Marco Scarnò [*] ***- Donatella Sforzini*** [*]

Caratterizzazione delle abitudini degli utenti mediante metodologie di analisi statistica applicate ai file log degli accessi

L'impiego a fini di consultazione e/o Ricerca dei Periodici elettronici implica un'interazione tra Uomo e macchina, nella misura in cui l'oggetto –l'articolo- cercato deve essere "raggiunto" seguendo delle regole imposte da un "motore di ricerca". Questa interazione si è evoluta nel tempo, sia per via di un miglioramento del modo in cui si presenta il motore di ricerca, sia per una più efficiente "strategia" d'uso di questo da parte degli utenti. E' da notare, tuttavia, che la quantità di articoli per argomento risulta essere in continua crescita, ed a volte essa può essere sovrabbondante rispetto le reali esigenze dell'utente stesso. Da questa osservazione nasce il progetto di monitorare il comportamento di ricerca, ossia il modo attraverso il quale avviene l'interazione tra Uomo e macchina. A questo scopo sono analizzati, quindi, i "weblog files", ossia le tracce delle azioni svolte dagli utenti, onde verificare se vi siano Strategie di Ricerca che consentono di raggiungere nel più breve tempo possibile la visualizzazione di un articolo. Nell'ipotesi che tale risultato possa poi essere identificato come un elemento di efficienza del sistema considerato

Premessa [**]

Il raggio di luce si insinuava sinuosamente dalla finestra socchiusa, poggiandosi sul foglio dove era quello strano indirizzo. La mia ricerca era quasi giunta a conclusione, avevo individuato tutti gli elementi significativi dopo mesi di esperimenti nel laboratorio. Eppure quel discorso nella biblioteca era ancora nei miei pensieri. Felice del lavoro fatto, ma desideroso anche di verificare se il suggerimento della mia interlocutrice fosse giusto. "Perché non cerchi, tra i lavori fatti dagli altri studiosi nel mondo, se esiste qualcuno che ha raggiunto gli stessi tuoi risultati?". Così quello scritto rapido su di un foglietto. Un indirizzo per un appuntamento virtuale, una serie di lettere che componevano la parola da scrivere sul computer per accedere all'emeroteca, anch'essa virtuale, dove avrei potuto verificare se il mio solitario lavoro era stato condiviso da altri. Il raggio di luce scivolava in basso, dal foglietto alla tastiera del mio computer… è l'ora di prepararsi all'appuntamento, pensavo, che preferirei, però, fosse non con una macchina…

1. Introduzione[**]

Ricordo la prima volta che usai Internet. Era il 1993; lo schermo del mio computer (un trentesimo meno *potente* di quello che adesso ho avanti) mostrava una pagina nera con una lista di files contenuti in un lontano computer in Svizzera. Anche le mail erano difficili da spedire; una serie di operazioni e di caratteri strani dovevano essere usati (@, ad esempio).
Qualche mese dopo vidi le prime pagine Web, che risultarono successivamente essenziali per lo sviluppo del lavoro legato alla mia tesi. Ricordo che cercavo articoli o indicazioni da parte di quanti avevano già affrontato temi simili al mio e, seppur con qualche difficoltà, riuscii a ottenere un discreto numero di scritti che, tuttora, conservo in un raccoglitore nell'armadio. Oggetti preziosi per la storia che essi hanno con me.
Qualche anno è passato e oramai il Web è divenuto lo strumento tramite il quale l'Informazione si diffonde nella maniera più rapida possibile. Tuttavia proprio questa sua

[*] CASPUR, Gruppo "ADAMS" (Analisi dati e metodologie statistiche); e-mail: <mscarno@caspur.it>.
[*] CASPUR, Gruppo "ADAMS" (Analisi dati e metodologie statistiche); e-mail: <donatella.sforzini@caspur.it>.
[**] Paragrafo a cura di M. Scarnò

caratteristica tende a farlo divenire *meno efficiente*, sia per causa della sovrabbondante quantità della stessa Informazione che in esso è disponibile, sia per la necessità che l'utente debba *imparare* sempre più a interagire con le possibilità legate a fare ricerche al suo interno.
Questo lavoro si riferisce ad uno studio relativo alle interazioni tra *ricercatore* e un *motore di ricerca* che fornisce, in risposta a delle interrogazioni, una serie di articoli contenuti nei maggiori periodici elettronici a carattere scientifico.

2. Il contesto

Nel 2003 una media di 80 mila interazioni giornaliere sono state quelle che hanno interessato gli utenti delle Università italiane o dei Centri di Ricerca dell'Italia Centrale-Meridionale, con il sito dell'Emeroteca virtuale gestito dal Consorzio interuniversitario CASPUR: http://periodici.caspur.it.
Elemento fondamentale dell'interazione è il *motore di ricerca*, messo a disposizione per facilitare la selezione di pochi articoli da una collezione di diversi milioni.
Apposite pagine del sito consentono agli utenti di fornire al sistema le *chiavi* (parole) che servono ad estrarre gli articoli che rispondono alle esigenze della ricerca stessa. Un processo di *richiesta-risposta* senza dubbio assai poco *intelligente* dal lato del motore di ricerca, il quale fornisce semplicemente le indicazioni di tutti quegli articoli in cui si trovano le chiavi desiderate.
"Searching for relevant information on the World Wide Web is often a laborious and frustrating task for casual and experienced users[1]", scrivono Holscher e Strube (C. Holscher e G. Strube, 2000), mentre risulta oramai noto che, se le attitudini necessarie per *navigare* i singoli siti Web sono accessibili alla maggioranza degli utenti dopo un semplice addestramento (si veda, ad esempio, J. Hurtienne e H. Wandke, 1997), è necessaria una maggiore *esperienza* e l'aver sviluppato una appropriata *strategia* per effettuare delle ricerche basate sull'impiego di motori di ricerca (si veda, ad esempio, A. Pollock e A. Hockey, 1997).
A tale proposito, in questo lavoro, sono state investigate proprio tali *strategie*, verificandone l'efficacia in termini sia dei passi necessari al raggiungimento di un articolo, sia in relazione alla strategia scelta per condurre il processo di ricerca stesso.
Prima di dettagliare i risultati raggiunti è opportuno osservare che non è propriamente corretto definire come *efficace* una strategia perché ha consentito all'utente di visualizzare un articolo. Infatti non è dato sapere, con le informazioni a disposizione, il livello di soddisfazione dell'utente rispetto all'articolo stesso.

3. Le informazioni a disposizione

Nel nostro lavoro sono stati impiegate le registrazioni delle operazioni condotte dagli utenti così come *memorizzate* dal motore di ricerca. Si tratta di files molto complessi (logfiles), dove ogni riga contiene sia l'indirizzo del computer dal quale è partita la richiesta (con l'indicazione dell'istante temporale di riferimento), sia il contenuto della richiesta stessa (o l'eventuale risposta fornita dal sistema). Ad esempio:
XXX.XXX.XXX.XXX - - [01/Nov/2002:07:30:28 +0100] "GET /cgi-bin/search.pl?
Database=journals&search_field=oncogenes+and+bladder+cancer&GetSearchResults=
Search&fields=Any&canned_search= HTTP/1.1" 200 42145

[1] La ricerca dell'informazione sul World Wide Web è, spesso, un compito laborioso e frustrante per gli utenti casuali ed esperti.

In questo caso, dal computer identificato con XXX.XXX.XXX.XXX, è stata inoltrata la richiesta di visualizzare tutti gli articoli che contenessero le parole *oncogenes*, *bladder* e *cancer*.
Al fine di effettuare un'analisi statistico-comportamentale degli utenti è, quindi, risultato necessario trasformare tali files in apposite tabelle dati dove vi sia una corretta valorizzazione dei campi possibili con le chiavi inserite dagli utenti stessi.
A tale scopo è stata realizzata un'applicazione appropriata, che non verrà nel prosieguo illustrata, in grado di costruire dinamicamente tale tabella dati (si veda M. Scarnò e D. Sforzini, 2002).
Un'importante e preliminare osservazione riguarda il fatto che tutti gli utenti sono identificati tramite i computer che impiegano; ciò implica che vi possono essere più individui distinti che risultano essere attributi allo stesso *indirizzo* (si pensi a quanti impiegano le postazioni delle biblioteche per consultare gli articoli, o quanti hanno computer che presentano, all'uscita della rete locale di riferimento, lo stesso indirizzo). Si ha, quindi, che nella maggioranza dei casi il comportamento di ricerca sarà *mediato*.
Un'altra rilevante osservazione riguarda l'impossibilità di definire delle *risposte* standard da parte del motore di ricerca; infatti ognuna di esse è strutturata sulla base delle caratteristiche di quanto scritto dall'utente nelle apposite sezioni delle pagine a disposizione.
Tuttavia, ai fini del presente lavoro, sono state classificate le strategie di ricerca mediante l'identificazione della scelta fatta all'utente rispetto le varie possibilità fornite dal motore stesso; in particolare sono state considerate:

- Ricerche di una o più parole contenute in una qualunque parte dell'articolo (ANY);
- Ricerche di una o più parole contenute nel solo abstract dell'articolo (ABSTRACT);
- Ricerche di una o più parole che identificassero l'autore (gli autori) (AUTHOR);
- Ricerche che fossero un *misto* delle precedenti (AVANZATA);
- Ricerche dirette dell'articolo scegliendo prima la rivista, poi l'anno, il numero, ecc. (BROWSING).

4. Trattamento preliminare dei dati

Prima di procedere all'analisi dei dati raccolti in riferimento al periodo Gennaio-Settembre 2003 (oltre 20 milioni d'interazioni tra utenti e sito Web di riferimento), si è proceduto all'identificazione della *sessione di ricerca*, ossia del momento in corrispondenza del quale un utente inizia a cercare un articolo.
Nell'ipotesi che questi può *fermarsi* per leggere i risultati di quanto trovato, si è studiata la distribuzione dei *tempi d'inattività* e si è verificato che questa, dopo una lenta discesa, intorno ai 27 minuti tendeva a risalire nuovamente. Da tale osservazione si è introdotta l'ipotesi che le ricerche, in media, tendevano a cominciare nuovamente dopo tale tempo, identificando così l'inizio di nuove fasi di *navigazione* del sito. Altri problemi affrontati hanno riguardato:

- La necessità di eliminare i *doppi click*, ossia di tutti quei record, nei dati a disposizione, che identificavano le medesime azioni compiute dagli utenti, ripetute per via dell'eventuale lentezza del sistema (o dell'impazienza dell'utente), che induceva la pressione ripetuta dei pulsanti presenti nelle pagine Web;
- L'individuazione e l'eliminazione dei comportamenti *anomali*, ossia di tutti quei record che erano da attribuire non a degli utenti reali bensì a programmi in grado d'interagire con il motore di ricerca stesso onde scaricare automaticamente gli articoli.

5. I risultati

Una preliminare indicazione riguardante il comportamento degli utenti rispetto all'interazione con il sito è data dal numero medio di azioni successive, o passi (da intendersi come le ricerche o la visualizzazioni di articoli), condotti dagli utenti. Il grafico 1 ne riporta l'andamento, da cui si evidenzia che sono, circa, 14 le azioni successive condotte dagli stessi utenti all'interno del sito.

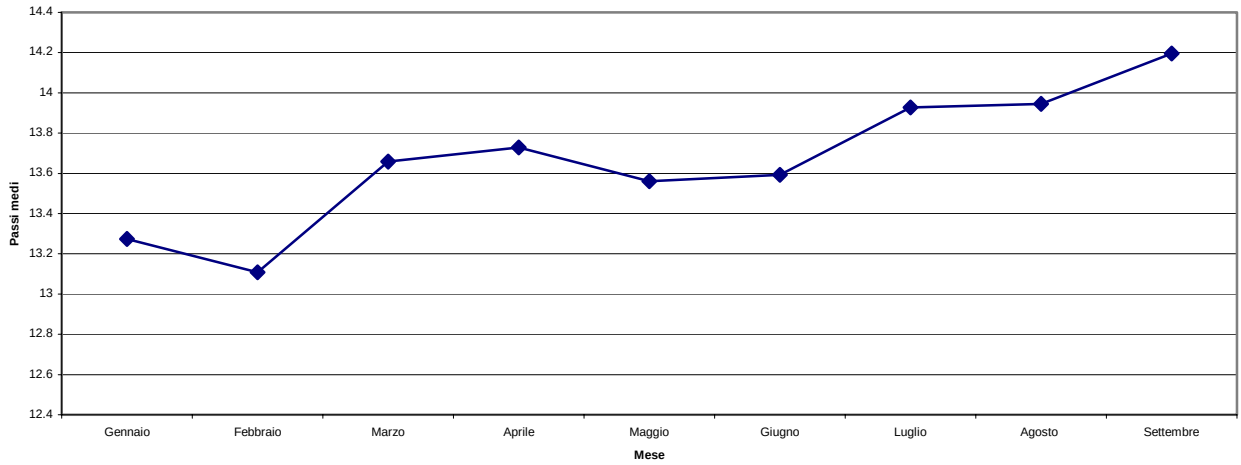

Grafico 1: numero di passi medi degli utenti del sito http://periodici.caspur.it impegnati in *Ricerche* o *Visualizzazione di articoli* (periodo: Gennaio-Settembre 2003)

Ovviamente tale numero può essere suddiviso a seconda di quale sia la *strategia* scelta per arrivare alla visualizzazione stessa di un articolo; il grafico 2 ne fornisce il valore a seconda della modalità di ricerca scelta dall'utente (si veda il paragrafo 4).

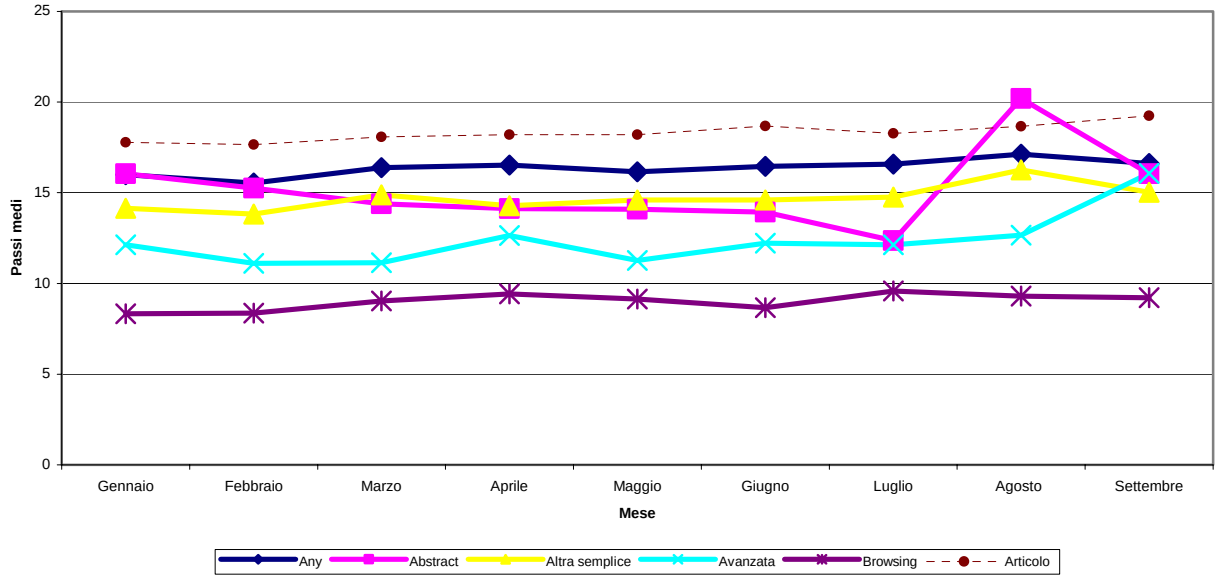

Grafico 2: numero di passi medi necessari agli utenti del sito http://periodici.caspur.it per giungere alla visualizzazione di un articolo a seconda della strategia di Ricerca impiegata (periodo: Gennaio-Settembre 2003)

Dal grafico 2 si osserva che il browsing risulta essere la strategia più rapida per accedere alla visualizzazione di un articolo, confermando i risultati dell'articolo citato di Hurtienne e Wandke. Le strategie meno *rapide* riguardano le ricerche su parole generiche (ANY) e quelle relative alle parole contenute nell'abstract dell'articolo stesso.
Altra indicazione interessante risulta essere il numero di articoli ai quali segue la visualizzazione di un altro articolo; tale valore, il più alto nel periodo, è di circa 18, il che vuol dire che gli utenti tendono a vedere, consecutivamente, circa 18 articoli. Come accennato risulta difficile sapere se tale comportamento derivi dalla *non soddisfazione* di quanto trovato o dall'interesse per tutti i titoli proposti; certo è che l'utente stesso rimane più a lungo *coinvolto* dai risultati che non dal percorso impiegato per raggiungerli.
Un ulteriore risultato ricavato dall'analisi dei dati a disposizione deriva dalla distribuzione percentuale delle modalità d'inizio delle ricerche (grafico 3), dove si osserva che, circa, il 60 % di esse comincia proprio con il browsing, mentre solo il 10% con la modalità ANY.
Ciò lascerebbe supporre che gli utenti del sito fanno uso dell'emeroteca come se fosse una *biblioteca*: si collegano e *sfogliano* gli ultimi numeri delle riviste arrivate.

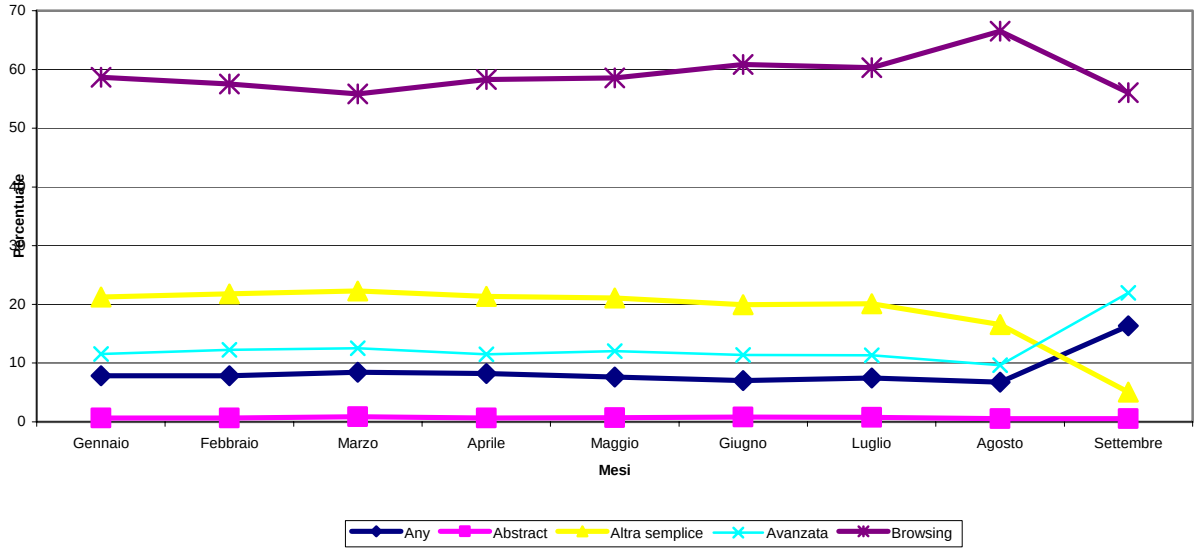

Grafico 3: distribuzione (%) dell'inizio delle ricerche degli utenti del sito http://periodici.caspur.it per strategia di Ricerca impiegata (periodo: Gennaio-Settembre 2003)

In ultimo appare interessante verificare la tabella doppia, espressa in percentuale sul totale, che riporta le indicazioni della generica strategia scelta e della successiva (tabella 1).

		Strategia al tempo t+1						
		Any	Abstract	Altra semplice	Avanzata	Browsing	Articolo	**Totale**
Strategia al tempo t (escluso l'inizio della ricerca)	Any	5.7	0.0	0.3	0.1	0.3	2.5	**9.1**
	Abstract	0.0	0.4	0.0	0.0	0.0	0.2	**0.7**
	Altra semplice	0.3	0.1	8.6	0.5	0.5	3.8	**13.7**
	Avanzata	0.1	0.0	0.6	5.6	0.5	1.4	**8.1**
	Browsing	0.3	0.0	0.4	0.4	14.1	8.3	**23.5**
	Articolo	2.5	0.2	3.1	1.1	4.8	33.3	**44.9**
	Totale	**8.9**	**0.7**	**13.0**	**7.7**	**20.3**	**49.4**	**100.0**

Tabella 1: frequenze percentuali (sul totale delle ricerche) a seconda della strategia al tempo t e al tempo successivo, degli utenti del sito: http://periodici.caspur.it (periodo: Gennaio-Settembre 2003)

Dalla tabella 1 si osserva che ad una ricerca generica (ANY) segue, più probabilmente, un'altra ricerca generica (magari avendo inserito una o più parole chiave onde *raffinare* la ricerca stessa), mentre la percentuale di articoli scaricati è senza dubbio proprio dell'aver effettuato un browsing delle riviste contenute nel sistema. Si osservi, a tale proposito, che anche il browsing consta delle necessità di operare diverse operazione successive; ciò è dovuto alla necessità d'identificare la rivista (o l'area) desiderata, poi l'anno, il numero, ecc.

Conclusioni

Le analisi condotte sui dati a disposizione hanno evidenziato l'esistenza di diverse tipologie comportamentali che potrebbero essere ricondotte a differenti tipologie di utenti:

- *Specializzati*, ossia legati ad uno specifico oggetto (una rivista, ad esempio), che tendono a raggiungere nel più breve tempo possibile;
- *Generici*, i quali compongono delle ricerche impiegando le potenzialità offerte dal motore di ricerca e che tendono a ripetere l'azione stessa di ricerca sino a restringere il numero di articoli ottenuti come risposta a quelli desiderati.

Certo è, tuttavia, che l'aver visualizzato un articolo non è l'ultima azione registrata, in quanto l'azione seguente è la visualizzazione di un articolo ulteriore.

Riferimenti bibliografici

C. Holscher, G. Strube, *Web Search Behaviour of Internet experts and newbies*, Computer Networks 33 (2000), 337-346

J. Hurtienne, H. Wandke, *How effectively and efficiently do users navigate in the WWW? An empirical study*, in D. Janetzko, B. Batanic, D. Schoder, M. Mattingley and G. Strube (Editors), Cam-97: Workshop on Cognition and Web, Freiburg 1997, 93-104

A. Pollock, A. Hockley, *What's wrong with Internet Searching*. D-Lib Magazine (documento WWW disponibile presso http://www.dlib.org/Dlib/march97/bt/03pollock.html)

M. Scarnò, D. Sforzini, *La diffusione della conoscenza via Internet: acquisizione ed elaborazione dei comportamenti degli utenti*, in Data Mining, Web Mining e CRM (a cura di F. Camillo e G. Tassinari), Franco Angeli Editore, Milano 2002, 116-131

M. Scarnò, D. Sforzini, *Does the Web dominate Web Users?*, lavoro non pubblicato presentato in occasione del Convegno GFKL 2003, Cottbus.