Interfacce Intelligenti a Banche di Dati Bibliografici

*Giorgio Brajnik, Stefano Mizzaro e Carlo Tasso**

Viene discussa l'importanza e il ruolo di interfacce intelligenti a sistemi di information retrieval nel reperimento di informazioni che effettivamente soddisfino i bisogni informativi degli utenti. La linea di ricerca FIRE ha avuto come obiettivo sviluppo e sperimentazione di un'interfaccia intelligente (basata su rappresentazione esplicita della conoscenza) ad un sistema di information retrieval di tipo booleano. Nello sviluppo di tale interfaccia un ruolo rilevante è rivestito dalle tecniche di intelligenza artificiale, mediante le quali è possibile modellare i ragionamenti e le conoscenze utilizzati da un intermediario di ricerca.

1. Introduzione

L'Information Retrieval è la disciplina che studia i sistemi di gestione e accesso a grandi quantità di dati non strutturati, tipicamente documenti scritti in linguaggio naturale. La natura non strutturata dei dati considerati differenzia i sistemi di information retrieval dai sistemi di gestione di basi di dati. Tale differenza rende assai complesso il problema del reperimento che non si può basare su precise chiavi d'accesso come avviene invece per le basi di dati. La soluzione tradizionale di tale problema si basa sull'utilizzo di una figura professionale, denominata intermediario di ricerca, che funge da tramite tra utente finale e sistema di information retrieval. Tale soluzione, sebbene efficace, non si adatta

* Laboratorio di Intelligenza Artificiale, Dipartimento di Matematica e Informatica, Università di Udine

all'enorme espansione - in numero e complessità - delle banche dati oggi disponibili e all'accresciuta necessità di accedere all'informazione in esse contenuta da parte di un numero sempre maggiore di utenti.

Per tale motivo lo sviluppo di sistemi di interfaccia a sistemi di information retrieval assume particolare rilevanza, in quanto può fornire concreti ed utili strumenti volti a facilitare il reperimento di informazioni che effettivamente soddisfino i bisogni informativi degli utenti. Nello sviluppo di tali interfacce un ruolo rilevante è rivestito dalle tecniche di intelligenza artificiale, mediante le quali è possibile modellare i ragionamenti e le conoscenze utilizzati da un intermediario di ricerca.

In tale quadro, presso il Laboratorio di Intelligenza Artificiale dell'Università di Udine, sono svolte da anni ricerche volte allo studio ed alla realizzazione sperimentale di tali interfacce a sistemi di information retrieval di tipo commerciale. Il progetto FIRE in particolare è dedicato allo sviluppo e sperimentazione di un'interfaccia intelligente (basata su rappresentazione esplicita della conoscenza) ad un sistema di information retrieval di tipo booleano.

Scopo del presente lavoro è illustrare i risultati ottenuti nella realizzazione del prototipo FIRE. La trattazione è divisa in 2 parti: una prima di carattere generale che inquadra il settore in cui il progetto FIRE si colloca, ed una seconda dedicata alla descrizione del progetto vero e proprio e del sistema realizzato.

2. L'information retrieval

2.1. *L'information retrieval tradizionale*

L'*Information Retrieval* (*IR*) [van Rijsbergen, 1979; Salton, 1989; Ingwersen, 1992] è la disciplina che si occupa di studiare, progettare e realizzare sistemi informatici, denominati *Sistemi di Information Retrieval* (*SIR*), che consentano la memorizzazione, il reperimento e la manutenzione di grosse quantità di dati non strutturati. Tipicamente tali dati sono descrizioni di libri, articoli, atti di conferenze, rapporti tecnici, ed in generale qualsiasi tipo di documento che contiene parti strutturate e parti non strutturate, costituite di solito da testo libero in linguaggio naturale. La col-

lezione di dati gestiti da un SIR è comunemente denominata *banca dati*. Talvolta le banche dati contengono l'intera copia di un documento (banche dati di tipo *full text*), mentre più frequentemente includono una descrizione bibliografica del documento (banche dati di tipo bibliografico), costituita da una parte strutturata che include informazioni quali il titolo, il nome dell'autore, la casa editrice, la data di pubblicazione, alcune parole chiave, ed una parte non strutturata in testo libero, costituita ad esempio da un sommario o dall'indice, che ne descrive sinteticamente il contenuto.

Come osservato in [Lancaster, 1968], il termine 'information' in 'information retrieval', sebbene diventato ormai una consuetudine assodata, è in qualche modo infelice, in quanto un SIR non 'informa' l'utente riguardo all'oggetto della richiesta (ossia non cambia la conoscenza dell'utente su tale argomento), come invece farebbe un sistema in grado di comprendere una domanda, di consultare qualche risorsa informativa e di fornire una risposta diretta. Un SIR si limita a restituire i documenti in cui l'utente può, in un secondo tempo, trovare l'informazione desiderata. Nel caso di banche di dati bibliografici, inoltre, i risultati restituiti dal sistema sono unicamente delle schede bibliografiche dei documenti che potrebbero risultare utili all'utente, e quindi in tale caso il SIR si limita ad informare sull'esistenza (o non esistenza) di documenti correlati alla sua richiesta, a fornirne l'indicazione bibliografica, e poco più.

Le tecnologie adottate per la realizzazione dei SIR sono varie e permettono diversi stili di interazione con i loro utilizzatori (utenti finali o intermediari); si veda [Belkin e Croft, 1987] per una rassegna. Nel seguito di questo articolo ci si limiterà a considerare SIR booleani (in cui la richiesta viene formulata usando operatori logici) di gran lunga i più diffusi ed usati.

La produzione e l'utilizzo di una banca dati bibliografici gestita da un SIR tradizionale prevedono l'esistenza di quattro attori fondamentali che operano (direttamente o indirettamente) sul SIR: l'*autore*, l'*indicizzatore*, l'*utente* e l'*intermediario*. La fig. 1 illustra lo scenario concettuale dell'IR, esplicitando processi, input, output ed esecutori.

Nella parte in alto a sinistra della figura è schematizzato il processo che dà origine alla banca dati: un autore, in base alle proprie conoscenze, idee, convinzioni produce un *documento*

(processo di *produzione del documento*); l'indicizzatore costruisce una rappresentazione di tale documento, denominata *surrogato* (processo di *indicizzazione*): vengono estratte dal documento le informazioni per i campi (strutturati e non) e vengono aggiunte a questi alcune chiavi di classificazione, che facilitano le successive operazioni di reperimento. Il surrogato viene poi inserito, insieme ad altri surrogati, nella banca dati sfruttando le funzioni messe a disposizione dal SIR (processo di *caricamento*).

Il resto della figura rappresenta l'utilizzo del SIR per il reperimento di informazioni contenute nella banca dati; in tale fase vengono coinvolti gli altri due attori, l'utente e l'intermediario, che funge da mediatore fra utente e sistema. La situazione in cui si trova un utente di un SIR è descritta in modo esauriente da un punto di vista cognitivo in [Ingwersen, 1992]: egli si trova in uno stato di conoscenza *incompleto* e ha necessità di informazioni che gli permettano di colmare una sua lacuna conoscitiva; tale necessità è denominata *bisogno informativo* (*BI*).

Si può pensare che il BI sia originato da un problema che l'utente vuole risolvere (o quantomeno da uno scopo che egli vuole raggiungere): il BI è inconscio, implicito nella mente dell'utente ed è il risultato della percezione che l'utente ha del proprio problema (o scopo). Il processo di *percezione* è senz'altro soggettivo e può dare origine ad un BI che rispecchi in modo più o meno corretto e completo il reale problema dell'utente. Si possono distinguere tre tipi di BI [McAlpine e Ingwersen, 1989; Ingwersen, 1992]:

- *bisogno verificativo* (verificative need): l'utente vuole trovare o verificare l'esistenza di documenti di cui già conosce l'autore, il titolo, ecc.;
- *bisogno conscio* (conscious topical need): l'utente vuole reperire documenti, di cui non ha riferimenti precisi, riguardanti un argomento a lui noto;
- *bisogno confuso* (muddled topical need): l'utente vuole reperire documenti relativi ad un argomento con cui non ha familiarità allo scopo di colmare una lacuna conoscitiva.

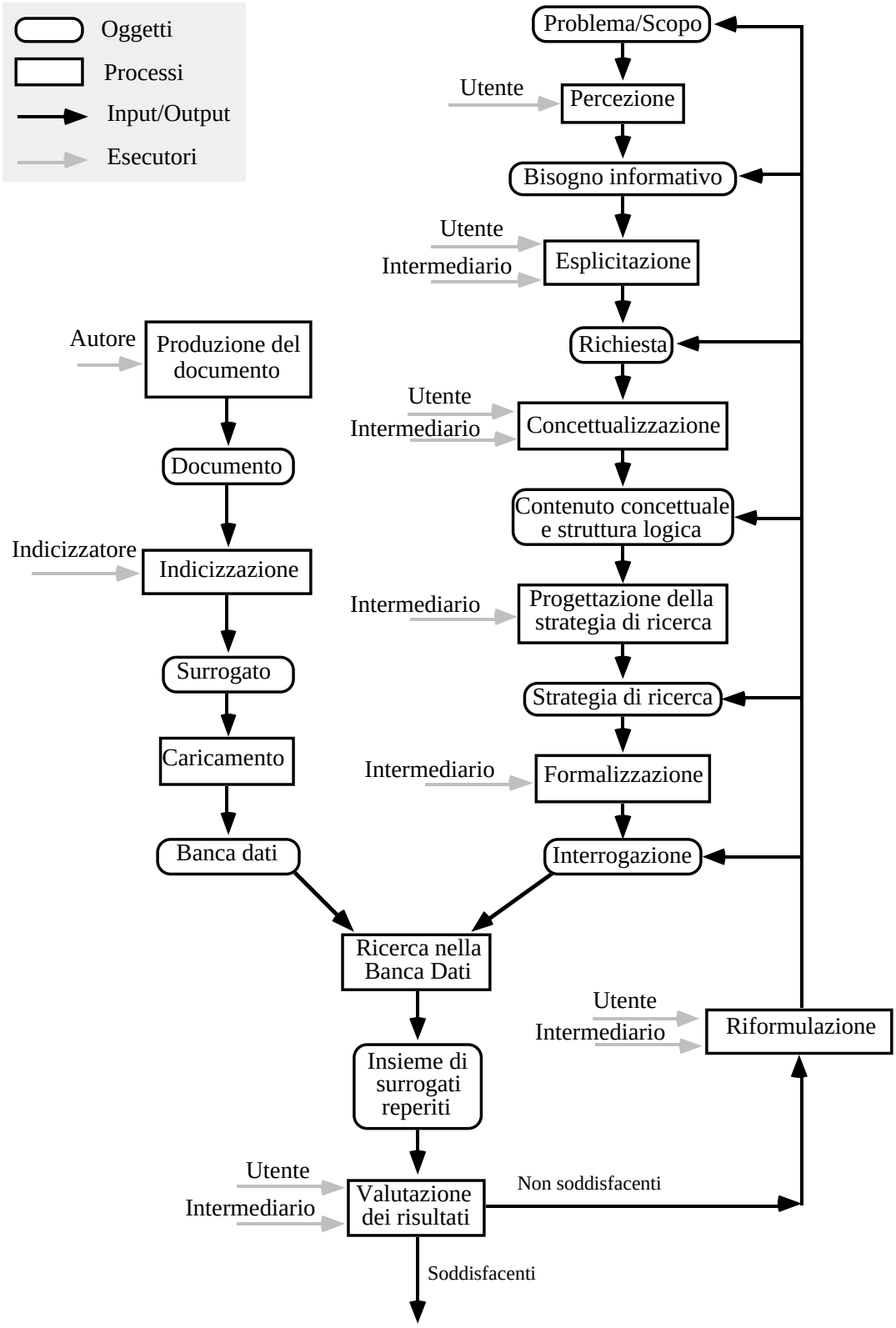

Figura 1. Produzione e utilizzo di una banca dati gestita da un SIR tradizionale

Inoltre, nel BI si possono riconoscere tre componenti:

- l'*argomento*, che indica l'area disciplinare ed i concetti che devono essere trattati dai documenti interessanti per l'utente; ad esempio "sistemi intelligenti per la supervisione di impianti";
- il *compito*, che è relativo al tipo di attività che l'utente deve svolgere con i documenti reperiti; ad esempio "predisporre una rassegna di diversi approcci sperimentali";
- il *contesto*, tutto ciò che non rientra nell'argomento né nel compito, ma che influenza comunque la conduzione della ricerca e che, in generale, dipende dall'utente e dalla sua situazione specifica. Il contesto include, ad esempio, la conoscenza di documenti rilevanti e/o utili, che l'utente può avere prima della ricerca, o di determinati autori o fonti bibliografiche, l'esperienza con la specifica banca dati, la quantità di tempo e il budget a disposizione per risolvere il problema.

Il BI, come detto, esiste fondamentalmente a livello inconscio; la sua *esplicitazione* è volta ad identificarne una rappresentazione esplicita, denominata *richiesta*. Solitamente, la richiesta è espressa in linguaggio naturale, scritto o orale, quindi facilmente comprensibile dall'utente, ma non utilizzabile direttamente dal sistema per accedere alla banca dati. L'esplicitazione del BI può essere più complessa di quanto sembri a prima vista, si pensi ad esempio al caso di un utente con un BI di tipo confuso. Il primo compito dell'intermediario risulta quindi quello di aiutare l'utente ad esplicitare il più correttamente e completamente possibile il suo BI.

Il secondo compito dell'intermediario riguarda il processo di *concettualizzazione* della richiesta, ossia quel processo svolto in collaborazione con l'utente e mirante all'identificazione del *contenuto concettuale e della struttura logica* della richiesta. Si tratta di *(i)* individuare i concetti presenti nella richiesta, ciascuno dei quali viene denominato solitamente *faccetta*, *(ii)* di esprimerli mediante un'opportuna terminologia, e di *(iii)* individuare la struttura logica della richiesta, espressa usualmente mediante operatori logici (AND, OR, NOT) applicati alle faccette.

Il compito e il contesto inclusi nel bisogno informativo, il contenuto concettuale e la struttura logica della richiesta sono considerati dall'intermediario per il processo di *progettazione*

della strategia di ricerca, volta ad individuare secondo quale approccio deve essere organizzata la ricerca nella banca dati, quali concetti devono essere ricercati per primi, secondo che modalità, quando deve intervenire una fase di valutazione dei risultati, e così via.

La strategia di ricerca, il contenuto concettuale e la struttura logica della richiesta sono espressi in una forma indipendente dal SIR su cui verrà poi effettuata la ricerca. È necessario quindi un processo di *formalizzazione*, che trasforma la richiesta in un'*interrogazione*, ossia una sequenza di comandi espressi nel linguaggio formale di interrogazione dello specifico SIR. Tale trasformazione è solitamente effettuata dall'intermediario, al fine di evitare all'utente le difficoltà tecniche e pratiche connesse all'uso del SIR.

Quando l'interrogazione viene sottoposta al SIR, viene attivato un processo di *ricerca nella banca dati* dei surrogati che la soddisfano. I SIR oggi diffusi commercialmente sono realizzati mediante i cosiddetti *indici invertiti*, che permettono, mediante la semplice specificazione di un dato termine (detto *termine di ricerca*), di accedere rapidamente a tutti i surrogati che contengono il termine stesso.

Spesso la prima formulazione del BI non dà risultati soddisfacenti: ciò viene individuato nel processo di *valutazione dei risultati*, che viene effettuato congiuntamente da utente ed intermediario al fine di giudicare se i singoli documenti reperiti sembrano soddisfare il BI. Si può rendere quindi necessario un processo di ri-espressione del BI, o di *riformulazione* (si veda [Meadow e Cochrane, 1981]), che può essere anche iterata più volte fino al raggiungimento dell'obiettivo prefissato: in prima istanza l'identificazione di documenti potenzialmente utili per soddisfare il BI, ed in ultima analisi l'effettiva raccolta delle informazioni necessarie a risolvere il problema che ha originato il BI. L'intermediario interviene anche in questo processo, in quanto la fase di riformulazione avviene sotto il suo controllo. La riformulazione può avere anche un altro effetto, ossia aiutare l'utente a comprendere meglio il proprio BI e quindi ad esprimerlo in modo più accurato. Le attività di percezione, esplicitazione, concettualizzazione, progettazione della strategia di ricerca e formalizzazione non sono quindi da ritenersi terminate una volta effet-

tuata la formulazione della prima richiesta: esse continuano durante tutto il processo di riformulazione.

Dato che l'intermediario gioca un ruolo essenziale per la buona riuscita della ricerca, è opportuno analizzare più in dettaglio la sua attività. Riassumendo quanto detto poc'anzi, si può dire che l'intermediario ha sei *compiti* principali: aiutare l'utente ad esplicitare il BI e a concettualizzare la richiesta, individuare la miglior strategia di ricerca, tradurre il contenuto concettuale e la struttura logica in interrogazione, aiutare l'utente a valutare i risultati e guidare l'utente durante la fase di riformulazione dell'interrogazione. Inoltre l'intermediario provvede a isolare l'utente da tutte le difficoltà tecniche legate all'utilizzo di un SIR. Per lo svolgimento di tali compiti l'intermediario ha a disposizione strumenti di ausilio alla ricerca contenenti conoscenze terminologiche sui domini trattati nelle banche dati, quali ad esempio thesaurus [Danesi, 1990], schemi di classificazione, vocabolari controllati, e soggettari. Inoltre l'intermediario conosce - almeno a grandi linee - il contenuto delle banche dati e i criteri utilizzati per l'indicizzazione dei documenti, si fa un'idea dell'utente con cui interagisce, sa organizzare l'intero processo di interazione fra utente e SIR, possiede alcune conoscenze sul dominio e sul particolare SIR utilizzato, conosce le tecniche più adatte per accedere alla banca dati, ed ha buone capacità di interazione con gli utenti.

2.2 *Limitazioni dei sistemi di information retrieval tradizionali*

Lo scenario illustrato nel paragrafo precedente non è esente da problemi che limitano le prestazioni che ci si può attendere da un SIR tradizionale. Si considerino in particolare:

- La scarsa conoscenza che l'intermediario e l'utente posseggono sui criteri usati per indicizzare i documenti, ossia le modalità con cui vengono decise le parole chiave inserite nella descrizione del documento: ad un certo termine possono essere attribuiti significati diversi, rendendo così più complicato il problema del reperimento dei documenti rilevanti relativamente ad un certo BI. L'utente finale cioè non sa con esattezza come i concetti da lui cercati siano stati citati o descritti al momento dell'indicizzazione dei docu-

menti. È ovvio quindi che accedendo al sistema in modo diretto, utilizzando come termini di ricerca solamente quelli che l'utente finale specifica in una prima sintetica formulazione della sua richiesta, si ottengono dei risultati insoddisfacenti: è facile infatti ottenere documenti che contengono i termini di ricerca utilizzati in un contesto non rilevante per il bisogno informativo, oppure può accadere che non vengano estratti documenti rilevanti semplicemente perché i riferimenti ai concetti di interesse in essi contenuti non sono espressi mediante i termini di ricerca utilizzati dall'utente, bensì in forma diversa (ad esempio mediante sinonimi, parafrasi o, peggio, mediante una o più frasi il cui significato fa riferimento indiretto ai concetti di interesse). Tale problema è noto con il nome di problema del vocabolario [Furnas et al., 1987].

- Gli enormi volumi di dati archiviati: banche dati di dimensioni considerate limitate possono contenere parecchie decine di migliaia di documenti e banche dati di grosse dimensioni raggiungono diversi milioni di documenti. L'interrogazione relativa ad una singola richiesta può portare quindi al reperimento di un numero eccessivo di documenti, di cui solo pochi risultano utili: è necessario quindi un dispendioso e difficile processo di analisi e raffinamento del BI, che molto spesso non può venir condotto a causa di limiti di tempo o di budget.
- L'eterogeneità dei documenti contenuti in una banca dati: essi possono riguardare argomenti diversi, più o meno correlati fra di loro, trattati a differenti livelli di specificità e con vari gradi di completezza. Ciò rende praticamente impossibile conoscere (sia da parte dell'intermediario che dell'utente abituale) il contenuto di una banca dati, se non in modo approssimato, e porta anche ad una maggiore inconsistenza di indicizzazione e al reperimento di documenti non attinenti.
- La difficoltà per l'utente di identificare ed esplicitare adeguatamente il proprio BI (ossia, la difficoltà ad eseguire le operazioni di percezione, esplicitazione e concettualizzazione). Vari aspetti di questo problema fondamentale dell'IR sono stati identificati da diversi studiosi ed etichettati in vari modi: in [Taylor, 1968] si parla di *visceral need,* in

[Belkin et al., 1982] di *anomalous state of knowledge* e in [Ingwersen, 1992] di *muddled need* e di *label effect*. I metodi che vengono usati per eliminare, o quantomeno limitare, i danni causati da tale problema si basano pesantemente sull'interazione uomo-uomo (fra utente e intermediario), e gli intermediari efficaci sono in grado di comprendere il BI dell'utente, il contesto in cui esso sorge, come questo cambia durante la ricerca, come la sua descrizione può mutare nel tempo, ecc.

- Il comportamento poco strutturato dell'utente durante la ricerca di informazione: egli può seguire differenti strategie, da vari punti di partenza e con differenti obiettivi. Ad esempio, l'utente può sapere dell'esistenza di alcuni documenti, o alcuni autori; egli può conoscere qualche documento contenente parecchie indicazioni bibliografiche utili; egli può non avere bisogno di certi documenti potenzialmente utili perché già conosciuti; egli può giudicare interessante un documento che non è attinente all'argomento della ricerca ma è comunque utile per risolvere il suo problema [Harter, 1992].

È chiaro che le possibilità di guidare le operazioni di reperimento mediante la ricerca e successiva combinazione booleana dei termini di ricerca non è adeguata a risolvere i problemi appena citati ed è invece preferibile un approccio che utilizzi tecniche che considerino la semantica dei documenti e delle richieste formulate dagli utenti del sistema. Ed è infatti proprio questo il motivo per cui gli utenti finali sono solitamente costretti a ricorrere ad un intermediario; questa soluzione d'altro canto è insoddisfacente per almeno tre motivi:

- È costosa ed impraticabile su larga scala (e vista la diffusione dell'IR e delle banche dati testuali questa è una limitazione pesante).
- Non sempre permette il contatto diretto tra utente e banca dati. Talvolta ciò può essere positivo, ma in altri casi può risultare un serio ostacolo per l'utente [Bates, 1990].
- Comunque l'intermediario umano è soggetto ad alcune limitazioni intrinseche nella propria natura umana: non può conoscere tutti i potenziali domini di interesse, può non essere aggiornato sugli ultimi cambiamenti intervenuti sulla specifica banca dati e sul linguaggio di interrogazione del

SIR, può dimenticarsi dell'esistenza di opportuni strumenti di ricerca, e così via.

Pertanto l'obiettivo di realizzare almeno alcune delle funzioni dell'intermediario in un agente artificiale informatizzato riveste un notevole interesse dal punto di vista applicativo ed è stato uno degli obiettivi principali del settore dell'Information Retrieval Intelligente.

2.3 *L'information retrieval intelligente*

Le considerazioni illustrate nel paragrafo precedente hanno dato origine al settore di ricerca noto con il termine di *Information Retrieval Intelligente* (*IRI*) [Croft, 1987], dedicato specificamente allo studio, allo sviluppo ed alla sperimentazione di *Sistemi di Information Retrieval Intelligenti* (*SIRI*), ossia quella categoria di sistemi informatici che gestiscono banche dati e che sono basati, per qualche loro componente, su tecnologie sviluppate nell'ambito dell'Intelligenza Artificiale (IA) o, più specificamente, dei Sistemi Basati sulla Conoscenza [Rich e Knight, 1991; Fum, 1994; Guida e Tasso, 1994], quali ad esempio una base di conoscenza usata per inferenze terminologiche, oppure una rete neurale per confrontare un documento con una richiesta. Il ruolo dell'IA sembra essere necessario in quanto i problemi da risolvere sono definibili con difficoltà (non è perfettamente chiaro quali siano l'input e l'output del problema, né si può definire in maniera formale, completa e corretta la relazione che intercorre tra essi). Pertanto, dato che le tecniche di IA permettono di rappresentare e ragionare su conoscenza euristica e di realizzare processi di adattamento e apprendimento, esse risultano le più adatte ad essere incluse in sistemi sofisticati di IR.

La ricerca nel settore dell'IRI è multidisciplinare, in quanto coinvolge discipline e problematiche che vanno da aspetti tecnici legati all'IR ad altri legati all'interazione uomo-macchina, all'elaborazione del linguaggio naturale, ai sistemi basati su conoscenza. La classificazione che segue, sebbene non esaustiva, fornisce una breve descrizione di quelle che sono le direzioni principali di ricerca che la comunità scientifica ha seguito nell'ultimo decennio.

Una prima direzione è relativa alla definizione di *modelli concettuali* di riferimento di quelle che dovrebbero essere le funzionalità ideali di un SIRI. Nel corso degli anni vari gruppi di ricerca hanno analizzato in dettaglio il comportamento tipico degli utenti durante la ricerca di informazioni, in contesti generali oppure in situazioni più specifiche (ad esempio la ricerca di informazioni che avviene nell'ambito di un team di progettisti durante l'attività di progettazione di qualche artefatto; oppure la ricerca di immagini da una base di dati pittorica), al fine di evidenziare e caratterizzare obiettivi, problemi, metodi, strumenti e competenze che permettano di comprendere il problema della ricerca delle informazioni così come esso viene percepito da un utente finale. In queste ricerche vengono identificati e descritti i comportamenti di un utente, i problemi in cui esso si imbatte e il tipo di supporto di cui egli necessita da parte di un intermediario. I più noti modelli proposti sono MONSTRAT [Belkin et al., 1987] e MEDIATOR [Ingwersen, 1992]. Entrambi propongono una serie di funzionalità del sistema che vanno dal sintetizzare un'interrogazione al costruirsi un modello dell'utente, dallo scegliere la banca dati più adatta ad un bisogno informativo al supportare l'utente nella valutazione dell'utilità dei documenti reperiti.

Un'altra direzione di ricerca mira allo studio e allo sviluppo di sistemi che svolgano il ruolo di *intermediari artificiali* fornendo aiuto all'utente in modo da risolvere, almeno in parte, i problemi che affliggono l'IR. Le ricerche in questa direzione, complementari a quelle precedentemente descritte, si articolano in vari sotto-obiettivi. C'è chi ha studiato il problema di realizzare una biblioteca virtuale in cui gli utenti hanno a disposizione la possibilità di scegliere le "stanze" in cui trovare libri di determinate categorie (ad esempio romanzi d'avventura per ragazzi, fantascienza, e così via). Nell'ambito di questo progetto (denominato Bookhouse, [Mark Pejtersen, 1989]) sono stati studiati a fondo i comportamenti e le strategie di ricerca adottate dagli utenti, nonché l'efficacia di vari strumenti grafici e di manipolazione messi a disposizione degli utenti finali. Bookhouse di per sé non cerca di comprendere il bisogno informativo, ma offre all'utente un insieme di strumenti che gli permettono di raggiungere le informazioni che sta cercando.

Un altro sotto-obiettivo di ricerca concerne la realizzazione di sistemi basati sulla conoscenza che emulano, almeno in parte, le attività di un intermediario umano, inclusa la capacità di comprendere (almeno grossolanamente) il problema informativo. È il caso del progetto I3R [Thompson e Croft, 1989], del sistema proposto in [Gauch e Smith, 1989], di RUBRIC [Tong et al., 1987] e di FIRE [Brajnik et al, 1991b]. In questi tipi di sistemi le conoscenze rappresentate codificano le strategie e le tattiche tipicamente utilizzate dagli intermediari umani per risolvere problemi informativi. Tali conoscenze possono essere indipendenti dal dominio (e quindi possono venir applicate dal sistema a qualsiasi problema informativo), oppure possono riguardare aspetti specificamente legati al dominio (ad esempio, a RUBRIC è possibile fornire regole che permettono di definire il concetto "information retrieval" in base ai termini "information" e "retrieval" in modo che esso possa venir inferito, magari con diversi gradi di certezza, a partire da espressioni presenti nel testo come "retrieval of information", "information retrieval" o "retrieval from unstructured information sources").

Questo approccio, incentrato sullo sviluppo di basi di conoscenza di dominio, viene seguito non solo da coloro che intendono realizzare dei meccanismi per riconoscere determinati concetti a partire dal testo (come in RUBRIC), ma anche nei sistemi in cui si desidera rappresentare le definizioni dei concetti in base alle relazioni semantiche che possono intercorrere tra essi. È questo il caso di EPX [Smith, Shute e Galdes, 1989], un sistema che utilizza una base di conoscenza concettuale del dominio (la chimica) allo scopo di ottenere una chiara definizione del problema informativo.

Un'ulteriore categoria di approcci all'IRI è costituita da ricerche aventi come obiettivo quello di affrontare il problema dell'interazione uomo–macchina nel contesto dell'IR. Sistemi come Grundy [Rich, 1979] o UMT [Brajnik, Guida e Tasso, 1990; Brajnik e Tasso, 1994] sono stati sviluppati per studiare metodi automatici per acquisire dall'utente informazioni che possono venir usate per migliorare l'interazione con esso. Il sistema cerca di costruirsi un modello dell'utente in termini delle conoscenze che egli possiede, delle sue preferenze di interazione o di soluzione del problema. Il modello viene poi usato per decidere in

che modo formulare una determinata domanda, o se fornire una spiegazione all'utente e a che livello di dettaglio.

Come si può notare, molti dei lavori precedentemente richiamati possono essere considerati, più in generale, come un contributo alla realizzazione di interfacce intelligenti verso sistemi di information retrieval. In tale direzione si muove, anche se a livello più astratto, il lavoro di [Bates, 1990], dedicato ad identificare il confine più opportuno tra il ruolo dell'uomo e quello della macchina in un'interfaccia di tale tipo. In particolare, Bates propone di usare due dimensioni indipendenti per analizzare il problema:

- il *livello di coinvolgimento*: che specifica come il controllo dell'interazione viene distribuito tra i due partecipanti. Bates prevede cinque livelli: *(i)* l'utente ha il controllo totale dell'interazione e l'interfaccia si limita ad eseguire i semplici comandi che le vengono impartiti; *(ii)* l'interfaccia è in grado, su richiesta, di proporre in maniera contestuale un insieme di possibili attività; *(iii)* l'interfaccia è in grado di svolgere, su richiesta, anche attività relativamente complesse; *(iv)* l'interfaccia "osserva" il comportamento dell'utente e fornisce suggerimenti; infine *(v)* l'interfaccia risolve in maniera completamente automatica tutto il problema di ricerca.
- il *livello di astrazione* delle attività svolte dall'interfaccia, le quali possono essere suddivise in: *(i)* mosse, (azioni atomiche quali, ad esempio l'inserimento di un operatore booleano), *(ii)* tattiche (insiemi di mosse volte a migliorare la ricerca), *(iii)* stratagemmi (insiemi di tattiche e di azioni atomiche che sfruttano la struttura della specifica banca dati) e *(iv)* strategie (piani globali per affrontare la ricerca).

Quali combinazioni di scelte possano essere fatte sulle due dimensioni al fine di ottimizzare l'efficacia delle prestazioni di un SIRI rimane un problema aperto e rappresenta un difficile passo del progetto dell'interfaccia. Non va dimenticato infatti che l'accesso *diretto* dell'utente al sistema, non mediato cioè dalla presenza di un intermediario umano, comporta una serie di limitazioni significative:

- il linguaggio di comunicazione con un intermediario artificiale è assai meno espressivo e flessibile del linguaggio naturale utilizzato con l'intermediario umano;

- la "larghezza di banda" del canale di interazione tra utente e calcolatore è assai più ristretta di quella tra uomo e uomo e ciò limita l'efficacia della comunicazione, sia in termini qualitativi che quantitativi;
- il dialogo risulta assai impoverito a causa delle limitate capacità cognitive del calcolatore ed in particolare del modello molto semplificato (se non addirittura assente) che il calcolatore possiede sull'utente.

3. Il Progetto FIRE

Il progetto illustrato in questo paragrafo si inserisce specificamente nel filone di ricerca riguardante lo studio e la realizzazione di interfacce intelligenti per sistemi di information retrieval che siano in grado di emulare alcune delle prestazioni tipiche degli intermediari umani.

3.1. *Introduzione*

Il progetto *FIRE* (*Flexible Information Retrieval Environment*) [Brajnik et al., 1990; 1991a; 1991b; 1991c] è stato svolto in collaborazione con la Datamat S.p.A. di Roma nell'ambito del Progetto Finalizzato "Sistemi Informatici e Calcolo Parallelo" del Consiglio Nazionale delle Ricerche. In generale le tematiche affrontate nel progetto riguardano i sistemi intelligenti di interfaccia cooperativa e flessibile per utenti finali di sistemi di basi di dati bibliografici. Il supporto può essere fornito all'utente a vari livelli: a livello concettuale l'interfaccia può coadiuvare l'esplicitazione, la concettualizzazione e la riformulazione del bisogno informativo, a livello più operativo può essere di supporto nelle specifiche operazioni di accesso, interpretazione, elaborazione e valutazione delle informazioni estratte.

In particolare gli obiettivi del progetto FIRE si focalizzano su tre aspetti:

- Aspetti generali. Essi riguardano: *(i)* la definizione, la realizzazione e la sperimentazione di un'interfaccia evoluta per basi di dati bibliografici; e *(ii)* la proposta di criteri metodo-

logici e tecniche specifiche per la sperimentazione e la valutazione di interfacce cooperative e flessibili per sistemi di basi di dati bibliografici.

- Supporto alla ricerca delle informazioni. Tali obiettivi riguardano: *(i)* la modellizzazione della conoscenza di cui è dotato il sottosistema dedicato a svolgere le funzioni di comprensione delle richieste formulate dall'utente e di riformulazione delle richieste stesse al fine di renderle più aderenti agli effettivi scopi informativi dell'utente; e *(ii)* il progetto e la realizzazione di tale sottosistema mediante tecniche di intelligenza artificiale.
- Modellizzazione di utenti individuali. In particolare tali aspetti riguardano: *(i)* lo sviluppo di una metodologia generale per la modellizzazione di utenti individuali e la sua validazione sperimentale; *(ii)* lo studio e la definizione del ruolo del sottosistema di modellizzazione; *(iii)* il progetto e la realizzazione del sottosistema di modellizzazione; e *(iv)* la verifica sperimentale dell'efficacia del sottosistema di modellizzazione. Questo aspetto del progetto non è trattato in questo lavoro.

Per realizzare le capacità di FIRE, sono utilizzati esplicitamente diversi tipi di conoscenze:

- Conoscenza esperta, che modella la conoscenza esperta dell'intermediario umano, ossia le tecniche che un intermediario usa per interagire con l'utente e per progettare la strategia di ricerca.
- Conoscenza terminologica dello specifico dominio della banca dati, che consente di scegliere i termini nelle fasi di concettualizzazione del BI e di riformulazione dell'interrogazione.
- Conoscenze morfologica, utilizzata per inserire nell'interrogazione termini troncati o varianti morfologiche dei termini già presenti.
- Conoscenza sulle modalità di accesso alla banca dati, che include informazioni sui campi esistenti in ciascun surrogato di documento, sui comandi d'accesso, e così via.

3.2. *Architettura del prototipo FIRE*

L'architettura complessiva di FIRE è illustrata in fig. 2. Il sistema è composto dai seguenti 4 moduli principali:

- l'*Interfaccia Utente* (*User Interface - UI*), dedicata a fornire agli utenti finali tutti gli strumenti di accesso al sistema;
- il *Sistema di Sviluppo delle Basi di Conoscenza* (*Knowledge Base Development System - KBDS*), dedicato a fornire un supporto interattivo e possibilmente (semi)automatico agli sviluppatori del sistema, ed in particolare agli ingegneri della conoscenza[1];
- il *Sistema Esperto di Information Retrieval* (*Information Retrieval Expert System - IRES*), dedicato all'emulazione dell'intermediario umano nelle sue tipiche attività di supporto alla formulazione delle richieste, di progetto della strategia di ricerca, di accesso vero e proprio al SIR e di valutazione dei risultati ottenuti;
- il *Sistema di Information Retrieval* (*Information Retrieval System - IRS*), dedicato alla memorizzazione dei documenti ed all'accesso agli stessi mediante un sistema tradizionale di tipo booleano basato su indici invertiti.

L'intero sistema è implementato su una rete di workstation Unix. Nell'implementazione di FIRE si sono utilizzati diversi linguaggi di programmazione: il modulo DBMAN è implementato in C, parte del modulo IRES in XPSE, una libreria LISP per la programmazione basata su regole di produzione sviluppata presso il Laboratorio di Intelligenza Artificiale dell'Università di Udine, l'Interfaccia Utente in CLIM, una libreria di routine grafiche per interfacciare LISP e X-Window, e il resto del sistema in Lucid Common LISP.

I quattro moduli principali e le loro componenti sono illustrati in dettaglio nei prossimi sottoparagrafi.

---

[1] Con il termine *ingegnere della conoscenza* si intende la figura professionale dedicata all'analisi ed alla modellizzazione della conoscenza degli esperti umani che svolgono l'attività che si desidera emulare con il sistema esperto.

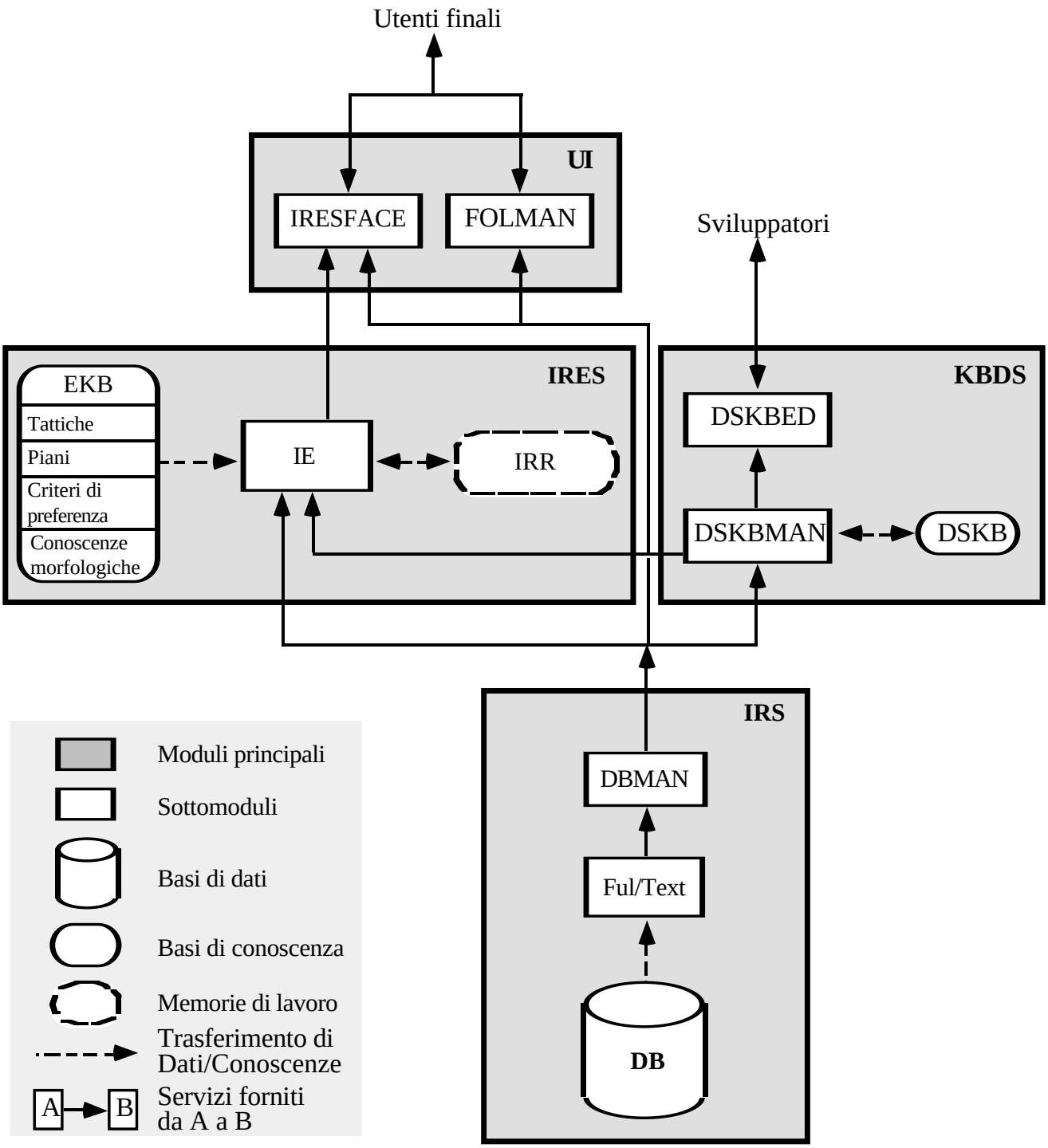

Figura 2. Architettura generale di FIRE.

3.2.1. L'Interfaccia Utente

Lo scopo di questo modulo è di permettere l'interazione fra FIRE e l'utente finale. L'interfaccia utente è composta da due sottomoduli:

- *IRESFACE* (*IRES interFACE*), che costituisce il principale canale di comunicazione con l'utente, a cui permette di:
  - comunicare al sistema la propria richiesta sotto forma di interrogazione booleana e di *vincoli sulla ricerca* (costituiti dal numero di documenti desiderati, espresso mediante un intervallo, e dal cosiddetto *obiettivo della ricerca*, che può essere 'alta precisione' se l'utente è interessato a reperire *unicamente* documenti attinenti e 'alto richiamo' se l'utente è interessato a reperire *tutti* i documenti attinenti contenuti nella banca dati);
  - visualizzare i surrogati reperiti;
  - selezionare termini da aggiungere all'interrogazione durante la fase di riformulazione.

- *FOLMAN* (*FOLder MANager*), che permette all'utente di classificare i documenti reperiti in categorie (prefissate o definite ad hoc dal singolo utente) quali 'attinente', 'non attinente', utile', 'non utile', e così via.

La comunicazione utente-sistema è progettata secondo le seguenti caratteristiche: *(i)* è basata sul modello orientato a finestre, menu e mouse, guidato da eventi; *(ii)* permette all'utente di ottenere il controllo sul sistema qualora egli lo desideri; *(iii)* permette all'utente di svolgere attività di monitoraggio, debugging e visualizzazione di varie informazioni relative al funzionamento di FIRE.

3.2.2. Il Sistema di Sviluppo delle Basi di Conoscenza

Scopo del sottosistema KBDS è di permettere all'ingegnere della conoscenza di sviluppare agevolmente e in modo interattivo le varie basi di conoscenza di cui FIRE dispone.

L'attività di inserimento e/o modifica delle conoscenze è fondamentale al fine di realizzare un'agevole sperimentazione di FIRE. Il ruolo di tale sottosistema è quindi considerato molto importante per un'effettiva realizzazione di un sistema di interfaccia evoluto, in quanto le conoscenze, esperte ed in particolar modo terminologiche sullo specifico dominio, sono presenti in notevolissima quantità. Per tale motivo, in versioni future di FIRE è previsto l'utilizzo di strumenti per l'acquisizione (semi) automatica della conoscenza (ad esempio editor intelligenti, moduli di elaborazione automatica di conoscenze terminologiche da testi in linguaggio naturale, ecc.).

Le componenti del KBDS sono:

- La base di conoscenza *DSKB* (*Base di Conoscenza Specifica del Dominio*), che include conoscenza terminologica riguardante i possibili argomenti trattati durante le sessioni con gli utenti. Tale base di conoscenza è essenzialmente basata sul contenuto di un thesaurus ossia un dizionario di termini in relazione tra loro riguardanti il settore di interesse della banca dati. In particolare le relazioni tra termini sono le seguenti: 'Termine-più-specifico' ('*NT*', da Narrower Term), 'Termine-più-generale' ('*BT*', da Broader Term), 'Termine-correlato' ('*RT*', da Related Term), 'Termine-usato-per' ('*UFT*', da Used For Term), e 'Termine-da-usare' ('*UT*', da Use Term). La DSKB contiene anche altre informazioni - solitamente non contenute in un thesaurus - quali il *posting count* dei termini (il numero delle citazioni presenti nella banca dati) ed il loro livello di specificità**.**
- Il modulo *DSKBED* (*DSKB EDitor*), un'interfaccia a menu e finestre che consente all'utente sviluppatore l'accesso alla DSKB, al fine di visionarne il contenuto e di aggiungere nuovi termini e nuove relazioni.
- Il modulo *DSKBMAN* (*DSKB MANager*), che implementa le funzioni di gestione (lettura e modifica) della DSKB.

Nella versione attuale del prototipo, il modulo KBDS è quindi limitato allo sviluppo ed alla gestione della DSKB e permette di esaminarla, definirne nodi e archi, e modificare le informazioni associate ai nodi.

3.2.3. Il Sistema Esperto di Information Retrieval

IRES, descritto in dettaglio in [Mizzaro, 1994], è un sistema basato sulla conoscenza che emula alcune delle operazioni normalmente svolte dall'intermediario umano; questo modulo si occupa in particolare di fornire supporto all'utente durante la riformulazione, suggerendogli le opportune modifiche da effettuare all'interrogazione in modo da esprimere in modo più preciso e completo il BI e produrre quindi risultati migliori.

IRES si occupa anche delle operazioni di formalizzazione e ricerca nella banca dati (operazioni che IRES è in grado di eseguire autonomamente). La riformulazione è comunque l'attività peculiare di IRES, e quindi la descrizione fornita in questa sede riguarda le componenti di IRES dedicate a tale attività.

Le componenti di IRES sono:

- la base di conoscenza *EKB* (*Expert Knowledge Base*), contenente conoscenze che rappresentano la competenza e le abilità di un intermediario umano;
- la memoria di lavoro *IRR* (*Internal Request Representation*) che racchiude lo stato corrente del problema, ossia una rappresentazione dell'interrogazione, dei risultati reperiti e degli obiettivi dell'utente;
- il motore inferenziale *IE (Inference Engine)* che in base al contenuto della Rappresentazione Interna della Richiesta seleziona ed applica le conoscenze più appropriate, procedendo in tal modo nell'attività di emulazione dell'intermediario.

IRES necessita anche di conoscenza terminologica sul dominio, reperita nella base di conoscenza DSKB descritta nel paragrafo precedente.

La EKB può essere suddivisa in: tattiche, piani, criteri di preferenza e conoscenze morfologiche. Queste conoscenze, descritte qui di seguito, sono sfruttate per supportare l'utente durante la riformulazione, la quale risulta essere una sequenza di operazioni di modifica dell'interrogazione, che vengono proposte dal sistema e confermate o meno dall'utente.

Una *tattica* è un'operazione atomica di modifica dell'interrogazione. Le tattiche implementate in IRES sono derivate da quelle illustrate in [Bates, 1979a; 1979b] e in [Brajnik et al.,

1991c]; esempi di tattiche sono: aggiungere a una faccetta i termini 'RT' (o 'BT' o 'NT') di un termine già appartenente a tale faccetta, aggiungere a una faccetta il termine troncato di un termine già incluso in tale faccetta, disattivare un termine, eccetera. L'esecuzione delle tattiche può richiedere quindi l'accesso alle conoscenze terminologiche contenute nella DSKB e l'utilizzo delle conoscenze morfologiche.

Un *piano* è una sequenza predefinita di una o più tattiche. Il raggruppamento delle tattiche in piani è giustificato da due motivi. Da un punto di vista concettuale, non sempre esistono criteri per poter scegliere in modo chiaro fra due tattiche alternative, oppure può essere sensato far seguire in ogni caso l'applicazione di una tattica ad un'altra. Da un punto di vista pragmatico, l'utilizzo dei piani permette di ottenere una maggiore efficienza temporale del sistema. Un esempio di piano è la successione delle 3 tattiche che, partendo da un dato termine in una faccetta, vi aggiungono i termini troncati, i termini correlati da una relazione di tipo RT nella DSKB, ed i termini morfologicamente simili. Si osservi che un piano (e una tattica) vengono applicati ad un termine alla volta dell'interrogazione (denominato *focus*).

Le possibili operazioni di modifica dell'interrogazione sono quindi individuate univocamente dal piano da applicare e dal focus a cui applicarlo. L'individuazione del focus è ovviamente importante quanto la scelta del piano da applicare: la coppia <focus,piano> è denominata *candidato.*

Per scegliere quale piano eseguire in una data situazione, vengono ordinati i candidati secondo un ordine di preferenza (che in generale è parziale) e quindi viene scelto uno dei massimali rispetto a tale ordinamento. I *criteri di preferenza* della EKB si basano su parametri del candidato quali: tipo del piano (alcuni piani saranno, in certe situazioni, preferibili ad altri), caratteristiche del focus (ad esempio, il suo posting count), caratteristiche della faccetta a cui il focus appartiene (ad esempio, il posting count della faccetta) e così via. Due esempi di criteri sono i seguenti:

- se l'interrogazione ha due faccette, ognuna contenente un unico termine, e si deve reperire un numero di documenti maggiore di quello reperito dall'interrogazione attuale, è preferibile aggiungere termini sinonimi inizialmente nella faccetta con il posting count minore fra le due (in quanto tale faccetta è il 'collo di bottiglia' della situazione);

- se l'utente sta effettuando una ricerca ad alta precisione, preferire i piani che non contengono operazioni di troncamento dei termini.

Le *conoscenze morfologiche* sono basate su algoritmi di troncamento quali quello di Porter [Porter, 1980] e sono utilizzate per individuare la radice morfologica di una parola e per trovare termini morfologicamente simili, cioè con la stessa radice.

3.2.4. Il Sistema di Information Retrieval

Il modulo IRS è incaricato di gestire le banche dati a cui FIRE può accedere ed è costituito da due sottomoduli:

- *Ful/Text*, il vero e proprio gestore degli archivi sottostanti, costituito da un pacchetto commerciale di Information Retrieval fornito al progetto FIRE dalla Datamat S.p.A.
- *DBMAN* (*Data Base MANager*), che fornisce agli altri moduli i servizi necessari per l'accesso alle banche dati, rendendo indipendente il resto di FIRE dal particolare sistema di IR utilizzato.

Una banca dati è costituita da una serie di archivi che rappresentano i documenti e le loro informazioni accessorie (indici, ecc.). FIRE può lavorare su più banche dati bibliografiche; nella versione attuale, ne sono disponibili 3:

- INSPEC, ricavata dalla banca dati INSPEC e relativa alle applicazioni industriali delle tecniche di intelligenza artificiale, contenente 20.000 documenti**;**
- USENET, ricavata raccogliendo circa 5.000 messaggi di Usenet, estratti da vari gruppi di interesse nell'area informatica;
- BSF (Bibliografia Storica Friulana), costituita da circa 5.500 documenti relativi ad aspetti storici, economici e giuridici della vita montana del Friuli-Venezia Giulia [Tasso, 1994].

3.3. *Funzionamento del prototipo FIRE*

Il modulo con cui l'utente interagisce è l'Interfaccia Utente: all'inizio della sessione egli può scegliere fra la modalità di lavoro 'utente finale' e 'utente progettista'. Dopodiché può inserire le informazioni che caratterizzano la sua richiesta. In base a ciò il

modulo IRES progetta e costruisce una strategia di ricerca adeguata alle richieste ricevute. In tale fase, il funzionamento del sistema prevede diverse modalità di interazione con l'utente al fine di garantire il massimo grado di comprensione dei suoi bisogni e di adattamento alle specifiche caratteristiche (terminologiche) della banca dati interrogata. Quando la strategia di ricerca è sufficientemente completa, questa viene formalizzata in un'interrogazione per il Sistema di Information Retrieval che provvede ad effettuare le ricerche nella banca dati, ottenendo così i primi risultati. Questi vengono presentati all'utente che può fornire delle indicazioni correttive al fine di raffinare la vecchia strategia, di permettere nuove riformulazioni e di interrogare nuovamente la banca dati. La sessione termina quando l'utente ritiene soddisfacenti i risultati ottenuti.

La fig. 3 presenta due finestre di IRESFACE. Quella grigio scuro è la FIRE Main Window, in cui l'utente comunica al sistema la propria richiesta sotto forma di struttura logica e contenuto concettuale: questa poi viene tradotta dal sistema in un'interrogazione in forma booleana.

È quindi compito dell'utente individuare i concetti che esprimono nel modo migliore (almeno in prima approssimazione) il proprio BI, specificare i termini con cui tali concetti possono essere espressi e raggruppare opportunamente tali termini in faccette organizzate fra loro da una certa struttura logica. Per fare ciò, l'utente ha a disposizione il Query Panel, mostrato a sinistra in figura (zona grigio scuro) e contenente la specifica della richiesta esemplificativa "Neural Networks AND (Aerospace Control OR Flight Control)". Accanto ad ogni termine, sono specificati il grado di interesse (deciso dall'utente), il posting count e un indicatore che segnala se il termine è incluso o meno nella DSKB.

Nella parte alta della FIRE Main Window sono a disposizione dell'utente cinque bottoni, che permettono di lanciare la fase di riformulazione, effettuare una ricerca diretta sulla banca dati, richiamare FOLMAN per classificare i documenti, inizializzare il sistema e terminare la sessione. A destra sono visualizzati i titoli dei documenti reperiti (che l'utente può consultare selezionandone il titolo).

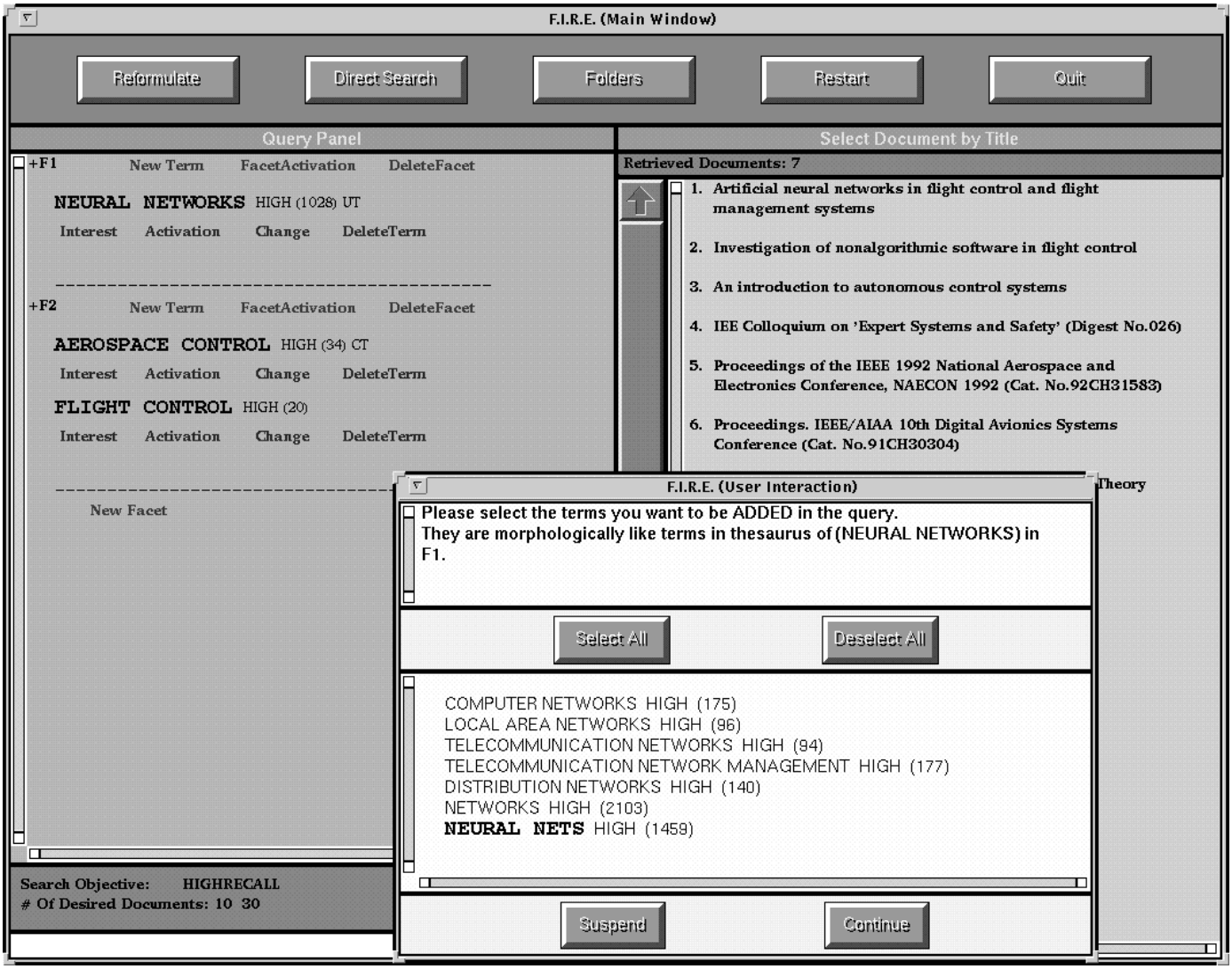

Figura 3. Finestre di IRESFACE.

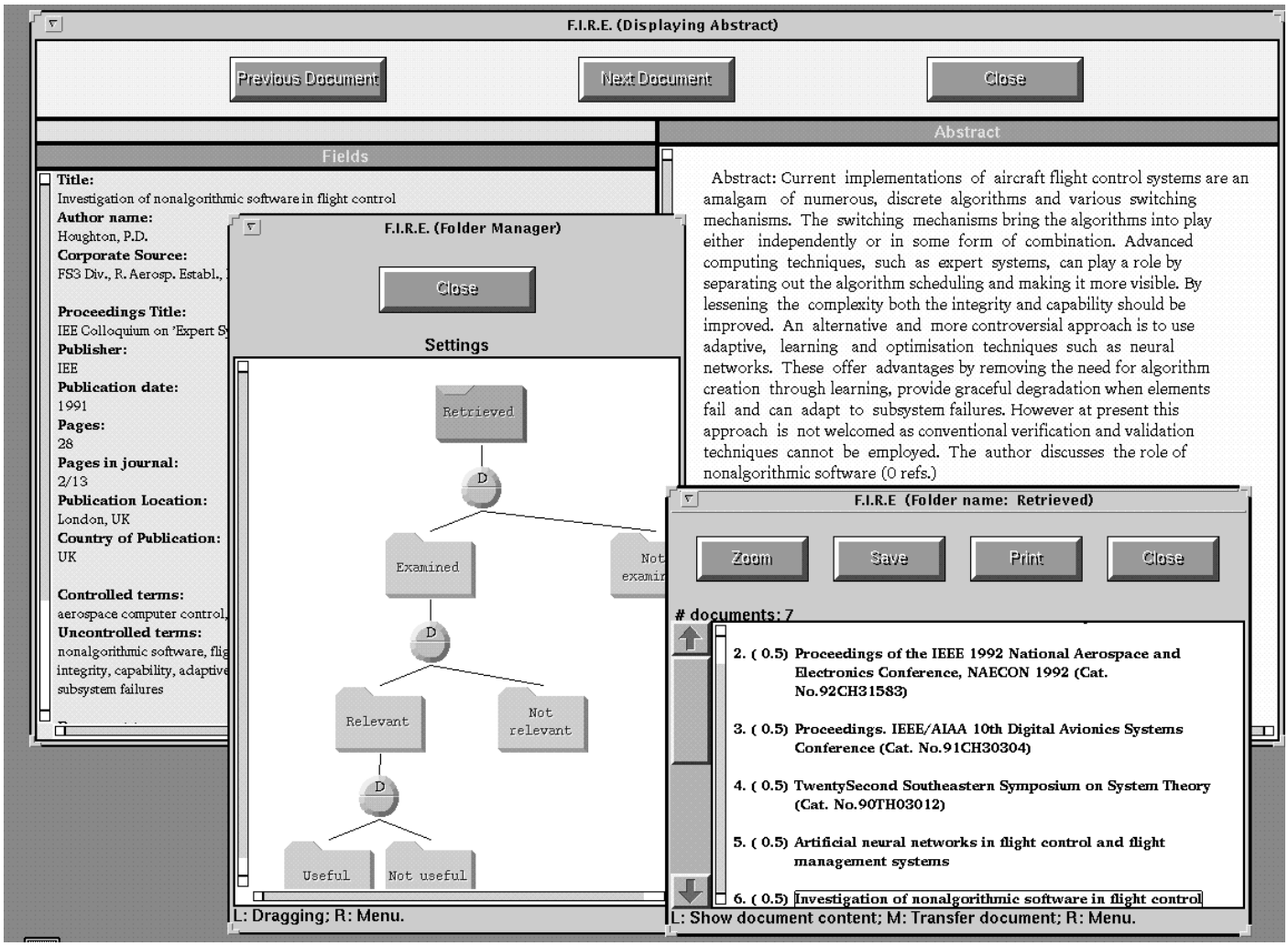

Figura 4. Finestre del FOLMAN.

La finestra in grigio chiaro mostra invece il tipo di supporto che FIRE è in grado di fornire all'utente durante la riformulazione: in tale finestra FIRE visualizza una lista di termini (ricavati tramite l'applicazione di una tattica) che possono essere usati per raffinare l'interrogazione.

In fig. 4 è mostrato il FOLMAN, lo strumento che permette all'utente la classificazione dei documenti. Nell'angolo in basso a destra sono visualizzati i titoli di alcuni documenti; l'utente può trascinare un documento in una delle cartelle della parte sinistra della figura, organizzate gerarchicamente, oppure consultare il documento (finestra sullo sfondo).

L'interazione FIRE-utente può avvenire in due modalità, *manuale* e *semiautomatica*. Nel primo caso, l'utente utilizza FIRE come un normale sistema booleano (sfruttando le potenzialità fornite dall'interfaccia a finestre): esprime il proprio BI, effettua una ricerca, esamina i titoli dei documenti reperiti, ne visiona il contenuto e li classifica usando il FOLMAN.

Se i risultati sono soddisfacenti, l'interazione con FIRE termina, altrimenti l'utente deve modificare l'interrogazione iniziale. Se ad esempio la ricerca precedente ha reperito pochi documenti l'utente può espandere l'interrogazione aggiungendo termini sinonimi a quelli utilizzati in precedenza. Per individuare tali sinonimi, l'utente può consultare le descrizioni dei documenti reperiti, o utilizzare un thesaurus cartaceo. Una volta modificata l'interrogazione, il procedimento viene iterato: si effettua una nuova ricerca, si valuta l'insieme di documenti reperiti dal sistema e così via.

Le potenzialità di FIRE vengono sfruttate appieno con la modalità di interazione semiautomatica, innescata dal bottone Reformulate: FIRE guida il processo di riformulazione, suggerendo le modifiche da effettuare all'interrogazione. È importante notare il carattere *supportivo* di FIRE: il sistema non modifica autonomamente l'interrogazione, ma si limita a suggerire tali modifiche all'utente, il quale accetta o meno le proposte del sistema.

3.4. *Valutazione del prototipo FIRE*

Fin dagli anni sessanta con l'esperimento Cranfield [Cleverdon et al., 1966], la valutazione di un SIR è stata uno dei principali oggetti di studio nel settore dell'IR.

Valutare un SIR è un compito allo stesso tempo necessario e complicato, in quanto è indispensabile poter quantificare le prestazioni di un SIR, ma a tutt'oggi non vi è accordo né sulla metodologia da seguire, né sulle metriche di valutazione da adottare [van Rijsbergen, 1979; Saracevic et al., 1988; Ingwersen, 1992; Tague-Sutcliffe, 1992; Robertson e Hancock-Beaulieu, 1992].

La necessità e la difficoltà della valutazione aumentano nel caso dei SIRI che interagiscano direttamente con l'utente finale, quali le interfacce intelligenti che emulano l'intermediario. Se tali sistemi sono realizzati mediante sistemi basati sulla conoscenza, le difficoltà della valutazione aumentano ulteriormente [Guida e Tasso, 1994].

Queste considerazioni hanno portato a sviluppare un esperimento volto a valutare, tramite un confronto con un intermediario umano, l'utilità dell'aiuto concettuale fornito da FIRE durante la riformulazione. L'aiuto che l'utente può ricevere è stato suddiviso in due categorie: terminologico (quali termini utilizzare per modificare l'interrogazione in una data situazione) e strategico (quali operazioni effettuare, in generale, per modificare l'interrogazione e reperire risultati migliori). Sulla base di tale suddivisione dei tipi di aiuto si è progettato l'impianto sperimentale schematizzato in tab. 1 e illustrato nel seguito.

	Gruppo 1	Gruppo 2	Gruppo 3
Sessione I	FIRE/M + Thesaurus	FIRE/M + Esperto T	FIRE/M + Esperto S
Sessione II	FIRE/A	FIRE/A	FIRE/A

Tabella 1. Organizzazione dell'esperimento.

Per l'esperimento (che è un esperimento di laboratorio, con BI artificiali, ossia indotti dagli sperimentatori e non realmente percepiti dagli utenti) si sono realizzate due versioni di FIRE, la

prima senza la possibilità di effettuare riformulazioni automatiche (indicata con il termine FIRE/M in tab. 1), la seconda con tale possibilità (indicata con il termine FIRE/A in tab. 1). Si sono utilizzati 45 studenti del terzo e quarto anno del Corso di Laurea in Scienze dell'Informazione, divisi in tre gruppi. Ogni soggetto ha effettuato due sessioni di uso di FIRE su due BI artificiali identici per tutti i soggetti. Nella prima sessione è stato utilizzato FIRE/M e un aiuto supplementare differente a seconda del gruppo: un thesaurus cartaceo, o un intermediario umano che fornisce esclusivamente aiuto terminologico (*Esperto T*), o un intermediario umano che fornisce esclusivamente aiuto strategico (*Esperto S*), e la seconda con FIRE/A. Le variabili indipendenti risultano quindi essere il tipo di sistema, con riformulazione o senza, e il tipo di aiuto, differente a seconda del gruppo.

Le variabili dipendenti sono state scelte al fine di ottenere una valutazione della soddisfazione globale dell'utente e quindi si sono misurate sia grandezze oggettive che soggettive. Per le grandezze del primo tipo, oltre alle classiche misure di prestazioni *recall*, *precision* ed *E-measure*, si sono utilizzate altre misure, quali il tempo impiegato, il numero di ricerche, il numero di documenti visionati; tutto ciò è stato rilevato tramite una traccia, generata automaticamente, delle operazioni effettuate dall'utente. Quali grandezze soggettive si sono utilizzate la soddisfazione per i risultati ottenuti, per il tipo di interazione, per l'aiuto ricevuto, ecc.; tali grandezze sono state rilevate con questionari, preparati ad hoc per l'esperimento, e con l'analisi delle videoregistrazioni delle sessioni.

I risultati di questa sperimentazione sono attualmente in fase di elaborazione e saranno oggetto di un futuro lavoro.

4. Conclusioni e sviluppi futuri

In questo lavoro si è illustrato il settore dell'information retrieval intelligente, nell'ambito del quale si affrontano con tecniche di intelligenza artificiale le limitazioni dei sistemi di information retrieval tradizionali. In particolare è stato descritto il prototipo FIRE, un'interfaccia intelligente a sistemi di information retrieval e di tale prototipo sono state presentate l'architettura, il funzionamento e la valutazione.

Numerosi sono gli sviluppi futuri pianificati nel progetto FIRE, ed in particolare:

- una modellizzazione più completa del bisogno informativo;
- l'individuazione di strumenti adatti ad aiutare l'utente nel processo di concettualizzazione;
- l'utilizzo di strumenti grafici per la navigazione nella rete terminologica dei termini relativi al dominio e l'integrazione di tali strumenti con i meccanismi utilizzati nel modulo IRES;
- la possibilità per l'utente di lavorare autonomamente e di essere osservato e corretto dal sistema, che agisce quindi da tutore.

Ringraziamenti

Numerose persone hanno collaborato negli anni al progetto FIRE e ad esse va il nostro più sentito ringraziamento. Si vogliono qui menzionare in modo particolare il dr. Pierluigi Marchetti dell'ESA/IRS di Frascati per aver fornito le parti della banca dati INSPEC utilizzate nelle sperimentazione, la dr.ssa Anna Floreanini per la realizzazione della banca dati BSF, ed il sig. Alessandro Bortuzzo per la parte sperimentale di valutazione.

Riferimenti

Bates, M.J. (1979a). Information Search Tactics. *Journal of the American Society for Information Science*, July 1979, pp. 205-214.

Bates, M.J. (1979b). Idea tactics. *Journal of the American Society for Information Science*, September 1979, pp. 280-289.

Bates, M.J. (1990). Where should the person stop and the information interface start?. *Information Processing and Management* 26(5), pp. 575-591.

Belkin, N., Croft, B. (1987). Retrieval Techniques. *Annual Review of Information Science and Technology* (*ARIST*) 22, pp. 109-145.

Belkin, N., et al. (1987). Distributed expert-based information systems: an interdisciplinary approach. *Information Processing and Management* 23(4), pp. 249-254.

Belkin, N.J., Oddy, R.N., Brooks, H.M. (1982). ASK for information retrieval. *Journal of Documentation* 38(2), 61-71 e 38(3), pp. 145-164.

Brajnik, G., Guida, G., Mastrodonato, L., Scaroni, C., Tasso, C. (1990). *Progetto generale del sistema FIRE: un ambiente flessibile per lo sviluppo di interfacce esperte a sistemi di basi di dati bibliografici*. Progetto finalizzato CNR Sistemi Informatici e Calcolo Parallelo, Sottoprogetto 5. Rapporto Stato di Avanzamento n. 5/43.

Brajnik, G., Guida, G., Tasso, C. (1990). User Modeling in Expert Man-Machine Interfaces: a Case Study in Intelligent Information Retrieval. *IEEE Trans. on SMC* 20(1), p. 166-185.

Brajnik, G., Mastrodonato, L., Scaroni ,C., Tasso, C. (1991a). *Progetto dell'interfaccia utente del sistema FIRE*. Progetto finalizzato CNR Sistemi Informatici e Calcolo Parallelo, Sottoprogetto 5. Rapporto Stato di Avanzamento n. 5/62.

Brajnik, G., Tasso, C., Mastrodonato, L., Scaroni, C. (1991b). FIRE: un prototipo di ambiente di sviluppo per interfacce intelligenti e cooperative per l'accesso a banche di dati bibliografici. In B. Fadini (a cura di), *Sistemi informatici e Calcolo parallelo*, Franco Angeli, Milano, pp. 258-266.

Brajnik, G., Guida, G., Mastrodonato, L., Scaroni, C., Tasso, C. (1991c). *Specifiche di progetto del modulo per il retrieval intelligente IRES nell'ambito del sistema FIRE*. Rapporto Tecnico CNR, n. 5/61, sottoprogetto Sistemi evoluti per Basi di Dati, Progetto Finalizzato CNR Sistemi Informatici e Calcolo Parallelo.

Brajnik, G., Tasso, C. (1994). A shell for developing non-monotonic user modeling systems. *International Journal of Human-Computer Studies* 40, pp. 31-62.

Cleverdon, C.W., Mills, J., Keen, M. (1966). *Factors Determining the Performance of Indexing Systems, Vol. 1: Design, Vol. 2: Test results*. College of Aeronautics, Cranfield, UK.

Croft, W.B. (1987). Approaches to intelligent information retrieval. *Information Processing and Management* 23(4), pp. 249-254.

Danesi, D. (1990). *Le variabili del thesauro - Gestione e struttura*. IFNIA, Laboratorio Thesauri, Firenze.

Fum, D. (1994). *Intelligenza artificiale*. il Mulino, Bologna.

Furnas, G.W., Landauer, T.K., Gomez, L.M., Dumais, S.T. (1987). The Vocabulary Problem in Human-System Communications. *Communications of the ACM* 30(11), pp. 964-971.

Gauch, S.E., Smith, J.B. (1989). Query reformulation strategies for searching in full–text. *Information Processing and Management* 25(3), pp. 253–263.

Guida, G., Tasso, C. (1994). *Design and Development of Knowledge-Based Systems – From Life Cycle to Methodology*. John Wiley & Sons, Chichester, UK.

Harter, S.P. (1992). Psychological Relevance and Information Science. *Journal of the American Society for Information Science* 43(9), pp. 602-615.

Ingwersen, P. (1992). *Information Retrieval Interaction*. Taylor Graham, London.

Lancaster, F.W. (1968). *Information Retrieval Systems: Characteristics, Testing and Evaluation*. John Wiley & Sons, New York.

Mark Pejtersen, A. (1989). A library system for information retrieval based on a cognitive task analysis and supported by an icon–based interface. *Proc. of the 12th ACM SIGIR International Conference on Research and Development in Information Retrieval*. Cambridge, MA, pp. 40–47.

McAlpine, G. Ingwersen, P. (1989). Integrated information retrieval in a knowledge worker support system. *ACM SIGIR Forum*, June, pp. 48-57.

Meadow, C.T., Cochrane, P.A. (1981). *Basics of Online Searching*. John Wiley & Sons, Chichester, UK.

Mizzaro, S. (1994). *La componente esperta del sistema FIRE*. Progetto finalizzato CNR Sistemi Informatici e Calcolo Parallelo, Sottoprogetto 5. Rapporto Stato di Avanzamento n. R/5/137.

Porter, M.F., 1980. An algorithm for suffix stripping. *Program* 14 (3), pp. 130-137.

Rich, E. (1979). User modelling via stereotypes. *Cognitive Science* 3, pp. 329-354.

Rich, E., Knight K. (1991). *Artificial Intelligence* (2nd edition). McGraw-Hill, New-York, NY.

Robertson, S.E., Hancock-Beaulieu, M.M. (1992). On the evaluation of IR systems. *Information Processing and Management* 28(4), pp. 457-466.

Salton, G. (1989). *Automatic Text Processing – The Transformation, Analysis, and Retrieval of Information by Computer*, Addison-Wesley, Reading, MA.

Saracevic, T., Kantor, P., Chamis, A., Trivison, D. (1988). A Study of Information Seeking and Retrieving, I, Background and Methodology. *Journal of the American Society for Information Science* 39(3), pp. 161-176.

Smith P.J., Shute S.J., Galdes D. (1989). In search of knowledge–based search tactics. *Proc. of the 12th ACM SIGIR International Conference on Research and Development in Information Retrieval*. Cambridge, MA, pp. 3–10.

Tague-Sutcliffe, J. (1992). The pragmatics of information retrieval experimentation, revisited. *Information Processing and Management* 28(4), pp. 467-490.

Tasso, C. (1994). L'intelligenza artificiale applicata alla ricerca bibliografica: una sperimentazione nell'ambito di una banca dati di interesse regionale. In A. De Cillia (a cura di),*Cultura e informatica – L'intelligenza artificiale nella ricerca bibliografica*, Arti Grafiche Friulane, Udine, pp. 53-77.

Taylor, R.S. (1968). Question Negotiation and information seeking in libraries. *College and Research Libraries* 29, pp. 178-194.

Thompson, R.H., Croft, W.B. (1989). Support for browsing in an intelligent Text Retrieval System. *International Journal of Man–Machine Studies* 30, pp. 639-668.

Tong, R.M., Appelbaum, L.A., Askman V.N., Cunningham, J.F. (1987). Conceptual information retrieval using RUBRIC. *Proc. of the 10th ACM SIGIR International Conference on Research and Development in Information Retrieval*. New Orleans, LU, pp. 247–253.

van Rijsbergen, C.J. (1979). *Information Retrieval* (2nd edition). Butterworths, London.