L'innovazione tecnologica nella documentazione

La mediazione documentaria come intersezione di linguaggi diversi

di Carla Basili

ELEMENTI DOMINANTI NELLO SCENARIO TECNOLOGICO

Tendenze di sviluppo nell'hardware e nel software

Le attività di ricerca e sviluppo nel settore dell'IT (Information Technology), come vedremo, si diramano in più direzioni, ma il substrato tecnologico alla base di ogni sviluppo — volendo escludere dalla nostra trattazione dispositivi quali le macchine per fotocopie o i proiettori di diapositive — è il computer inteso come sistema di componenti hardware e software.

L'evoluzione del computer, per sé, influenza pertanto ogni altra linea di sviluppo tecnologico, sia questa la multimedialità, l'intelligenza artificiale o l'elaborazione in rete. Le tendenze generali oggi in atto nella tecnologia di base si possono sintetizzare come segue:

— costante diminuzione di dimensione e costo dei componenti, cui fa riscontro un costante aumento della velocità di elaborazione;[1]

— progressivo potenziamento delle capacità di elaborazione, che ispira lo sviluppo di applicazioni sempre più sofisticate;

— in particolare, il computer viene utilizzato per progettare nuovo hardware e software, in un effetto a catena, secondo il quale ogni nuova generazione di sistemi facilita la progettazione della generazione successiva; questo processo continuerà fino all'esaurimento di tutte le alternative tecnologiche e fino al limite delle possibilità fisiche offerte dall'hardware ed il raggiungimento di tali limiti non appare imminente.

Elaborazione parallela

La natura dei problemi computazionali che oggi si affrontano nelle applicazioni avanzate — quali riconoscimento di forme o ricerca di un termine in vasti depositi di documenti elettronici testuali — coinvolge molto spesso operazioni relativamente semplici, applicate ad una grande quantità di dati. Un elaboratore convenzionale — progettato, cioè, secondo il modello di von Neumann — con un'unica unità di elaborazione, esegue una operazione alla volta, in maniera seriale. Quando l'algoritmo richiede di applicare una stessa operazione a milioni di elementi, il modello di elaborazione seriale è impraticabile [Hillis 1986]. Esempio: dati due nodi A e B in un grafo esteso, trovare il percorso più breve che li congiunge. Le operazioni coinvolte in questo algoritmo sono semplici e ripetitive, ma necessitano di essere attivate su tutti i nodi del grafo compresi tra A e B.

Una efficace alternativa all'architettura seriale è l'architettura parallela, dove un elaboratore è dotato di più (a volte migliaia) unità di elaborazione che operano in parallelo. Risolvere un *problema* in parallelo significa decomporlo in attività che cooperano contemporaneamente. Esistono due criteri generali di scomposizione: scomposizione in domini e scomposizione funzionale. La scomposizione in domini opera sui soli dati: un problema viene suddiviso in modo tale che la stessa attività venga esplicata simultaneamente su diversi insiemi di dati. Questo tipo di decomposizione realizza la forma più semplice di parallelismo. La decomposizione funzionale vede la soluzione di un problema scomposta in attività indipendenti. È la forma più complessa di parallelismo, che richiede di stravolgere il modo di concepire un algoritmo.

Le strategie di risoluzione in parallelo adottano di frequente

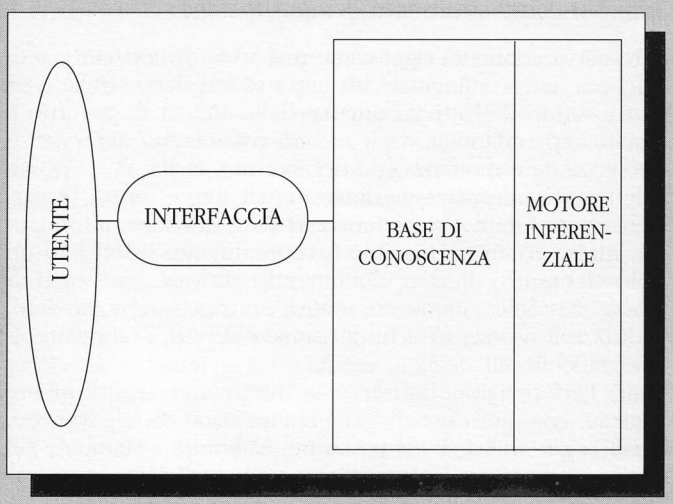
[1] Il computer, arrivando alla soglia del costo personale, tende a divenire una tecnologia sempre più pervasiva, entra in casa. Ciò ha grande rilevanza sia per la concezione di "società dell'informazione", sia per l'autonomia da parte dell'utente finale (*end-user computing*).

entrambi i criteri, pertanto un generico algoritmo struttura il proprio parallelismo su due fronti: dati e funzioni. Una applicazione in campo documentario del modello di elaborazione parallela (a scomposizione in domini) è il sistema WAIS (Wide Area Information Server), un sistema ideato e sviluppato da quattro organizzazioni: Dow Jones & Company Inc., Thinking Machines Corporation, Apple Computer e KPMG. La consultazione di diverse fonti di informazione implica la formulazione della stessa interrogazione più volte e con diversi linguaggi di interrogazione. Uno degli obiettivi primari di WAIS è quello di rendere trasparente l'accesso ad una molteplicità di database locali e remoti. Con WAIS l'utente sceglie un insieme di fonti e formula una interrogazione che viene attivata in parallelo sulle fonti scelte. Il sistema automaticamente inoltra la richiesta a tutti i server selezionati, ordina e consolida in un unico insieme i risultati provenienti dalle diverse fonti, così da facilitarne la manipolazione da parte dell'utente.

Intelligenza artificiale

La ricerca nel settore dell'intelligenza artificiale si articola in due filoni:
— la modellazione e la comprensione dell'intelligenza umana (filone teorico);
— la soluzione di problemi specifici che richiedono elaborazione di conoscenza, attraverso lo sviluppo di "sistemi esperti" (filone pragmatico o euristico).
Il filone teorico tende a definire l'intelligenza, a comprendere i processi cognitivi, a rappresentare conoscenza e significato attraverso reti semantiche, regole di produzione, reti neuronali etc... Il filone pragmatico o euristico dell'intelligenza artificiale applica i risultati del filone teorico a casi specifici, dando luogo a sistemi noti come "sistemi esperti". La configurazione classica di sistema esperto (Fig. 1) comprende:
— una interfaccia: è il modulo che gestisce il colloquio tra mondo esterno e sistema;
— una base di conoscenza: contiene tutti i fatti noti al sistema, le leggi che li governano e le regole per la loro manipolazione;
— un motore inferenziale: è la componente che consulta e manipola la base di conoscenza; è il nucleo operativo del sistema.

Fig. 1



Al di là delle definizioni di carattere tecnico, un sistema esperto tende a replicare in un computer un "comportamento intelligente" nella soluzione di un particolare problema. Vickery ci ricorda che anche i calcolatori elettronici della prima generazione furono chiamati "cervelli giganti" o — peggio ancora — "cervelloni", il che oggi fa sorridere, in quanto i criteri di valutazione sono cambiati [Vickery 1990]. Quei calcolatori ed i loro successori operano in base a programmi "convenzionali" che agiscono su dati attraverso algoritmi. Un algoritmo è una specificazione dettagliata di operazioni che devono essere svolte secondo una successione predefinita, per produrre un risultato predefinito.
Un sistema intelligente, invece, opera su conoscenza attraverso deduzioni. Ovviamente, come ogni programma di elaborazione elettronica, un programma intelligente opera comunque con dati ed algoritmi; ciò che lo differenzia dai programmi di elaborazione tradizionali è la natura dei dati (ove prevale la rappresentazione di associazioni semantiche) e la natura deduttiva dell'algoritmo. La conoscenza può essere rappresentata secondo svariati modelli, quali una rete di concetti, un insieme di regole, un sistema di oggetti. L'applicabilità dei sistemi esperti nel campo della documentazione è limitata a servizi tecnici, quali la catalogazione, la classificazione, la gestione della raccolta o comunque è circoscritta ad ambiti ove è possibile individuare un insieme definito di regole.

Ipertesti e multimedialità: ipermedia

Logica lineare vs *logica associativa*

Ipertesti ed ipermedia sono strumenti per organizzare informazione in un elaboratore elettronico in modo che sia consultabile in maniera associativa. Questi strumenti si differenziano da forme di rappresentazione dell'informazione a noi più familiari proprio in quanto si fondano su una logica associativa.
La natura dei supporti sui quali tradizionalmente è stata registrata la conoscenza — i libri — è lineare: un libro è organizzato in capitoli, sezioni, paragrafi, sottoparagrafi, che generalmente sono consultati sequenzialmente. Un caso diverso è il materiale di tipo enciclopedico, dove la conoscenza è registrata in unità informative indipendenti, accessibili in maniera diretta e dove è frequente il rinvio ad altra conoscenza, attraverso formule del tipo "vedi" o "vedi anche". Ancor più del libro, l'elaborazione elettronica nasce e si mantiene per decenni rigidamente sequenziale (si pensi alla logica di un diagramma di flusso che descrive un algoritmo, scomponendolo in una successione di passi elementari). Si ricordi che l'elaborazione parallela è uno sviluppo tecnologico piuttosto recente.
Tutto ciò costituisce una forzatura alla natura della mente umana, che invece è tendenzialmente associativa, come ricor- ➤

da Vannevar Bush, consigliere scientifico del presidente Roosvelt, che nel 1945 nel suo articolo seminale *As we may think* (apparso sul numero di luglio di "The Atlantic Monthley") afferma che: "La mente umana... opera in base ad associazioni. Dinanzi ad un elemento informativo, istantaneamente salta ad un altro elemento che gli viene suggerito per associazione di pensiero, seguendo una intricata ragnatela (web) di sentieri percorsi dalle cellule cerebrali. Non si può sperare di riuscire a duplicare appieno, con mezzi artificiali, questo processo mentale, ma certo si può studiarlo con profitto.
La più importante idea che si può trarre da questa analogia concerne la selezione. La selezione per associazione, contrapposta a quella per indicizzazione, può essere meccanizzata".
Nello stesso articolo Bush presenta l'idea di una macchina — il Memex — per scorrere testi e prendere appunti in un voluminoso sistema di testi e di grafici in linea; la macchina Memex contiene una vasta libreria, corredata di note personali, oltre che disegni e fotografie memorizzati nella forma di microfilm. Bush, auspicando per gli anni del dopoguerra un grande impegno per meccanizzare il sistema della letteratura scientifica, avverte nello stesso tempo l'esigenza di forme più naturali di indicizzazione.
Dovranno passare due decenni prima che la tecnologia per realizzare l'idea di Bush sia disponibile. Solo nel 1965, infatti, Ted Nelson decide di attuare in concreto i dettami teorici di Bush e conia il termine "ipertesto" *per indicare l'idea di un testo scritto e letto in maniera non sequenziale, dove il contenuto non è confinato e limitato dalla organizzazione del supporto fisico di registrazione.*

Iperdocumento

Un iperdocumento è una rete di nodi interconnessi tramite collegamenti (link). Ogni nodo contiene un blocco di informazione, ed è accessibile in maniera diretta. Questo per quel che riguarda la rappresentazione dell'informazione all'interno del sistema di elaborazione. Ciò che invece è visibile all'utente sullo schemo del computer è un testo, che presenta qua e là dei termini o delle frasi sottolineate. La modalità di consultazione dell'informazione ipertestuale è la *navigazione*, che vede l'utente passare da un nodo all'altro, seguendo i collegamenti associativi contenuti nei singoli nodi. La navigazione esclude la presenza di un percorso predeterminato nella fruizione dell'informazione: l'utente è libero di scegliere di volta in volta le connessioni associative che lo interessano, attraverso l'uso di un dispositivo di "point and click" quale il *mouse*.
Oltre che come testo sottolineato o evidenziato, i collegamenti tra nodi possono essere rappresentati in forma iconografica, per esempio come pulsanti attivabili.
Un documento ipermediale è un iperdocumento multimediale ed il termine "ipermedia" include quindi quello di "ipertesto". Il termine "ipermedia" si riferisce alla creazione, presentazione e all'accesso ad informazione registrata in un computer come una ragnatela di unità associate, utilizzando la combinazione di differenti media quali testo, grafici, suoni, immagini in movimento.
Si noti che i termini ipertesto e ipermedia indicano concetti e non prodotti.

Connettività

Internet come strumento di lavoro

Accedere ad Internet ed utilizzare i suoi servizi è facile; non ci sono restrizioni sul tipo di calcolatore che può essere connesso, il software per l'uso è reperibile gratuitamente dalla rete stessa, la modalità di interazione è semplice, avvalendosi di pulsanti attivabili mediante dispositivi "point-and-click" (*mouse*). Questo insieme di facilitazioni colloca la rete tra gli strumenti della scrivania, accanto al telefono, alla penna, al sistema di videoscrittura.

Internet come strumento di comunicazione

Una delle applicazioni più utilizzate, fin dalle origini di Internet, è la posta elettronica, strumento che si è rivelato comodo quanto il telefono e anzi, nel caso di grandi differenze di fuso orario, addirittura più economico e funzionale. Attorno alla posta elettronica, poi, si è sviluppato il servizio di conferenza elettronica o lista di discussione, che costituisce una risorsa particolarmente originale della rete.
La lista è sede di argomentazioni specializzate ad un ambito disciplinare, il che implica che quanti vi partecipano condividono un certo livello di competenza in quel settore. Molto spesso, inoltre, emergono dei "leader" nella discussione, che "sostengono la conversazione", ed in generale sono dei veri esperti nell'argomento. Ciò determina un alto livello di qualità nei messaggi che circolano entro la conferenza, i quali possono assumere la forma di veri e propri articoli (che vanno sotto il nome di *discussion papers*). La validità del contenuto informativo delle liste, insieme con l'esigenza di mantenere traccia dell'evolvere del discorso, ha dato luogo ad un insieme di funzioni di corredo al servizio di lista, che realizzano sia l'archiviazione dei messaggi, sia l'interrogazione secondo alcuni campi di ricerca. In questo modo è possibile avere l'elenco ed il testo completo di tutti i messaggi scambiati entro la conferenza a partire, per esempio, da una certa data e relativamente ad un certo argomento [Basili 1995].

Internet come strumento di informazione

Internet si configura oggi come una sorta di notiziario, o di bacheca, intercontinentale, su tutti i settori dello scibile e su tutti i settori dell'attività umana, dalle attività di governo a quelle imprenditoriali, dagli accordi commerciali ai servizi di promozione e monitoraggio del mercato. Nella rete è possibile trovare informazione di ogni contenuto e forma. In particolare, in Internet si preferisce parlare di risorsa informativa, intesa come un insieme (eventualmente ridotto ad un solo elemento) di: dati, documenti, software, messaggi di posta elettronica, immagini, grafici, immagini in movimento, indirizzi di risorse, elenchi di risorse e servizi, riferimenti bibliografici [Basili 1995].
Tutta l'informazione in rete è — ovviamente — in formato digitale, con tutti i benefici che scaturiscono da tale formato, quali la possibilità di memorizzare, elaborare e stampare sul proprio calcolatore l'informazione "catturata" dalla rete.

Internet strumento di documentazione?

Il volume dell'informazione in rete, pur notevole, è relativamente esiguo se confrontato con il volume dell'informazione stampata, di cui siamo abituati ad usufruire quotidianamente. Nel caso dell'informazione pubblicata, tuttavia, sono stati costruiti strumenti che consentono di utilizzare effettivamente questa mole di conoscenza e che nel loro complesso costituiscono la *mediazione documentaria*. L'obiettivo dell'apparato documentario, come è noto, è permettere di individuare e reperire tutta e sola l'informazione utile ad una specifica esigenza informativa.
Se analizziamo l'offerta dell'informazione in rete, invece, troviamo che questa risulta:
— priva di meccanismi di coordinamento del ciclo di vita dell'informazione, atti ad assicurare procedure universali di pubblicazione, di aggiornamento e di cancellazione della singola risorsa; da ciò deriva una caratteristica molto "scomoda" dell'informazione in rete: la sua volatilità; allo stesso modo non esiste un meccanismo di validazione e di certificazione di qualità;
— non organizzata, né strutturata; numerosi sono i tentativi di organizzare l'informazione in rete sia di matrice pratico/tecnologica — quali le guide all'informazione in Internet, i *subject trees* nel *gopher*, le guide attive e le biblioteche virtuali nel web — sia di matrice bibliotecaria — quali gli sforzi normativi per un ampliamento del formato bibliografico atto a descrivere le risorse informative di rete, o il progetto CATRIONA dell'Università di Bath che vuole estendere il catalogo della biblioteca per includere le risorse di rete;
— orientata al testo completo e non alla descrizione bibliografica.
In questo senso Internet non è ancora uno strumento per la documentazione, anche se divengono sempre più numerose le iniziative di sistematizzazione dell'informazione in rete [Basili 1995b].

SCENARI TECNOLOGICI NELLA DOCUMENTAZIONE

I momenti dell'introduzione di una tecnologia

L'introduzione di una nuova tecnologia in una organizzazione si attua attraverso passaggi successivi:
1) l'obiettivo iniziale è la *replica* automatica di processi manuali; a tal fine si sviluppa una applicazione o si acquisisce un pacchetto applicativo preconfezionato;
2) segue una fase di *parallelo* tra processi manuali e processi automatici; è tipico delle tecnologie dell'informazione il concetto di *momento zero* inteso come la data di attivazione della procedura automatizzata; tale momento zero ha senso ed è efficace nel caso di processi con un ciclo di vita periodico (per esempio la contabilità) e senza memoria; alcune applicazioni, tuttavia, sono invece di carattere storico ed anzi hanno senso proprio in virtù della loro *memoria*; le applicazioni documentarie cadono — come è noto — in questa classe; in questi casi è tipicamente presente il *problema del pregresso*, inteso come la necessità di inserire nel sistema

tutti i dati prodotti dall'organizzazione fin dalla sua nascita;
3) la *transizione in linea (online transition)* si attua quando si diffonde la consuetudine all'uso della nuova tecnologia, quando cioè la nuova tecnologia penetra nel tessuto operativo dell'istituzione; nel caso della documentazione la transizione in linea non ha obiettivi legati all'utilizzo del documento in formato elettronico da parte dell'utente finale, bensì è efficace ai fini della mediazione documentaria.
Esempio: il voluminoso rapporto dei lavori di una commissione o il testo di una legge parteciperanno comunque del tradizionale circuito cartaceo di distribuzione e saranno fruibili comunque nel formato a stampa. La transizione in linea di tale circuito, infatti, implicherebbe una trasformazione radicale dei metodi e dell'organizzazione del lavoro non solo della funzione di documentazione, ma anche dell'utente finale per il quale la tecnologia dovrebbe divenire parte integrante dello stile di lavoro. Né è d'altra parte auspicabile la lettura di un testo — in particolare se voluminoso — da schermo. La disponibilità del testo completo in formato digitale, tuttavia, apre la via a nuove forme di mediazione documentaria, dove l'enfasi si sposta, come vedremo, dalla indicizzazione alla navigazione ipertestuale nel testo completo;
4) l'*innovazione* è la trasformazione ed il miglioramento dell'operatività e dell'organizzazione del lavoro e si realizza quando l'applicazione della tecnologia amplia le funzionalità del processo manuale.

Evoluzione del concetto di documento

La rete ha ulteriormente dilatato il concetto di documento che già la ipermedialità aveva affermato, aggiungendo la caratteristica della distribuzione geografica a quella di ipermedialità [Basili 1996].
Una conseguenza fondamentale — ai fini del controllo di qualità dell'informazione — è che la responsabilità del documento è distribuita: si tratta di risorse dislocate in siti diversi, con responsabilità sulla risorsa distinte, con meccanismi di aggiornamento distinti, con controlli di qualità distinti.
L'informazione in rete, infatti, è stereotipamente rappresentata dalla pagina web, che consiste di un insieme di elementi informativi eterogenei, quali: testi; notizie; depositi di dati; riviste elettroniche; archivi di conferenze elettroniche; informazione secondaria (TOC, bibliografie); servizi; software; altre pagine/siti web; siti gopher; immagini, grafici, suoni; che riguardano uno specifico argomento o settore disciplinare; connessi da legami ipertestuali, che pongono in relazione fonti diverse, generalmente residenti in siti geograficamente distinti.
La pagina web non è dunque esattamente la replica elettronica di un documento cartaceo. ➤

Tecnologia e mediazione documentaria

Ricerca a testo libero vs vocabolario controllato[2]

I sistemi di recupero dell'informazione (*Information retrieval* — IR) esistono dagli anni Sessanta [Smeaton 1995] e si sono via via evoluti di pari passo con:
— l'evolvere della disponibilità tecnologica;
— i cambiamenti nelle tecniche di produzione, raccolta e memorizzazione dell'informazione, ove l'elaboratore elettronico è andato assumendo un ruolo sempre più determinante per tutti gli aspetti della gestione dell'informazione.
La prima applicazione dei sistemi per il recupero dell'informazione è stata l'interrogazione di cataloghi di biblioteca [Smeaton 1995]. Successivamente, negli anni Settanta si sono imposti i fornitori commerciali di basi di dati bibliografiche e negli stessi anni si è andato affinando e potenziando lo sviluppo del thesaurus come strumento per la ricerca post coordinata. Il grande passaggio si è tuttavia verificato alla fine degli anni Settanta, quando la tecnologia ha reso possibile la costruzione e l'affermazione delle basi di dati a testo completo. Il primo approccio nel recupero dell'informazione a testo completo è stato il (brutale) utilizzo della potenza di calcolo per ricercare — attraverso il flusso invertito — termini in voluminose basi di dati testuali, applicando le combinazioni della logica booleana; in questo modo ogni termine veniva utilizzato come chiave di ricerca [Dubois 1987]. Di qui la grande diatriba tra i sostenitori del linguaggio controllato e quelli della ricerca a testo libero [Dubois 1987]. I fondamentali criteri di valutazione dei sistemi di IR sono, come è noto:
— *richiamo:* inteso come la capacità di recuperare il maggior numero possibile di documenti rilevanti;
— *precisione:* intesa come la capacità di prevenire il recupero di documenti non rilevanti.
Svariati studi di comparazione delle due tecniche di recupero — attraverso vocabolari controllati o a testo libero — sono giunti alla conclusione che il vocabolario controllato aumenta la precisione ed il richiamo, ma non relativamente a soggetti nuovi o interdisciplinari. Riportiamo di seguito le conclusioni tratte da [Dubois 1987] sul confronto tra le due tecniche, sottolineando che queste sono state prodotte nel 1987, quindi con la disponibilità tecnologica di quel momento:

VANTAGGI DELLA RICERCA A TESTO LIBERO

— basso costo: eliminato l'alto costo della soggettazione e della costruzione di thesauri;
— semplificazione della interrogazione: abolita la necessità di conoscere il vocabolario controllato;
— possibilità di ricerca sull'intero contenuto: non solo sul titolo, sull'abstract e per soggetto;
— ogni termine ha lo stesso valore ai fini del recupero: è una chiave di ricerca;
— eliminazione dell'errore umano nella soggettazione;
— nessun ritardo nell'integrare termini nuovi.

SVANTAGGI DELLA RICERCA A TEMPO LIBERO

— maggiore responsabilità all'utente;
— possibilità di non raggiungere informazione implicita;
— assenza di relazioni semantiche tra termini;
— necessità di conoscere il linguaggio della disciplina.

VANTAGGI DEL VOCABOLARIO CONTROLLATO

— capacità di risolvere problemi semantici (disambiguare);
— capacità di identificare relazioni tra termini.

SVANTAGGI DEL VOCABOLARIO CONTROLLATO

— alto costo per personale qualificato;
— aggiornamento del vocabolario non garantito;
— possibilità di errore umano.
Nonostante lo sforzo di analisi di Dubois resta la difficoltà di attribuire un peso ed una relativa importanza a ciascuna delle caratteristiche elencate, che siano di base ad una valutazione oggettiva.
Un elemento molto importante nella scelta tra vocabolario controllato e ricerca a testo libero è l'ambito disciplinare della base documentaria; in particolare l'estensione e i confini della disciplina e, ancor più importante, l'esattezza della terminologia. Nei prodotti commerciali si tende a fondere le due possibilità, il vocabolario controllato e la ricerca a testo libero. Una alternativa per limitare i costi di soggettazione può essere quella di utilizzare un thesaurus solo come guida alla ricerca e non come strumento di indicizzazione.

Elaborazione del linguaggio naturale

Una branca del filone teorico dell'intelligenza artificiale interessa in particolare le applicazioni documentarie ed è l'elaborazione del linguaggio naturale (Natural Language Processing — NLP). Un sistema ideale per il recupero dell'informazione è un meccanismo che, una volta interpretata la domanda, restituisce in risposta non solo quanto è stato chiesto, ma anche quanto si intendeva chiedere. Ciò implica che il sistema non deve limitarsi a recepire i singoli termini della domanda, ma deve anche interpretare ciò che il complesso di più termini vuole significare. Questo richiede capacità di elaborazione intelligente di testi, intesa come la capacità di cogliere l'essenza del contenuto di informazione testuale ed è l'obiettivo della ricerca nel settore dell'elaborazione del linguaggio naturale.
L'interpretazione del linguaggio naturale, vuoi per realizzare traduttori automatici (linguistica computazionale), vuoi per automatizzare funzioni di soggettazione, vuoi per ottimizzare il recupero dell'informazione, vuoi, in generale, per l'interpretazione di testi letterari è uno tra i più ambiziosi obiettivi dell'intelligenza artificiale.
La ricerca in questo settore si scontra con due problemi fondamentali: replicare la capacità di compredere la semantica di un testo in un particolare contesto (evitando di interpretare "spirito forte" con "alcool ad alta gradazione") applicare le tecniche di interpretazione a ingenti volumi di documenti.
La ricerca nel campo della elaborazione del linguaggio natu-

[2] Linguaggi controllati sono i sistemi di classificazione, i soggettari, i thesauri.

rale, inoltre, è dilaniato da due vincoli contrastanti: l'interpretazione di un testo è tanto più precisa quanto più voluminoso è il testo (avrebbe poca efficacia applicare queste tecniche, per esempio, al solo titolo di un testo); tanto più vasta è la base documentale tanto meno è praticabile l'interpretazione del testo.
Da questo contrasto discende il senso di tentativi di combinare l'elaborazione del linguaggio naturale con l'elaborazione parallela.
La conferenza annuale TREC (Text REtrieval Conference) — sostenuta dall'agenzia statunitense ARPA — è una sede di riferimento internazionale per la discussione e la sperimentazione di tecnologie per la gestione di ingenti volumi di dati.

Navigazione ipertestuale in basi documentarie strutturate

Il volume dell'informazione testuale disponibile in formato elettronico è enorme e in continuo aumento [Smeaton 1995]: i quotidiani sono generati in formato elettronico; gli editori stampano libri a partire da fonti elettroniche; gli articoli scientifici vengono prodotti dall'autore con sistemi di videoscrittura e in questo formato vengono spediti alle riviste; gli uffici usano sistemi di videoscrittura per generare lettere e documenti; la trascrizione delle sedute parlamentari è registrata per mezzo di sistemi elettronici, come pure la documentazione di processi. Di qui deriva il costante impulso alla ricerca nel settore delle basi documentarie a testo completo, dette anche basi testuali.
La tecnologia delle basi documentarie a testo completo richiede specifici modelli di rappresentazione del documento. Uno degli sviluppi più interessanti è il meta-linguaggio SGML (Standard Generalised Markup Language) che consente di descrivere la struttura di un testo, indipendentemente dal sistema con il quale questo viene prodotto. In altri termini, attraverso lo standard SGML si rende esplicita la strutturazione di un testo il quale — così strutturato — può essere elaborato da un sistema di fotocomposizione per essere stampato, oppure può alimentare una base di documenti ipermediali (per esempio gestita da Hypercard), oppure può essere inserito in una base di dati testuali. Il formato SGML, dunque, è trasportabile attraverso più ambienti di elaborazione di documenti.
I sistemi tradizionali per il recupero dell'informazione a testo completo (*text retrieval systems*) presentano le seguenti caratteristiche [Macleod 1990]:
— operano attraverso termini di indicizzazione assegnati al documento;
— registrano la frequenza dei termini utilmente alla costruzione di algoritmi ponderati basati sulla frequenza dei termini;
— operano sull'intero testo; solo alcuni sistemi consentono di ripartire il testo in paragrafi, permettendo in questo modo di limitare la ricerca ad una parte del documento;
— presentano semplici possibilità di aggiornamento, di solito limitate all'inserimento e alla cancellazione di documenti;
— distinguono tra contenuto (il testo del documento) ed attributi (informazione secondaria).
Una visione più avanzata di organizzazione documentaria vede l'applicazione dello standard SGML abbinata alla tecnologia ipertestuale.
Il documento viene generato con una struttura gerarchica definita da chi progetta la base documentaria, ed attuata attraverso lo standard SGML, nella quale si distinguono i seguenti elementi: il titolo del documento, l'autore, altre informazioni sul documento di carattere complessivo, sezioni, sottosezioni, paragrafi e/o figure, frasi, termini.
Ciascun elemento della struttura del documento è candidato a divenire un nodo ipertestuale collegato ad altre parti dello stesso documento o di altri documenti. Si pensi ad un articolo scientifico, organizzato in sezioni e paragrafi, con citazioni di altri articoli:
— il testo dell'articolo può essere strutturato secondo lo standard SGML;
— parti del documento possono essere collegate da legami ipertestuali;
— le citazioni bibliografiche possono essere collegate tramite legami ipertestuali ai rispettivi documenti, in particolare al brano di interesse per la citazione; a partire da un documento, in questo modo, è possibile consultare tutto il materiale da questo citato;
— se l'articolo fa riferimento ad un archivio di dati scientifici, è possibile stabilire un legame ipertestuale con tale archivio, per la visualizzazione dei dati.
L'abbinamento SGML - ipertesto estende e migliora il concetto di documento, fornendo al contempo la possibilità di collegare gruppi arbitrari di documenti e di parti di documenti in più modi. Figurativamente potremmo dire che l'uso combinato di SGML e di legami ipertestuali trasforma un "testo piatto" in un "testo multidimensionale".
Per quanto riguarda le modalità di consultazione della base documentaria, questo modello consente una duplice modalità di interazione:
— la ricerca a testo libero nelle parti del documento o, ancor più significativamente, nei titoli delle parti (in genere esplicative del contenuto della parte stessa);
— la navigazione nella base documentaria, una volta posizionati su un documento di interesse.

Disseminazione selettiva dell'informazione: strumenti di filtering

Mai in passato così tanta informazione è stata disponibile al pubblico e mai come ora la velocità di produzione di nuova informazione è stata così alta. Mai come ora, infine, è stata disponibile così tanta informazione in formato elettronico.
Nel 1989 Wurman scrive un libro [Wurman 1989] dal titolo *Information Anxiety* dove afferma che questo disturbo "è prodotto dallo scollo sempre più ampio tra ciò che possiamo capire e ciò che riteniamo di dover capire. È il buco nero tra dato e conoscenza e accade quando l'informazione non ci dice ciò che vogliamo o dobbiamo sapere".
D'altra parte già nel 1982 il presidente dell'ACM, Peter Dennings, affermava che: "La visibilità dei personal computer, delle stazioni di lavoro individuali e delle reti locali ha focalizzato l'attenzione per lo più sulla generazione di informazione — il processo di produrre e disseminare documenti. È ora tempo di focalizzare l'attenzione sulla ricezione di informazione — il processo di controllare e filtrare l'informa- ➤

zione che raggiunge quanti debbono usarla".
L'obiettivo degli strumenti di *filtering* è realizzare l'estrazione sistematica di informazione importante per un utente, a partire da un più vasto flusso di nuova informazione che viene continuamente prodotta.
Un problema centrale nel funzionamento dei sistemi di *filtering* è determinare e rappresentare l'interesse del lettore, ciò che viene chiamato "profilo di interesse". La soluzione più utilizzata è quella di richiedere all'utente stesso un insieme di termini che descrivano la sua esigenza informativa; il complesso di tali termini viene quindi a comporre il profilo di interesse per quell'utente, profilo che, tendenzialmente, resta a lungo immutato.
Information retrieval e *information filtering* sono due approcci distinti della ricerca documentaria:
— nel *retrieval* esiste una *collezione statica di informazione*, sulla quale vengono attivate le più *diverse interrogazioni* da parte dell'utente;
— nel *filtering* l'*esigenza informativa è statica*, ma è attivata su un *flusso dinamico di informazione*.[3]
Le prime iniziative erano sistemi manuali di *alerting* e realizzavano la disseminazione selettiva dell'informazione (SDI).
L'esplosione dell'informazione disponibile in rete pone ora l'accento sulla realizzazione di sistemi automatici di *filtering*. Un elenco di questi sistemi è disponibile all'indirizzo: http://www.enee.umd.edu//medlab/filter/filter.html.
In letteratura i sistemi di filtering sono noti anche come "agenti intelligenti" o "knowbot"; questa terminologia testimonia e sottolinea alcuni tentativi — ancora in fase sperimentale — di applicazione delle tecniche di elaborazione del linguaggio naturale a questo tipo di applicazione.

Evoluzione dei sistemi documentari: una traccia

Nella Tabella 1 viene tracciato un quadro descrittivo di una sorta di evoluzione (non sempre lineare a causa dell'intersecarsi di sviluppi tecnologici diversi) nei sistemi di organizzazione e recupero di informazione.

Un modello per la ricerca di informazione

La formulazione della domanda nella ricerca documentaria varia come variano gli atteggiamenti e le motivazioni dei ricercatori. La situazione più semplice (non la più frequente) si verifica quando l'utente sa cosa cercare: può trattarsi di un autore, un titolo o un soggetto ben noto. In questo caso la formulazione della domanda è precisa e otterrà sempre una risposta dal sistema, negativa o positiva che essa sia. Più spesso il suo interesse è solo scorrere la base documentaria: non ha in mente un'idea precisa, ma vuole solo acquisire informazioni sulla sua area d'interesse. Il suo atteggiamento è recettivo alle informazioni utili che può trarre dalla base documentaria, ma piuttosto passivo.
Quasi sempre si sposta da un documento all'altro, senza una precisa strategia di ricerca, ma piuttosto incuriosito da ciò che via via incontra e disposto a cambiare percorso per approfondire un argomento in cui si imbatte nel suo vagolare. Questo atteggiamento è descritto dal termine inglese *serendipity*.
Nel 1986 Heyer definisce un modello che descrive tre modalità di ricerca di informazione:
— la ricerca puntuale (*hunting*): si riferisce alla ricerca su uno specifico argomento; l'utente formula la sua interrogazione, la affina in base alla risposta del sistema e, ottenuta la risposta, in genere scarta l'interrogazione che l'ha generata;
— la navigazione (*browsing*): è il vagolare senza una meta precisa nell'informazione; è la più diffusa (almeno in fase di primo approccio) modalità di consultazione dell'informazione in rete, nella sua tipica forma ipermediale;
— *filtering* (*grazing*-pastura): l'utente viene regolarmente "alimentato" di informazione corrente, in base ad un profilo di interesse che in genere resta tendenzialmente costante nel tempo.
Heyer ritiene che questo modello possa descrivere qualunque modalità di ricerca e che non si possono individuare altri modi con i quali un utente raccoglie informazione. Questa opinione è condivisa da molti [Hawkins 1996], essendo un modello sufficientemente ampio e descrive gran parte delle modalità di interazione dell'utenza quando cerca infor-

Tab. 1

BASE INFORMATIVA	INDICIZZAZIONE	INTERFACCIA
surrogato del documento	vocabolario controllato soggettazione manuale	ricerca a comandi o a menu
documento a testo completo	vocabolario controllato soggettazione manuale	ricerca con interfaccia grafica
documento a testo completo	ogni termine è indicizzato indicizzazione automatica (pseudo soggettazione basata sulla frequenza)	ricerca con interfaccia grafica
documento a testo completo strutturato (SGML)	indicizzazione automatica nelle intestazioni e nel testo	ricerca e navigazione ipertestuale con interfaccia grafica

[3] In questo contributo parliamo del "*filtering* attivo o positivo" e non del "*filtering* esclusivo o negativo" che funge da censura nei riguardi di informazione della quale si vuole inibire l'uso.

mazione elettronica. In questo contributo abbiamo cercato di descrivere applicazioni tecnologicamente avanzate per tutte queste forme di ricerca.

CONCLUSIONI

Nuove tecnologie dell'informazione e nuove specializzazioni si sviluppano più velocemente di quanto possano essere utilizzate, portando con sé ogni volta una nuova terminologia. Il prodotto della tecnologia dell'informazione si configura sempre più come la confluenza di più sviluppi indipendenti che si applicano — secondo le più svariate combinazioni — a domini applicativi specifici. Allo stesso modo la funzione di documentazione, il cui obiettivo prioritario è il recupero dell'informazione, è la convergenza di più attività e richiede una molteplicità di competenze, ciascuna con un proprio linguaggio o, per meglio dire, con un proprio gergo.
La mediazione documentaria, ed in particolare il recupero dell'informazione, viene così ad essere l'intersezione di linguaggi diversi, provenienti da diversi ambiti: il documento, la classificazione, la tecnologia, l'intermediario, l'utente e l'indicizzazione [Brier 1996].
Il connubio tecnologia e documentazione non è dunque limitato a due termini, ma si amplia per includere un complesso di attività, strumenti e metodologie, che cooperano per raggiungere una gamma diversificata di obiettivi. Tale complessità implica che la funzione di documentazione è il risultato del lavoro di un gruppo di specialisti in ambiti diversi, il cui maggiore sforzo per rendere efficiente il servizio deve essere quello di saper comunicare.
Ci piace concludere con una insolita e bizzarra definizione di documentalista, data *involontariamente* da Arthur Conan Doyle, l'autore di Sherlock Holmes, nel seguente brano dove il famoso investigatore parla del fratello Mycroft:
[...] Mycroft possiede un cervello ordinato e metodico, con un'enorme capacità di incamerare nella memoria fatti che riguardino chiunque. Le stesse grandi facoltà che io ho dedicato alla scoperta dei crimini, lui le ha dedicate a questa sua attività particolare.
[...] Supponiamo che un ministro abbia bisogno di informazioni su un argomento che coinvolgela Marina, l'India, il Canda e il problema del bimetallismo; potrebbe riceverle separatamente, da vari dicasteri, ma Mycroft è l'unico in grado di sintetizzarle e di dire subito in quale modo ciascun fattore può influenzare l'altro.
In un primo tempo si servirono di lui come di una scorciatoia, una comodità; ora si è reso indispensabile. In quel suo gran cervello, ogni cosa è incasellata e può essere tirata fuori in un istante.
Più di una volta la sua parola è stata decisiva per la politica nazionale.
(Arthur Conan Doyle (1859-1930), *L'ultimo saluto di Sherlock Holmes)* ■

RIFERIMENTI BIBLIOGRAFICI

[Basili 1995] C. BASILI (a cura di), *Internet e informazione scientifica: opportunità e problemi aperti*, Note di bibliografia e di documentazione scientifica LXI, CNR-ISRDS, settembre 1995, p. 164.
[Basili 1995b] C. BASILI, *Searching for information by subject: what does it mean in today's Internet environment?*, "The Electronic Library", Ott. 1995, 459-466.
[Basili 1996] C. BASILI, *Il problema della qualità per l'informazione scientifica in rete*, Rapporto Tecnico CNR/ISRDS/RT-30/96, ottobre 1996.
[Brier 1996] S. BRIER, *Cybersemiotics: a new interdisciplinary development applied to the problems of knowledge organisation and document retrieval in Information Science*, "Journal of documentation", 52 (1996), 3, p. 296-344.
[Dubois 1987] C.P.R. DUBOIS, *Free text vs. controlled vocabulary; a reassessment*, "Online review", 11 (1987), 4, p. 243-253.
[Hawkins 1996] D.T. HAWKINS, *Hunting, Grazing, Browsing: a model for Online Information Retrieval*, "Online", Jan./Feb. 1996, p. 71-73.
[Heyer 1986] M. HEYER, *The creative challenge of CD-ROM*, in *CD-ROM: the new papyrus*, Lambert, S. and Ropiequet, S. (eds), "Microsoft Press", 1986, p. 347-357.
[Hillis 1986] W.D. HILLIS, *Parallel computers for AI databases*, in Brodie, L. and Mylopoulus, J. (eds), *On knowledge base management systems*, Springer-Verlag, New York, 1986n p. 660.
[Ingwersen 1995] P. INGWERSEN, *Information and Information Science*, "Encyclopedia of Library and Information Science", vol. 56, 1995, p. 137-174.
[Macleod 1990] I.A. MACLEOD, *Storage and retrieval of structured documents*, "Information Processing and Management", 26 (1990), 2, p. 197-208.
[Smeaton 1995] A.F. SMEATON, *Natural language processing used in information retrieval tasks: an overview of achievements to date*, "Encyclopedia of Library and Information Science", vol. 55, 1995, p. 220-238.
[Vickery 1990] B. VICKERY, A. VICKERY, *Intelligence and information systems*, "Journal of Information Science", 16 (1990), p. 65-7.
[Wurman 1989] R.S. WURMAN, *Information Anxiety*, Doubleday, New York, 1989.
[IFLA 1996] Inventory of Full-Text Information Retrieval Software Vendors http://www.nlc-bnc.ca/ifla/VII/s21/p1996/full-text.htm.