Consiglio Nazionale delle Ricerche

ISTITUTO DI ELABORAZIONE
DELLA INFORMAZIONE

PISA

UN'INTERFACCIA PER L'ESPLORAZIONE DELLE CONOSCENZE IN UN SISTEMA DI RICUPERO DELL'INFORMAZIONE

M.B. Baldacci, A.M. Nardelli, M.C. Parise

Nota interna B83-20

Dicembre 1983

UN'INTERFACCIA PER L'ESPLORAZIONE DELLE CONOSCENZE IN UN SISTEMA DI RICUPERO DELL'UNFORMAZIONE^

M.B Baldacci A.M. Nardelli M.C. Parise

1. La formulazione delle richieste

Nei sistemi di ricupero dell'informazione, un problema che non sembra trovare una facile risposta è quello dell'espressione, da parte dell'utente, del suo "reale" bisogno di informazione. Il concetto di "bisogno di informazione", che può essere inteso come il riconoscimento, da parte di chi lo esprime, di una situazione problematica, è stato analizzato in modo approfondito poiché ci si è resi conto che su questo concetto sono state fatte molte assunzioni che hanno limitato la comprensioni dei problemi e la possibilità di risolverli [1]. A volte l'utente può aver bisogno di informazioni che egli riesce chiaramente a descrivere, altre volte invece il suo bisogno gli si presenta in modo talmente vago che può rivelarsi utile ricercare stimoli per la sua definizione: è il caso in cui l'utente non riesce a formulare una richiesta ma è in grado di riconoscere ciò che gli può essere utile fra le cose che gli vengono mostrate.

Tutta questa tematica, che è oggetto di ricerca dei lavori [1-3], ha portato nel passato alla realizzazioni di sistemi di ricupero dell'informazione nei quali è privilegiata la fase di

---

^Lavoro eseguito in preparazione della dissertazione orale per la laurea in Scienze dell'Informazione presso l'Università di Pisa. Candidate: A.M. Nardelli e M.C. Parise; Relatori: R. Sprugnoli e M.B. Baldacci. Anno Accademico 1981-82.

"esplorazione" delle rappresentazioni dei documenti, al fine di permettere all'utente un passaggio meno problematico dai bisogni di informazione alla loro espressione in richieste da fare al sistema [4-5].

2. Il sistema e l'ambiente.

Se si vuole tenere conto dei fattori umani e ambientali che influenzano l'efficienza di un sistema di documentazione, e' necessario ampliare il concetto di sistema considerando anche l'ambiente circostante [6].

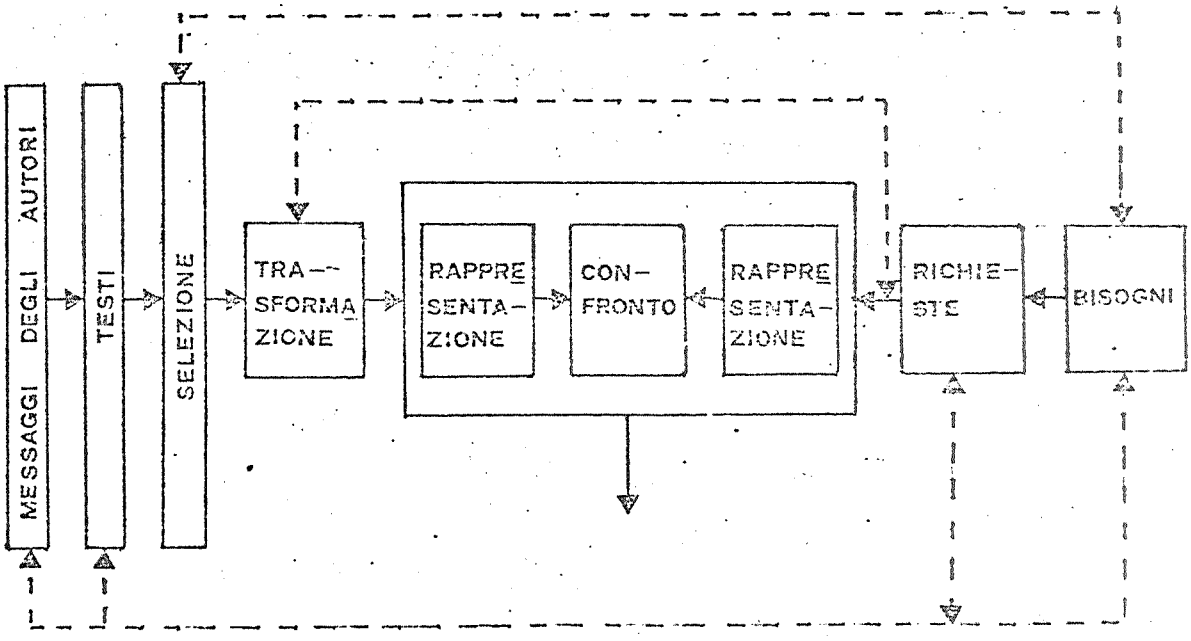

FIG. 1

Come infatti e' messo in evidenza dalla fig. 1, oltre che dalle componenti interne (rappresentazioni, confronto), l'efficienza del sistema e' influenzata anche dalle relazioni che esistono fra esse e in particolare da quella che lega la selezione dei documenti ai bisogni degli utenti. Se da una parte e' difficile evitare l'influenza dei fattori umani sulla descrizione dei documenti, dall'altra si possono creare delle condizioni perche' sia agevolata la formulazione della richiesta da parte dell'utente.

Tenendo conto di queste osservazioni e' stato proposto un sistema provvisto di un'interfaccia che permette ad un qualsiasi utente, anche occasionale, di avere dal sistema informazioni che lo guidino nella definizione della strategia di ricerca [7]. Si fa uso di un linguaggio libero per facilitare l'approccio dell'utente al sistema, ma nello stesso tempo si fornisce un vocabolario controllato di dimensioni modeste in modo che sia sempre possibile iniziare una ricerca. Ad ogni documento, infatti, vengono associati uno o piu' descrittori appartenenti al vocabolario controllato e un certo numero di termini estratti liberamente dal testo. Ne risulta una descrizione "a due livelli" che si dimostra un ottimo strumento per poter ottenere un alto grado di specificita', con i termini liberi, in un contesto ben definito mediante l'uso di termini controllati. A partire da essa e' possibile creare in maniera automatica, una rete semantica nella quale i

termini sono messi in correlazione sia quando sono associati allo stesso documento, sia quando sono associati allo stesso descrittore generale. Inoltre la registrazione di termini utilizzati dall'utente e non ritrovati all'interno del sistema, puo' fornire delle indicazioni sulla relazione esistente fra bisogni dell'utente e selezione dei documenti. In base a questa proposta e' stata realizzata l'interfaccia che e' oggetto del presente lavoro. I principi su cui poggia tale interfaccia sono esposti nel par.3; nel par.4 e' descritta la sua realizzazione.

3. L'esplorazione delle conoscenze

3.1 La rappresentazione dei documenti.

La base di dati che si considera e' costituita da un insieme di record ciascuno dei quali rappresenta la descrizione di un documento.

Ad ogni documento vengono associati nella sua descrizione uno o piu' termini cosiddetti "generali", appartenenti ad un vocabolario controllato di dimensioni limitate, e un insieme di termini "analitici" (li chiameremo anche specifici o particolari), liberamente tratti dal testo del documento.

La funzione del termine generale e' quella di individuare un ben determinato campo di indagine a cui il documento puo' essere associato.

I termini analitici costituiscono invece una vera e propria descrizione del contenuto semantico del documento.

E' da notare che i termini generali danno un'indicazione del settore di conoscenza a cui e' legato un documento, relativamente al contenuto informativo dell'intera base di dati.

La presenza di un piccolo vocabolario controllato costituito da questi termini, si rivela utile al momento della ricerca

per avere informazioni sulle aree di conoscenza disponibili. Esso rappresenta, tra l'altro, un mezzo sicuro per poter dare inizio ad una ricerca.

Una delle maggiori difficolta' che si incontrano nel rapporto utente-sistema sta, infatti, nella formulazione, da parte dell'utente, di richieste che utilizzino termini noti al sistema. In genere si cerca di risolvere tale problema proponendo all'utente la consultazione di thesauri appositamente costruiti che suggeriscono possibili termini di ricerca. E' possibile iniziare la ricerca se l'utente riesce ad esprimersi con almeno un termine presente nel thesaurus.

Con l'interfaccia proposta, invece, l'utente, nel formulare la sua richiesta, viene lasciato libero di esprimere i suoi bisogni con la terminologia che piu' gli sembra adatta.

La ricerca ha inizio solo se i termini da lui usati sono presenti tutti o in parte, nella rappresentazione di almeno un documento della base di dati.

Se pero' questo non si verifica, la disponibilita' dell'insieme di termini controllati puo' permettere ugualmente di dare l'avvio alla ricerca.

Tale insieme e' infatti sufficientemente piccolo da poter essere proposto interamente all'utente che puo' scegliere senza grosse difficolta' quali termini utilizzare per una nuova formulazione della richiesta.

3.2 La rappresentazione della conoscenza.

La descrizione "a due livelli" che abbiamo visto, realizza implicitamente un legame fra documento e documento: la presenza in piu' descrizioni dello stesso termine generale indica che i documenti relativi trattano la stessa branca di conoscenza; allo stesso modo, le descrizioni in cui compare lo stesso termine analitico sono relative a documenti che trattano quello stesso argomento. Un altro tipo di legame si ha poi fra termini generali e termini specifici quando questi sono presenti nella descrizione di uno stesso documento.

Tutta la conoscenza presente nella base di dati puo' essere quindi sinteticamente rappresentata mediante una "rete-semantica" in cui termini generali, termini analitici e documenti costituiscono i nodi e le connessioni fra essi, gli archi (vedi fig.2).

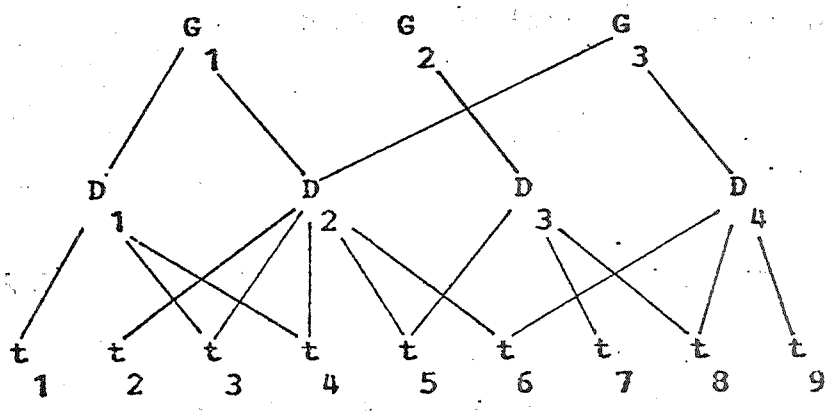

Fig.2

In tale rete e' possibile "navigare" da un nodo ad un altro mediante l'uso di alcune funzioni che sfruttano le relazioni esistenti fra i nodi.

Indicando con:

$D=\{d_1, d_2, \ldots, d_m\}$ l'insieme totale delle descrizioni dei documenti che costituiscono la collezione.

$G=\{g_1, g_2, \ldots, g_k\}$ l'insieme totale dei termini generali presenti all'interno delle descrizioni.

$T=\{t_1, t_2, \ldots, t_h\}$ l'insieme totale dei termini specifici presenti all'interno delle descrizioni.

con $m, k, h \in N$

si definiscono le seguenti funzioni:

$t1: D \longrightarrow \mathcal{B}(T)$

$t1(d_i) = T' \subseteq T$ con $d_i \in D$

$t1^*(D') = \bigcup_{d_{j_i} \in D'} t1(d_{j_i})$ con $D' = \{d_{j_1}, d_{j_2}, \ldots, d_{j_i}\} \subseteq D$

$t2: G \longrightarrow \mathcal{B}(T)$

$t2(g_i) = T' \subseteq T$ con $g_i \in G$

$t2^*(G') = \bigcup_{g_{i_j} \in G'} t2(g_{i_j})$ con $G' = \{g_{i_1}, g_{i_2}, \ldots, g_{i_k}\} \subseteq G$

$G1: D \longrightarrow \mathcal{B}(G)$

$G1(d_i) = G' \subseteq G$ con $d_i \in D$

$G1^*(D') = \bigcup_{d_{j_i} \in D'} G1(d_{j_i})$

La funzione t1 (che per una migliore comprensione indicheremo con t(Doc)), a partire da un insieme di descrizioni di documenti, individua i termini analitici presenti in esse.

Considerando, ad esempio, le descrizioni d_1 e d_2 nella fig.3.1, il risultato dell'applicazione di tale funzione sara' dato dall'insieme $\{t_1, t_2, t_3, t_4, t_5, t_6\}$.

La funzione t2 (che indicheremo con t(g)), individua l'insieme dei termini analitici collegati ai termini generali in considerazione (per la loro presenza simultanea all'interno di una stessa descrizione).

Se applichiamo la funzione t(g) ai termini g_2 e g_3 di fig.3.1, si ottiene come risultato l'insieme $\{t_2, t_3, t_4, t_5, t_6, t_7, t_8, t_9\}$.

La funzione G1 (che indicheremo con g(Doc)), applicata ad un insieme di descrizioni di documenti, individua tutti i termini generali che compaiono in tali descrizioni.

Considerando, per esempio, le descrizioni d_2 e d_3 in fig.3.1, si ottiene l'insieme $\{g_1, g_2, g_3\}$.

L'uso di queste funzioni permette di esplorare la base di dati secondo criteri logici diversi a seconda delle diverse

necessita' che, di volta in volta, gli utenti esprimono.
Applicando la funzione t(Doc) si ricavano tutti i termini analitici presenti nelle descrizioni degli ultimi documenti individuati; operando poi una scelta fra essi, che sia significativa per la ricerca, si puo' poi procedere al recupero di quei documenti che sono descritti con i termini scelti. Con buona probabilita' i documenti recuperati saranno rilevanti e su di essi puo' essere riapplicato lo stesso procedimento.
Ricorrendo all'altra funzione, G(Doc), possiamo allargare le nostre conoscenze sui campi disponibili alle indagini e a cui sono legati gli ultimi documenti considerati.
La funzione t(g) permette, poi, di ricavare dai nuovi termini generali altri termini analitici mediante i quali e' possibile formulare una nuova richiesta.

4. Organizzazione dell'interfaccia

L'interfaccia è stata progettata per un sistema di ricupero dell'informazione implementato con il S.G.B.D. RESP ed è stata realizzata servendosi delle funzioni che tale sistema mette a disposizione.^

L'interfaccia realizzata vuol essere in parte un modo per facilitare l'utente nella sua ricerca evitandogli l'uso di un qualsiasi tipo di manuale o dizionario, e in parte uno strumento di controllo affinche' ad ogni passo la ricerca risulti significativa.

La logica su cui si basa l'uso delle funzioni nell'interfaccia e' che l'alternarsi del loro impiego deve essere tale da permettere un'esplorazione graduale e non ripetitiva della base di dati.

La ricerca e' di tipo interattivo, e questo per poter utilizzare ad ogni passo il giudizio dell'utente sul risultato temporaneo e, in seguito ad esso, decidere come proseguire la ricerca.

L'interazione permette inoltre all'utente di chiarire a se stesso le proprie necessita' e portare avanti, man mano, una ricerca sempre meglio definita.

L'interfaccia aiuta l'utente ad accedere ad almeno un documento dal momento che nella richiesta fatta esiste almeno un termine presente in qualche rappresentazione di documento (questo per la presenza del dizionario a dimensione limitata di cui si e' gia' parlato).

---

^Si ringrazia vivamente R. Sprugnoli per l'aiuto che ci ha dato per la realizzazione dell'interfaccia.

Le possibili strategie di ricerca di cui si compone l'interfaccia sono:

-Esplorazione di tipo E1: e' imperniata sulla funzione t(Doc); si considerano i termini specifici dei documenti individuati e si propongono all'utente che sceglie tra questi quelli che meglio descrivono cio' che sta cercando. Tali termini possono individuare particolari aspetti dell'argomento di ricerca a cui l'utente poteva, in un primo momento, non aver pensato oppure non sapere in che modo fare riferimento.
Mediante l'operazione booleana di unione (OR) applicata ai termini scelti, viene recuperato un certo numero di documenti.

Esplorazione di tipo E2: si attua mediante un allargamento della conoscenza dell'utente sui termini generali legati a documenti risultati significativi.
Viene applicata la funzione G(Doc) agli ultimi documenti recuperati e si fa eseguire una scelta all'utente dei termini cosi' ottenuti.
Poiche' i termini generali individuano altri

campi, applicando la funzione t(g) a quelli scelti, possono essere suggeriti nuovi termini particolari da usare per il recupero di altri documenti. Tale recupero viene effettuato mediante l'operazione booleana di unione (OR) applicata ai termini particolari scelti dall'utente.

-Esplorazione di tipo E3: fa uso della funzione G(Doc) ampliando cosi' le conoscenze sugli argomenti generali presenti nelle descrizioni dei documenti di cui l'utente sembra essere soddisfatto.
Il recupero dei documenti si realizza con l'applicazione dell'operazione booleana di unione (OR) ai termini generali scelti.

-Riduzione di tipo R1: perche' si possa realizzare e' necessaria la presenza sia dei termini generali g, sia dei termini specifici t, poiche' il recupero si basa sull'operazione booleana di intersezione (AND) fra l'insieme unione dei termini generali e l'insieme unione dei termini specifici scelti.

-Riduzione di tipo R2: riguarda il caso particolare in cui l'utente desidera recuperare soltanto quei documenti che trattano ciascuno tutti gli argomenti indicati. Tale ricerca viene effettuata mediante l'operazione di intersezione fra l'insieme unione dei termini generali scelti e l'insieme intersezione dei termini specifici scelti.

Partendo da un certo numero di documenti ricavati dalla richiesta iniziale dell'utente, nell'eventualita' che la risposta non sia soddisfacente, si procede o riducendo il numero dei documenti o ricercandone di nuovi a seconda dei desideri dell'utente. A questi viene anche data la possibilita' di vedere quali sono i documenti recuperati e dare un giudizio di rilevanza che permetta di continuare la ricerca in base ai documenti giudicati positivamente.
Nel caso l'utente voglia ottenere nuovi documenti, si applica l'esplorazione di tipo E1 poiche' questa propone all'attenzione dell'utente una sequenza di termini specifici che e' probabile si rivelino significativi per la ricerca dell'utente in quanto presenti nella descrizione di documenti che questi ha ritenuto soddisfacenti.
Su tale esplorazione viene effettuato un ciclo finche' fra i termini scelti dall'utente ce ne sono di nuovi rispetto a

tutti i precedenti considerati.

Quando l'utente non trova termini nuovi fra quelli proposti, o perche' non ne esistono piu' o perche' non ce ne sono altri che si addicono alla sua ricerca, si rende necessaria una esplorazione della terminologia che prenda in considerazione i termini generali legati agli ultimi documenti recuperati, si passa percio' all'esplorazione di tipo E2.

Se questa porta al recupero di nuovi documenti significativi, si ritorna all'applicazione di E1 sempre che il desiderio dell'utente sia quello di individuare nuovi documenti.

Si ritiene infatti che la modalita' di ricerca piu' significativa sia quella che si riconduce all'uso di termini specifici utilizzati come descrittori di documenti giudicati rilevanti.

Quando l'applicazione di E2 non porta alla scoperta di nuovi documenti, si passa all'esplorazione di tipo E3 in modo da avere una visione piu' ampia della collezione dei documenti. Anche su E3 e' possibile operare un ciclo finche' la presenza dei termini generali nuovi porta al recupero di documenti diversi. Quando questo non e' piu' possibile viene mostrata la struttura dei termini generali in modo che l'utente abbia la possibilita' di scegliere fra questi i termini che piu' rispondono alle sue necessita' e con questi poter riformulare la domanda.

Ogni volta che all'utente viene mostrato il numero dei documenti recuperati ad un certo passo, gli viene data la possibilita' non solo di ricercare documenti nuovi, ma anche di ridurre il numero di documenti se questo risulta troppo grande.

In tal caso si applica la riduzione R1 sulla quale e' anche possibile ciclare quando l'utente desideri continuare a ridurre il numero di documenti. Per la realizzazione del ciclo gli viene imposto di effettuare una scelta sull'ultimo insieme di termini specifici considerati.

Se ci si accorge che il desiderio dell'utente non e' tanto quello di diminuire il numero dei documenti quanto piuttosto di individuare dei documenti ben precisi, si ricorre all'applicazione della riduzione R2.

Se, durante un'operazione di riduzione, l'utente manifesta il desiderio di ricercare nuovi documenti, si passa all'applicazione dell'esplorazione E2. questo perche' si suppone che un tale tipo di esplorazione, individuando altri campi d'interesse, possa fornire all'utente, in base a questi, termini specifici diversi da quelli da lui scartati durante la riduzione.

La rappresentazione grafica delle connessioni fra le strategie è mostrata in fig.3.

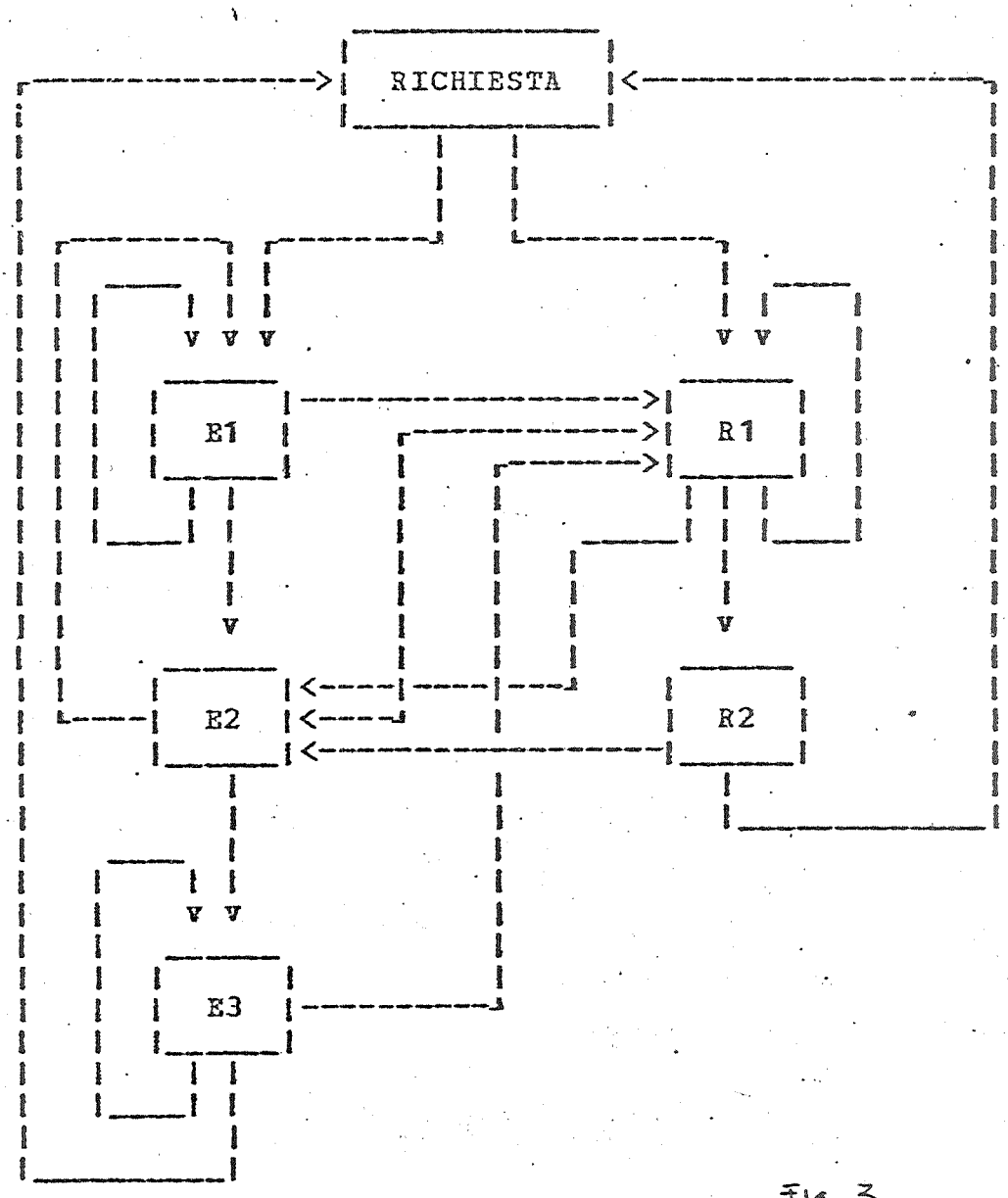

Fig.3

Il generico passo di ricerca, dopo il recupero di documenti dovuto alla richiesta iniziale, si presenta nel modo seguente:

- dinamica del query: decisione da parte dell'utente sulla continuazione o meno della ricerca; l'utente puo' anche decidere di vedere i documenti recuperati e intervenire cancellando quelli non rilevanti per continuare la ricerca in base ai rimanenti.

- esecuzione di una delle strategie di ricerca qualora l'utente voglia continuare.

La rappresentazione grafica della dinamica del query e' mostrata in fig. 4.

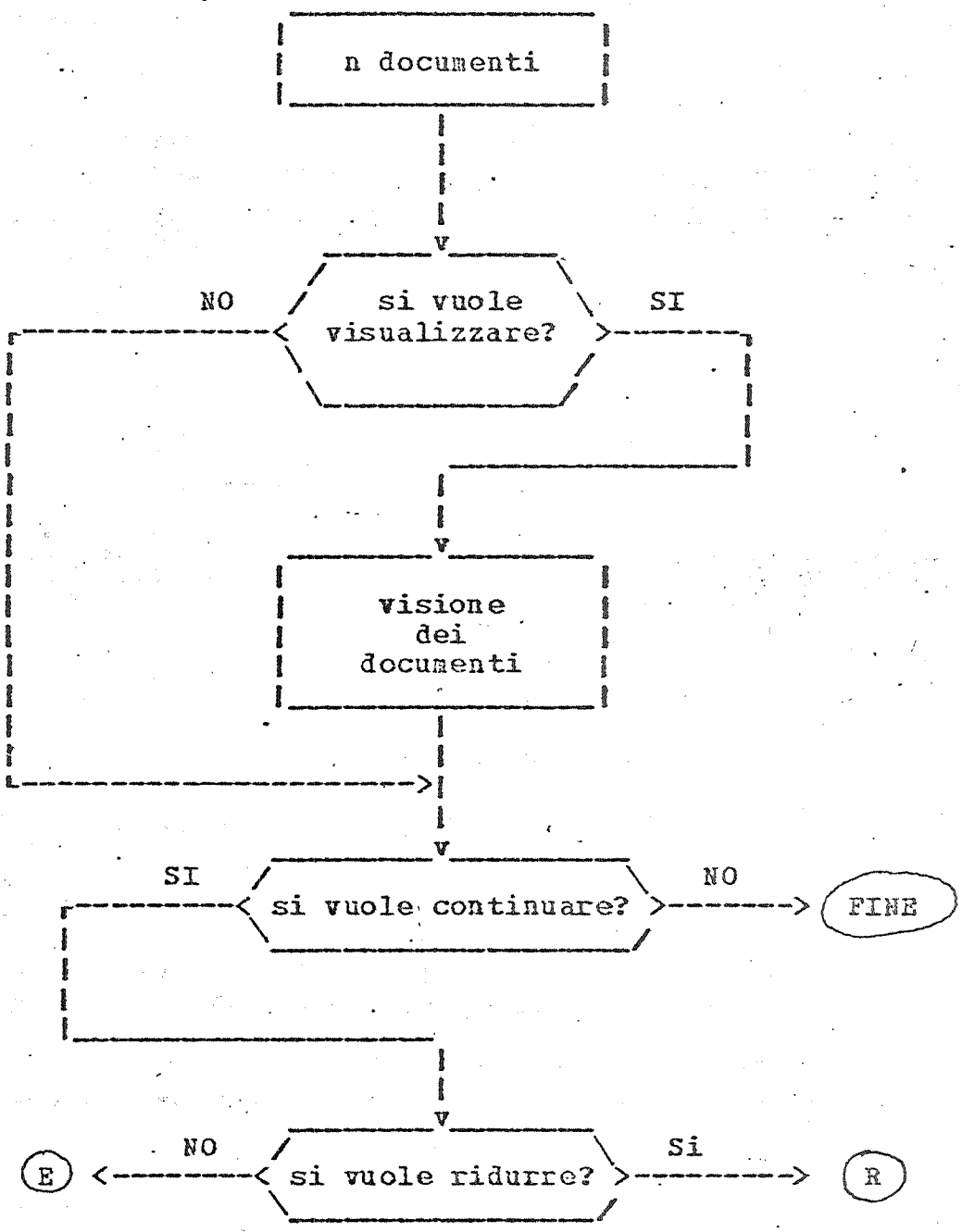

Fig. 4

Si e' ritenuto opportuno memorizzare man mano i termini scelti dall'utente per evitare che in un passo successivo possa essere fatta una scelta di termini gia' avuta in precedenza.

Ad ogni scelta e' quindi necessario verificare l'esistenza di termini nuovi in essa, per decidere se insistere sul ciclo in questione o se e' il caso di passare ad una esplorazione di tipo diverso quando nell'ultima scelta effettuata non sonopresenti termini nuovi.

E' stato inserito anche un controllo per verificare che il ciclo fornisca nuovi documenti; in caso negativo si salta anche da qui ad un nuovo tipo di ricerca.

E' stata data la possibilita' all'utente di vedere gli eventuali nuovi documenti recuperati per poter ottenere un giudizio di rilevanza su questi.

La possibilita' di alternare esplorazioni e riduzioni permette di assecondare ad ogni passo i desideri dell'utente il cui compito e' semplicemente quello di accettare o respingere i termini che gli vengono mostrati e/o i documenti.

Se da una parte il numero dei documenti recuperati sembra avere una certa importanza quando si rivela troppo alto e si rende quindi necessario operare in modo da ridurlo (si parla percio' di riduzione), dall'altra, quando l'insieme dei documenti recuperati non e' troppo grande, si e' ritenuto che, in generale, quello che piu' interessa all'utente non e'

tanto la quantita' del materiale recuperato quanto la sua specificita' riguardo al problema in esame.

E' per questo che nei punti in cui e' richiesto un giudizio dell'utente sulla situazione, una volta accertata la sua volonta' di proseguire nella ricerca, se egli non vuole ridurre il numero dei documenti dell'ultimo recupero, se ne deduce che vuole migliorare i risultati e per questo si procede mediante esplorazioni del linguaggio di descrizione. Tutto cio' senza alcuna preoccupazione per quello che sara' il numero dei documenti recuperati.

Per la realizzazione di questa interfaccia non si e' reso necessario l'uso di alcuna struttura semantica precostituita. Tutta la ricerca si basa sulla presenza dei termini indicati dall'utente all'interno delle descrizioni dei documenti.

I termini che l'utente utilizza per iniziare la sua ricerca possono talvolta non essere ritrovati all'interno dei dizionari del sistema. In tal caso essi vengono memorizzati poiche' possono fornire informazioni o su modi alternativi di indicizzazione dei documenti, o su particolari bisogni dell'utente che la collezione a disposizione non riesce a soddisfare. In quest'ultimo caso si puo' pensare di intervenire, eventualmente, sulla selezione dei documenti.

4.2 Conclusioni

L'interfaccia realizzata offre la possibilita' ad un qualsiasi tipo di utente di iniziare e portare a termine una ricerca. Il compito dell'utente viene ridotto a semplici scelte fra varie alternative che gli vengono proposte; risulta molto piu' semplice, infatti, riconoscere qualcosa di utile fra cio' che viene mostrato, che non formulare una richiesta ben definita, cosa che implica, talvolta, una conoscenza dettagliata sia dell'argomento di ricerca, sia degli strumenti forniti dal sistema. Il fatto che tutta la ricerca sia imperniata sulla continua interazione fra utente e sistema fa si' che l'ambito dell'indagine risulti sempre meglio delineato per lo stesso utente che, in alcuni casi, puo' trovarsi di fronte ad un argomento completamente nuovo. I collegamenti stabiliti fra le diverse strategie sembrano guidare bene l'utente nell'esplorazione della base di dati e, nello stesso tempo, seguono i suoi desideri in modo soddisfacente.

In appendice viene presentato un esempio di seduta a terminale in cui e' possibile osservare la "navigazione" permessa dall'interfaccia all'interno della base di dati durante un'operazione di ricerca.

R i f e r i m e n t i

1 Belkin, N.J., "The problem of 'matching in information retrieval", Proc. of the 2nd. Information Retrieval Forum, Copenhagen, 1977.

2 Belkin, N.J., "Anomaoulus states of knowledge as a basis for information retrieval", Canadian J. of Information Science, 5 (1980), &33-143.

3 Belkin, N.J., Brooks, H.H., Oddy, R.N., "Representing and classifying anomalous states of knowledge", Proc. ASLIB, London, 1979, pp.227-238.

4 Oddy, R.N., "Information retrieval through man-machine dialogue", J. Documentation, 33 (1977), 1-14.

5 Baldacci, M.B., Sprugnoli, R., "Recupero dell'informazione bibliografica: un'interfaccia utente-sistema per la definizione delle strategie di ricerca". AICA 1977, pp.129-30.

6 Baldacci, M.B., "Dalla documentazione automatica alla scienza dell'informazione", AICA 1980, Bologna, 1980, pp.1346-1363.

7 Baldacci, M.B., "L'esplorazione delle conoscenze per il ricupero dell'informazione". Lavoro non pubblicato, presentato al Convegno "L'informatica sanitaria nel contesto della informatizzazione degli Enti Locali", CNR, PFI, Firenze, 1981.

8 Sprugnoli, R., RESP: User's manual. I.E.I., Nota Interna B80-9, Giugno 1980.