

# Le differenti *relevance* in *information retrieval*: una classificazione

Stefano Mizzaro

Dipartimento di Matematica e Informatica, Università di Udine  
V. delle Scienze, 206 – Loc. Rizzi – 33100 Udine – Italy  
e-mail: mizzaro@dimi.uniud.it

**Sommario:** Il concetto di *relevance*, di fondamentale importanza per il settore dell'*information retrieval*, non è interamente compreso, a mio avviso a causa di una confusione di fondo. Questo lavoro ha l'obiettivo di chiarire tale concetto: viene presentato un *framework* formale che permette di definire le varie *relevance* esistenti e le relazioni fra di esse. È inoltre illustrato come tale *framework* possa essere utilizzato per classificare le definizioni di *relevance* proposte da altri autori e quali conseguenze abbia sulla realizzazione di sistemi di *information retrieval*.

## 1 Introduzione

Nel settore dell'*information retrieval* (IR) e nelle attività di progettazione, realizzazione e valutazione di *sistemi di information retrieval* (IRS, da *Information Retrieval System*), il concetto di *relevance* (in prima approssimazione, quanto un documento sia effettivamente adatto a soddisfare le necessità dell'utente) è un concetto fondamentale, ma ancora oggi dibattuto e ben lontano dall'essere interamente compreso. A supporto di queste affermazioni riporto qui due (fra le molte possibili) citazioni, tratte rispettivamente dall'articolo introduttivo di un numero monografico sulla *relevance* del *Journal of American Society for Information Science* e da uno degli ultimi articoli di rassegna su tale concetto:

... *relevance*, acknowledged as the most fundamental and much debated concern for information science... Early on, information scientists recognized that the concept of *relevance* was integral to information system design, development and evaluation. However, there was little agreement as to the exact nature of *relevance* and even less that it could be operationalized in systems or for the evaluation of systems. ... this lack of agreement continues to an extent at the present [9], p. 124

Since information science first began to coalesce into a distinct discipline in the forties and early fifties, *relevance* has been identified as its fundamental and central concept ... an enormous body of information science literature is based on work that uses *relevance*, without thoroughly understand what it means... without an understanding of what *relevance* means to users, it seems difficult to imagine how a system can retrieve relevant information for users [25], p. 755-756.

L'*attinenza*<sup>1</sup> ha ormai quasi 40 anni, risalendo alla International Conference for Scientific Information del 1958. Non si può dire che in questi 40 anni sia stata trascurata: soprattutto negli anni '60 e '70 sono infatti stati pubblicati parecchi lavori, sia di carattere sperimentale che teorico (ad esempio, in ordine cronologico: [19, 6, 20, 17, 8, 5]). Qui evito di presentare una rassegna completa della letteratura sull'*attinenza*, per la quale si possono consultare [25, 23] o anche [1, 14], ma un rapido accenno è utile per chiarire gli obiettivi di questo lavoro.

I lavori sull'*attinenza* hanno evidenziato la difficoltà a definire tale concetto e lo spostamento dell'attenzione da un punto di vista orientato al sistema ad una visione più orientata all'utente. Il grande numero di studi effettuati ha portato all'adozione, forse non sempre motivata, di numerosi termini differenti per indicare concetti simili: si è parlato, oltre che di *relevance*, di *utility*, *usefulness*, *pertinence*, *topicality*, *user satisfaction* e altri ancora. Negli anni '70, Saracevic ha cercato di riunire gli sforzi compiuti fino ad allora, ottenendo il risultato seguente:

Relevance is the (A) of a (B) existing between a (C) and a (D) as determined by an (E).

[Where:]

A: Measure, estimate, judgment, ...

B: Utility, matching, satisfaction, ...

C: Document, document representation, information provided, ...

D: Question, question representation, information need, ...

E: Requester, intermediary, expert, ...

[23], p. 328.

Questa definizione (o, più propriamente, algoritmo per generare tutte le definizioni) di *attinenza* è quanto di meglio si abbia oggi a disposizione, almeno per quanto ne so. Secondo me, l'ordine che Saracevic sembra porre nel settore è però solo apparente: la sua definizione ha sì il pregio di evidenziare il fatto che esistono più *attinenze*, ma contiene il seme della confusione che si può riscontrare ancora oggi, in quanto Saracevic mescola l'*attinenza* di per sé (punti C e D) con la *misura* dell'*attinenza* (punti A ed E), non individua in modo chiaro il

<sup>1</sup> Si osservi che il termine inglese '*relevance*' va tradotto in italiano con '*pertinenza*' o '*attinenza*' e non con '*rilevanza*' (che, in realtà, significa '*importanza*'), come comunemente si fa: basta consultare un qualsiasi vocabolario italiano-inglese per convincersene! La mia proposta, adottata in questo lavoro, è di usare il termine '*attinenza*', in quanto '*pertinenza*' ricorda '*pertinence*', che è stato usato con un significato particolare (si veda ad esempio [23]).

tipo di accoppiamento, di relazione fra le entità dei punti C e D e non considera le relazioni fra le varie attinenze.

Riassumendo, si può quindi dire che l'attinenza non è ben compresa, e che ciò è senz'altro in parte dovuto al fatto che è un concetto complesso, la cui definizione in termini formali è problematica, se non impossibile [18]. Ma, a mio avviso, le difficoltà derivano anche da una confusione di fondo, dovuta in buona parte all'esistenza di varie attinenze, alla mancanza della comprensione delle relazioni fra di esse e all'uso di una terminologia non sempre coerente; quindi, una presentazione organica e analitica della situazione può senz'altro aiutare a chiarire il quadro. L'obiettivo di questo lavoro è delineare un *framework* formale, che permetta di individuare le varie attinenze, riunirle in un'unica definizione, classificarle e indicare quali relazioni vi siano fra di esse.

Questo lavoro è strutturato nel modo seguente. Essendo l'attinenza strettamente connessa con la situazione in cui si trova l'utente di un IRS che necessita di informazioni, appunto, attinenti, i paragrafi 2 e 3 affrontano e descrivono alcune problematiche dell'interazione utente-IRS e gettano le basi per il paragrafo 4, che presenta una definizione formale delle varie attinenze e delle relazioni fra di esse. La validità e l'utilità di questo lavoro vengono poi saggiate nei paragrafi 5, in cui classifico alcune definizioni di attinenza proposte in passato da vari autori, e 6, in cui illustro le conseguenze che lo scenario del paragrafo 4 ha sulla costruzione di IRS. Nel paragrafo 7, infine, accenno ai possibili (e numerosi) sviluppi futuri.

## 2 Rappresentazioni del problema dell'utente

In questo paragrafo descrivo da un punto di vista concettuale la situazione di un utente che si reca in un centro di documentazione e interagisce con un intermediario per ottenere dei servizi da un IRS. La presentazione proposta qui è una mia elaborazione di spunti ricavati da [13, 3, 12]; inoltre, qui ipotizzo la presenza di un intermediario umano, ma i punti essenziali della trattazione restano pressoché inalterati nel caso in cui l'utente interagisca direttamente con un IRS per utenti finali.

L'interazione fra utente e IRS è stata, ed è tuttora, oggetto di studio; i punti da tenere in considerazione in questa sede sono schematizzati in figura 1 ed elencati qui di seguito:

- l'utente che necessita dei servizi di un IRS ha un *bisogno informativo (BI)*, originato da un *problema* (o da uno *scopo*) che l'utente deve o vuole risolvere (o raggiungere). Il *BI* è inconscio, è implicito nella mente dell'utente e può a mio avviso essere considerato la percezione che l'utente ha del proprio *problema/scopo* (in seguito, per brevità, *P*);
- l'operazione di passaggio da *P* a *BI*, che denomi *intrinsecazione*, non è semplice come può sembrare di primo acchito: l'utente si viene spesso a trovare in una condizione in cui vuole colmare una propria lacuna conoscitiva su un settore del quale non ha una buona conoscenza (altrimenti non avrebbe tale lacuna!), quindi spesso egli non comprende esattamente ciò di cui ha bisogno, e di conseguenza l'*intrinsecazione* del *P* avviene in modo non corretto;
- il *BI* può essere esplicitato (operazione di *esplicitazione*), ottenendo come risultato la *richiesta (R)*. Nei centri di documentazione, tale operazione è molto frequente e avviene ogni volta che l'utente spiega all'intermediario umano 'di che cosa ha bisogno' o 'che documenti vuole'. In questi casi il linguaggio adottato per la rappresentazione della *R* è il linguaggio naturale, scritto o orale, ma si potrebbe pensare anche a linguaggi differenti, ad esempio un linguaggio grafico, o il linguaggio a gesti dei sordomuti, e così via;
- così come l'*intrinsecazione*, anche l'operazione di *esplicitazione* non è semplice, in quanto intervengono i fenomeni noti col nome di *label effect* [13] e *vocabulary problem* [10];
- la *R* non è adatta ad essere compresa da un IRS tradizionale e deve essere tradotta in un linguaggio (formale) di interrogazione: l'operazione è denominata *formalizzazione* e il risultato *query (Q)*. Tale

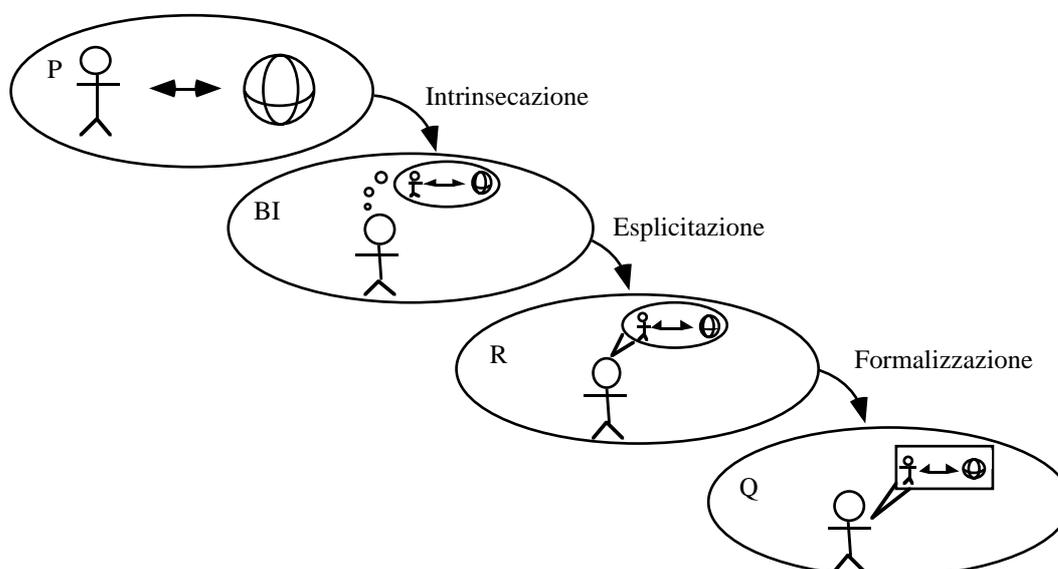


Fig. 1. Rappresentazioni del P dell'utente.

operazione è solitamente effettuata dall'intermediario, che conosce il linguaggio di interrogazione e anche essa non è semplice in quanto il linguaggio da utilizzare è, rispetto al linguaggio in cui è espressa la  $R$ , meno comprensibile dall'utente e meno espressivo.

In seguito indicherò con *rappresentazioni del problema* ( $RP$ ) l'insieme  $\{P, BI, R, Q\}$ .

La descrizione dello scenario presentata finora è incompleta, in quanto statica; per avvicinarsi alla realtà occorre considerare i punti seguenti:

- (i) talvolta, se non spesso, la prima  $Q$  non porta ad ottenere i risultati (cioè i documenti reperiti) desiderati ed è necessario modificarla (e questa affermazione risulta senza dubbio sensata alla luce della problematicità delle operazioni di intrinsecazione, esplicitazione e formalizzazione illustrata poc'anzi). Tale modifica è solitamente effettuata dall'intermediario, con l'eventuale coinvolgimento dell'utente: basandosi sui documenti reperiti dalla prima  $Q$ , ed eventualmente utilizzando thesauri e tecniche morfologiche, l'intermediario *modifica* la  $Q$ ;
- (ii) durante l'interazione fra utente, intermediario e sistema, anche la  $R$  dell'utente può risultare modificata (operazione di *ri-espressione* della  $R$ ). Lo è (quantomeno in modo implicito) ogniqualvolta utente e intermediario si rendono conto che, ad esempio, bisogna aggiungere un termine sinonimo nella  $Q$ , ma può esserlo, in modo esplicito, se l'utente si accorge di non aver ben esplicitato il  $BI$ , lo fa presente all'intermediario e gli comunica la nuova  $R$ ;
- (iii) infine, durante l'interazione fra utente, intermediario e sistema, l'utente può rendersi conto di aver male percepito il proprio  $P$ , e la visione di documenti (o surrogati) e termini relativi al  $P$  possono indubbiamente facilitare questo processo. Ne risulta quindi che anche il  $BI$  può mutare durante l'interazione (operazione di *migliore comprensione* del  $BI$ ).

Queste considerazioni portano a modificare il quadro di figura 1, rendendolo più dinamico; la situazione ottenuta è schematizzata in figura 2, in cui è illustrato l'*albero dinamico delle trasformazioni delle RP*, che generalizza la catena di trasformazioni di figura 1 introducendovi la variabilità nel tempo.

In figura 2 si possono individuare 4 livelli, rappresentati graficamente dalle ellissi, denominati *livello P*, *livello BI*, *livello R* e *livello Q*. Dal  $P$  dell'utente,  $p_0$ , per mezzo di operazioni di intrinsecazione, esplicitazione, e formalizzazione (le 3 frecce scure più a sinistra in figura ed etichettate rispettivamente con  $i$ ,  $e$ ,  $f$ ), si arriva ad un  $BI$  iniziale  $b_0$ , ad una  $R$  iniziale  $r_0$  e ad una  $Q$  iniziale  $q_0$  (e questa è la situazione di figura 1). Come descritto nel punto (i) precedente, tale  $Q$  può venire modificata ('puntini' al livello  $Q$  in figura), raffinando così la rappresentazione del  $P$ . Durante il processo di modifica della  $Q$ , come indicato nel punto (ii) precedente, l'utente può ri-esprimere la  $R$  (freccia chiara etichettata con  $f^{-1}$  in figura dal livello  $Q$  al livello  $R$  e freccia da  $r_0$  a  $r_1$ ), dopodiché vengono ripetute le operazioni di formalizzazione e modifica della  $Q$ . L'utente può anche (punto (iii)) comprendere meglio il proprio  $BI$  (freccia chiara etichettata con  $e^{-1}$  dal livello  $R$  al livello  $BI$  e freccia al livello  $BI$ ), e poi ripetere il procedimento.

Riassumendo: col passare del tempo, per mezzo di intrinsecazione, esplicitazione, formalizzazione, modifiche della  $Q$ , ri-espressioni della  $R$  e maggiore comprensione del  $BI$ , si passa da una  $Q$  iniziale  $q_0$  che non rappresenta correttamente il  $P$  ad una  $Q$  finale  $q_n$  che lo rappresenta in modo più corretto. In base all'albero dinamico di figura 2 è possibile definire un insieme discreto e totalmente ordinato di istanti di tempo:

$$T = \{t(p_0), t(b_0), t(r_0), t(q_0), t(q_1), t(r_1), t(q_2), \dots, t(b_p), \dots, t(r_m), \dots, t(q_n), t(f)\}.$$

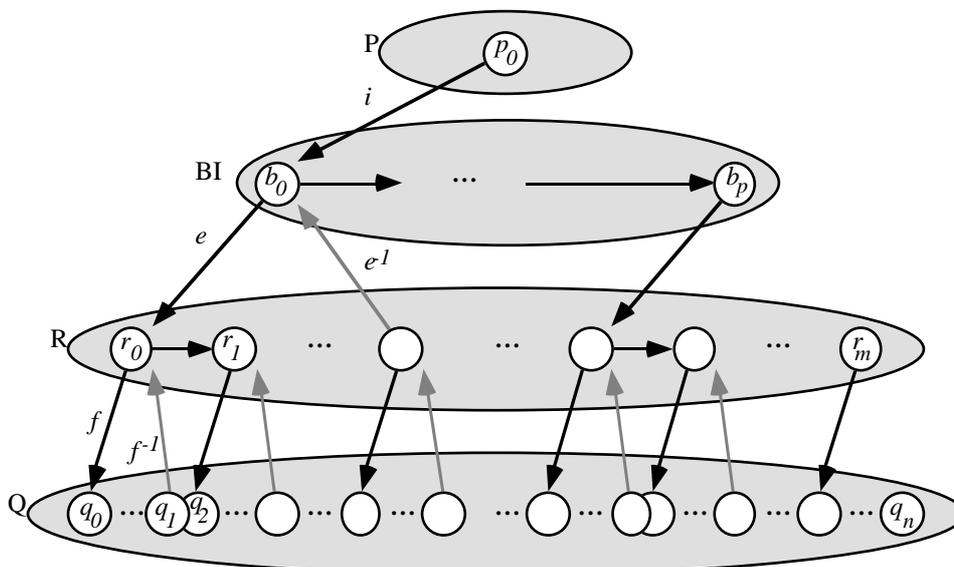


Fig. 2. Albero dinamico delle trasformazioni delle  $RP$ .

Tale insieme contiene gli istanti di tempo corrispondenti a tutte le  $RP$  di figura 2, più  $t(f)$  che rappresenta l'istante di tempo in cui l'utente ha risolto il suo  $P$ . L'ordinamento su  $T$ , che è quello che si ha leggendo gli elementi da sinistra a destra, è indotto dal normale ordinamento temporale e sarà indicato con  $\langle_T$ . L'utilità di questo insieme sarà chiara più avanti.

Si osservi che l'utente può rendersi conto della sua errata percezione del  $P$  (punto (iii)) anche quando, *off-line*, usa i documenti reperiti: in tal caso, egli può recarsi nuovamente al centro di documentazione e ripetere le operazioni di esplicitazione e formalizzazione, con tutto quel che segue. L'uso dei documenti può però portare anche ad un nuovo  $P$ , correlato con quello originale, ma differente. Lo schema di figura 2 comprende la prima di queste due possibilità (è il caso del punto (iii) illustrato poc'anzi), mentre per tenere in conto la seconda possibilità bisogna pensare ad un nuovo albero.

Così come vi sono varie  $RP$ , anche i risultati ottenuti da un IRS possono essere classificati: si possono individuare i *surrogati* ( $S$ ), i *documenti* ( $D$ ) e l'*informazione* ( $I$ ). In seguito, indicherò con *risorse informative* ( $RI$ ) l'insieme  $\{S, D, I\}$ .

L'utilità degli argomenti illustrati qui sarà chiara fra poco; prima è però necessario introdurre altri concetti, compito svolto nel prossimo paragrafo.

### 3 Argomento, compito e contesto

Gli elementi degli insiemi  $RP$  ( $P$ ,  $BI$ ,  $R$  e  $Q$ ) e  $RI$  ( $I$ ,  $D$  e  $S$ ) possono a mio avviso essere scomposti in tre componenti: *argomento*, *compito* e *contesto*. In breve, la componente *argomento* (o *topic*) riguarda l'area, i concetti che devono essere trattati dai documenti che l'utente vuole reperire; il *compito* (o *task*) è relativo al tipo di attività che l'utente deve svolgere con i documenti reperiti; il *contesto* comprende gli aspetti non contenuti nelle altre due componenti, quali ad esempio la conoscenza che l'utente ha del settore o del compito che deve svolgere.<sup>2</sup> Per chiarire la scomposizione in argomento, compito e contesto, presento due esempi di  $P$  che si possono avere in situazioni reali.

**Esempio 1.** Si pensi ad uno studente universitario col  $P$  di fare una tesina sul concetto di attinenza in IR.<sup>3</sup> In tal caso, l'argomento è 'attinenza e IR', il compito è 'fare una tesina' e il contesto è rappresentato dalle conoscenze che lo studente ha sull'argomento, dalla sua concezione di tesina e da altri fattori.

Le 3 componenti possono essere individuate nella  $R$  e nella  $Q$ : l'individuazione nella richiesta è già stata effettuata (la frase 'Si pensi ad uno studente universitario col  $P$  di fare una tesina sul concetto di attinenza in IR' è in realtà una possibile  $R$ , si veda la nota 3, e in essa ho appena individuato le 3 componenti); per quanto riguarda la  $Q$ , è sensato ipotizzare (nel caso di una  $Q$  booleana) che vi sia, ad esempio, una faccetta negativa che scarti i documenti troppo specialistici o difficili, o quelli che già si conoscono. In realtà le componenti del  $P$  non devono necessariamente finire interamente nella  $Q$ : alcuni aspetti possono ad esempio influenzare il numero di documenti di cui l'utente ha bisogno, altri possono essere usati per effettuare un ranking dei documenti reperiti, e così via.

Le 3 componenti si riflettono anche sulle caratteristiche di documenti e surrogati: i documenti reperiti dovranno essere attinenti non solo per l'argomento ma anche per il compito e per il contesto. Infatti, l'argomento eliminerà tutti i documenti della banca dati che non parlino di 'attinenza e IR', il compito eliminerà quelli che non siano adatti a scrivere una tesina (magari perché troppo specialistici o difficili) e il contesto eliminerà ad esempio i documenti già noti allo studente, o scritti in una lingua non conosciuta, e così via.

L'individuazione delle 3 componenti in  $P$ ,  $BI$  e  $I$  è ovviamente più problematica, in quanto entità molto meno "osservabili" rispetto a  $R$ ,  $Q$ ,  $D$  e  $S$ . Comunque, si può ipotizzare una proiezione sul  $BI$  e  $P$  da  $R$  e sull' $I$  da  $D$ .

**Esempio 2.** Un progettista software ha il  $P$  di reperire informazioni, e redigere una relazione, su applicazioni di intelligenza artificiale, già sviluppate ed in uso, relative alla diagnosi di guasti per pompe di calore in reattori nucleari a fissione.

In questo caso, l'argomento è 'diagnosi di guasti per pompe di calore in reattori nucleari a fissione' e il compito è 'scrivere una relazione'. Si noti come il contesto vari a seconda di *chi* ha il  $P$ : è ben diverso il caso di un progettista software rispetto ad un esperto di fisica nucleare, a causa delle differenti conoscenze possedute.

Indicherò con  $A$  l'argomento, con  $C$  il compito e con  $c$  il contesto. L'insieme delle componenti sarà

$$C_o = 2^{\{A, C, c\}} - \emptyset,$$

in cui  $2^x$  sta per l'insieme delle parti di  $x$  e  $\emptyset$  è l'insieme vuoto.

### 4 L'insieme parzialmente ordinato delle attinenze

Esistono varie attinenze, come già affermato nel paragrafo 1. Lo scenario dinamico dell'interazione utente-IRS descritto nel paragrafo 2 e l'individuazione delle componenti compito, argomento e contesto del paragrafo 3 possono essere combinati per ottenere una classificazione unitaria delle varie attinenze e per individuare le relazioni fra di esse.

<sup>2</sup> La componente contesto può probabilmente essere ulteriormente scomposta a sua volta e individuata in modo più preciso, ma la scomposizione in argomento, compito e contesto proposta qui è comunque sufficiente per ragionare sul concetto di attinenza, come risulterà chiaro nel prossimo paragrafo.

<sup>3</sup> Ovviamente, quella nel testo è una rappresentazione scritta del  $P$ , non il  $P$  stesso!

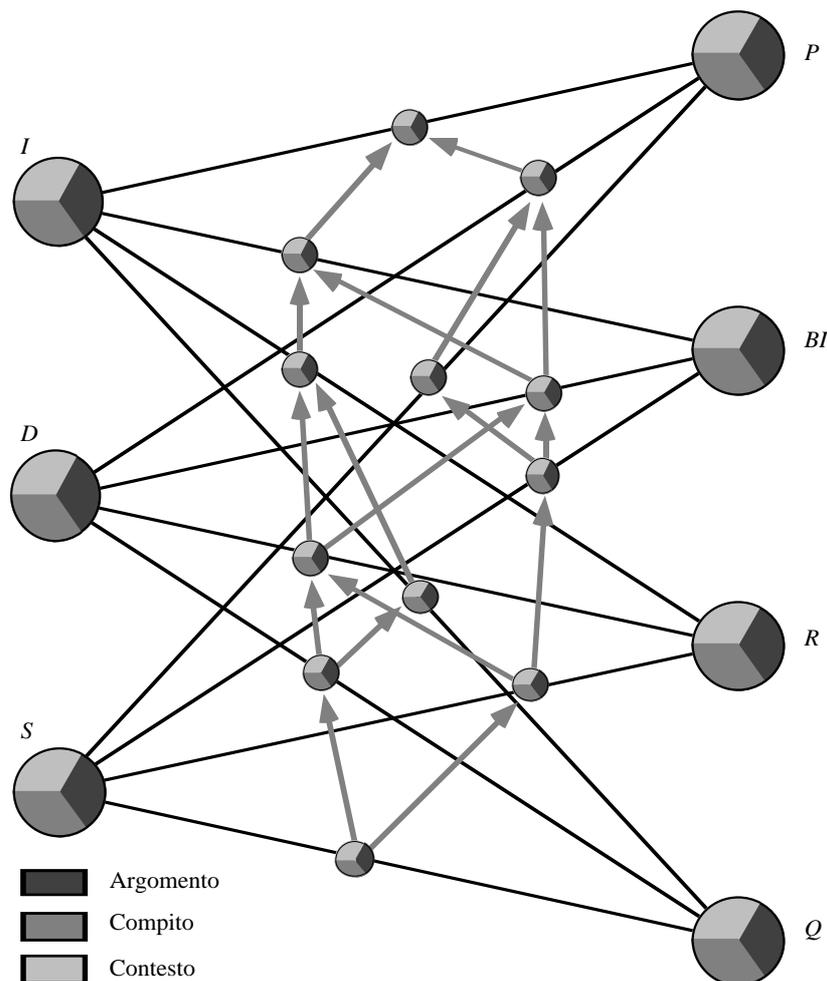
L'attinenza è una relazione fra 2 entità, una delle quali è un elemento dell'insieme  $RP$  ( $P$ ,  $BI$ ,  $R$  o  $Q$ ) e l'altra è una  $RI$  ( $I$ ,  $D$  o  $S$ ). Essendo varie le rappresentazioni del  $P$  e dell'informazione fornita dall'IRS, combinandole si ottengono attinenze differenti: l'attinenza di un documento rispetto ad una  $Q$  e l'attinenza dell'informazione fornita dal documento rispetto al  $P$  dell'utente sono due attinenze diverse. Inoltre, anche la scomposizione in argomento, compito e contesto individua attinenze differenti: un conto è, ad esempio, l'attinenza rispetto all'argomento, un altro l'attinenza rispetto al compito. Infine, l'attinenza è in funzione del tempo: in due istanti di tempo diversi vi possono essere due attinenze differenti.

La situazione è schematizzata in figura 3: a sinistra vi sono le  $RI$ , a destra le  $RP$  e ogni linea che collega due di questi oggetti rappresenta una particolare attinenza (graficamente evidenziata da un cerchio sulla linea). Tale attinenza può inoltre essere relativa ad una qualsiasi combinazione delle tre componenti (argomento, compito o contesto) e ciò è rappresentato graficamente dai differenti livelli di grigio impiegati per le 3 componenti. Per semplificare la figura, l'evoluzione nel tempo delle  $RP$  (si veda la figura 2) non è rappresentata graficamente. Le varie attinenze possono essere ordinate con un ordinamento parziale (che chiamerò *ordine delle attinenze*), che indica quale sia l'attinenza importante per l'utente ed è rappresentato graficamente dagli archi di colore grigio: si ottiene così l'*insieme parzialmente ordinato delle attinenze*.

A questo punto, presento la definizione formale delle attinenze e dell'ordine delle attinenze. Definisco dapprima l'insieme delle *attinenze statiche* (indicato con  $AS$ ), come il prodotto cartesiano dei tre insiemi  $RI$ ,  $RP$  e  $Co$ :

$$AS = RI \times RP \times Co;$$

una particolare attinenza statica è quindi una tripla appartenente a tale insieme. Per individuare univocamente l'attinenza statica a cui mi riferisco di volta in volta, indicherò con  $as(x,y,z)$  l'attinenza statica di  $x$  rispetto a  $y$  per quanto concerne le componenti  $z$ . Ad esempio, l'estremo inferiore del reticolo di figura 3 è indicato con  $as(S,Q,\{A,C,c\})$ : l'attinenza statica di un surrogato  $S$  rispetto alla query  $Q$  su tutte e tre le componenti;  $as(S,Q,\{A\})$  è la *topicality* tradizionale, mentre  $as(I,P,\{A,C,c\})$  è l'attinenza che interessa all'utente. Userò il simbolo '?' per indicare un valore indefinito; così  $as(?R,\{A\})$  indicherà un'attinenza statica fra una risorsa informativa non nota e la  $R$  per quanto concerne la componente argomento. Il simbolo '\_' sarà utilizzato per indicare un valore qualsiasi: così, ad esempio,  $as(\_,\_,\{A,C,c\})$  indica l'insieme delle attinenze di figura 3.



**Fig. 3.** L'insieme parzialmente ordinato delle attinenze.

Definendo su  $RI, RP$  e  $Co$  i tre rispettivi ordini (i primi due totali, il terzo parziale)  $<_{RI}, <_{RP}$  e  $<_{Co}$  nel modo seguente

$$S <_{RI} D <_{RI} I; Q <_{RP} R <_{RP} BI <_{RP} P; <_{Co} = \subset$$

(in cui  $\subset$  è l'inclusione stretta di insiemi), si può poi definire l'ordine  $<_{AS}$ :

$$as(x_1, x_2, x_3) <_{AS} as(y_1, y_2, y_3) \text{ sse } \exists i(x_i < y_i) \text{ e } \neg \exists j(y_j < x_j)$$

(dove con  $<$  si intende l'ordinamento opportuno fra  $<_{RI}, <_{RP}$  e  $<_{Co}$ ), che sarà denominato *ordine parziale delle attinenze statiche*. Si osservi che tale ordine è indicato solo parzialmente in figura 3, in quanto gli archi grigi non rappresentano  $<_{Co}$ .

Infine, a partire dall'attinenza statica e dall'insieme degli istanti di tempo  $T$  definito nel paragrafo 2, si può definire l'insieme delle *attinenze dinamiche* (in seguito, semplicemente 'attinenze') come  $A = AS \times T$ . Gli elementi di tale insieme verranno indicati con  $a(x, t)$ , in cui  $x$  è un'attinenza statica, e i simboli '?' e '\_' continueranno ad avere lo stesso significato. Su  $A$  si può definire l'ordinamento parziale delle attinenze  $<_A$  come:

$$a(x, t_1) <_A a(y, t_2) \text{ sse } x <_{AS} y \text{ e } t_1 <_T t_2.$$

La situazione dovrebbe essere ora chiara: vi sono varie attinenze, gli elementi di  $A$ ; l'ordine delle attinenze  $<_A$  permette di confrontarle fra loro; per avvicinarsi ad  $a(as(I, P, \{A, C, c\}), t(f))$ , che è quella che interessa all'utente, si deve risalire in  $A$  lungo l'ordinamento  $<_A$ ; in genere, risalendo lungo  $<_A$ , si incontrano sì attinenze più interessanti per l'utente, ma anche più difficili da trattare.

Le considerazioni esposte finora in questo paragrafo sono a mio avviso importanti per comprendere correttamente il concetto di attinenza; nei prossimi due paragrafi illustro come possano essere utilizzate per classificare i vari studi sull'attinenza e quali conseguenze esse comportino sulla realizzazione di IRS per utenti finali.

## 5 Definizioni di attinenza: una classificazione

In questo paragrafo illustro come l'analisi del paragrafo precedente possa essere usata come griglia per classificare le definizioni di attinenza presentate in letteratura.

Cuadra e Katter [6] intendono l'attinenza come  $a(as(D, R, \{A\}), ?)$ :

Relevance is the correspondence in context between an *information requirement statement* and an *article*; i. e., the extent to which the article covers *material* that is appropriate to the requirement statement

[6], citato in [23], il corsivo è mio.

Ancora più in basso nell'ordine delle attinenze si colloca  $a(as(S, R, \{A\}), ?)$ , l'attinenza usata nei classici test di valutazione di IRS, quali Cranfield [4] e TREC [11].

Parecchi studiosi hanno cercato di avvicinarsi ad  $a(as(I, P, \{A, C, c\}), ?)$ . È il caso, ad esempio, sia di Sandore [22], che ha cercato di valutare sperimentalmente  $a(as(I, P, \{A, C, c\}), t(f))$ :

The actual test of whether an online search printout truly addressed the user's search query takes place after the user leaves the search session and begins to follow up on the information retrieved in the search [22], p. 34,

sia di Cooper [5], che ha distinto fra  $a(as(?, ?, \{A\}), ?)$  e  $a(as(?, ?, \{A, C, c\}), ?)$ .

... relevance has to do with *aboutness* (or pertinence or *topic-relatedness*) and is ultimately defined in terms of logical implication, whereas *utility* is a catch-all concept involving not only topic relatedness but also quality, novelty, importance, credibility and many other things [5], citato in [23], p. 334.

Infine, altri studiosi hanno comparato due differenti attinenze. Saracevic et al. [24] hanno confrontato  $a(as(S, R, \{A\}), ?)$  con  $a(as(S, P, \{A, C, c\}), ?)$  (da loro denominate rispettivamente 'relevance' e 'utility'). E Regazzi [21] ha studiato le correlazioni fra  $a(as(D, R, \{A\}), ?)$  e  $a(as(D, R, \{A, C, c\}), ?)$ , trovandole simili:

Half of the judges were asked to rate the documents on how relevant the documents were to the search *topic*; the other half rated the same documents on the basis of the document's perceived *utility* for the individual judge. ...

The study finds that there is *no operational difference* between the relevance-theoretic and the utility-theoretic model of evaluation [21], p. 235, il corsivo è mio.

E anche nella valutazione sperimentale del sistema FIRE (descritto in [7, 16]), in corso di svolgimento, si usano due differenti attinenze, ossia  $a(as(S, R, \{A\}), ?)$  e  $a(as(S, R, \{A, C\}), ?)$ .

## 6 La realizzazione di IRS per utenti finali

In questo paragrafo illustro quali conseguenze abbia lo scenario del paragrafo 4 sull'attività di realizzazione di IRS per utenti finali.

Per costruire sistemi che cerchino di massimizzare l'attinenza che interessa all'utente,  $a(as(I, P, \{A, C, c\}), t(f))$ , si deve risalire lungo l'ordinamento  $<_T$ . Per fare ciò, si può procedere in quattro direzioni indipendenti: o risalire lungo  $<_{RI}$  e quindi, ad esempio, passare da banche dati bibliografiche a banche dati *full-text*, o risalire lungo  $<_{RP}$

e quindi passare ad esempio da sistemi che accettano una  $Q$  in linguaggio formale a IRS che siano in grado di trattare una  $R$  in linguaggio naturale, o risalire lungo  $\langle C_o$  e quindi ad esempio costruire IRS che non lavorino unicamente sull'aspetto argomento ma trattino anche le altre componenti del  $P$ , o infine, risalire lungo  $\langle T$  e quindi preferire un IRS che consenta e stimoli un'effettiva interazione fra utente e sistema, e che porti ad un'iterazione dei processi di intrinsecazione, esplicitazione, formalizzazione, modifica della  $Q$ , ri-espressione della  $R$  e maggiore comprensione del  $BI$ , rispetto ad un IRS che, per dirla con Bates "takes a request in natural language, goes off and searches the information store, and returns to the user the ideal best retrieved set of documents or information" ([2], p. 575). A mio avviso, la terza direzione merita un'attenzione particolare. Infatti, secondo me, realizzare un IRS per utenti finali che sia in grado di costruire un modello del compito dell'utente è un'attività relativamente semplice e, allo stesso tempo, con alti benefici potenziali; inoltre, per quanto ne so, non è ancora stato realizzato un IRS di questo tipo.

Per avere un'idea quantitativa del miglioramento delle prestazioni ottenibile con IRS con modello del compito, si riconsideri l'esempio 1 del paragrafo 3: uno studente deve preparare una tesina e quindi necessita di documenti di un certo tipo (ad esempio, non è interessato ad articoli su atti di conferenze, in quanto troppo specialistici, mentre gli sarebbero molto utili articoli introduttivi o di rassegna). Si supponga che nella banca dati vi siano 3 tipi di documenti (ad esempio, articoli su rivista, articoli di conferenze e libri), che per ogni documento sia memorizzato il tipo a cui appartiene e che i 3 tipi dei documenti siano equiprobabili. Ora, se l'utente non specifica all'IRS il tipo di documento desiderato (informazione ricavabile dalla componente compito), almeno i 2/3 degli articoli reperiti risulteranno non attinenti al  $P$  dell'utente, mentre con un IRS con modello del compito, le prestazioni potrebbero risultare 3 volte superiori. Ho usato il condizionale, in quanto non è affatto detto che la componente compito del  $BI$  possa essere rappresentata in un IRS senza l'incertezza intrinseca nei processi di IR; per alleviare questo problema, secondo me bisogna far sì che l'utente comunichi all'IRS il proprio compito (e lo comprenda in modo migliore) durante tutta l'interazione con il sistema: la riformulazione deve quindi servire a meglio definire (e comprendere) tutte le componenti del  $BI$ , e non solo la componente argomento.

La rappresentazione del compito è possibile negli IRS tradizionali, ma con grande fatica dell'utente. Egli dovrebbe infatti conoscere il formato e il linguaggio d'interrogazione della banca dati che sta interrogando per poter specificare nella query la componente compito. È quindi senz'altro desiderabile e utile un IRS che si limiti a chiedere all'utente il tipo di compito da svolgere (fare una tesina, una ricerca bibliografica, una tesi, ecc.), che derivi autonomamente le caratteristiche desiderabili (e non) dei documenti (ed eventualmente le proponga all'utente per una conferma) e che traduca il tutto in un linguaggio adatto alla particolare banca dati. Per poter fare ciò, l'IRS deve possedere una base di conoscenze contenente i possibili stereotipi dei vari compiti, in modo da poter dedurre da ogni compito le caratteristiche dei documenti. Tali caratteristiche possono poi venire sfruttate o per inserire nuovi termini nella query o per effettuare un *ranking* dei documenti reperiti, e quindi l'IRS deve possedere anche le conoscenze per poter scegliere l'alternativa migliore a seconda della situazione.

Dalle considerazioni del paragrafo 4 si possono anche dedurre alcuni *postulati di impotenza* (prendendo a prestito il termine da Swanson, [26]) per l'IR; ad esempio:

- $a(as(\_,R,\_),t(q_n))$  è la massima attinenza trattabile da un IRS senza utilizzare conoscenza incerta, in quanto né  $BI$  né futuro possono essere conosciuti con certezza;
- $a(as(S,\_,\_),\_)$  è la massima attinenza trattabile in IRS per banche dati bibliografiche. Per poter andare oltre (più in alto), sono necessari IRS *full-text*.

Un'altra osservazione è che il carattere dinamico dell'attinenza porta a riconsiderare il *relevance feedback* (ossia l'utilizzo dei giudizi di attinenza espressi dall'utente durante la riformulazione): l'attinenza espressa durante l'interazione è un'attinenza differente da  $a(as(I,P,\{A,C,c\}),t(f))$  e quindi non è detto che un documento giudicato attinente durante l'interazione lo sia poi effettivamente.

## 7 Conclusioni e lavoro futuro

In questo lavoro ho indagato sul concetto di attinenza; tale concetto, come illustrato nel paragrafo 1, è allo stesso tempo importante e non pienamente compreso. Sfruttando la presentazione di alcune problematiche relative all'interazione utente-IRS (paragrafi 2 e 3), ho definito l'insieme delle attinenze  $A$  e la relazione di ordine  $\langle A$  (paragrafo 4). Tale lavoro è poi stato utilizzato per classificare alcune definizioni di attinenza reperite in letteratura (paragrafo 5) e ha portato ad alcune conseguenze sulla costruzione di IRS per utenti finali (illustrate nel paragrafo 6).

Lo studio presentato qui è senz'altro in una fase preliminare e lascia aperte numerose possibilità per lavori futuri. Senz'altro la scomposizione in argomento, compito e contesto necessita di ulteriori studi, per verificare innanzi tutto se sia corretta, e poi se possa essere estesa, raffinata e definita in modo più formale. Un'altra ricerca interessante riguarda i legami fra le varie attinenze: nel paragrafo 4 ho definito un ordinamento su di esse; è pensabile definire anche una *metrica* per misurare quanto sono vicine attinenze differenti? Come sostenuto da un anonimo revisore, con cui concordo pienamente, l'importanza di questa linea di ricerca è dovuta alle immediate ripercussioni pratiche; infatti, essa permetterebbe di scegliere in modo oggettivo fra le possibili direzioni per lo sviluppo di IRS più efficaci illustrate all'inizio del paragrafo 6. Tale metrica va a mio avviso definita sulla base di un'attività sperimentale, che permetta di confrontare le differenze fra attinenze adiacenti. In questa direzione, come già evidenziato nel paragrafo 5, vanno ad esempio [5, 21, 24].

Anche la classificazione del paragrafo 5 va completata, includendovi altri lavori reperibili in letteratura sull'attinenza, e la realizzazione di un IRS con modello del compito e una sua valutazione sperimentale (così come le indagini lungo le altre direzioni indicate nel paragrafo 6) sono senza dubbio attività importanti.

Una volta compresa appieno la natura del concetto di attinenza, ha senso affrontare il problema del *giudizio di attinenza*, e considerare i parametri che lo influenzano: *da chi, quando e come* viene effettuato.

Infine, come rientra nello scenario delineato la relatività dell'informazione rispetto al ricevente [15]? E cosa succede considerando non solo l'attinenza dei singoli documenti, ma l'attinenza dell'*insieme* dei documenti (e quindi le influenze fra documenti differenti)?

## Ringraziamenti

Desidero ringraziare Carlo Tasso, Giorgio Brajnik e Alessandro Bortuzzo, con cui ho avuto modo di discutere gli argomenti qui esposti e da cui ho ricevuto preziosi suggerimenti.

## Riferimenti bibliografici

- [1] Barry, C.L., 1994. User-Defined Relevance Criteria: An Exploratory Study, *J. of the American Society for Information Science*, 45(3), 149-159.
- [2] Bates, M.J., 1990. Where should the person stop and the information interface start?, *Information Processing and Management*, 26(5), 575-591.
- [3] Belkin, N.J., Oddy, R.N., e Brooks, H.M., 1982. ASK for information retrieval, *J. of Documentation*, 38(2), 61-71 e 38(3), 145-164.
- [4] Cleverdon, C.W., Mills, J. e Keen, M., 1966. *Factors Determining the Performance of Indexing Systems, Vol. 1: Design, Vol. 2: Test results*, Cranfield, UK, College of Aeronautics.
- [5] Cooper, W., 1973. On selecting a measure of retrieval effectiveness, *J. of the American Society for Information Science*, 24(2), 87-100 (part 1) e 24(6), 413-424 (part 2).
- [6] Cuadra, C.A. e Katter, R.V., 1967. *Experimental Studied of Relevance Judgements*, TM-3520/001, 002, 003/00, Systems Development Corporation, Santa Monica, California.
- [7] Floreanini, A., 1993. *Il progetto FIRE: studio e realizzazione di un sistema di intelligent information retrieval*, Research report UDMI/17/93/RR del Dip. di Matematica e Informatica dell'Univ. di Udine, Italy.
- [8] Foskett, D.J., 1972. A note on the concept of "relevance", *Information Storage and Retrieval*, 8, 77-78.
- [9] Froehlich, T.J., 1994. Relevance Reconsidered – Towards an Agenda for the 21st Century: Introduction to Special Topic Issue on Relevance Research, *J. of the American Society for Information Science*, 45(3), 124-133.
- [10] Furnas, G.W., Landauer, T.K., Gomez, L.M., e Dumais, S.T., 1987. The Vocabulary Problem in Human-System Communications, *Communications of the ACM*, 30(11), 964-971.
- [11] Harman D., 1993. Overview of the first TREC conference, *Proc. of the 16th Annual Int'l ACM/SIGIR Conf. on Research and Development in Information Retrieval*, Pittsburg, PA, ACM Press, 36-47.
- [12] Harter, S.P., 1992. Psychological Relevance and Information Science, *J. of the American Society for Information Science*, 43(9), 602-615.
- [13] Ingwersen, P., 1992. *Information Retrieval Interaction*, Taylor Graham, London.
- [14] Janes, J.W., 1994. Other People's Judgments: A Comparison of User's and Other's Judgments of Document Relevance, Topicality, and Utility, *J. of the American Society for Information Science*, 45(3), 160-171.
- [15] Longo, G.O., 1994. *Teoria dell'informazione*, inedito.
- [16] Mizzaro, S., 1995. *La componente esperta del sistema FIRE*, Progetto finalizzato CNR Sistemi Informatici e Calcolo Parallelo, Sottoprogetto 5. Rapporto Stato di Avanzamento n. R/5/137, e Research report UDMI/1/95/RR del Dipartimento di Matematica e Informatica dell'Università di Udine, Udine, Italy.
- [17] O'Connor, J., 1968. Some questions concerning "information need", *American Documentation*, 19, 200-203.
- [18] Park, T.K., 1994. Toward a Theory of User-Based Relevance: A Call for a New Paradigm of Inquiry, *J. of the American Society for Information Science*, 45(3), 135-141.
- [19] Rees, A.M. e Saracevic, T., 1966. The measurability of relevance, *Proceedings of the American Documentation Institute*, 3, 225-234.
- [20] Rees, A.M. e Schultz, D.G., 1967. *A field experimental approach to the study of relevance assessments in relation to document searching*, Vol. I, Final report (NSF Contract No. C-423). Case Western Reserve University, Cleveland, OH.
- [21] Regazzi, J.J., 1988. Performance Measures for Information Retrieval Systems - An Experimental Approach, *J. of the American Society for Information Science*, 39(4), 235-251.
- [22] Sandore, B., 1990. Online Searching: What Measures Satisfaction?, *Library and Information Science Research*, 12, 33-54.
- [23] Saracevic, T., 1975. RELEVANCE: A Review of and a Framework for the Thinking of the Notion in Information Science, *J. of the American Society for Information Science*, 26(6), 321-343.
- [24] Saracevic, T., Kantor, P., Chamis, A. e Trivison, D., 1988. A Study of Information Seeking and Retrieving. I. Background and methodology, *J. of the American Society for Information Science*, 39(3), 161-176.
- [25] Schamber, L., Eisenberg, M.B. e Nilan, M.S., 1990. A re-examination of relevance: Toward a dynamic, situational definition, *Information Processing and Management*, 26(6), 755-776.
- [26] Swanson, D.R., 1988. Historical Note: Information Retrieval and the Future of an Illusion, *J. of the American Society for Information Science*, 39(2), 92-98.