L'evoluzione dei metodi per la valutazione dei sistemi di information retrieval

Claudio Carpineto e Stefano Mizzaro

Fondazione Ugo Bordoni,
Via B. Castiglione 59, 00142 Rome, Italy

Sommario

In questa relazione si esaminano i metodi tradizionali per la valutazione di laboratorio dei sistemi di "information retrieval" e si discute la loro estensione a problemi di larga scala, con particolare riferimento all'esperienza TREC. Successivamente si considerano i requisiti dei metodi interattivi per la selezione delle informazioni e i problemi connessi alla loro valutazione. In particolare, si presentano i risultati relativi alla valutazione di FIRE, un sistema interattivo che fornisce supporto all'utente nella riformulazione di interrogazioni booleane. Infine si delineano i nuovi scenari di valutazione, più orientati al ruolo e agli obiettivi dell'utente e alla dinamica dell'interazione utente-sistema nella ricerca delle informazioni.

INDICE

1. Introduzione 3

2. Metodi di valutazione tradizionali 4

2.1 Curve richiamo-precisione (recall-precision) 5

2.2 Richiamo e precisione a parità di documenti recuperati 6

2.3 Altre misure di prestazione 7

3. L'esperienza TREC 8

3.1 I dati degli esperimenti TREC 9

3.2 L'interattività in TREC 11

4. FIRE: information retrieval con supporto alla riformulazione delle interrogazioni 12

4.1 La valutazione di sistemi interattivi con metodi quantitativi e qualitativi 13

4.2 Il prototipo FIRE 14

4.3 Valutazione di FIRE: tipi di aiuto 16

4.4 Valutazione di FIRE: obiettivi 17

4.5 Valutazione di FIRE: disegno sperimentale 17

4.5.1 Soggetti e gruppi sperimentali 18

4.5.2 Variabili indipendenti e dipendenti 20

4.5.3 Materiale e procedure 20

4.6 Valutazione di FIRE: risultati 21

4.6.1 Soddisfazione dell'utente 21

4.6.2 Valore aggiunto della riformulazione automatica 22

4.6.3 Importanza di tipi e modalità degli aiuti 24

4.7 Conclusioni della valutazione con FIRE 26

5. Verso nuovi paradigmi di valutazione 28

Ringraziamenti 31

Riferimenti bibliografici 32

1. Introduzione

I grandi progressi tecnologici che sono stati registrati negli ultimi anni nella raccolta, memorizzazione e diffusione delle informazioni digitali, dai CD-ROM, alle reti locali di calcolatori fino ad Internet, hanno creato anche una robusta offerta di nuovi servizi e strumenti per il reperimento e la selezione delle informazioni. Basti pensare al World-Wide Web e ai motori di ricerca per l'accesso diretto alle informazioni su Internet, che oggi sono adoperati da milioni di persone. È anche per questo che da alcuni anni si guarda con un forte e rinnovato interesse al campo della valutazione dei sistemi di information retrieval, che aveva conosciuto un forte impulso e ottenuto risultati importanti fino ai primi anni '70, seguiti poi da un periodo di sostanziale stasi. Prima di cominciare ad analizzare i metodi di valutazione cerchiamo di definire sinteticamente il processo di information retrieval da valutare.

In generale, si tratta di consentire ad un utente di accedere alle informazioni di suo interesse. Il punto di partenza è quindi quello che si chiama bisogno informativo, o stato anomalo di conoscenza, che spinge un utente ad impegnarsi attivamente nella ricerca di informazioni, ad esempio formulando un'interrogazione ad un sistema di information retrieval mediante un linguaggio che può essere compreso dal sistema, oppure navigando attraverso un ipertesto. L'interrogazione viene confrontata dal sistema con le basi documentali che esso gestisce, o meglio con le rappresentazioni di esse che sono state costruite manualmente o automaticamente dal sistema stesso. Il risultato del confronto fra l'interrogazione e le rappresentazioni dei documenti, oppure dell'ispezione dell'ipertesto da parte dell'utente, consiste di solito di un insieme di documenti di potenziale interesse per l'utente, la cui valutazione può spingere l'utente a formulare una nuova interrogazione, o a continuare nella navigazione, e occasionalmente può comportare una modifica del bisogno informativo o anche un cambiamento della rappresentazione dei documenti.

E' chiaro che assumendo una caratterizzazione del processo di information retrieval di questo tipo, peraltro comparabile con la situazione riscontrabile in ambienti operativi, le variabili che entrano in gioco nella sua valutazione sono molteplici. Esse riguardano principalmente la collezione documentale, la sua rappresentazione, il tipo di utente, il bisogno informativo, la trasformazione di quest'ultimo in un'interrogazione da immettere nel sistema, il metodo di ricerca dei documenti pertinenti, la misura di prestazione e l'analisi dei risultati. E' difficile valutare l'efficacia di un processo così complesso. Una possibilità è quella di cercare di misurare le prestazioni di un sistema in ambiente operativo, quindi con utenti, bisogni e procedure di ricerca reali. Questo tipo di valutazione è sicuramente utile però produce risultati di natura globale, che sono difficilmente imputabili ai fattori specifici determinanti. Per converso, le valutazioni di laboratorio consentono in linea di principio di controllare un maggior numero di

variabili e in modo maggiormente mirato agli obiettivi dell'esperimento. In effetti, esiste una tradizione largamente prevalente che ha privilegiato questo tipo di valutazione. Si veda ad esempio la discussione contenuta in Tague-Sutcliffe 1992 relativamente alle scelte che è possibile fare per ciascuna delle variabili indicata precedentemente, con l'obiettivo di accrescere la validità e l'affidabilità dell'esperimento.

In questo articolo consideriamo solamente valutazioni di laboratorio e ci riferiamo, almeno inizialmente, ad un contesto più ristretto, in cui si vogliono valutare le prestazioni di un sistema di information retrieval adoperato in modo non interattivo, relativamente ad un insieme di richieste prefissato e disponendo di una collezione bibliografica con documenti pertinenti alle richieste noti. Questa scelta è molto comune e consente di semplificare drasticamente la complessità del disegno sperimentale. Nell'articolo si esaminano i metodi di valutazione di laboratorio più diffusi e si discute la loro estensione per valutazioni su larga scala, con particolare riferimento all'esperienza TREC. Successivamente si considerano i requisiti dei metodi interattivi per la selezione delle informazioni e i problemi connessi alla loro valutazione. In particolare, si presentano i risultati relativi alla valutazione di FIRE, un particolare sistema interattivo che fornisce supporto all'utente nella riformulazione di interrogazioni booleane. L'ultima sezione dell'articolo è dedicata ai nuovi scenari di valutazione, più orientati al ruolo e agli obiettivi dell'utente e alla dinamica dell'interazione utente-sistema nella ricerca delle informazioni.

2. Metodi di valutazione tradizionali

Per valutare un sistema di information retrieval ci vogliono due ingredienti principali: i dati di prova e una misura dell'efficacia della ricerca fatta dal sistema. I dati di prova consistono di una collezione bibliografica, di un insieme di richieste da soddisfare e di un insieme di giudizi di pertinenza (quali documenti sono pertinenti a ciascuna richiesta). L'utilizzazione di questo tipo di dati presuppone poi l'adozione di due ipotesi, più implicite, altrettanto importanti: che ogni documento possa essere giudicato pertinente o non pertinente relativamente a ciascuna richiesta, e che ciascun documento pertinente sia ugualmente pertinente, nel senso che invece di ipotizzare un'utilità decrescente dei documenti pertinenti recuperati si assume che la capacità che un documento pertinente ha di concorrere al soddisfacimento di una richiesta d'utente non dipenda da quanti documenti pertinenti sono già stati recuperati o da quanti ce ne sono ancora da recuperare.

La misura delle prestazioni cerca di stabilire la similarità fra l'output del sistema di information retrieval, costituito di solito da un insieme di documenti ordinati per pertinenza, e quello dei giudizi di rilevanza noti a priori. La grandezza fondamentale, che sta alla base della maggior

parte delle misure proposte, è quella di documento recuperato pertinente; in altri termini, in conseguenza anche delle ipotesi espresse precedentemente, ogni ragionevole misura di valutazione deve essere proporzionale al numero di documenti recuperati dal sistema che sono pertinenti.

Le due variabili di riferimento che sono state definite per misurare le prestazioni di un sistema sono precisione (precision) e richiamo (recall), ed entrambe sono proporzionali al numero di documenti pertinenti recuperati; quello che cambia è il peso assegnato a ciascun documento addizionale pertinente. Più precisamente, la precisione è data dal rapporto fra il numero di documenti recuperati pertinenti e il numero di documenti recuperati. Il richiamo è il rapporto fra il numero di documenti recuperati pertinenti e il numero di documenti pertinenti. La precisione, pertanto, misura l'abilità del sistema di recuperare *solo* documenti pertinenti; il richiamo misura l'abilità del sistema di recuperare *tutti* i documenti pertinenti. Sperimentalmente si è visto che queste due variabili, comprese fra 0 e 1, crescono in modo inverso. Per bassi valori del richiamo si hanno buoni valori di precisione; se si vuole aumentare il richiamo si ha di solito una diminuzione della precisione.

Avendo definito questi concetti, quello che serve è un meccanismo basato su di essi per esprimere sinteticamente le prestazioni di un sistema. Nel prosieguo si descrivono le due tecniche principali adoperate per quantificare in termini di precisione e richiamo l'efficacia di un insieme di documenti ordinato per pertinenza, discutendone i relativi meriti e lacune.

2.1 Curve richiamo-precisione (recall-precision)

Il modo più diffuso per rappresentare e misurare le prestazioni è mediante una curva richiamo-precisione, che unisce in un grafico una serie di punti rappresentativi di coppie di valori precisione-richiamo. Per calcolare questi punti si procede nel modo seguente. Per ogni richiesta, si considera l'insieme ordinato dei documenti recuperati dal sistema e si sceglie come variabile indipendente il richiamo. Per ogni valore di richiamo compreso nell'intervallo 0,1 - 1, con passo 0,1, si calcola il valore di precisione associato. In pratica, occorre determinare per ogni punto il sottoinsieme di documenti recuperati sufficiente a garantirsi i livelli di richiamo desiderati e poi calcolare il valore di precisione relativamente ad esso. Tracciando la precisione in funzione del richiamo, quello che tipicamente si ottiene è una classica curva monotona discendente a gradino, con concavità verso l'alto o verso il basso. E' possibile anche effettuare un calcolo leggermente più sofisticato, detto metodo di interpolazione, che consente di gestire in modo uniforme le approssimazioni coinvolte nel calcolo descritto e di eliminare eventuali piccole oscillazioni che producono una curve a dente di sega invece che a gradino, a prezzo però di ritoccare verso l'alto la rappresentazione complessiva delle prestazioni del sistema (Salton and McGill, 1983). Una volta costruita la curva precisione-richiamo relativa a ciascuna richiesta, il

passo successivo è di fondere i risultati in una curva di prestazioni complessiva, facendo la media dei valori di precisione in ciascun punto relativamente all'insieme di richieste. A questo punto si possono confrontare le prestazioni di sistemi differenti sovrapponendo le relative curve richiamo-precisione e adottando metodi statistici per valutare se le differenze riscontrate siano significative o meno (Hull 1993).

Poiché la misura di prestazioni è relativa ad un insieme di richieste, un parametro molto importante da considerare è la sensibilità della misura rispetto alla variabilità delle richieste, in particolare per quanto riguarda il numero di documenti pertinenti. Nella curva richiamo-precisione, per assicurare la consistenza di scala rispetto a richieste differenti si normalizza rispetto al valore di richiamo. In altri termini, si assume che un particolare livello di richiamo venga raggiunto per qualsiasi richiesta, dopodiché il sistema migliore è quello che raggiunge questo livello con il minor numero di documenti non pertinenti. Mentre questa può essere in molti casi un'assunzione corretta, ci sono applicazioni in cui il richiamo non sembra una buona misura di equivalenza fra richieste. Se ad esempio una richiesta ha 20 documenti pertinenti e un'altra ne ha 200, allora ottenere il 50% di richiamo può essere per un utente un obiettivo perfettamente ragionevole per la prima richiesta e di scarso interesse nella seconda. In generale, quindi, la validità di una curva richiamo-precisione può essere seriamente inficiata quando livelli simili di richiamo corrispondono a numeri molto differenti di documenti pertinenti. Nel paragrafo successivo vedremo una tecnica alternativa che non risente di questa limitazione.

2.2 Richiamo e precisione a parità di documenti recuperati

Un approccio diverso alla presentazione dei risultati della valutazione basata su richiamo e precisione consiste nel calcolare, relativamente a ciascuna richiesta, il richiamo e/o la precisione per un valore prefissato di documenti recuperati, e poi nel fare la media rispetto alle richieste. Seguendo la terminologia di Hull 1993, ci riferiremo ad esso come metodo DCV (document cut-off value).

Il metodo DCV assume quindi che l'utente sia disposto ad esaminare solo un numero fissato di documenti per ogni richiesta. Dal punto di vista della valutazione dei risultati, questo equivale a dire che il metodo DCV normalizza sulla base dello sforzo equivalente invece che sulla base della prestazione equivalente, come avviene invece per la curva richiamo-precisione. Le due variabili calcolate col metodo DCV - valore medio di richiamo e valore medio di precisione - consentono di discriminare fra due filosofie di ricerca differenti da parte dell'utente. Basarsi sulla precisione significa assegnare a ciascun documento pertinente la stessa importanza, indipendentemente dalla richiesta alla quale è associato, mentre con il richiamo si assume che i documenti pertinenti relativi a richieste con un maggior numero di documenti pertinenti siano meno importanti di quelli relativi a richieste con un insieme di documenti pertinenti più piccolo.

Un parametro cruciale del metodo DCV è costituito dalla scelta del valore di soglia, anche per la possibile insorgenza di fenomeni inattesi. Si consideri che se il valore di soglia è maggiore del numero di documenti pertinenti la precisione non potrà mai essere uguale ad 1 e il suo valore, se tutti i documenti pertinenti sono stati recuperati, diminuirà al crescere del valore di soglia. Inoltre, è possibile che due metodi siano giudicati ugualmente efficaci anche se presentano forti differenze di precisione all'interno dell'insieme ordinato dei documenti recuperati determinato dal valore di soglia, ad esempio perché uno recupera subito tutti i documenti pertinenti mentre per l'altro bisogna aspettare la seconda parte dell'ordinamento. Quando il valore di soglia è minore del numero di documenti pertinenti è il richiamo che non potrà mai essere uguale ad 1. In questo caso, inoltre, può accadere che un metodo che si comporta molto bene ottenga un giudizio in termini di richiamo che in apparenza è solo leggermente più alto di quello di un metodo nettamente inferiore, solo perché per entrambi il valore del richiamo è comunque in assoluto molto basso.

In generale quindi si può dire che un valore di soglia basso è più adatto per misurare la precisione, mentre un valore di soglia alto è più adatto per misurare il richiamo. Per evitare di fare scelte drastiche e potenzialmente anomale, è stato suggerito di calcolare il metodo DCV su un insieme di valori di soglia, analogamente a quanto viene fatto per le curve richiamo-precisione che adoperano un insieme di valori di richiamo. Hull 1993, in particolare, suggerisce di adoperare l'insieme fra 1 e 20, calcolando richiamo e precisione per ogni richiesta per ciascuno dei 20 valori di soglia, e poi mediando sull'insieme di richieste.

Per completare la discussione sui due diversi modi di utilizzare le variabili richiamo e precisione, cioè curve richiamo-precisione e metodo DCV, è utile sottolineare il fatto che ciascun metodo presenta vantaggi e svantaggi. Il comportamento effettivo dell'utente è probabilmente basato su una combinazione degli aspetti considerati in ciascuno. L'utente sarà probabilmente disponibile a esaminare nuovi documenti con tolleranza ridotta (misurata dal DCV), ma questa tolleranza dipenderà anche dall'ammontare di informazione pertinente già recuperata (misurata dalla curva richiamo-precisione). In ogni caso, considerando anche l'affermazione di strumenti quali i motori di ricerca su Internet, mediare sulla base dello sforzo equivalente invece che della prestazione equivalente sembra oggi più sensato in molte applicazioni reali.

2.3 Altre misure di prestazione

In aggiunta alle misure basate su richiamo e precisione sono state proposti vari altri metodi di valutazione, come ad esempio il cosiddetto "fallout", che cerca di caratterizzare l'abilità del sistema nel respingere documenti non pertinenti, o la misura E introdotta da Swets, che combina, mediante rappresentazioni probabilistiche, richiamo e fallout. Queste tecniche,

descritte in modo approfondito in Salton and McGill 1983, non hanno però guadagnato il consenso necessario per scalzare la supremazia dei metodi basati su recall e precision, principalmente a causa della difficoltà della loro interpretazione da parte dell'utente. Fra le tecniche di valutazione alternative che sono state proposte si distingue per semplicità ed eleganza la lunghezza attesa della ricerca (expected search length) di Cooper, la quale, a differenza di altre misure, può applicarsi direttamente anche in quei casi in cui l'output del sistema di information retrieval è costituito da un insieme di documenti parzialmente ordinato invece che totalmente ordinato, come ad esempio succede nell'ordinamento basato semplicemente sul numero di collisioni fra documento e richiesta. Si tratta di una misura a valore singolo che cerca di misurare la differenza fra l'ordinamento prodotto dal sistema e l'ordinamento ideale in cui tutti gli elementi pertinenti sono recuperati prima dei non pertinenti. Più in particolare, la lunghezza attesa della ricerca viene definita come il numero medio di elementi non pertinenti che bisogna esaminare prima che si riesca a recuperare il numero desiderato di elementi. Questo calcolo nel caso di ordinamenti parziali può diventare alquanto complicato perché bisogna considerare tutti i possibili modi in cui gli elementi rilevanti di un sottoinsieme non ordinato di elementi possono distribuirsi all'interno del sottoinsieme. Questa misura di prestazione è stata recentemente rivisitata da Dunlop 1997, il quale ha sostenuto che le due variabili coinvolte - numero di documenti da esaminare, numero di documenti da recuperare - sembrano modellare in modo più rispondente alle esigenze dell'utente le caratteristiche prestazionali di un sistema di information retrieval, e ha proposto di adoperare, invece di una misura singola come Cooper, un grafico che illustri l'andamento del numero di documenti da esaminare in funzione del numero di documenti da recuperare.

3. L'esperienza TREC

I metodi di valutazione discussi nel paragrafo precedente sono stati adottati universalmente e sono rimasti sostanzialmente invariati per quasi trent'anni, sin dai primi esperimenti con la collezione Cranfield negli anni '60 (Cleverdon, 1962). Una delle ragioni principali del loro successo è che si tratta di valutazioni "oggettive" che possono essere svolte in un ambiente di laboratorio semplice, consistente in una collezione bibliografica di prova, un insieme di richieste informative e un insieme di giudizi di pertinenza. La validità di questi metodi tuttavia si basa su alcune assunzioni che sono intrinsecamente deboli, e che sono progressivamente diventate meno realistiche col mutare delle condizioni operative dei sistemi moderni di information retrieval, in particolare quelli per la selezione delle informazioni disponibili in rete. Un aspetto cruciale è rappresentato dalle limitazioni di scala e dimensione degli esperimenti di laboratorio rispetto al funzionamento dei sistemi in ambiente operativo. In altri termini, si tratta di capire se risultati sperimentali ottenuti su collezioni bibliografiche molto contenute, sia per il numero di documenti

sia per la lunghezza degli stessi, e utilizzando richieste molto precise, siano trasportabili anche in situazioni sperimentali reali. Il problema della dimensione degli esperimenti viene affrontato in modo approfondito nel ciclo di esperimenti TREC (Text Retrieval Conference), iniziato nel 1992 e che si svolge con cadenza annuale (si veda ad esempio Voorhees and Harman 1996). TREC è un'iniziativa americana sponsorizzata da ARPA e NIST, il cui principale obiettivo è quello di consentire di effettuare valutazioni su larga scala dei sistemi di information retrieval, impiegando collezioni bibliografiche eterogenee e di grandi dimensioni.

3.1 I dati degli esperimenti TREC

Quali dati vengono impiegati negli esperimenti TREC? Più in particolare, come vengono dimensionate e gestite le tre variabili indipendenti principali delle condizioni sperimentali - documenti, richieste, giudizi di pertinenza - per consentire il cambiamento di scala degli esperimenti?

Per quanto riguarda i documenti, la collezione di prova comprende milioni di elementi (Gbytes di dati) provenienti da diverse sorgenti informative - come ad esempio articoli del Wall Street Journal e del Financial Times, registrazioni delle sedute del Congresso americano, raccolte di articoli scientifici, ecc. - e di lunghezza variabile (da poche centinaia ad alcune migliaia di termini per documento), laddove l'insieme di documenti adoperato in precedenza nelle prove di valutazione consisteva in collezioni omogenee di dimensione assai più ridotta (alcune migliaia di documenti, come nelle collezioni Cranfield, CACM, CISI, Lisa, ecc.) con documenti più concisi.

La scelta delle caratteristiche delle richieste ("topics") da impiegare negli esperimenti non è stata univoca nel ciclo di esperimenti TREC. Si è cominciato con richieste molto elaborate, contenenti una media di 130 parole (nel TREC-2), strutturate in campi (dominio, titolo, descrizione, un campo "narrative" per specificare quando un documento è pertinente, una lista di concetti collegati), con l'obiettivo di separare nettamente la richiesta informativa originaria dalla interrogazione immessa dall'utente nel sistema. Successivamente, soprattutto in considerazione del fatto che le interrogazioni di utenti reali sono generalmente molto concise, le descrizioni delle richieste sono state drasticamente ridotte, fino ad arrivare, in TREC-4, ad un campo singolo contenente una descrizione di una frase lunga in media 16 parole, in modo abbastanza simile a quanto avveniva con le collezioni campione pre-TREC. Più recentemente, nel TREC-5, si è individuato un bilanciamento fra questi due estremi, reintroducendo il campo titolo e il campo "narrative" e limitando la lunghezza complessiva della richiesta (ottenendo una media di 82 parole per richiesta). Ancora per quanto riguarda le richieste, è importante sottolineare che i loro inventori sono le stesse persone che saranno poi chiamate a valutare i documenti recuperati dal sistema relativamente a ciascuna di esse. Inoltre, le 50 richieste finali utilizzate in ciascun TREC

vengono selezionate a partire da 150 richieste candidato, scartando quelle che causano il recupero di un numero troppo grande o troppo piccolo di documenti o che sono ritenute più ambigue.

Veniamo ora all'ultima decisiva variabile coinvolta nella progettazione degli esperimenti, e cioè la valutazione della pertinenza dei documenti recuperati ai fini di calcolare le prestazioni dei sistemi attraverso le curve recall/precision descritte nel paragrafo precedente. Ci sono due ordini di problemi. Il primo è che le dimensioni delle collezioni rendono impossibile la valutazione preventiva ed esaustiva della pertinenza di ciascun documento rispetto a ciascuna richiesta, come invece avviene con le collezioni pre-TREC. Per superare questa difficoltà si adotta il metodo del "pooling", che consiste nel creare un bacino di possibili documenti pertinenti mettendo assieme un campione di documenti selezionati dai vari sistemi oggetto di misura e facendo poi le valutazioni solo sui documenti del bacino. In particolare, in TREC si prendono i primi 100 documenti recuperati da ciascun metodo relativamente a ciascuna richiesta, cosicché, considerando anche che la sovrapposizione risultante dalla fusione dei diversi campioni è alquanto pronunciata, gli esperti devono valutare complessivamente, per una edizione TREC, poche migliaia di documenti.

La seconda questione concerne la qualità e l'affidabilità dei giudizi di pertinenza espressi dagli esperti. In particolare, le grandezze di interesse sono la completezza del metodo scelto, che non garantisce il recupero di tutti i documenti pertinenti ad una certa richiesta, e la coerenza dei giudizi espressi sulla pertinenza o non pertinenza dei documenti, che potrebbero cambiare quando cambiano gli esperti. Per la completezza, si è visto in TREC-5 che nel bacino formato dai secondi 100 documenti recuperati da ciascun metodo si trovava in media soltanto un documento pertinente che non era contenuto nel bacino formato dai primi 100 documenti. Inoltre, la distribuzione dei nuovi documenti pertinenti era praticamente uniforme rispetto ai singoli sistemi, il che garantiva la mancanza di polarizzazioni indesiderate, mentre era fortemente variabile con le richieste, nel senso che richieste con più documenti pertinenti iniziali avevano anche più documenti pertinenti nuovi. Questi risultati, se da un lato consigliano l'utilizzazione di richieste con un numero limitato di documenti pertinenti, testimoniano della sostanziale completezza del metodo adottato. Per quanto riguarda la valutazione della coerenza, è importante sottolineare che lo scopo non è quello di stabilire la compatibilità fra le valutazioni espresse da esperti differenti, quanto invece di assicurarsi che ciascun giudizio rifletta i meriti relativi dei sistemi di information retrieval oggetto di valutazione. Nel TREC-4 sono stati condotti degli esperimenti per misurare questo aspetto. Per ciascuna delle richieste, è stato estratto in modo casuale un sottoinsieme dei documenti presenti nel bacino (200 pertinenti e 200 non pertinenti) che sono stati poi fatti valutare di nuovo da altri due esperti. In effetti, i due turni di valutazione sono risultati in forte disaccordo, in particolare per quanto riguarda il giudizio sulla pertinenza (30% di disaccordo), mentre la variazione sul giudizio di non pertinenza era

molto più contenuta (3%). A questo punto si è valutata l'incidenza delle variazioni degli insiemi dei giudizi di pertinenza sulle prestazioni dei sistemi. Il risultato, che conferma analoghi risultati trovati per test di dimensioni molto più ridotte, è stato che c'è una sostanziale stabilità di valutazione quando il confronto riguarda varianti di uno stesso sistema, mentre si può manifestare instabilità quando ad essere confrontati sono sistemi differenti. Così ad esempio può succedere che il sistema A abbia delle prestazioni migliori del sistema B adoperando i giudizi di pertinenza X e peggiori quando adopero i giudizi di pertinenza Y, anche se bisogna considerare che differenze di prestazioni apparentemente evidenti possono non essere statisticamente significative (Tague-Sutcliffe and Blustein, 1995).

3.2 L'interattività in TREC

La scala degli esperimenti non è né l'unico né il principale problema dei metodi di valutazione tradizionali. Un secondo importante aspetto riguarda l'incapacità di considerare il contributo umano al processo di recupero delle informazioni. Nel modello che soggiace ai metodi di valutazione considerati, il sistema di information retrieval da valutare include sia il meccanismo di recupero che le eventuali attività umane associate (indicizzazione, formulazione delle interrogazioni, ricerca, ecc.). Esso viene pertanto trattato come una scatola nera, il cui input è la richiesta dell'utente e il cui output l'insieme di documenti recuperati. Si tratta evidentemente di una rappresentazione grossolana e poco realistica del processo di recupero delle informazioni. Anche se ci limitiamo a considerare sistemi di information retrieval convenzionali, come quello booleano, la ricerca avviene tipicamente in modo interattivo ed iterativo: l'utente interroga il sistema, esamina i risultati e poi formula una nuova interrogazione, fino a quando abbandona, per soddisfazione o frustrazione. Con l'affermazione degli strumenti di ricerca delle informazioni su Internet, il problema dell'interattività è oggi ancora più sentito, perché il ruolo dell'utente è cresciuto e si è diversificato. Da una parte, l'utilizzazione dei motori di ricerca richiede la capacità per l'utente di riuscire a formulare interrogazioni precise per discriminare l'enorme massa di documenti recuperati dal sistema; in secondo luogo, con l'affermarsi di sistemi di information retrieval più orientati alla navigazione - valga per tutti l'esempio del World Wide Web - è l'utente che cerca e riconosce ciò che gli serve invece di avere il sistema che trova ciò che serve all'utente in risposta alla sua richiesta.

Il problema della valutazione delle componenti interattive si è cominciato a trattare anche in TREC, seppure in modo limitato e con caratteristiche molto peculiari. Nelle conferenze TREC, infatti, l'attenzione si concentra su due modalità principali di recupero delle informazioni, "routing" e "ad hoc". Nel compito "routing" si parte da un insieme di richieste predefinite e di documenti di addestramento e l'obiettivo è selezionare nuovi dati, come accade nei sistemi SDI (selective dissemination of information); nel compito "ad hoc", invece, i dati non cambiano

mentre le richieste da soddisfare sono nuove, analogamente a quanto avviene per l'interrogazione di un sistema di information retrieval online. Accanto a queste modalità classiche di recupero delle informazioni è stata recentemente introdotta una nuova sessione, denominata "interactive track". E' chiaro che considerare l'utente come un protagonista della ricerca significa abbandonare l'approccio alla valutazione centrato sul sistema, tipico dei metodi visti fino ad ora, per includere nelle variabili da controllare fattori legati all'utente e al suo comportamento. Si tratta evidentemente di un compito difficile, specialmente nella misura in cui ci si propone di isolare e valutare gli aspetti d'interazione che contribuiscono al risultato della ricerca. Una possibilità è quella di adoperare gli strumenti impiegati nelle scienze sociali, come questionari, interviste ecc. In TREC si è cercato, più ambiziosamente, di simulare una ricerca interattiva realistica delle informazioni preservando il contesto della valutazione di laboratorio effettuata sui metodi non interattivi, anche allo scopo di consentire una compatibilità di valutazione fra i due approcci. In pratica, la "interactive track" in TREC prevede che la formulazione originaria della richiesta possa essere pre-elaborata interattivamente attraverso l'uso di vari strumenti (tesauri, sessioni col sistema, ecc.) con l'obiettivo di produrre una interrogazione "ottimale" da sottoporre al sistema di information retrieval, in alternativa a quella costruita automaticamente. I risultati fino ad ora non sono stati incoraggianti, in quanto si è visto che di solito le interrogazioni costruite interattivamente non producono risultati migliori di quelle costruite automaticamente. Questi risultati, tuttavia, sembrano più legati al disegno sperimentale scelto, che è poco realistico e penalizza il contributo dell'interazione alla ricerca, piuttosto che ad una presunta incapacità da parte dell'intervento dell'utente di migliorare le prestazioni di un sistema di information retrieval.

Nella prossima sezione esamineremo una metodologia di valutazione per sistemi interattivi che diverge da quella utilizzata un TREC, e cerca di combinare metodi quantitativi e qualitativi. La discussione verrà svolta con riferimento ad un particolare sistema interattivo per la selezione delle informazioni (FIRE), mediante il quale è stata condotta la sperimentazione.

4. FIRE: information retrieval con supporto alla riformulazione delle interrogazioni

In questo paragrafo affrontiamo il problema della valutazione dei Sistemi Interattivi per il Reperimento delle Informazioni (a cui in seguito ci riferiremo per brevità con SIRI), ossia di sistemi che dialogano direttamente con l'utente ed in cui l'interattività gioca un ruolo fondamentale.

4.1 La valutazione di sistemi interattivi con metodi quantitativi e qualitativi

La valutazione di SIRI (Tague-Sutcliffe, 1992; Robertson e Beaulieu, 1992) può fare luce su parecchi punti ancora non compresi. I risultati degli esperimenti che studiano il comportamento di utenti che usano il sistema possono portare sia a una migliore comprensione del ruolo dei SIRI sia a identificare importanti tematiche nell'interazione uomo-macchina per l'information retrieval. E valutazioni diagnostiche sono essenziali per ottenere linee guida per il progetto e per determinare le scelte di progetto più promettenti.

La valutazione di SIRI è però difficile. Essa eredita le difficoltà della valutazione dei sistemi di information retrieval: trattare il concetto sfuggevole di pertinenza (Mizzaro, 1997a; Mizzaro, 1997b; Mizzaro, 1998), avere a che fare con i problemi classici dell'information retrieval, quali il problema del vocabolario, lo stato anomalo di conoscenza, l'effetto etichetta, l'esplicitazione di bisogni informativi confusi, e così via (Ingwersen, 1992). Ma la valutazione di SIRI porta in primo piano l'utente, i suoi bisogni informativi, la sua interazione con il sistema. La conseguenza è la necessità di considerare misure non solo quantitative (come precisione e richiamo) ma anche aspetti qualitativi che sono più complessi da definire, acquisire e codificare, come la percezione che l'utente ha dei risultati ottenuti e del processo di ricerca di informazioni. Inoltre, le valutazioni diagnostiche, il cui scopo è di determinare la bontà di scelte alternative (ad esempio, funzionalità e stili di interazione differenti), hanno bisogno di caratterizzare e misurare aspetti cognitivi che potrebbe essere difficile focalizzare e isolare da disturbi ambientali non desiderati.

La valutazione della soddisfazione dell'utente nel settore dell'information retrieval è stata indagata in alcuni lavori precedenti, ma spesso le tecniche di acquisizione dei dati adottate non sembrano le più appropriate: domande dirette (come "è stato facile usare il sistema?") portano di solito a risposte non attendibili; le interviste hanno bisogno di metodi complessi e di una grande mole di lavoro per codificare il ricco insieme di dati ricavato dalle risposte dei soggetti; e le tecniche di think-aloud non garantiscono che la verbalizzazione rispecchi il comportamento, enche non lo influenzi. I differenziali semantici, le scale di valutazione per categorie, i metodi di stima della magnitudine, quando usati in modo appropriato, sembrano più accurati ed efficaci.

La metodologia di valutazione descritta qui (e in Brajnik, Mizzaro e Tasso, 1996 e Mizzaro, 1997a) si differenzia dalle proposte classiche basate su precisione e richiamo, e cerca di risolvere questi problemi. In questo paragrafo vengono discussi e valutati i differenti tipi di aiuto che un SIRI (o un intermediario umano) può fornire agli utenti e viene descritto un esperimento, progettato sulla base di una metodologia originale, che determina importanza ed efficacia di vari tipi di aiuto e la migliore modalità per fornirli all'utente. L'esperimento è condotto utilizzando FIRE (Flexible Information Retrieval Environment), un SIRI intelligente (ossia un SIRI basato su tecniche di intelligenza artificiale), ma la metodologia seguita e alcuni

dei risultati ottenuti sono validi anche per SIRI generici. Grazie all'adozione di una classificazione degli aiuti e ad un insieme di variabili osservate che copre non solo l'efficacia della ricerca ma anche la soddisfazione dell'utente,e possibile analizzare il comportamento dell'utente, valutare FIRE in modo più approfondito di quanto si potrebbe fare con una metodologia di valutazione classica, e identificare linee guida generali per l'implementazione di SIRI.

Nel paragrafo 4.2 viene brevemente descritto il SIRI FIRE usato nell'esperimento; nel paragrafo 4.3 viene descritta una tassonomia di tipi di aiuto che un SIRI può fornire all'utente e, su questa base, vengono delineati gli obiettivi dell'esperimento di valutazione effettuato. L'esperimento viene descritto nei paragrafi successivi: nel paragrafo 4.4 sono illustrati gli obiettivi dell'esperimento; nel paragrafo 4.5 il disegno sperimentale e nel paragrafo 4.6 i risultati ottenuti. Infine, nel paragrafo 4.7 viene presentata un'analisi dei risultati dell'esperimento, da cui si ricavano alcune linee guida generali per il progetto di SIRI e alcune osservazioni sulla valutazione di sistemi di questo tipo.

4.2 Il prototipo FIRE

Viene qui presentata una breve descrizione del funzionamento di FIRE (per ulteriori dettagli si vedano Mizzaro, 1997a; Brajnik, Mizzaro e Tasso, 1995). FIRE è un prototipo di SIRI che utilizza alcune basi di conoscenze per migliorare l'interazione con l'utente; è implementato su una rete di workstation SUN in vari linguaggi di programmazione ed accede ad una collezione ottenuta estraendo 20.000 documenti dalla banda dati INSPEC.

La figura 1 presenta due finestre dell'interfaccia utente di FIRE. Quella grigio scuro è la FIRE Main Window, in cui l'utente comunica al sistema la propria richiesta. E' quindi compito dell'utente individuare i concetti che esprimono nel modo migliore (almeno in prima approssimazione) il proprio bisogno informativo, specificare i termini con cui i concetti possono essere espressi e raggruppare i termini in faccette organizzate in 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)". Nella parte alta della FIRE Main Window sono a disposizione dell'utente cinque bottoni:

- Direct Search, che reperisce i (titoli dei) documenti che soddisfano l'interrogazione;
- Reformulate, che avvia la fase di riformulazione;
- Folders, che consente all'utente di catalogare i documenti reperiti, suddividendoli in pertinenti, non pertinenti, utili, non utili, ecc.;
- Restart, che consente di inizializzare il sistema;

- Quit, per terminare la sessione e uscire da FIRE.

A destra sono visualizzati i titoli dei documenti reperiti (che l'utente può consultare selezionandone il titolo). La finestra in grigio chiaro mostra invece il tipo di supporto che FIRE è in grado di fornire all'utente durante la riformulazione: in questa finestra FIRE visualizza una lista di termini che possono essere usati per raffinare l'interrogazione. L'utente può accettare alcuni termini, selezionandoli col mouse (in grassetto in figura) ed aggiungendoli all'interrogazione.

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 bisogno informativo, effettua una ricerca, esamina i titoli dei documenti reperiti, ne visiona il contenuto e li classifica. 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 i 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 potenzialit\`a 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.

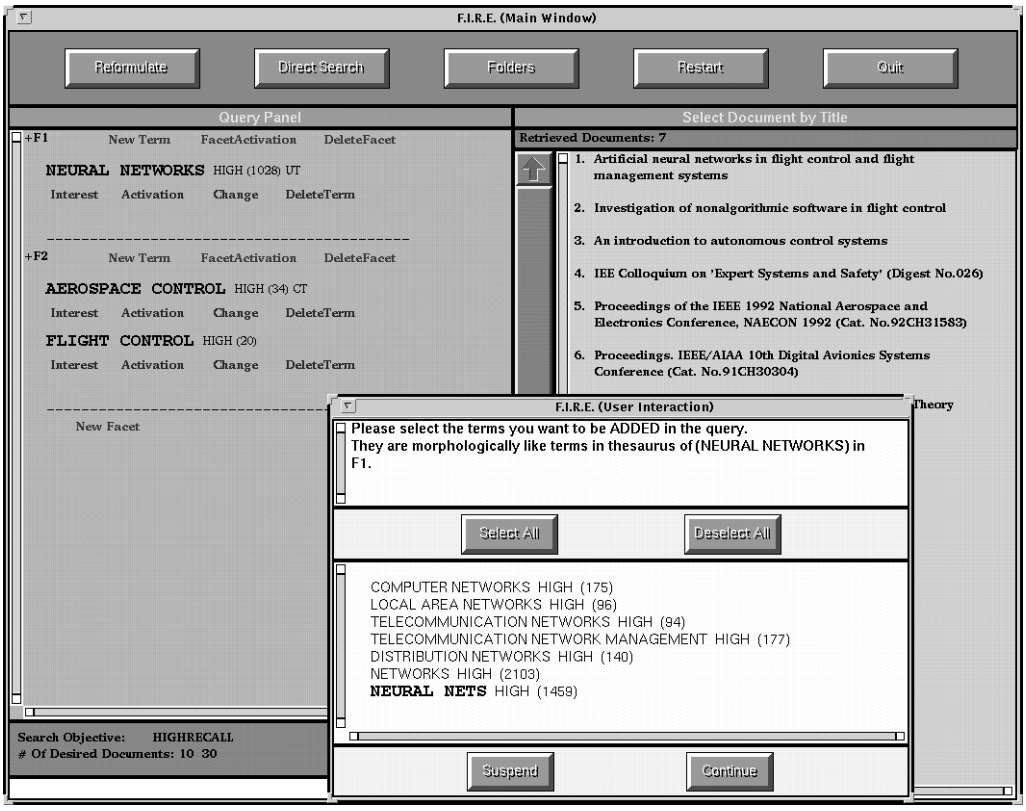

Figura 1: L'interfaccia utente di FIRE: IRESFACE.

4.3 Valutazione di FIRE: tipi di aiuto

Gli aiuti che un SIRI può fornire all'utente durante l'interazione possono essere classificati secondo la seguente tassonomia (denominata spazio degli aiuti). Per quanto concerne la natura dell'aiuto, si può distinguere fra:

- aiuto tecnico, che permette all'utente di interagire con il SIRI in modo efficace e soddisfacente. Ad esempio, evidenziando il ruolo di una certa opzione di controllo, o di un certo comando.
- aiuto concettuale, che supporta l'utente nella risoluzione di problemi correlati al processo di ricerca di informazioni. L'aiuto concettuale si può scomporre in:
    - aiuto terminologico, per arricchire il vocabolario che l'utente adotta quando esplicita il bisogno informativo. Ad esempio, suggerendo liste di termini sinonimi;
    - aiuto strategico, per migliorare la capacità dell'utente di condurre una ricerca efficace. Ad esempio, per uscire da situazioni avverse, quali il reperimento di zero documenti.

L'aiuto è fornito tramite una sorta di dialogo, ossia una specifica interazione fra utente e SIRI, che può avvenire in differenti modalità:

- contestuale o generica, a seconda del contesto da cui dipende l'aiuto. Se dipende dallo specifico comportamento dell'utente e dalla situazione in cui egli si trova (ad esempio, suggerire un termine sinonimo di un termine appena aggiunto all'interrogazione dall'utente), allora è contestuale. Se si riferisce ad aspetti o suggerimenti generali (ad esempio, suggerire che la riduzione di concetti in ande un modo per allargare l'insieme di documenti reperiti) è generico.
- su richiesta o autonomo, a seconda di chi (utente o SIRI) avvia il dialogo di aiuto. E' su richiesta se l'aiuto segue un'esplicita richiesta dell'utente (ad esempio, una richiesta di termini sinonimi). E' autonomo se è fornito automaticamente dal sistema al verificarsi di una certa situazione (ad esempio, azioni dell'utente incoerenti con gli obiettivi, come ridurre il numero di sinonimi per un'interrogazione che ha reperito zero documenti).
- controllato dal sistema o controllato dall'utente, a seconda di chi controlla l'evoluzione del dialogo. Ad esempio, un dialogo di aiuto controllato dall'utente potrebbe essere un'interazione in cui il SIRI offre all'utente la possibilità di navigare in una rete di termini correlati, e l'utente decide se navigare o meno, e in che direzioni. Un dialogo controllato dal sistema potrebbe presentare all'utente una lista di termini, chiedere all'utente di confermarne o selezionarne alcuni e continuare.

FIRE quindi fornisce aiuto terminologico in modalità contestuale, su richiesta e controllato dal sistema (ma l'utente può riprendere il controllo in ogni momento cliccando sul bottone Suspend).

4.4 Valutazione di FIRE: obiettivi

Due sono gli obiettivi dell'esperimento di valutazione di FIRE: (i) determinare il valore aggiunto della riformulazione di FIRE, e (ii) valutare l'importanza dei differenti tipi di aiuto e indagare sulle migliori modalità per fornirli. Il primo obiettivo è motivato dalla mancanza di un metodo efficace per determinare i miglioramenti nella fase di sviluppo di SIRI. Il secondo nasce dalla necessità di valutare differenti scelte di progetto riguardanti non solo l'aiuto terminologico, ma anche, più in generale, l'aiuto concettuale.

4.5 Valutazione di FIRE: disegno sperimentale

L'esperimento, progettato ed eseguito in collaborazione con un gruppo di psicologi, può essere classificato come (Robertson e Beaulieu, 1992):

- Un esperimento di laboratorio (in contrapposizione a un esperimento operativo), in cui quindi sono usati bisogni informativi indotti dagli sperimentatori (e non bisogni informativi

reali dei soggetti) e la collezione usata (20.000 documenti concernenti l'intelligenza artificiale) è più piccola delle collezioni reali.

- Un esperimento diagnostico (in contrapposizione a un esperimento black-box), teso quindi a ottenere informazioni utili per il progetto del SIRI, e non solo a valutare le prestazioni del sistema.
- Un esperimento in cui sono usate misure sia quantitative sia qualitative, allo scopo di considerare le prestazioni della ricerca, la soddisfazione dell'utente e il suo comportamento (che cosa gli utenti chiedono, quanto spesso, quali strategie usano, e così via).

4.5.1 Soggetti e gruppi sperimentali

I soggetti coinvolti nell'esperimento sono 45 studenti del terzo e quarto anno del corso di laurea in informatica, con limitata conoscenza delle problematiche dell'information retrieval e del dominio della collezione. Nessuno di essi ha in precedenza usato o visto usare FIRE. Dopo una fase di addestramento, ogni soggetto effettua due ricerche in due sessioni: la prima usando FIRE con la riformulazione disabilitata e la seconda usando il sistema completo. I soggetti vengono divisi in tre gruppi sulla base del tipo di aiuto ricevuto nella prima sessione, come illustrato nella tabella 1, nella quale:

- TH sta per un ausilio fornito per mezzo di un thesaurus cartaceo (ottenuto dalla DSKB di FIRE) che il soggetto può consultare manualmente;
- ET sta per un esperto umano che fornisce solo aiuto terminologico in modalità contestuale su richiesta;
- ES sta per un esperto umano che fornisce aiuto strategico in modalità generica su richiesta;
- FIRE- sta per FIRE senza riformulazione;
- FIRE+ sta per il sistema completo.
-

	Gruppo TH	Gruppo ET	Gruppo ES
Sessione I	TH	ET	ES
	+	+	+
	FIRE-	FIRE-	FIRE-
Sessione II	FIRE+	FIRE+	FIRE+

Tabella 1: Organizzazione dei soggetti in gruppi sperimentali.

Nelle due sessioni, ai soggetti vengono dati (in ordine casuale) due bisogni informativi (etichettati A e B, si vedano le figure 2 e 3). Ogni bisogno informativo è costituito da un argomento e da un compito da (fingere di) svolgere con i documenti reperiti. I due bisogni informativi sono progettati in modo da non reperire troppi documenti: per ognuno di essi, nella collezione sono presenti meno di 20 documenti.

Immagina di lavorare per un'azienda che si occupa di reattori a fissione. Il laboratorio di ricerca ha dei problemi nello sviluppo di un progetto riguardante l'argomento:

Fault diagnosis of cooling systems in fission reactors

Il tuo superiore ti chiede di documentarti sulle applicazioni già esistenti in ambiti reali, per vedere come sono stati affrontati quei problemi. Sulla base della tua esperienza precedente, immagini che nella collezione di FIRE ci siano non più di 20 documenti su questo argomento. Decidi quindi di effettuare una ricerca bibliografica per trovare sei articoli che riguardino in dettaglio sistemi reali o prototipi collaudati in condizioni reali.

Figura 2: Il bisogno informativo A.

Immagina di essere un dottorando presso il Dipartimento di Informatica. Il tuo professore ti chiede con un preavviso di soli due giorni di preparare una lezione per gli studenti del primo anno. Il titolo della lezione è:

The role of human factors in interfaces for monitoring systems

Non hai alcuna preparazione su questo argomento e decidi quindi di usare FIRE per trovare riferimenti bibliografici. Il tuo professore ti suggerisce che dovrebbero esserci una quindicina di documenti su questo argomento, su cui tu non hai alcuna preparazione. Siccome hai poco tempo, leggerai solo i documenti più generali e introduttivi, scartando atti di conferenze e altri documenti troppo specifici.

Figura 3: Il bisogno informativo B.

4.5.2 Variabili indipendenti e dipendenti

Due variabili indipendenti sono dunque usate per creare i vari scenari: il tipo di sistema (con o senza riformulazione) e il tipo di aiuto (TH, ET, ES). Le variabili dipendenti sono scelte per ottenere tre categorie di informazioni: sulla soddisfazione dell'utente, sulle prestazioni e sul comportamento dell'utente. La soddisfazione dell'utente è misurata per mezzo di questionari in cui i soggetti valutano: la qualità dell'informazione reperita; l'efficacia del sistema; la complessità del bisogno informativo; l'aiuto ricevuto; l'influenza della riformulazione nel comprendere e applicare conoscenze strategiche; e la relativa forza delle fonti di aiuto (thesaurus, esperto umano, riformulazione).

Gli indici delle prestazioni includono l'efficacia della ricerca, valutata in termini di precisione e richiamo sulla base di due insiemi di documenti. Il primo insieme è costruito a priori dagli sperimentatori, che hanno selezionato i documenti considerati pertinenti; l'altro è determinato a posteriori, selezionando i documenti pertinenti sulla base dei giudizi espressi dai soggetti durante le loro ricerche.

Le informazioni sul comportamento dell'utente, ottenute tramite il logging automatico e la videoregistrazione delle sessioni, comprendono: il numero totale di ricerche effettuate dal sistema; il numero totale di ricerche effettuate dal soggetto; il numero totale di riformulazioni; il numero di faccette e termini per ogni interrogazione; gli obiettivi della ricerca definiti dal soggetto; il numero di termini suggeriti dal sistema; il numero di termini accettati dal soggetto; la durata della sessione; il tempo impiegato dal soggetto per costruire la prima espressione del bisogno informativo; il numero totale di documenti trovati dal sistema; domande e azioni del soggetto.

4.5.3 Materiale e procedure

Per l'esperimento sono stati progettati tre questionari: Q1 per raccogliere, tramite domande chiuse e scale di Likert (un particolare tipo di scala di valutazione per categorie), genere, età, esperienza sull'information retrieval con e senza calcolatori, conoscenza della lingua inglese e attitudine nei confronti dell'uso dei calcolatori nell'information retrieval; Q2 per ottenere (con differenziali semantici e scale di Likert) la soddisfazione del soggetto rispetto al sistema, ai risultati della ricerca e al tipo di aiuto disponibile; Q3 per misurare (con line-length) la valutazione del soggetto sull'attività di ricerca globale e sul tipo di aiuto preferito.
L'esperimento è diviso in tre fasi:

1. Fase di addestramento: ogni soggetto viene introdotto all'esperimento con una presentazione informale, compila il questionario Q1, legge le istruzioni riguardanti il sistema senza riformulazione (FIRE-), osserva lo sperimentatore che effettua una ricerca dimostrativa ed

effettua una ricerca di prova per proprio conto (entrambi questi bisogni informativi, per le due ricerche di addestramento, sono gli stessi per tutti i soggetti).

2. Fase sperimentale: si comunica al soggetto quale tipo di aiuto ha a disposizione e gli viene sottoposto il primo bisogno informativo. Egli ha 30 minuti per effettuare la ricerca, alla fine della quale compila il questionario Q2. Successivamente, gli vengono consegnati il bisogno informativo per la seconda sessione e le istruzioni per l'uso del sistema con la riformulazione (FIRE+). Dopo un'altra mezz'ora, il soggetto compila un nuovo questionario Q2.
3. Fase post-sperimentale: il soggetto compila il questionario Q3.

4.6 Valutazione di FIRE: risultati

Questo paragrafo è diviso in tre parti in cui sono riportati i risultati più importanti emersi dall'analisi di questionari, log e materiale videoregistrato.

4.6.1 Soddisfazione dell'utente

Le analisi statistiche dei dati derivati da Q1 mostrano che il campione dei soggetti è omogeneo, così come lo sono i tre gruppi sperimentali. La soddisfazione dei soggetti è determinata tramite Q2 sulla base di quattro variabili: soddisfazione per l'informazione ricevuta, per il sistema usato, per il processo di ricerca e per l'aiuto ricevuto. Molto spesso queste quattro variabili risultano strettamente correlate; i questi casi si parlerà di soddisfazione. Durante la prima sessione, i bisogni informativi non influenzano la soddisfazione: non c'e nessuna correlazione fra prestazioni e soddisfazione. Invece, durante la seconda sessione c'è una forte dipendenza della soddisfazione dal bisogno informativo: i soggetti che lavorano sul bisogno informativo A risultano più soddisfatti dei soggetti che lavorano su B.

Se si osserva l'andamento della soddisfazione dell'utente fra la prima e la seconda sessione, si ha che per il bisogno informativo A essa cresce (FIRE+ porta a una maggiore soddisfazione per l'informazione ricevuta e per il processo di ricerca), mentre per il bisogno informativo B la soddisfazione dell'utente descresce. Una domanda del questionario Q3 chiede quale sia il sistema preferito per eventuali ricerche future. I soggetti che hanno usato FIRE+ sul bisogno informativo A (più facile), lo preferiscono. Indipendentemente dal bisogno informativo, i soggetti più autonomi (quelli che hanno posto meno domande nella prima sessione) e i soggetti del gruppo ET affermano che FIRE+ ha dato un aiuto determinante per soddisfare il loro bisogno informativo.

4.6.2 Valore aggiunto della riformulazione automatica

Le misure dell'efficacia della ricerca effettuate sulla base dei due insiemi di documenti (a seconda di giudizi a priori o a posteriori) hanno un'alta correlazione, e portano a conclusioni analoghe. I dati riportati qui si riferiscono ai giudizi a priori.

Le tabelle 2 e 3 contengono media (μ), deviazione standard (σ), mediana (M) e intervallo interquartile (IQR) di precisione e richiamo per i due bisogni informativi e per i vari gruppi nelle due sessioni; le figure 4 e 5 illustrano in forma grafica le medie di precisione e richiamo per i vari gruppi e sessioni.

Precisione		SessioneI				SessioneII			
		μ	σ	M	IQR	μ	σ	M	IQR
A	TH	37.3	26	43	38.5	29.8	23	30	32.8
	ET	29.5	22	31	29.7	45.1	29	40	22
	ES	36.8	21	46.5	31.5	35.9	20	33	20.5
	Tutti	36.4		40	34.5	36.6		35	31.5
B	TH	23.1	30	17	4.2	33.9	32	33	35
	ET	18	17	10	17	26.9	31	17	14.8
	ES	11	10	17	20	15	33	05	
	Tutti	17.6		17	20	24.9		17	35

Tabella 2: Distribuzione della precisione.

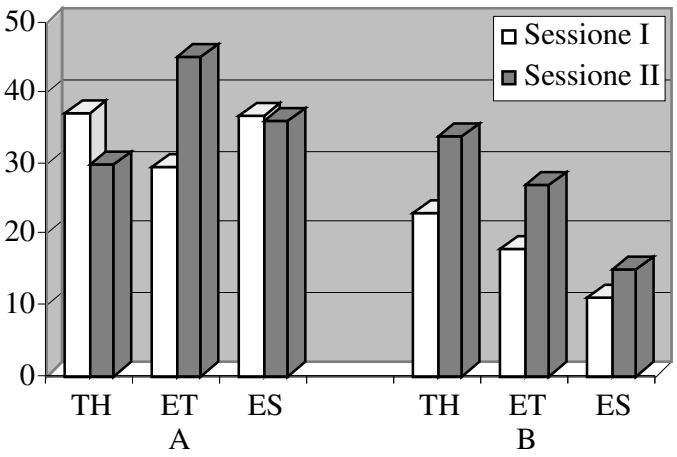

Figura 4: Precisione media.

Richiamo		SessioneI				SessioneII			
		μ	σ	M	IQR	μ	σ	M	IQR
A	TH	31	23	50	41.5	33.5	26	33.5	41.5
	ET	29.4	22	33.5	25.2	42.9	25	34	25.5
	ES	39.9	22	42	24.5	35.9	14	34	25
	Tutti	33.5	34	33	37.2	3433			
B	TH	12.8	7	17	4.2	19.3	17	17	17
	ET	14.4	11	17	17	14.9	10	17	4.2
	ES	9.7	8	17	17	6.4	12	0	4.2
	Tutti	12.3	17	17	13.3	17	17		

Tabella 3: Distribuzione del richiamo.

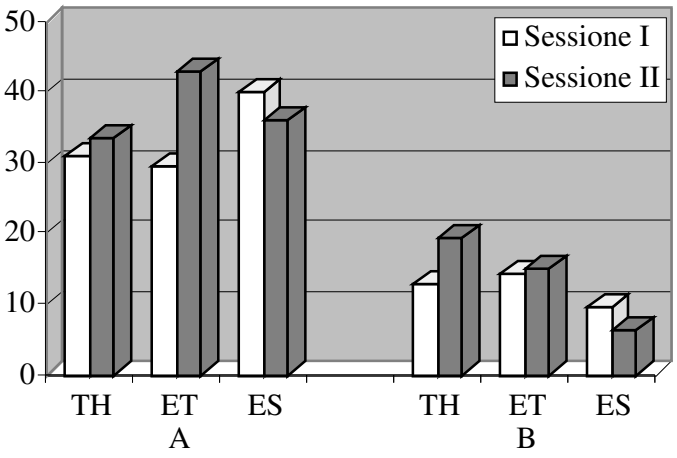

Figura 5: Richiamo medio.

Analizzando questi dati emergono differenti livelli di difficoltà per i due bisogni informativi: il bisogno informativo B ha sistematicamente causato livelli di prestazioni inferiori rispetto ad A. Una possibile spiegazione di questo fenomeno è che B sia più difficile da concettualizzare (identificare le faccette appropriate). Questa ipotesi è supportata dal fatto che mentre in A i

soggetti di solito identificano tre faccette, in B ne identificano solo due (mentre i migliori risultati si hanno con tre). Un'ipotesi alternativa è che per il bisogno informativo B sia più difficile per l'utente giudicare la pertinenza dei documenti reperiti (perché l'argomento trattato è più ostico). Sorprendentemente, non c'e nessuna correlazione significativa fra la complessità dei bisogni informativi percepita dai soggetti (ottenuta con tre domande di un differenziale semantico nel questionario Q2) e gli indici di prestazione. L'unico dato in questa direzione è che i soggetti percepiscono una maggiore difficoltà a scoprire buone strategie nel bisogno informativo B. Inoltre, nella prima sessione non c'è differenza fra i gruppi per quanto concerne la difficoltà nel trovare parole chiave, nell'includerle nella faccetta appropriata e nell'identificare buone strategie di ricerca. Infine, la frequenza delle richieste di aiuto non è influenzata dal bisogno informativo.

Inoltre si nota come la riformulazione porti a un miglioramento (statisticamente non significativo) delle prestazioni, con effetti differenti sui vari gruppi: in ET le prestazioni sembrano migliorare, in ES peggiorare. Un'interpretazione possibile è un effetto di apprendimento: durante la prima sessione i soggetti del gruppo ET apprendono come usare in modo efficace un aiuto terminologico, e ciò viene sfruttato nella seconda sessione. D'altro canto, i membri del gruppo ES non sono stati esposti all'aiuto terminologico e trovano quindi maggiori difficoltà.

Nella prima sessione non c'e correlazione fra il numero medio di termini nell'interrogazione e le prestazioni. Questa correlazione diventa positiva nella seconda sessione per il bisogno informativo A. Ciò suggerisce che quando i soggetti concettualizzano in modo appropriato il bisogno informativo (caso del bisogno informativo A, come detto nel paragrafo 4.6.1), la riformulazione ha un effetto sinergico. In casi più complessi (come in B), in cui i soggetti non riescono a ottenere una buona concettualizzazione, la riformulazione non li aiuta.

4.6.3 Importanza di tipi e modalità degli aiuti

L'aiuto tecnico è talvolta richiesto (in modalità contestuale) o talvolta non richiesto, ma comunque necessario:

1. Il 40% del numero totale di richieste dei soggetti (115 in totale) è relativo all'aiuto tecnico; tutte queste richieste sono contestuali.
2. In molte situazioni i soggetti non utilizzano in modo appropriato i comandi di FIRE. Spesso queste situazioni possono essere individuate automaticamente, e un aiuto appropriato può essere fornito in modo automatico all'utente: ad esempio, spesso i soggetti sottopongono per due volte consecutive la stessa interrogazione al sistema, senza modificare nulla.
3. I soggetti tendono a non utilizzare tutti i comandi disponibili. Ad esempio, essi raramente modificano gli obiettivi della ricerca, usando il valore di default (alto richiamo).

4. I soggetti manifestano problemi con l'uso degli operatori booleani, anche se questi vengono presentati da FIRE in modo grafico e con alcune restrizioni. Questo avviene per soggetti con prestazioni di qualsiasi livello e, come detto in precedenza, questi problemi non vengono percepiti.

L'aiuto terminologico è richiesto esplicitamente e in forma contestuale. Infatti, le richieste di aiuto sono per la maggior parte (58%) richieste di termini specifici da includere nell'interrogazione.

L'aiuto strategico non è richiesto, ma è necessario:

1. L'aiuto strategico è raramente richiesto. Solo 3 richieste (su 115) sono di aiuto strategico; nella prima sessione, il numero di richieste terminologiche nei gruppi ET ed ES è significativamente superiore a quello delle richieste strategiche. I soggetti tendono a non percepire i problemi strategici che incontrano durante il processo di ricerca di informazioni.
2. I soggetti si trovano talvolta in situazioni avverse di tre tipi: "niente" (interrogazioni con nessun documento reperito), "troppo" (interrogazioni con troppi documenti reperiti) e "interrogazioni anomale" (interrogazioni con, ad esempio, faccette duplicate). Il punto non è tanto che i soggetti si imbattano in questi problemi, ma che (talvolta) essi non siano in grado di uscirne: dai log si vede infatti come queste situazioni si ripetano su interrogazioni consecutive.
3. Per i soggetti con risultati peggiori è particolarmente difficile espandere l'insieme dei documenti reperiti. La tabella 4 mostra come le situazioni "niente" siano molto più comuni per i soggetti peggiori (definiti qui come gli ultimi cinque soggetti ottenuti ordinando in modo decrescente tutti i soggetti sulla base delle prestazioni) che per i soggetti migliori (i primi cinque). Situazioni di tipo "niente" sono anche molto più comuni, per i soggetti peggiori, delle situazioni di tipo "troppo".
4. I soggetti peggiori (definiti come sopra) sono incapaci di diagnosticare il loro comportamento e rivelano scarsa flessibilità nel cambiarlo. Essi adottano un piccolo insieme di azioni per modificare l'interrogazione (ad esempio, troncare un termine) e continuano ad applicarle anche quando esse sono non efficaci.
5. I soggetti migliori e peggiori effettuano un numero simile di azioni incoerenti, coerente e ambigue. Un'azione incoerente è qui definita come una modifica dell'interrogazione che porta a un incremento (decremento) del numero di documenti reperiti quando l'interrogazione precedente aveva già reperito troppi (troppo pochi) documenti; un'azione coerente ha la naturale definizione speculare e un'azione ambigua è non facilmente classificabile. La tabella 5 riporta il numero di azioni coerenti (+), incoerenti (-) e ambigue (?) per entrambi i tipi di soggetti. Si vede come le percentuali di azioni coerenti, incoerenti e

ambigue siano simili per i due tipi di soggetti; la differenza fra soggetti migliori e peggiori è nel numero di azioni effettuate (che per i soggetti migliori è circa la metà).

	Niente	Troppo
Migliori	7	5
Peggiori	33	9

Tabella 4: Soggetti migliori e peggiori: situazioni di tipo "niente" e "troppo".

	Valori				Percentuali		
	+	-	?	Tot.	+	-	?
Migliori	19	6	6	31	61	19	19
Peggiori	35	14	18	67	52	21	27
Tot.	54	20	24	98	55	20	25

Tabella 5: Soggetti migliori e peggiori: tipi di modifiche dell'interrogazione.

L'interazione controllata dall'utente è preferita e richiesta:

1. I soggetti preferiscono mantenere il controllo durante la riformulazione. Essi sospendono spesso la riformulazione (45% delle volte) allo scopo di riprendere il controllo e seguire una strada differente. Inoltre, c'e una correlazione positiva fra la soddisfazione dell'utente per l'aiuto ricevuto e il numero di operazioni di sospensione.
2. Il 78% dei soggetti alterna riformulazioni, ricerche dirette e classificazioni di documenti. I soggetti rimanenti esaminano e classificano i documenti reperiti solo alla fine delle sessioni. Ciò suggerisce che gli utenti non vogliano "un sistema che accetta una richiesta in linguaggio naturale, va a cercare nel negozio di informazioni e restituisce all'utente l'insieme ideale dei documenti migliori" (Bates, 1990, pag. 575).

4.7 Conclusioni della valutazione con FIRE

L'esperimento di valutazione di FIRE descritto in questo paragrafo fornisce informazioni utili per comprendere alcuni principi alla base del progetto di buoni SIRI. Lo spazio degli aiuti, la divisione in gruppi sperimentali, il ricco insieme di variabili misurate e lo specifico modello di interazione implementato dal SIRI FIRE costituiscono un modello esplicito del processo di

ricerca di informazioni effettuato usando SIRI. Questo modello permette non solo la valutazione black-box dell'efficacia del sistema, ma anche la valutazione diagnostica della sensibilita degli indici di prestazione a differenti modalità per fornire differenti tipi di aiuto.

I risultati ottenuti mostrano come tutti e tre i tipi di aiuto siano importanti e necessari in SIRI e come debbano essere adottate modalità differenti per fornirli all'utente. Più in dettaglio, si possono derivare le seguenti conclusioni, valide in generale per i SIRI:

- L'aiuto terminologico è importante; esso è esplicitamente richiesto in modalità contestuale. La riformulazione di FIRE, sebbene migliori la soddisfazione dell'utente nel caso di semplici problemi, non migliora le prestazioni in modo significativo.
- L'aiuto strategico è anch'esso importante. Esso dovrebbe essere fornito dal SIRI in modalità autonoma e sotto controllo dell'utente. Più in particolare, dovrebbe supportare l'utente nella concettualizzazione e nell'espansione dell'interrogazione e diagnosticare comportamenti non efficaci o incoerenti.
- L'aiuto tecnico è necessario per migliorare l'usabilità del SIRI e migliorare l'efficacia della ricerca. Dovrebbe essere fornito sia su richiesta sia autonomamente, in modalità contestuale.
- L'interazione controllata dall'utente è preferita. Il SIRI dovrebbe coadiuvare gli utenti nell'esplorazione autonoma dello spazio delle informazioni (termini, documenti, relazioni fra termini, contenuto dei documenti, e così via), nell'alternare attività differenti, e nello sviluppare strategie di ricerca adeguate.

Il modello di interazione e l'esperimento presentati qui trascurano comunque alcuni punti importanti. Ad esempio, non sono stati investigati i livelli di astrazione e di coinvolgimento della Bates (Bates, 1990), che potrebbero fornire informazioni aggiuntive molto utili per estendere le linee guida per il progetto presentate poc'anzi. Inoltre, non sono stati evidenziati aspetti relativi all'informazione presentata all'utente e al modo di presentarla. Gli spazi di informazioni potrebbero essere rappresentati da ricche reti di relazioni fra informazioni, e l'esplorazione interattiva di questi spazi potrebbe essere basata su un'attività di filtraggio eseguita dal SIRI per ridurre ciò che viene presentato all'utente: gli utenti manterrebbero il controllo dell'esplorazione e il SIRI limiterebbe la quantità di informazione presentata loro per evitare tipici effetti di sovraccarico. Probabilmente questo sarebbe un metodo per fornire aiuto terminologico efficace e soddisfacente per l'utente.

Comunque, esperimenti di questo tipo possono fornire dati utili per meglio comprendere aspetti generali e problemi specifici di SIRI. Ma questi esperimenti possono anche fornire indicazioni interessanti sull'attività di valutazione di SIRI. Sulla base dell'esperimento qui descritto, si può infatti dedurre che:

- Il comportamento soggettivo e la sua registrazione sono aspetti cruciali della valutazione. L'uso di differenziali semantici, scale di Likert e stima della magnitudine a lunghezza di segmento si è dimostrato un metodo efficace per acquisizione, validazione e analisi di questi importanti dati.
- Esprimenti di questo tipo sono molto complessi e richiedono parecchie risorse e parecchio tempo: il sistema valutato non può essere un semplice prototipo, ma deve essere ingegnerizzato per ottenere affidabilità, robustezza ed efficienza; l'esecuzione dell'esperimento di per sé ha richiesto circa tre mesi uomo di lavoro; e l'enorme massa di dati derivati (che vengono tutt'ora analizzati) richiede elaborazioni statistiche sofisticate. E' quindi necessario un gruppo di lavoro interdisciplinare: informatici, psicologi (per progettare i questionari ed effettuare l'esperimento senza influenzare i soggetti) e statistici (per elaborazioni statistiche).
- Vale la pena di fare sforzi di questa portata? Probabilmente sì. Sono state derivate molte linee guida su FIRE e sui SIRI in generale; alcuni di questi risultati erano previsti e prevedibili prima dell'esperimento (ad esempio, la riformulazione di FIRE propone troppi termini agli utenti), ma altri sono inaspettati (ad esempio, l'aiuto strategico va fornito in modalità automatica). Un esperimento più povero non avrebbe probabilmente fornito queste informazioni impreviste.
- Il livello di difficoltà differente dei due bisogni informativi A e B, non osservato dopo un esperimento pilota, né compreso in modo completo dopo l'esperimento, è un problema che dovrebbe essere preso in considerazione in esperimenti simili.

5. Verso nuovi paradigmi di valutazione

E’ opinione comune (Salton 1992) che i test di laboratorio tradizionali, anche quelli condotti su collezioni di piccola taglia, continueranno a giocare una parte importante nella valutazione dei sistemi di information retrieval, specialmente nella misura in cui non verranno individuati metodi ugualmente produttivi o non emergeranno specifiche e comprovate obiezioni, come ad esempio la presenza di polarizzazioni indesiderate nei dati. Contemporaneamente però bisogna prendere atto che accanto alle critiche tradizionali che accompagnano questi metodi, delle quali abbiamo in parte discusso precedentemente, si pone il problema di un ripensamento generale della filosofia stessa di valutazione, che consideri i grandi mutamenti tecnologici e comportamentali di questi ultimi anni nel campo dell’accesso alle informazioni. Poiché le nuove applicazioni di information retrieval hanno caratteristiche molto diverse da quelle precedenti, c’è bisogno di nuovi concetti e metodi di valutazione. Questi temi, discussi nel seguito in modo più

approfondito, sono al centro anche di un'azione di ricerca di base della Comunità Europea denominata MIRA.

L'information retrieval è stato concepito e si è sviluppato come soluzione al problema di trovare i documenti di una collezione pertinenti alla richiesta di un utente, il suo campo d'applicazione essendo sostanzialmente quello delle ricerche bibliografiche. Oggi la ricerca di informazione ha enormemente dilatato le sue coordinate. La ricerca viene effettuata su basi di dati distribuite multimediali, e coinvolge attori diversi, i quali possono avere obiettivi differenziati. Vediamo meglio come è cambiata la tipologia d'utente e quali sono le nuove funzionalità dei sistemi di information retrieval.

Le persone che fanno ricerca di informazioni non sempre sono mosse da un bisogno informativo preciso. Spesso hanno un bisogno informativo vago, che precisano o correggono durante l'interazione con il sistema, oppure sono spinte dalla semplice curiosità. Da questo punto di vista il World-Wide Web ha favorito tendenze eccentriche, dalla serendipità alla navigazione fine a se stessa. Una conseguenza particolarmente rilevante di ciò è che la metafora dell'interrogazione come specchio del bisogno informativo da soddisfare è oggi quanto mai imperfetta. Bisogna considerare inoltre che anche quando la ricerca è finalizzata al raggiungimento di un obiettivo primario, raramente si tratta di acquisire un'informazione in modo fine a se stesso. L'informazione è sempre più spesso parte integrante di un'attività più complessa svolta con l'ausilio del calcolatore, come ad esempio la stesura di un rapporto, o la preparazione di un bilancio, o la soluzione di un problema di programmazione. Questa dimensione è stata fino ad ora pressoché ignorata, con qualche lodevole eccezione (Mizzaro 1998), mentre potrebbe essere un aspetto decisivo dei futuri servizi per il reperimento (e la riutilizzazione) delle informazioni. Per molti utenti, infine, l'obiettivo non è raggiungere un documento che colmi una lacuna conoscitiva o sia utilizzabile per svolgere un'attività professionale, quanto invece soddisfare bisogni di altra natura, come ad esempio vedere un film o ascoltare un brano musicale. Soddisfare questo tipo di obiettivo può essere altrettanto importante per un servizio di "information" retrieval. È chiaro che la profonda diversificazione delle caratteristiche e degli obiettivi dell'utente dei sistemi di information retrieval complica la definizione di criteri di valutazione precisi e universali; sicuramente essa rende ancora più preziosi i risultati di studi effettuati con utenti reali, spinti da obiettivi reali.

Anche se limitiamo la nostra indagine all'utilizzazione convenzionale più ristretta dei sistemi di information retrieval, vediamo che c'è una proliferazione di nuovi servizi e funzionalità. La diversità e la complessità dell'informazione accessibile, specialmente alla luce dei progressi della tecnologia multimediale e della crescita dell'interconnessione, richiedono lo sviluppo di metodi di indicizzazione e recupero più intelligenti, nonché di nuove metafore e tecniche di interazione. Vanno nella prima direzione, ad esempio, alcuni progetti in corso di svolgimento nel contesto

dell'iniziativa americana "digital libraries" che mirano allo sviluppo di un servizio "news on demand", in cui l'utente possa indicizzare e recuperare in modo diretto servizi giornalistici di interesse messi in onda dai telegiornali, senza passare attraverso una scansione seriale delle trasmissioni ma semplicemente comunicando a voce quello che gli interessa (si veda ad esempio il progetto Informefia Digital Video Library della Carnegie Mellon University). Parallelamente, poiché le possibilità d'interazione per l'utente diventano sempre più complesse e un numero crescente di utenti inesperti si "connettono", diventa essenziale disporre di strumenti di aiuto all'interazione in ogni fase del processo di ricerca delle informazioni. Esempi di sistemi ritagliati su aspetti specifici dell'interazione utente-sistema sono il già visto FIRE, che supporta la chiarificazione e la formulazione del bisogno informativo, Profile (Amati et al 1997), che utilizza il "relevance feedback" nella selezione probabilistica delle informazioni, e Ulisse (Carpineto e Romano 1996), che consente di integrare varie modalità interattive per la selezione delle informazioni (interrogazione, browsing e potatura) in un singolo spazio informativo di tipo ipertestuale. Queste ricerche rappresentano dei passi avanti nella direzione di aumentare e favorire il ruolo dell'utente nella ricerca delle informazioni, ma la distanza da colmare è grande. Le funzionalità di supporto all'utente dovranno essere molto più potenti e versatili, e dovranno includere la possibilità di comunicazioni multimodali fra utente e sistema, la presentazione integrata e coordinata di media differenti e una gestione complessiva dell'interazione fra utenti e sistema, secondo la metafora degli agenti software.

I cambiamenti delineati rimandano, seppure con modalità ed intensità differenti, alla necessità di effettuare valutazioni "in linea" invece che su collezioni campione. In altre parole, i nuovi metodi di valutazione dovranno controllare il comportamento dinamico sia del sistema sia dell'utente, stimando lo sforzo dell'utente in relazione alla sua soddisfazione attraverso l'intera sessione di ricerca. Da questo punto di vista non basta identificare e valutare il contributo che le principali componenti d'interazione, come ad esempio strategia di ricerca, caratteristiche dell'interfaccia e "relevance feedback", forniscono alle prestazioni complessive; occorre altresì studiare la loro integrazione ottimale e investigare l'adeguatezza delle misure di efficacia alla luce delle mutate condizioni progettuali. Un aspetto importante da considerare, che ostacola il raggiungimento di un simile obiettivo, è che la percezione del bisogno informativo da parte dell'utente, così come la strategia di ricerca e i vincoli esterni (tempo, risorse computazionali, costi di accesso all'informazione), possono cambiare durante l'interazione. Non si può assumere infatti che il problema iniziale rimanga immutato durante la ricerca e che ci sia una soluzione ottimale che può essere raggiunta dopo un certo numero d'iterazioni. Le misure tradizionali 'oggettive" (richiamo, precisione, lunghezza attesa della ricerca, tempo d'interazione, ecc.), così come quelle soggettive, potranno essere adoperate, ma con riferimento ai singoli stadi dell'interazione fra utente e sistema.

Un'ultima osservazione concerne l'utilizzazione di materiale multimediale o di altra natura non testuale, come tabelle, programmi sorgente, ecc. Le unità informative di questo tipo hanno una complessità intrinseca elevata, un sistema di relazioni con contesti diversi molto ricco, possibilità di percezione da parte dell'utente peculiari. Così ad esempio, recuperare una procedura software non basta a risolvere un problema di programmazione se non c'è la possibilità di accedere alla sua documentazione. Questi aspetti caratteristici dell'informazione non testuale fanno emergere nuove funzionalità da considerare e stimolano la revisione dei concetti tradizionali, inclusi quelli fondamentali di richiamo e precisione.

Ringraziamenti

Questo lavoro è stato eseguito nell'ambito della Convenzione in atto tra Fondazione Ugo Bordoni e Telecom Italia. Il sitema FIRE è stato sviluppato al Laboratorio di Intelligenza Artificiale del Dipartimento di Matematica e Informatica dell'Università di Udine da Stefano Mizzaro, sotto la guida di Carlo Tasso e Giorgio Brajnik e in collaborazione con un gruppo di ricercatori del Dipartimento di Psicologia dell'Università di Trieste guidato da Danilo Fum.

Riferimenti bibliografici

Amati, G., Crestani, F., Ubaldini, F., & De Nardis, S. (1997). Probabilistic learning for information filtering. *Proceedings RIAO'97*.

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

Brajnik, G., Mizzaro, S. e Tasso, C. (1995). Interfacce intelligenti a banche di dati bibliografici, in D. Sacca (ed.), Sistemi evoluti per basi di dati, Franco Angeli, Milano, pp. 95-128.

Brajnik, G, Mizzaro, S. e Tasso, C. (1996). Evaluating User Interfaces to Information Retrieval Systems: A Case Study on User Support, in SIGIR96, 19th International Conference on Research and Development in Information Retrieval, Zurigo, Svizzera, pp. 128-136.

Carpineto, C. & Romano, G. (1996). Information retrieval through hybrid navigation of lattice representations. *International Journal of Human-Computer Studies*, 45, pp. 553-578.

Cleverdon, C.W. (1962). The testing and analysis of an investigation into the comparative efficiency of indexing systems, Cranfield, U.K.: College of Aeronautics.

Dunlop, M. (1997). Time, relevance and interaction modelling for information retrieval. *Proceedings of the 20th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. pp. 206-213, Philadelphia, USA.

Hull, D. (1993). Using statistical testing in the evaluation of retrieval experiments. *Proceedings of the 16th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. pp. 329-338, Pittsburgh, USA.

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

Mizzaro, S. (1997a). Il Reperimento delle informazioni: analisi teorica e sperimentazione, Tesi di dottorato in Ingegneria dell'Informazione (IX Ciclo), Universita di Trieste.

Mizzaro, S. (1997b). Relevance: The whole history*, Journal of the American Society for Information Science* 48(9): 810-832.

Mizzaro, S. (1998). How many relevances in information retrieval?, Interacting With Computers . In corso di stampa.

Robertson, S. E. e Beaulieu, M. M. H. (1992). On the evaluation of IR systems, Information Processing & Management 28(4): 457-466.

Salton, G., & McGill, M. (1983). *Introduction to Modern Information Retrieval*. New York: McGraw Hill.

Salton, G. (1992). The state of retrieval system evaluation. Information Processing and Management. 28(4), 441-449.

Tague-Sutcliffe, J. (1992). The Pragmatics of Information Retrieval Experimentation, Revisited. *Information Processing & Management*, 28(4), 467-490.

Tague-Sutcliffe, J., & Blustein, J. (1995). *A statistical analysis of TREC-3 data*. In Harman, D. (Ed.), Proceedings of the Third Text REtrieval Conference (TREC-3).

Voorhees, E., & Harman, D. (1996). *Overview of the Fifth Text REtrieval Conference (TREC-5)*. In Voorhees, E. & Harman, D. (Eds.), Proceedings of the Fifth Text REtrieval Conference (TREC-5) .