UNIVERSITA' DEGLI STUDI DI PADOVA

FACOLTA' DI SCIENZE STATISTICHE

CORSO DI LAUREA IN STATISTICA E TECNOLOGIE INFORMATICHE

IMPLICAZIONI NELL'USO DELLA RETROAZIONE BASATA SULLA RILEVANZA NEI SISTEMI DI REPERIMENTO DELL'INFORMAZIONE

Relatore: Prof. Massimo Melucci

Laureando: Luca Giuseppe Grifò
N.Matr: 449762

ANNO ACCADEMICO
2006-2007

INDICE GENERALE:

1.1 Una visione d'insieme

Il compito di un sistema di *Information Retrieval* è quello di reperire un insieme di documenti: un utente esprime in linguaggio naturale una richiesta, il sistema si occupa di tradurla per essere applicata allo specifico algoritmo di reperimento dell'informazione, sia esso probabilistico, vettoriale o frequentistico, e presenta infine una lista di documenti ordinati per grado di rilevanza. Il sistema ha lo scopo di reperire tutti e solo quei documenti contenenti informazione rilevante all'esigenza informativa dell'utente. L'insieme dei documenti proposti rappresenta la migliore risposta possibile alla richiesta fatta dall'utente.

Nella pratica, infatti, l'esatta corrispondenza tra la richiesta dell'utente e l'effettiva rilevanza dei documenti reperiti non è sempre perfetta. Questo fenomeno accade per vari motivi: in primo luogo, l'utente esprime la propria richiesta in termini di linguaggio naturale pertanto possono verificarsi differenze tra al semantica delle parole usate nei documenti e quella usata dall'utente [Terra e Warren, 2005] che il sistema di *Information Retrieval* non è sempre in grado di gestire in modo autonomo; in secondo luogo, l'informazione portata dall'utente non sempre risulta essere sufficiente per un funzionamento efficace del sistema.

Partendo da queste ultime considerazioni, sono state sviluppate tecniche di modifica della richiesta. La più diffusa di queste tecniche [Jansen, Spink e Saracevic, 1999] è il sistema di *Relevance Feedback*.

Tale tecnica è strutturata in due fasi: a seguito della formulazione della richiesta da parte dell'utente il sistema presenta un insieme di documenti elencati in ordine decrescente di grado di rilevanza e ciò è il risultato dell'informazione apportata con la semplice richiesta; successivamente viene richiesto l'intervento dell'utente che dovrà marcare come rilevanti (*Positive Relevance Feedback*: PRF) o non pertinenti (*Negative Relevance Feedback*: NRF) i

documenti proposti; di conseguenza, il sistema propone un nuovo insieme di documenti ordinati che terranno conto della nuova informazione apportata dall'intervento diretto dell'utente. Sebbene tale tecnica sia ben consolidata, è altresì noto come i risultati non siano sempre soddisfacenti. Come sottolineato nell'articolo *The Effect of Accessing Nonmatching Documents on Relevance Feedback* di M.D. Dunlop (1997), il marcare come rilevanti documenti con un grado di rilevanza elevato, cioè quelli con un rango elevato, equivale a confermare il risultato proposto dal sistema; è quindi lecito attendersi che nella successiva fase, l'elenco dei documenti non si discosti da ciò che già era stato proposto senza l'intervento del *feedback* dell'utente.

Viceversa, marcare come non rilevanti documenti con un rango elevato dovrebbe modificare in modo significativo l'ordine dei documenti reperiti dopo il *Feedback* dato che si sta smentendo il risultato proposto dal sistema. Si è invece osservato nelle applicazioni che un sistema così strutturato non ha il comportamento previsto col risultato che il *feedback* fornito dall'utente non influisce in maniera significativa nell'ordinamento dei documenti ottenuti nella seconda fase. Più precisamente, nel meccanismo di *Negative Relevance Feedback* non vi è la necessaria informazione per poter modificare significativamente l'ordine dei documenti [Dunlop, 1997]. Questa mancanza deriva principalmente da come vengono trattati i documenti non reperiti dal sistema: sono essi marcabili a priori non rilevanti (anche se non sono stati visionati dall'utente)? Oppure devono non devono essere presi numericamente in considerazione? Partendo da queste basi, lo scopo della presente Tesi è di indagare circa la funzionalità del sistema di *Relevance Feedback* e fornire quindi una spiegazione matematica della differenza di tra il comportamento ipotizzato e quello effettivamente osservato negli esperimenti da Dulop.

1.2 Metodologie e approccio adottati

Prima di procedere alla trattazione dell'argomento in oggetto è consigliabile chiarire le metodologie e gli strumenti matematici adottati per raggiungere lo scopo di questa tesi.

Un aspetto importante della trattazione è riservato al tipo di approccio adottato: esso è mutuato dalle metodologie utilizzate in ambito statistico ed è chiamato "modello a scatola nera" [Azzalini e Scarpa, 2004].

Il modello "a scatola nera" permette di indagare il funzionamento del fenomeno oggetto di studio senza riprodurre neppure parzialmente il suo meccanismo. Di per sé tale approccio sembrerebbe controproducente in relazione al tipo di argomento trattato, ma in realtà presenta un vantaggio innegabile: permette la costruzione di un modello matematico senza la necessità di conoscere la formulazione del fenomeno in oggetto; tale modello avrà una semplice funzione operativa, cioè quella di "spiegare il funzionamento“, non di riprodurlo. Non è poi da sottovalutare un altro vantaggio: i possibili errori o le eventuali semplificazioni del sistema preso in esame vengono evitate visto che non si userà la stessa formulazione per verificarne il funzionamento.

Sebbene non si facciano ipotesi circa il sistema di *Information Retrieval* usato, è altresì fondamentale chiarire che si dà per certo il buon funzionamento dello stesso: l'algoritmo usato dal sistema si presuppone funzioni in maniera efficace e i risultati che propone siano al massimo delle capacità [*Probability Ranking Principle*, S.E Robertson, 1977].

Oltre a ciò si suppone che l'utente abbia accesso anche all'insieme dei documenti non reperiti dal sistema cosicché sia possibile esprimere un giudizio di rilevanza su di essi.

Quest'ultima ipotesi è necessaria ai fini della formulazione matematica adottata, ma servirà

anche per dimostrare come l'accesso a quei documenti fornisce un miglioramento del sistema di *Relevance Feedback*.

I modelli matematici utilizzati sono derivati dal Teorema di Bayes sulle probabilità condizionate: nel caso di PRF si userà il Rapporto a favore condizionato, un semplice rapporto tra le quantità in gioco usato in questo contesto per verificare come si discostano in termini numerici gli insiemi considerati; nel caso di NRF si adotterà una formulazione estesa del Teorema di Bayes con applicazione del concetto di Probabilità a Priori il cui obiettivo è verificare quale sia la relazione che contribuisca a modificare significativamente l'ordine dei documenti ottenuti a seguito del F*eedback* fornito.

1.3 Concetto di efficacia del sistema IR

Un sistema di *Information Retrieval* risponde all'esigenza informativa I in relazione ad una richiesta formulata da un utente. Tale sistema in quanto deterministico nel suo funzionamento reperisce un documento con certezza. Pertanto la nozione di probabilità non è applicabile. Partendo dal presupposto espresso ad inizio tesi, cioè che si adotta un metodo di indagine "a scatola nera", non si è a conoscenza delle modalità di funzionamento del sistema IR e si è quindi portati a ragionare in termini di probabilità di reperimento dei documenti.

Consideriamo l'esigenza informativa I: ad essa si associano tutte le possibili richieste osservabili che portano alla medesima informazione I; come è lecito attendersi, un sistema IR risponderà in modo diverso a diverse richieste, pertanto l'insieme dei documenti reperiti A varia anche se l'informazione I è la medesima. Definire quindi un sistema IR "ben funzionante" equivale ad ammettere che tale sistema tende a rispondere con lo stesso insieme A anche quando la richiesta varia per la medesima esigenza informativa I.

Si noti come in questa formulazione, l'insieme A non rappresenti l'evento in cui sia stata reperita informazione rilevante ma solo l'evento in cui siano stati reperiti alcuni documenti, cioè quell'insieme A. Di conseguenza, P[A] non indica la probabilità di rilevanza.

Secondo i principi di funzionamento appena esposti è lecito aspettarsi che un sistema di *Information Retrieval* risulti ben funzionante se si verifica la seguente relazione:

$$P[A] \geqslant P[\bar{A}]$$

con A sottoinsieme dei documenti reperiti e $\bar{A}$ il relativo sottoinsieme dei documenti non presentati all'utente. Con la precedente relazione si sta quindi ammettendo che un sistema è efficace se la probabilità dei documenti reperiti è maggiore di quella dei documenti non reperiti. E' bene comunque osservare che le probabilità in gioco nella relazione considerata

non siano ottenute secondo l'accezione classica della probabilità, casi favorevoli su casi possibili. Ciò significa che le probabilità considerate siano stabilite a priori e che la stessa relazione sia da ritenersi sempre verificata, cioè si assume che un sistema di *Information Retrieval* ben funzionante confermi la disuguaglianza.

2.1 Definizione del Rapporto a favore

Il Rapporto a favore di un certo evento A è definito da:

$$\frac{(P[A])}{(P[\bar{A}])} = \frac{(P[A])}{(1-P[A])}$$

Il Rapporto a favore di un evento A, pertanto, esprime di quanto è più probabile che l'evento A si realizzi rispetto al fatto che l'evento non si realizzi. Ad esempio, se:

$$P[A] = \frac{2}{3} \quad \text{allora} \quad P[A] = 2\,P[\bar{A}]$$

Risulta essere intuitivo come in questo caso il Rapporto a favore di A è uguale a 2. In generale quindi, si dice che se il Rapporto a favore è uguale ad ∂, le possibilità di realizzazione dell'ipotesi considerata sono di "∂ a 1" a favore di A.

2.2 Definizione di Rapporto a favore condizionato

Il Rapporto a favore può essere usato anche in un contesto che preveda l'introduzione di una ulteriore ipotesi a supporto. In questo caso si parla di Rapporto a favore condizionato. In prima istanza si considerino i semplici eventi condizionati.

Dato l'evento A, le probabilità condizionate all'evento B sono:

$$P[A|B]=\frac{(P[B|A]P[A])}{(P[B])} \qquad P[\bar{A}|B]=\frac{(P[B|\bar{A}]P[\bar{A}])}{(P[B])}$$

Sostituendo nella formulazione del Rapporto a favore semplice le probabilità condizionate appena sopra indicate si ottiene il Rapporto a favore condizionato:

$$\frac{(P[A|B])}{(P[\bar{A}|B])}=\frac{(P[A])}{(P[\bar{A}])}\frac{(P[B|A])}{(P[B|\bar{A}])} \qquad (2.1)$$

Il nuovo valore del Rapporto a favore di A è il suo valore non condizionato moltiplicato per il quoziente tra la probabilità condizionata in relazione alla nuova prova B supposto che A sia vera e la probabilità condizionata sempre in relazione con la nuova prova B ma questa volta supposto che A sia falsa.

E' interessante notare dalla semplice analisi della formulazione come la quantità appena esposta aumenti o diminuisca a seconda delle ipotesi in gioco.

Infatti, parafrasando dal linguaggio matematico:

> *il Rapporto a favore di A condizionato a B aumenti in maniera marcata quando A è vera (si trova a numeratore) piuttosto che quando A è falsa* (2.a)

mentre il viceversa può essere letto come:

> *il Rapporto a favore diminuisce "più velocemente" quando A è falsa (si trova a denominatore) piuttosto che quando A è vera.* (2.b)

2.3 Definizione del Teorema di Bayes e relativa formulazione

Consideriamo gli insiemi A e B. Dalla teoria insiemistica si ha che l'insieme A può essere scomposto nel seguente modo:

$$A=(A\cap B)\cup(A\cap\bar{B})$$

Per ipotesi si suppone anche che:

$$(A\cap B)\cup(A\cap\bar{B})\neq 0$$

Passando alle probabilità degli insiemi A e B e applicando le formulazioni precedenti si ottiene:

$$P[A]=P[A\cap B]+P[A\cap\bar{B}] \qquad (2.2)$$

Dall'applicazione diretta del Teorema di Bayes sulla probabilità condizionata si ha che:

$$P[A|B]=\frac{(P[A\cap B])}{(P[B])} \qquad (2.3)$$

Con un semplice passaggio di algebra insiemistica la formula di Bayes si riscrive:

$$P[B|A]=\frac{(P[B\cap A])}{(P[A])}$$

E dividendo numeratore e denominatore per P[A] si ottiene:

$$P[B\cap A]=P[B|A]\times P[A] \qquad (2.4)$$

Considerando le precisazioni appena esposte, è possibile dare una formulazione alternativa le Teorema di Bayes per il calcolo della probabilità condizionata. Infatti partendo da:

$$P[A|B]=\frac{(P[A\cap B])}{(P[B])}$$

si applica a denominatore la 2.2 per l'evento B e si ottiene:

$$P[A|B]=\frac{(P[A\cap B])}{(P[B\cap A]+P[B\cap \bar{A}])}$$

Infine si sostituiscono sia numeratore e denominatore sfruttando la 2.4 commutando inoltre A con B a numeratore:

$$P[A|B]=\frac{(P[B|A]\times P[A])}{(P[B|A]\times P[A]+P[B|\bar{A}]\times P[\bar{A}])} \qquad (2.5)$$

Si consideri P[A] la valutazione di verosimiglianza dell'ipotesi a priori, cioè senza l'introduzione di una nuova prova a supporto B. Quest'ultima rafforza l'evento A se lo rende più probabile, in particolare se:

$$P[A|B]\geqslant P[A] \qquad (2.6)$$

Dalla formula 2.5 si desume osservando il denominatore che la precedente diseguaglianza è vera solo se:

$$P[B|A]\geqslant P[B|A]\times P[A]+P[B|\bar{A}]\times P[\bar{A}]$$

o, equivalentemente, quando:

$$P[B|A]\geqslant P[B|A]\times P[A]+P[B|\bar{A}]\times(1-P[A])$$

e dopo semplici passaggi algebrici si ottiene quindi:

$$P[B|A] \geqslant P[B|\bar{A}] \tag{2.7}$$

In altri termini, sempre cercando di tradurre dal linguaggio matematico:

> *una nuova prova si può considerare a supporto di una particolare ipotesi solo se il suo realizzarsi è più probabile quando l'ipotesi è vera piuttosto che quando l'ipotesi è falsa.* (2.c)

3.1 Applicazione del Rapporto a favore con PRF

Tramite la formulazione proposta poc'anzi del Rapporto a favore condizionato si cerca ora di indagare circa il funzionamento di un sistema di *Information Retrieval* che preveda l'implementazione di un sistema di *Relevance Feedback*. In questa fase, però, si prende in esame solo il caso di *Positive Relevance Feedback*.

Prima di procedere alla trattazione, è necessario puntualizzare come va intesa la formulazione adottata. Come è possibile ricavare dalla Teoria di Bayes, gli eventi considerati nella trattazione sono consequenziali; nello specifico, posto l'evento A e la sua relativa probabilità P[A], l'evento a supporto B è da ritenersi non contemporaneo ad A ma consequenziale. Infatti, mutuando la terminologia del Calcolo delle Probabilità, A è l'evento "a priori" mentre B è l'evento "a posteriori" [dall'Aglio, 1987].

Quindi riportando nel contesto della tesi in oggetto, se A è l'insieme dei documenti reperiti dal sistema IR e B è l'insieme dei *feedback* rilasciati dall'utente allora si suppone che i due eventi A e B non siano contemporanei come anche l'evento condizionato A|B e la relativa probabilità P[A|B].

Cerchiamo ora di contestualizzare il Rapporto a favore introducendo le quantità necessarie per il modello considerato:

- A è l'evento "insieme dei documenti reperiti dai sistema IR" al tempo T = 1
- P[A] è la "probabilità di ottenere l'insieme A di documenti reperiti" al tempo T = 1
- $\bar{A}$ è l'evento "insieme di documenti non reperiti dal sistema IR" al tempo T = 1
- P[$\bar{A}$] è la "probabilità di ottenere l'insieme $\bar{A}$ di documenti non reperiti" al tempo T = 1
- B è l'evento "insieme di documenti di A marcati con *feedback* positivo dall'utente" al tempo T = 2
- P[A|B] è la probabilità di ottenere lo stesso insieme di documenti A al tempo T = 3 a seguito del *feedback* positivo B

La formulazione da applicare è la seguente:

$$\frac{(P[A|B])}{(P[\bar{A}|B])} = \frac{(P[A])}{(P[\bar{A}])} \frac{(P[B|A])}{(P[B|\bar{A}])}$$

Come già esposto in precedenza, la funzione matematica del Rapporto a favore ha tale andamento:

> *il Rapporto a favore di A condizionato a B aumenti in maniera marcata quando A è vera (si trova a numeratore) piuttosto che quando A è falsa*

e il viceversa:

> *il Rapporto a favore diminuisce "più velocemente" quando A è falsa (si trova a denominatore) piuttosto che quando A è vera.*

Ora ai generici eventi si sostituiscono le quantità appena elencate per adattare il significato al contesto in oggetto. Il risultato è il seguente:

> *la probabilità di ottenere lo stesso insieme di documenti A a seguito del feedback positivo B aumenta se la probabilità A è più elevata della probabilità di* $\bar{A}$ (3.a)

e, come prima, il caso contrario:

> *la probabilità di ottenere lo stesso insieme di documenti A a seguito del feedback positivo B diminuisce se la probabilità A è inferiore della probabilità di* $\bar{A}$ (3.b)

Analizziamo quindi le affermazioni precedenti, 3.a e 3.b: un sistema di *Information Retrieval* ben funzionante, come da ipotesi poste in avvio di tesi, reperisce i documenti in maniera efficiente se si verifica la seguente condizione:

$$P[A] \geqslant P[\bar{A}] \tag{3.1}$$

cioè si sta affermando che la probabilità dei documenti reperiti deve essere maggiore della probabilità dei documenti non reperiti, pena il funzionamento non corretto dell'algoritmo.

L'utente, nel primo caso, marcando con un *feedback* positivo elementi di A sta confermando il risultato proposto dal sistema; quindi come si evince dalla teoria aumenta la probabilità che lo stesso insieme di documenti venga riproposto sostanzialmente inalterato anche al tempo T = 3 se come da ipotesi la probabilità di A è coerentemente maggiore di quella di $\bar{A}$. E ciò è quello che intuitivamente ci si aspetta dal sistema di *Relevance Feedback*.

Il secondo caso risulta essere altrettanto interessante. Posta come ipotesi che l'utente abbia accesso anche all'insieme di documenti non reperiti, la probabilità di ottenere lo stesso elenco

di documenti A diminuisce se non si verifica la relazione 3.1, che per ipotesi è data per vera. Quindi analizzando la formulazione l'unica relazione che può far diminuire il Rapporto è la seguente:

$$P[B|A] \leqslant P[B|\bar{A}]$$

Questo significa che l'eventuale possibilità da parte dell'utente di marcare con *feedback* positivo elementi dell'insieme dei documenti non reperiti avrebbe l'effetto di far diminuire la probabilità dell'insieme A|B cosicché da far mutare l'ordine dei documenti reperiti al tempo T = 3.

3.2 Applicazione del Teorema di Bayes con NRF

Rispetto al paragrafo precedente, nel quale si indagava circa il funzionamento del *feedback* positivo, per capire la funzionalità di quello negativo è utile cambiare tipo modello passando all'applicazione del Teorema di Bayes non in maniera diretta ma mediante la formulazione già esposta. E' bene precisare che le ipotesi specificate in precedenza devono essere date per certe: l'utente ha accesso all'universo dei documenti per esprimere il suo giudizio sulla rilevanza e il sistema di *Information Retrieval* funziona al massimo della sua efficacia.
Analogamente al paragrafo precedente è opportuno in questa fase della trattazione specificare quali siano gli eventi da sostituire alla formulazione per adattarla al contesto in oggetto:

- A è l'evento "insieme dei documenti reperiti dai sistema IR" al tempo T = 1
- P[A] è la "probabilità di ottenere l'insieme A di documenti reperiti" al tempo T = 1
- $\bar{A}$ è l'evento "insieme di documenti non reperiti dal sistema IR" al tempo T = 1
- P[$\bar{A}$] è la "probabilità di ottenere l'insieme $\bar{A}$ di documenti non reperiti" al tempo T = 1
- B è l'evento "insieme di documenti di A marcati con *feedback* negativo dall'utente" al tempo T = 2
- P[B|A] è la probabilità che l'utente marchi al tempo T = 2 documenti dell'insieme A
- P[A|B] è la probabilità di ottenere lo stesso insieme di documenti A al tempo T = 3 a seguito del *feedback* negativo B

La formulazione, da utilizzare è la seguente:

$$P[A|B]=\frac{(P[B|A]\times P[A])}{(P[B|A]\times P[A]+P[B|\bar{A}]\times P[\bar{A}])}$$

Ora se si considera l'evento "ottenere lo stesso insieme di documenti A a seguito del *feedback* B" questo equivale alla probabilità condizionata di ottenere nuovamente A a seguito dell'introduzione di una nuova prova [Ross, 2002]. E la prova B a sostegno rafforza l'evento a priori A se vale, come già anticipato, la relazione:

$$P[A|B]\geqslant P[A]$$

Ma tale relazione si verifica se vale anche la seguente (già osservato nella trattazione teorica):

$$P[B|A]\geqslant P[B|\bar{A}] \qquad (3.2)$$

Ora, rapportandolo al caso in oggetto della tesi, la relazione 3.2 può essere parafrasata nel seguente modo:

> *la probabilità che l'utente marchi con feedback negativo documenti da A è maggiore rispetto alla probabilità che l'utente marchi elementi dall'insieme dei documenti non reperiti* $\bar{A}$ (3.c)

Quest'ultima affermazione potrà sembrare scontata visto che in genere l'utente non ha accesso all'insieme di documenti non reperiti cosicché $P[B|\bar{A}]$ risulti sempre uguale a 0. Ma considerando le ipotesi da cui si è partiti, il risultato ottenuto è assai interessante. Infatti, supposto che l'utente abbia accesso ai documenti non reperiti, la relazione è comunque sempre

verificata visto che l'altra ipotesi da cui si è partiti è che il sistema di *Information Retrieval* funzioni in maniera performante, cioè vale la seguente relazione:

$$P[A] \geqslant P[\bar{A}]$$

Quindi affinché la probabilità di ottenere lo stesso elenco di documenti al tempo T = 3 diminuisca, cioè che il *feedback* negativo dell'utente abbia influenzato il successivo risultato presentato dal sistema IR, si deve verificare l'inverso della relazione 3.2 :

$$P[B|A] \leqslant P[B|\bar{A}]$$

In sostanza l'utente deve avere accesso ai documenti non reperiti e ne deve marcare in maniera sufficiente da annullare la "forza" della probabilità stabilita dal sistema P[A], soglia stabilita per il reperimento dei documenti al tempo T = 1.

La conseguenza di queste affermazione appare ora logica: sebbene l'utente marchi con *feedback* negativo elementi di A l'informazione ottenuta dai sui *feedback* non sarà mai sufficiente da controvertere l'elenco dei documenti ottenuti al tempo T = 3.

4.1 Valutazioni sull'efficacia del *Relevance Feedback*

Il modello costruito per valutare un sistema di *Information Retrieval* basato sull'applicazione del *Relevance Feedback* fa emergere quali siano i limiti e le lacune di tale sistema. In primo luogo è interessante fare una considerazione di per sé banale: maggiore è la precisione del sistema di reperimento dei documenti, maggiore sarà l'efficacia della successiva applicazione del *Relevance Feedback*. In altri termini se il sistema di *Information Retrieval* riesce a tradurre tutta l'informazione presente nella richiesta posta dall'utente allora l'applicazione del *Relevance Feedback* sull'insieme dei documenti così ottenuti darà i risultati auspicati.

In secondo luogo è di cruciale importanza assegnare un ruolo preciso al sottoinsieme di documenti non reperiti dal sistema. Infatti tale sottoinsieme può essere considerato un due modi distinti ma non complementari:

- informazione ritenuta a priori non rilevante e quindi non considerata
- informazione semplicemente non valutabile dall'utente

Com'è però emerso dalla trattazione in oggetto, l'accesso ai documenti non reperiti per la loro valutazione in termini di rilevanza risulta essere il collo di bottiglia del sistema: essi hanno un peso notevole sia in termini informativi, cioè di informazione apportata, sia in termini meramente numerici (è logico supporre che siano un sottoinsieme densamente popolato della collezione) e l'algoritmo usato nell'applicazione delle tecniche di *Relevance Feedback* non può prescindere da essi.

4.2 Possibile applicazione algoritmica

Dalla trattazione della tesi in oggetto si evince che per un funzionamento efficace del sistema di *Information Retrieval* basato sul *Relevance Feedback*, l'attenzione deve essere concentrata su due aspetti: una coerente traduzione dell'informazione apportata dall'utente ed estendere la valutazione circa la rilevanza anche ai documenti non reperiti dal sistema. Se il primo dei due problemi è appannaggio esclusivo dell'algoritmo di reperimento dei documenti usato dal sistema di *Information Retrieval*, è il secondo aspetto che intendo trattare in questo paragrafo, proponendo un algoritmo di mia concezione per poter estendere la valutazione circa la rilevanza anche ai documenti non reperiti.

Nel dettaglio, si parte da una collezione di documenti indicizzata mediante l'utilizzo di descrittori: per ogni documento si contano i descrittori presenti e se ne stila una classifica decrescente in base alla frequenza.

La prima fase prevede l'intervento dell'algoritmo del sistema di *Information Retrieval* che presenta all'utente, autore della richiesta, un sottoinsieme di documenti della collezione analizzata. In seconda fase l'utente interviene marcando come rilevante o non rilevante documenti dal sottoinsieme proposto. Nella fase successiva si analizzano i documenti marcati dall'utente e si indicizzano come il resto della collezione, cioè mediante descrittori. La frequenza di determinati descrittori in un documento marcato, quindi, verrà confrontata con il sottoinsieme dei documenti non reperiti dal sistema: se tale documento è stato marcato come rilevante, il sistema cerca quei descrittori tra l'insieme dei documenti non reperiti e li presenta all'utente ordinandoli in maniera decrescente secondo la frequenza degli stessi descrittori; viceversa se il documento viene marcato come non rilevante il sistema esclude dall'insieme dei documenti proposti tutti quei documenti che presentano descrittori con frequenza simile.

Il risultato che si intende ottenere con un sistema così configurato è la possibilità di espandere l'informazione introdotta dall'utente mediante la sua richiesta e sfruttare quest'ultima per valutare anche i documenti non reperiti.

A titolo esemplificativo, consideriamo la seguente collezione di documenti nella quale si identifica con d_i il documento i-esimo e con x_j il descrittore j-esimo:

	DESCRITTORI			
DOCUMENTI	x_1	x_2	x_3	x_4
d_1	3	1	0	2
d_2	1	1	1	5
d_3	2	1	1	1
d_4	3	4	1	0
d_5	0	3	3	6
d_6	0	1	3	3

La richiesta espressa dall'utente contiene il descrittore x_1.

L'algoritmo di reperimento dei documenti presenta in maniera ordinata la seguente lista di documenti:

$$d_4\ [3, 4, 1, 0]$$

$$d_1\ [3, 1, 0, 2]$$

$$d_3\ [2, 1, 1, 1]$$

$$d_2\ [1, 1, 1, 5]$$

In questa fase interviene l'utente che mediante il suo giudizio circa la rilevanza marca positivamente i documenti d_1 e d_2 mentre marca negativamente il documento d_4.

Confrontando ora i descrittori relativi ai documenti marcati si nota come risulta essere rilevante in termine di informazione il descrittore x_4 presente invece nei documenti non reperiti d_5 e d_6.

I documenti marcati come non rilevanti, invece, pongono l'attenzione sul descrittore x_1: sebbene sia presente nella richiesta l'utente ritiene che esso non costituisca un'informazione necessaria ai fini della pertinenze del documento.

Confrontando quindi la frequenza dei descrittori nei documenti non reperiti il sistema propone all'utente un nuovo insieme tenendo conto del peso in termini di informazione dato al descrittore x_4.

Il nuovo elenco di documenti proposto sarà il seguente:

$$d_1 [\ 3, 1, 0, 2]$$

$$d_5 [\ 0, 3, 3, 6]$$

$$d_2 [\ 0, 1, 3, 3]$$

Bibliografia

Articoli Consultati:

M. D. Dunlop, 1997. The Effect of Accessing Nonmatching Documents on Relevance Feedback, University of Glasgow, Scotland.

I. Ruthven and M. Lalmas, 2005. A survey on the use of Relevance Feedback for information access system. Knowledge Engineering Review, pp. 95-145. University of Strathclyde, Glasgow

E. Terra and R. Warren, 2005. Poison Pills: Harmfull Relevant Documents in FeedBack, In Proceedings of the 14th Conference on Information and Knowledge Managemen, Bremen, Germany

B. J. Jansen, A. Spink, & Saracevic, T. 1999. The use of relevance feedback on the web: Implications for web IR systém design. 1999 World Conference on the WWW and Internet, Honolulu, Hawaii.

M. Melucci, 2006. Notes on Negative Relevance Feedback in the Probabilistic Model, University of Padova, Padova

S.E. Robertson, 1977. The probability Ranking Priciple in IR. Journal of Documentation pp. 294-304 University College of London, London

M. E. Maron and J. L. Kuhns, 1960 On relevance, probabilistic indexing and information retrieval. Journal of the Association for Computing Machinery, University College of London, London

V. Camagni, 2006. I motori di ricerca e il web semantico. Pc Professionale n°187, pp. 222-223 Arnoldo Mondadori editore, Milano

Testi Consultati:

S. Ross, 2002. A first course in probability, Prentice-Hall edition. New Jeresey

G. Dall'Aglio, 1987. Calcolo delle probabilità, edizione Zanichelli. Bologna

A. Azzalini e B. Scarpa, 2004. Analisi dei dati e Data Mining, edizione Springer. Milano

M. Agosti e M. Melucci. Reperimento dell'informazione. Libreria Progetto Editrice, Padova, 2007.

Fonti Elettroniche Consultate:

http://books.google.it/ *Dicembre 2006 / Gennaio 2007*

http://it.wikipedia.org/ *Dicembre 2006 / Gennaio 2007*

http://portal.acm.org/ *Dicembre 2006 / Gennaio 2007*

Un sentito grazie al Professor
Massimo Melucci
che mi ha seguito nella redazione
della presente tesi

Vicenza, 26 Febbraio 2007
Grifò Luca Giuseppe