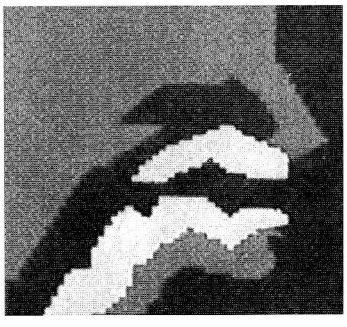
La ricerca ed il recupero dell'informazione

Verso la desktop library?

La biblioteca tradizionale dovrà diventare una *desktop library*, una biblioteca elettronica a cui si accede attraverso la rete. Questo ha implicazioni sul tipo di informazione da rendere disponibile nonché sugli strumenti per ricercarla e recuperarla. I sistemi automatizzati noti come *information retrieval* (Ir), in questa evoluzione della biblioteca, sono sempre più importanti. Gli sviluppi tecnologici più recenti non possono quindi essere ignorati dai bibliotecari.

Sinteticamente, si può dire che dall'information retrieval si è passati al Ftr (inteso dapprima come Free text retrieval e poi come Full text retrieval) e ai Nidr (Network information discovery and retrieval). La nuova generazione di sistemi di information retrieval, o genericamente, dei motori di ricerca (*search engine*), tende infatti a combinare le possibilità della ricerca con la gestione e la fornitura del documento a testo pieno. Oltre a ciò, l'architettura client-server ha reso disponibili interfacce per gli utenti più flessibili per una molteplicità di fonti informative in rete.



Recupero dei documenti elettronici

Gli attuali sistemi di recupero dell'informazione elettronica che gestiscono anche testo pieno (Ftr) hanno esteso le possibilità di ricerca al *contenuto* dell'intero documento. Dal più semplice al più sofisticato, essi consentono la ricerca sull'intero documento (testo libero o testo pieno) e la ricerca estesa ad immagini e suoni. Questa estensione di funzionalità ha aumentato la complessità dei sistemi di ricerca che sono oggi disponibili in commercio e che consentono non solo la semplice ricerca booleana ma anche la ricerca in linguaggio naturale (Natural language processing-Nlp); alcuni Ftr consentono inoltre la gestione dei documenti. Questi sistemi di ricerca vanno usati criticamente, cercando di sfruttare tutto quello che sanno fare meglio e non sottoutilizzandoli, cioè cercando di fargli fare quello che fanno meno bene. Ad esempio quello che questi sistemi non fanno ancora molto bene è la presentazione a video dei documenti. I primi sistemi Ftr consentivano la gestione del documento come immagine, soluzione che dava possibilità migliori di visualizzazione, oltre che garanzia di sicurezza sulla correttezza del testo, ma limitava le possibilità dell'indicizzazione, legata ad una citazione bibliografica che era una descrizione breve del documento, collegata all'immagine con un numero di riferimento. Successivamente, se il testo non veniva trattato come immagine, la scelta alternativa è stata di gestire testi in Ascii o testi formattati con Sgml. Nel caso di testi in Ascii i documenti sono visti come unico file sequenziale su cui si costruiscono gli indici di tutte le parole significative: è il caso della ricerca a testo libero. Nel caso di testi formattati (cioè con una struttura interna e non come unico archivio sequenziale), il testo è considerato e gestito o come testo narrativo o testo con tavole ed immagini collegate e l'indicizzazione segue la formattazione del documento: è il caso del testo pieno.

Quello che lo sviluppo dei nuovi motori di ricerca Ftr sa fare proprio bene è la costruzione di indici. Usando varie tecniche, si arriva anche alla classificazione automatica di parti del documento. È un tema di attualità ed è anche il tema che suscita la maggiore diffidenza dei bibliotecari, giusta-

Oltre Boole

I sistemi di recupero dell'informazione (IR) usano vari modelli. Il più noto è il modello di Boole. La ricerca in questo modello è un processo interattivo che avviene attraverso un colloquio tra utente e calcolatore, poiché, per i non esperti che non sono aiutati da un intermediario, è quasi impossibile formulare senza errori ed approssimazioni successive, la domanda che darà i risultati giusti. La ricerca avviene in tre fasi:

1) il termine scelto (concetti atomici o parole chiave) viene confrontato con i termini di tavole memorizzate nel calcolatore (*inverted file*);

2) utilizzando gli operatori logici sono applicate operazioni di insieme sui risultati ottenuti;

3) viene visualizzata la citazione del documento che contiene i dati selezionati con la successione prevista dagli operatori. Stairs è il più popolare di questi sistemi di ricerca, detti post-coordinati. Tra gli esperti è ormai assodato che il modello booleano è il peggior modello possibile per il recupero dell'informazione. Però da oltre trenta anni è usato dalle biblioteche, anche le più evolute, perché sperimentato e facile da usare.

Le alternative al modello di Boole sono state sviluppate in due linee di tendenza. I sistemi più avanzati si basano su un approccio statistico ed usano *modalità di ricerca fuzzy* che è l'opposto della ricerca esatta (o *match esatto*) poiché i risultati sono estesi a documenti in qualche modo collegati alla richiesta e si ha una risposta elencata in ordine di rilevanza; altri sistemi sono basati sull'intelligenza artificiale e sono stati molto popolari negli anni Ottanta ma sono ad un livello di sviluppo ancora insufficiente.

Il primo gruppo di sistemi di ricerca sono quelli basati sul modello vettoriale e la frequenza statistica di parole. La struttura centrale dei dati è il vettore. Le combinazioni di parole e documenti sono memorizzate come matrici o tavole ed ogni documento, o parola chiave, è rappresentata come un vettore, orizzontale per i documenti e verticale per le parole chiave. Ogni combinazione parole-documento ha un valore tra 0 e 1 e misura il valore dell'informazione di quella parola in quel documento. Più una parola è presente nel minor numero di

mente gelosi delle loro competenze professionali tradizionali. Premesso che occorre naturalmente cautela nell'affidarsi ai nuovi sistemi di classificazione automatica, c'è però anche da considerare che nel campo della ricerca di documenti a testo pieno i bibliotecari non possono pensare di trasportare acriticamente la loro esperienza, acquisita nell'indicizzazione di citazioni bibliografiche di documenti. Le difficoltà e le esigenze di ricerca sono completamente diverse. Non c'è nessuna esperienza precedente a cui rifarsi. Le biblioteche automatizzate indicizzano solo dati bibliografici (cioè citazioni bibliografiche che rimandano a documenti cartacei) che sono memorizzati negli opac e nelle banche dati; la ricerca avviene su campi brevi e strutturati; per il recupero dei dati usano quasi esclusivamente la ricerca booleana basata sulla ricerca esatta di parole chiave. Con il testo pieno ed i documenti multimediali, composti di entità diverse che comprendono anche suoni ed immagini, le difficoltà sono ben diverse. Andranno adattate le soluzioni realizzate per le problematiche tipiche della ricerca bibliografica alle nuove difficoltà che sono proprie della ricerca di documenti a testo libero e di documenti multimediali. Sarà opportuno studiare bene le nuove possibilità degli Ftr per usarle con creatività, integrando vecchie conoscenze e nuovi strumenti di lavoro. I programmi che utilizzano completamente la struttura dei documenti Sgml potranno in futuro rendere superata l'attività di indicizzazione operata da parte dei bibliotecari cioè sia la catalogazione che la compilazione di bibliografie.

Un'area di attività molto promettente è nell'uso dell'ipertesto, per aumentare le possibilità di ricerca o aggiungere collegamenti dal testo alle immagini. Uno strumento prima considerato alternativo all'information retrieval, come l'ipertesto, è ora perfettamente integrato nella ricerca dell'informazione, che si è così arricchita delle possibilità di navigazione. Concludendo, un Ftr deve effettuare:

— l'indicizzazione automatica del linguaggio naturale in testi o formattati o liberi;
— la ricerca con "and", "or", "not" e con il troncamento per comprendere tutte le variazioni del lemma ed i sinonimi oltre alle parole collegate al termine di ricerca dal tesauro o da reti semantiche;
— la ricerca non solo di parole ma anche di frasi nello stesso paragrafo o nell'intero testo con l'elencazione dei risultati in ordine di rilevanza;
— la navigazione tra legami ipertestuali;
— la visualizzazione selettiva di sezioni del documento ed il contesto in cui si trovano i termini di ricerca;
— il passaggio a stampa oppure la trasmissione del file per rielaborazioni successive;
— la memorizzazione della strategia di ricerca e/o la possibilità di costruire un profilo di ricerca per riutilizzarlo in altre sessioni.

Recupero dell'informazione in rete

Un problema per la ricerca dell'informazione nei documenti elettronici è quello connesso alla necessità di effettuare la ricerca in più calcolatori. Una biblioteca elettronica è infatti costituita da una molteplicità di server in rete e non è memorizzata in un solo calcolatore.

L'accesso alle biblioteche elettroniche è assicurato dai Nidr che organizzano l'informazione in rete e ne consentono il recupero. I Nidr utilizzano l'architettura client/server ed i protocolli di comunicazione. L'utente finale deve dotarsi di un "client" (cioè un personal computer equipaggiato di un software client) che farà la ricerca per lui. Una stessa richiesta, trasmessa da un client a vari server, riceverà più risposte, che il client presenterà all'utente finale come se la ricerca fosse stata effettuata in un solo archivio di documenti. Nell'architettura client/server la ricerca è effettuata dai calcolatori e sono quindi utilizzati protocolli standard di comunicazione tra calcolatori. Per i bibliotecari i protocolli più importanti per la ricerca ed il recupero dell'informazione in rete sono SR/Z39.50 e Http.

SR/Z39.50 è un protocollo standard internazionale per la ricerca realizzato negli anni Ottanta, Http è il protocollo usato da World wide web ed è molto più recente. Le installazioni di Z39.50 sono limitate, molto più diffuse le installazioni di Http, legate al successo di Web. Alcune differenze tra i due protocolli sono: ➤

documenti più grande è il valore dell'informazione di quella parola. Nel modello vettoriale sia le richieste dell'utente che i documenti sono "mappati" come vettori, comparati al momento della ricerca e i documenti sono disposti in ordine di probabilità di rilevanza ad una richiesta.

Una tecnica per aumentare la precisione della risposta è la retroazione della rilevanza (*relevance feedback*). Data l'esistenza di vettori dei documenti è possibile aggiustare automaticamente la domanda.

Un nuovo approccio dei sistemi vettoriali è l'applicazione dell'algoritmo genetico. Ogni individuo in una popolazione ha una combinazione di caratteri, ereditata dai suoi genitori. Allo stesso modo un programma che usa l'algoritmo genetico formula una serie di possibili soluzioni sulla base dei risultati della ricerca. L'utente deve indicare quali documenti sono i più rilevanti e la rappresentazione di questi documenti è quella di successo, che diventa la base per formulare la nuova richiesta.

I sistemi del secondo gruppo utilizzano i risultati dell'intelligenza artificiale e costruiscono delle basi di conoscenza che consentono una ricerca esperta. Un esempio di sistemi di questo modello è Topic. Questo sistema di ricerca è formato da quattro componenti: il modulo di indicizzazione, il motore di ricerca, l'interfaccia utente e la parte semantica. La parte semantica è la particolarità di Topic ed è appunto una collezione di "topici" come strumento di rappresentazione della conoscenza, simile ad un tesauro, implementato sulla base di regole. Ma mentre i tesauri o gli schemi di classificazione si riferiscono a concetti, i topici di Topic sono proprietà dei documenti, sono costruzioni ad hoc che servono per l'euristica della ricerca. I concetti nella parte semantica di Topic sono sistemati in alberi semantici o grafi, in cui le stringhe che si trovano nei documenti possono essere considerate come le foglie. L'occorrenza di tali stringhe, usando gli operatori di Boole o operatori di prossimità, è considerata prova che pesa per la rilevanza del concetto più alto nel grafo. I topici, o concetti più alti, sono costruiti da esperti, predisponendo una base di conoscenza al sistema. Successivamente l'utente del sistema può aggiungere anche i suoi topici, accessibili o no ad altri utenti. Il problema di Topic, e di tutti i sistemi fondati su basi di conoscenza, è che funziona bene solo in piccole collezioni di dati.

Finding aids

Per assistere l'utente nella individuazione dei documenti elettronici in rete, la Biblioteca dell'Università di Berkeley ha promosso il Finding aid project (http://sunsite.berkeley.edu/FindingAids). I finding aids sono documenti usati per descrivere, controllare e consentire l'accesso a collezioni di documenti elettronici, cioè sono le bibliografie della biblioteca del futuro. Come le bibliografie, i finding aids sono documenti che citano i documenti primari. Utilizzano lo standard Sgml che assicura che i finding aids rimarranno validi nel tempo, anche se cambierà hardware e software. Attualmente sono circa 200 le liste rese disponibili dalle biblioteche che aderiscono al progetto (tra cui la Library of Congress, la National Library di Australia, l'Università di California ed altre) per un totale di circa 5.000 pagine.

— Http lavora senza realizzare la connessione tra i calcolatori mentre Z39.50 attiva e mantiene la connessione fino alla fine della sessione di ricerca;
— Http è molto semplice. Essenzialmente consente l'accesso ad un documento con una Url. Z39.50 offre molti altri servizi;
— il client Z39.50 è meno passivo di quanto lo sia il client Http in cui è l'utente finale che formula la domanda da sottoporre al server remoto;
— Z39.50 è un protocollo specializzato per le biblioteche mentre Http è un protocollo generico.

Www e Z39.50 possono essere integrati per ricercare banche dati bibliografiche e banche dati a testo pieno e per consentire l'accesso a server remoti. Www consente di realizzare collegamenti ipertestuali, controllando l'interfaccia locale. Con una stessa interfaccia, che può essere anche piuttosto semplice, sarà possibile per l'utente abituale di una biblioteca interrogare fonti locali o fonti remote, senza differenze. Z39.50 consente di fare ricerche simultanee in una molteplicità di banche dati ma la visualizzazione dei documenti può essere quella di Www. In altre parole i due protocolli non sono in competizione ma si integrano per una migliore funzionalità.

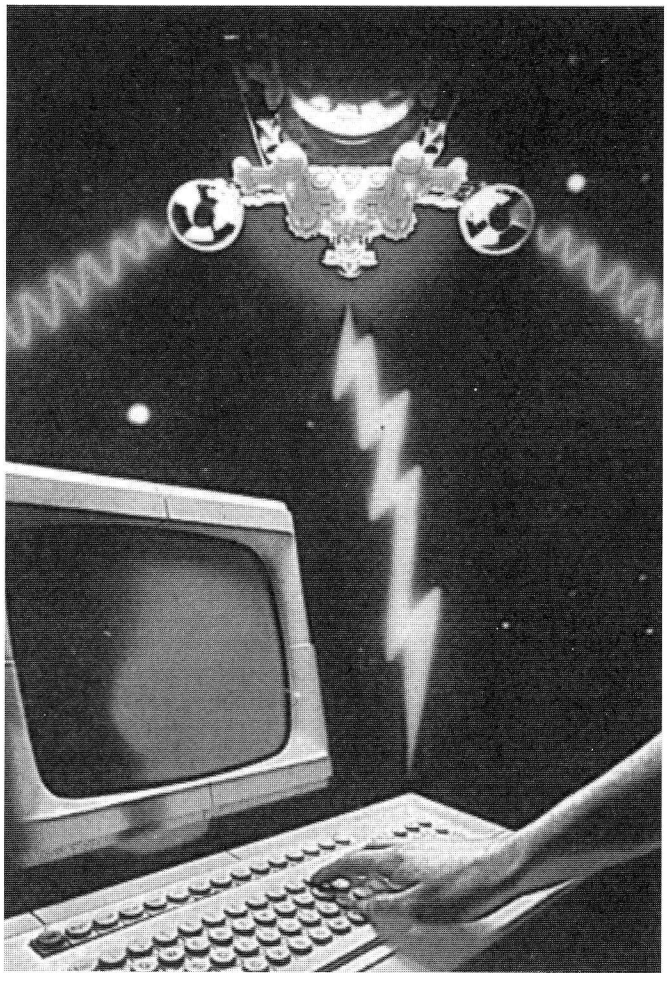
Altri strumenti di ricerca in rete sono gli indici automatici ai siti Web, noti come motori di ricerca (*search engine*). Nella prima fase dell'organizzazione delle fonti informative in rete, l'attività prevalente era navigare tra i siti Web, alla ricerca di quello che poteva essere interessante ma, attualmente, è sempre più frequente la ricerca di termini per una risposta veloce all'esigenza di sapere quel che c'è. Per l'estrema variabilità delle fonti in Internet, infatti, nessuna lista, in linea o su carta, riesce ad assicurare il controllo di tutto quel che è disponibile.

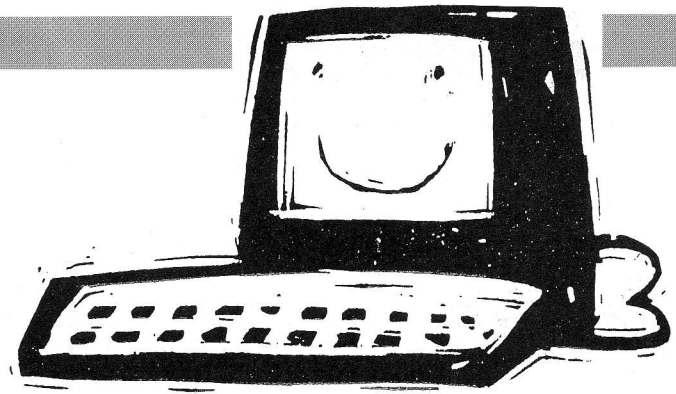
Questi motori di ricerca, comparati agli Ftr o anche ai Nidr, sono davvero molto limitati. I termini di ricerca, immessi dall'utente, sono semplicemente comparati ai termini delle pagine contenute nei siti Web. Non c'è alcun controllo in tesauri o dizionari controllati, non c'è la possibilità di combinare i risultati o di limitare la ricerca a certi campi, non si possono avere i risultati ordinati per rilevanza. Alcuni motori di ricerca offrono la possibilità di utilizzare gli operatori: "and", "or", "adjacent", "near", "not". Altri, più sofisticati, consentono la ricerca di frasi o di concetti.

La precisione dei motori di ricerca è quindi un problema. Per certe ricerche alcuni motori sono meglio di altri. Ad esempio Webcrawler è migliore quando il termine di ricerca è oscuro, poiché indicizza ogni parola dei siti Web. La precisione in questo caso potrà essere migliorata aggiungendo più termini. Lycos invece indicizza solo il titolo e le 100 parole più significative del sito Web. È quindi più veloce ma recupera meno informazioni.

L'interfaccia utente

Le caratteristiche più importanti dell'evoluzione dei sistemi di recupero dell'informazione sono lo sviluppo della ricerca in linguaggio naturale e l'aver spezzata la catena Dati-Indici-Interfacce che caratterizza i sistemi di automazione centralizzati in grandi calcolatori (*mainframe*) sostituita dai protocolli di comunicazione tra calcolatori. Gli Ftr ed i Nidr differiscono per le funzionalità possibili per l'utente. Le sofisticate ricerche degli Ftr non sono possibili per la navigazione e la ricerca di fonti in rete. I Nidr infatti offrono accesso ad un vasto numero di sistemi ma consentono una scelta più limitata di opzioni. D'altro canto, l'uso di tecniche sofisticate deve essere utilizzato con attenzione: i risultati, nella maggior parte delle ricerche d'informazione, po-





trebbero essere l'opposto di ciò che serve all'utente. C'è da bilanciare l'accettabilità dei risultati con la perdita di funzionalità sofisticate di ricerca oppure bisogna consentire all'utente la scelta di utilizzare o l'uno o l'altro dei sistemi di ricerca.
Il client dei Nidr si occupa in particolare dell'interfaccia utente ed è un'unica interfaccia per una pluralità di applicazioni. Rispetto al client, sarà necessario evitare una molteplicità di protocolli di comunicazione e formati di dati difformi. Questo implica che bisogna utilizzare il più possibile le soluzioni standard. Altri problemi di interfaccia per gli utenti riguardano la difficoltà di presentare nello schermo tutte le possibili azioni consentite, dati i limiti di spazio dello schermo. Alcuni Nidr utilizzano una interfaccia grafica (Graphical user interface - Gui). In particolare uno standard de facto è Wimp (*Windows, icons, mouse, pull-down up menus*). I vantaggi di Wimp sono:
— è facile da usare, sia con Mac che con Dos;
— c'è standardizzazione di chiavi di ricerca;
— consente sia una ricerca semplice che una complessa;
— mette più informazione sullo schermo con i bottoni.

I problemi ancora da risolvere

Gli attuali problemi del processo di ricerca dell'informazione riguardano l'atteggiamento dei bibliotecari. Il migliore motore di ricerca non serve se l'informazione disponibile è scarsa o è impossibile l'accesso ai documenti recuperati nella ricerca. Infatti il rischio è che l'utente possa essere incoraggiato ad usare i sistemi di ricerca da una buona interfaccia, ma con gli stessi risultati di ora.
Cioè l'utente recupererà scarsa informazione e, una volta individuato un documento, avrà difficoltà ad ottenerlo in lettura. È quindi di cruciale importanza costruire un paradigma della biblioteca elettronica, che deve considerare i seguenti aspetti:
— facilitare l'accesso all'informazione e non solo alle descrizioni di documenti. Migliori capacità di accesso ai documenti fanno parte dello sviluppo degli opac e quindi, quando possibile, devono essere digitalizzati i documenti;
— sperimentare possibili modelli di copyright che non ostacolino l'accesso ai documenti elettronici;
— avere abbastanza informazione nella citazione da consentire all'utente di decidere se il documento è utile o no (ad esempio aggiungendo gli indici dei capitoli);
— utilizzare sistemi che consentono la ricerca in linguaggio naturale, perché sono più efficaci per i documenti a testo pieno;
— applicare interfacce Gui per gli utenti, capaci di visualizzare sia testo che immagini.