto dalla loro mole, ma anche da una serie di fattori legati all'affidabilità dei server su cui tali dati vengono ospitati e dalle reti attraversate per raggiungerli.

La studio dei motori di ricerca è relativamente recente, e le pubblicazioni in letteratura su questo argomento, per quanto esso sia molto dibattuto, tendono ad essere molto poche e soprattutto molto poco approfondite; tra i motivi che spingono i ricercatori che se ne occupano a non divulgare tutti i dettagli del loro lavoro vi sono forti interessi economici legati all'uso commerciale dei motori di ricerca.

Lo scopo di questa tesi è stata la costruzione di un motore di ricerca scalabile, preciso e soprattutto flessibile e che potesse competere con i motori di ricerca commerciali — se non per numero di pagine trattate, almeno per qualità dei risultati. Per raggiungere questi obbiettivi si è partiti da una implementazione accurata dei migliori algoritmi noti in letteratura, a cui è seguita una fase di affinamento basata da un lato su considerazioni di tipo ingegneristico e dall'altro da un attento confronto dei risultati ottenuti con quelli restituiti dai motori di ricerca commerciali in risposta a varie interrogazioni.

[1] Si calcola che, compressa, la porzione di web ad oggi raggiungibile occupi circa 50 terabyte, per un totale di oltre tre miliardi di pagine.

Lo schema seguito da un motore di ricerca nel momento in cui un utente inserisce un'interrogazione è, in prima approssimazione, il seguente. Anzitutto il motore opera una scrematura delle pagine che soddisfano l'interrogazione sottopostagli, in maniera tale da eliminare subito un numero significativo di pagine che probabilmente non interessano all'utente; ciononostante, a causa delle dimensioni del web, il numero di pagine rimanenti è in generale molto elevato. Pertanto, l'ordinamento relativo di tali pagine è forse la cosa più importante che un motore di ricerca si deve occupare di fornire, in quanto è proprio tale ordinamento a dare all'utente la sensazione di aver trovato quel che stava cercando.

Una tecnica possibile, ad esempio, è quella di preferire le pagine a cui si riferiscono molte altre pagine, oppure quelle in cui i termini dell'interrogazione compaiono in posizioni ravvicinate. Altre tecniche, più sofisticate, si basano sulla struttura di interconnessione tra le pagine del web.

Nei primi capitoli vengono anzitutto presentati ed analizzati alcuni algoritmi noti proposti da tempo in letteratura e che si reputa vengano utilizzati dai più famosi motori di ricerca commerciali per decidere l'ordine di presentazione dei risultati. Alcuni di questi algoritmi sono stati utilizzati nel motore di ricerca che costituisce lo scopo di questa tesi.

Nei capitoli successivi si analizzano alcune euristiche tese anzitutto a migliorare la qualità dei risultati e che cercano di fornire risposte più mirate e precise alle interrogazioni sottoposte al motore di ricerca, in maniera da emulare e, per quanto possibile, migliorare lo stato attuale dell'arte, rappresentato in questo caso dai motori di ricerca commerciali; in seconda battuta, vengono presentate altre euristiche volte a diminuire i tempi di risposta alle interrogazioni, in maniera da rendere realistico l'utilizzo del motore di ricerca sviluppato.

Infine, è stata studiata ed implementata una tecnica per aggregare i risultati dei vari algoritmi tra di loro in maniera efficiente ed in modo altamente parametrico, così da rendere possibile la sperimentazione dell'impatto dei vari algoritmi sulla qualità della risposta.