I vestiti nuovi dell'imperatore

Analisi dell'algoritmo utilizzato per stabilire la rilevanza delle registrazioni bibliografiche nei cataloghi delle biblioteche italiane

DANILO DEANA
Università degli studi di Milano
danilo.deana@unimi.it

Metodologia

L'argomento di questo articolo si situa all'interno della disciplina conosciuta come *information retrieval* (IR). Nata a metà dello scorso secolo – il primo a utilizzare il termine fu Calvin Mooers in un articolo pubblicato nel 1950 – l'IR si occupa di come sia possibile trovare materiali (di norma documenti) di natura non strutturata (solitamente testo) che soddisfino un determinato bisogno informativo all'interno di collezioni di grandi dimensioni.[1]
Un sistema di IR è un'applicazione che riceve dall'esterno come *input* una query, ne confronta il contenuto con gli indici dei documenti che sono conservati al suo interno e restituisce come *output* un elenco di risultati solitamente ordinati per rilevanza.[2] Gli indici sono il modo in cui vengono rappresentati e organizzati i documenti conservati all'interno del sistema.
I temi legati al recupero dell'informazione possono essere affrontati da due punti di vista distinti e complementari: uno centrato sugli elaboratori, l'altro sull'uomo. Dal primo punto di vista, l'IR si occupa principalmente della creazione di indici efficienti, dell'elaborazione con prestazioni elevate delle query e dello sviluppo di algoritmi di classificazione dei documenti per fare in modo che i risultati corrispondano sempre di più alle esigenze informative. Nella visione centrata sull'uomo, l'IR consiste invece nello studio del comportamento dell'utente, nella comprensione dei suoi principali bisogni e nel determinare in che modo tale comprensione influisca sull'organizzazione e sul funzionamento del sistema di recupero.[3]
Noi ci concentreremo sul primo punto di vista, che è dominante nel mondo accademico e in quello delle aziende.[4] Intendiamo quindi esaminare l'algoritmo utilizzato per stabilire la rilevanza delle registrazioni nei cataloghi delle biblioteche italiane da un punto di vista oggettivo,[5] un approccio opposto a quello soggettivo adottato da Carlo Bianchini nella sua indagine sull'interazione tra gli utenti e il catalogo della biblioteca del Dipartimento di musicologia e beni culturali dell'Università degli studi di Pavia.[6]

L'articolo fa parte di uno studio più ampio dedicato alla rivoluzione tecnologica e alle biblioteche che sarà pubblicato a marzo nella collana "Biblioteconomia e scienza dell'informazione" dell'Editrice Bibliografica.

La rilevanza è un concetto complesso. Tefko Saracevic, docente presso la School of Communication & Information della State University del New Jersey, ne ha individuati cinque tipi base: rilevanza di sistema o algoritmica, rilevanza topica o di soggetto, rilevanza cognitiva o pertinenza, rilevanza situazionale o utilità, rilevanza motivazionale o affettiva.[7]
Quella che qui ci interessa è la prima, la rilevanza di sistema o algoritmica. I documenti conservati all'interno di un sistema di IR sono rappresentati, organizzati e associati alle query secondo modi e mezzi specifici del sistema stesso. Dato che un sistema di IR è progettato per recuperare l'insieme di documenti considerati più significativi rispetto ai bisogni informativi espressi attraverso una query, i modi e i mezzi hanno in sé una presunzione di rilevanza.[8] La difficoltà consiste nello stabilirne l'efficacia, senza confonderla con quella topica o di soggetto o con quella cognitiva o pertinenza, senza cioè coinvolgere l'utente che ha formulato la query.[9]
Un aiuto da questo punto di vista viene dal fatto che un catalogo di biblioteca è in grado di soddisfare solo un determinato numero di esigenze informative, che sono state tutte definite nell'*IFLA Library Reference Model* (IFLA LRM) attraverso cinque attività (Trovare, Identificare, Selezionare, Ottenere ed Esplorare) e i corrispondenti casi d'uso. I casi d'uso hanno reso possibile collegare le attività al modello. Essi infatti definiscono la ricerca di informazioni dell'utente nei termini delle entità, degli attributi e delle relazioni del modello.[10]
Valuteremo quindi l'algoritmo utilizzato per stabilire la rilevanza delle registrazioni nei cataloghi delle biblioteche italiane sulla base della sua capacità di individuare quale sia il caso o i casi d'uso cui la query ha maggior probabilità di essere ricondotta e di restituire i risultati ordinati sulla base di questo presupposto.

Campione

I cataloghi presi in considerazione sono quello dell'Indice del Servizio bibliotecario nazionale (SBN), gestito dall'Istituto centrale per il catalogo unico delle biblioteche italiane e per le informazioni bibliografiche (ICCU), e quello dell'Università degli studi di Milano, gestito dal Servizio bibliotecario di ateneo.
Il Catalogo dell'Indice SBN è il più grande e consultato d'Italia. Quello dell'Università degli studi di Milano è basato su SebinaYOU, un modulo di Sebina Open Library e Sebina Next, due dei sistemi di automazione per biblioteche più diffusi nel nostro paese. Sebina Open Library, Sebina Next e SebinaYOU sono prodotti di DM Cultura, un'azienda attiva da più di trent'anni nel settore delle biblioteche.
Nella pagina iniziale del catalogo dell'Indice SBN e di SebinaYOU, come in quasi tutti i cataloghi delle biblioteche italiane e straniere, è presente una maschera composta da una singola casella di testo dove è possibile inserire uno o più termini e avviare quindi una ricerca, esattamente come in Google. Il risultato è un elenco di registrazioni bibliografiche (che d'ora in poi chiameremo semplicemente registrazioni) ordinate per rilevanza e divise in pagine.
Questa modalità di ricerca, che nei cataloghi è definita libera o semplice, è la più utilizzata dagli utenti, anche perché per passare a quella avanzata è necessario selezionare un collegamento.[11]
La scelta di inserire nella pagina iniziale una maschera di ricerca composta da una singola casella di testo, oltre a risultare familiare a chiunque utilizzi i motori di ricerca, ha anche il vantaggio di risolvere, almeno in apparenza, il problema della struttura che dovrebbe guidare gli utenti dal linguaggio che conoscono a quello utilizzato nell'organizzazione delle informazioni bibliografiche. La mancanza di questa struttura era già stata rilevata da Elaine Svenonius all'inizio del secolo. Secondo Svenonius, l'introduzione dei cataloghi elettronici, che a partire dagli anni settanta del secolo scorso hanno iniziato a sostituire quelli cartacei, hanno addirittura aggravato il problema.[12]
La ricerca avanzata permette di svolgere ricerche più mirate, ma utilizza termini che hanno un significato diverso per gli utenti e per i bibliotecari.[13] Con "Autore", ad esempio, i bibliotecari indicano tutti coloro (persone, enti o congressi) che hanno contribuito a qualsiasi titolo alla creazione della risorsa.[14] Un utente, almeno nella nostra esperienza, identifica un autore con l'autore principale, che per lui è di solito una persona.
Gli studi sul comportamento degli utenti sono iniziati subito dopo l'introduzione dei primi cataloghi elettronici.[15] Da allora c'è stata una continua evoluzione, legata soprattutto alla nascita dei motori di ricerca, che ne hanno modificato le aspettative e le abitudini.[16]
Per quanto riguarda la nostra indagine, gli studi più recenti concordano sul fatto che nei cataloghi delle

biblioteche le parole chiave utilizzate per formulare una query raramente sono più di tre.[17] L'uso degli smartphone e dei tablet ne ha ulteriormente ridotto il numero.[18]
Le registrazioni ottenute attraverso una ricerca di questo tipo sono solitamente relative anche a risorse che non hanno nulla a che vedere con il bisogno informativo di chi ha formulato la query. L'algoritmo impiegato per stabilirne la rilevanza è quindi molto importante, in quanto può far sì che le più significative siano mostrate per prime.
Per rendersi conto di quanto sia importante comparire all'inizio di un elenco di risultati, è sufficiente considerare il numero delle pagine web dedicate all'argomento, che supera i due miliardi. È nata anche una nuova attività: l'ottimizzazione per i motori di ricerca (*Search Engine Optimization*, SEO), il cui scopo è quello di migliorare (o mantenere) il posizionamento di un sito web nelle *Search Engine Results Pages* (SERP).
Per valutare l'algoritmo impiegheremo le registrazioni ottenute come risultato di una query formulata attraverso la ricerca semplice per la quale è stato utilizzato un solo termine: "Gadda". Nel caso del Catalogo dell'Indice SBN le registrazioni restituite sono state 2.395, in quello dell'Università degli studi di Milano 214 (tutti i dati sono stati raccolti il 21 agosto 2018).

Catalogo dell'Indice del Servizio bibliotecario nazionale (SBN)

Una registrazione è rilevante se è relativa a un documento che soddisfa il bisogno informativo dell'utente, non perché contiene i termini della query (in questo caso abbiamo scelto di definirla "attinente"). Per valutare l'algoritmo di rilevanza si deve quindi convertire la query in un'espressione esplicita del bisogno e utilizzarla per suddividere le registrazioni restituite tra rilevanti o non rilevanti.[19]
Nel Catalogo dell'Indice SBN la ricerca semplice considera solo 4 canali (Autore, Titolo, Soggetto e Descrizione Dewey) dei 24 disponibili in quella avanzata: Autore, Titolo, Collezione, Titolo uniforme, Luogo di pubblicazione, Editore, Soggetto, Classificazione Dewey, Descrizione Dewey, Identificativo SBN, Codice ISBN (International Standard Book Number), Codice ISSN (International Standard Serial Number), Codice ISSN-L (usato per identificare le diverse versioni di uno stesso periodico), Codice ISNM (International Standard Music Number), Numero edizione lastra musica a stampa, Numero edizione registrazione sonora, Numero edizione matrice registrazione sonora, Codice ISRC (International Standard Recording Code), Numero videoregistrazione, Numero risorsa sonora, Codice EAN (European Article Number), Codice UPC (Universal Product Code), Codice BNI (Bibliografia nazionale italiana) e Codice CUBI (Catalogo cumulativo 18861957 del Bollettino delle pubblicazioni italiane).
Per descrivere i bisogni informativi che possono essere soddisfatti attraverso una ricerca semplice è possibile fare riferimento a una parte dei casi d'uso impiegati dall'IFLA LRM per illustrare nel dettaglio l'attività di Trovare, ossia "raccogliere informazioni su una o più risorse di interesse attraverso una ricerca svolta utilizzando un qualsiasi criterio rilevante".[20]

Canale	Caso d'uso
Autore	Trovare le risorse in relazione con un determinato agente
Titolo	Trovare le manifestazioni dell'espressione di un'opera attraverso un titolo associato all'opera, a una delle sue espressioni o delle sue manifestazioni
Soggetto	Trovare le risorse che incorporano un'opera che è in una relazione del tipo "ha come soggetto" con una data *res* (o un insieme di *res*) utilizzando un *nomen* (per una data *res*) impiegato nel *Nuovo Soggettario*
Descrizione Dewey	Trovare le risorse che incorporano un'opera che è in una relazione del tipo "ha come soggetto" con una data *res* (o un insieme di *res*) utilizzando un *nomen* (per una data *res*) impiegato nella *Classificazione Decimale Dewey*

Tabella 1

Nel Catalogo dell'Indice SBN ogni canale di ricerca corrisponde a uno o più campi del formato UNIMARC utilizzato per strutturare le registrazioni. In particolare, il canale Autore corrisponde ai campi 700, 701 e 702, Titolo al campo 200 e ai campi dal 500 al 577, Soggetto ai campi dal 601 al 608 e Descrizione Dewey al campo 676[21] (Tabella 2).
Con la query formulata attraverso la ricerca semplice per la quale è stato utilizzato solo il termine "Gadda" sono state ottenute 1.739 registrazioni attraverso

```
LDR    10607nam1 2203421 i 4500
001      IT\ICCU\CFI\0155806
003      http://id.sbn.it/bid/CFI0155806
005      20090926211703.0
100  \   \   $a 19910513g19881993||||0itac50 ba
101  \   \   $a ita
102  \   \   $a IT
181  \   1   $6 z01 $a i $bxxxe
182  \   1   $6 z01 $a n
183  \   1   $6 z01 $a nc $2 RDAcarrier
200  1   \   $a Opere di Carlo Emilio Gadda $f edizione diretta da Dante Isella
210  \   \   $a Milano $c Garzanti $d 1988-1993
215  \   \   $a 5 volumi $d 19 cm
410  \   \   $a I libri della Spiga
606  \   \   $a Gadda, Carlo Emilio $X Opere $X Bibliografia $2 FI
             $3 IT\ICCU\MILC\064224
676  \   \   $a 853.914 $c Narrativa italiana. 1945- $v 19
699  \   \   $a Bibliografia sistematica $y Bibliografia
700  \   1   $a Gadda $b , Carlo Emilio $3 IT\ICCU\CFIV\000056 $4 070
702  \   1   $a Isella $b , Dante $3 IT\ICCU\CFIV\020051
801  \   3   $a IT $b ICCU $c 20180827
```

Tabella 2

il canale Autore, 702 attraverso il canale Titolo, 210 attraverso il canale Soggetto e nessuna registrazione nel caso di Descrizione Dewey.

Canale di ricerca	Registrazioni recuperate	Percentuale sul totale
Autore	1.739	72,60
Titolo	702	29,31
Soggetto	210	0.87
Descrizione Dewey	0	0,00
Totale	**2.395**	**100,00**

Tabella 3

Il totale delle registrazioni recuperate è minore della somma dei risultati parziali. Il motivo è che nel momento in cui i quattro insiemi di registrazioni sono stati uniti a formarne uno solo, le registrazioni che facevano parte di più insiemi (quella relativa alle *Opere di Carlo Emilio Gadda* riportata nella Tabella 2, ad esempio, appartiene al primo, al secondo e al terzo) sono state considerate una sola volta.
Il basso numero di registrazioni che contengono soggetti in cui compare il termine "Gadda" rispetto al totale di quelle recuperate si spiega con la scelta dell'ICCU di "non richiedere un soggetto interno per le opere a carattere narrativo, le raccolte di poesie, i frammenti di opere, e anche le opere di autori classici".[22] Inoltre non tutte le registrazioni che dovrebbero avere un soggetto lo hanno effettivamente.[23] Le registrazioni recuperate attraverso il canale Soggetto che contengono il descrittore "Gadda, Carlo Emilio" sono 188, mentre quelle recuperate attraverso il canale Titolo relative a studi sullo scrittore milanese sono 451, più del doppio.
Prima di esaminare l'algoritmo utilizzato all'interno del Catalogo dell'Indice SBN per stabilire la rilevanza delle registrazioni, è opportuno considerare due misure che permettono di stabilire l'efficacia dei sistemi di *information retrieval*: il recupero e la precisione. Sarà così possibile mettere in evidenza alcune particolarità di questo catalogo.
Il recupero è il numero di registrazioni attinenti effettivamente recuperate da una ricerca diviso il numero totale di registrazioni che avrebbero dovuto esserlo.
La precisione è il numero di registrazioni attinenti recuperate da una ricerca diviso il numero totale di registrazioni recuperate dalla ricerca stessa.
Sono state considerate attinenti le registrazioni che contengono in uno dei campi associati ai canali Autore, Titolo, Soggetto o Descrizione Dewey il termine "Gadda".
I valori del recupero e della precisione possono variare da 0 (non sono state recuperate registrazioni attinenti) a 1 (sono state recuperate solo registrazioni attinenti).
Per stabilire il numero di registrazioni che avrebbero dovuto essere recuperate è stato utilizzato un canale di ricerca disponibile solo nella modalità avanzata, Parole chiave, attraverso cui è possibile ottenere un elenco completo delle registrazioni che contengono in uno qualsiasi dei campi in cui sono strutturate il termine "Gadda".

Canale di ricerca	Registrazioni recuperate	Registrazioni che avrebbero dovuto essere recuperate	Registrazioni attinenti	Recupero	Precisione
Autore	1.739	1.739	1.739	1,00	1,00
Titolo	702	702	552	1,00	0,78
Soggetto	210	210	210	1,00	1,00
Descrizione Dewey	0	0	0	-	-
Totale	**2.395**	**2.417**	**2.245**	**0,96**	**0,93**

Tabella 4

Il valore del recupero è 0,96. Se fosse stato utilizzato un altro termine che ha maggiori probabilità di comparire nei campi UNIMARC che non fanno capo ai canali Autore, Titolo, Soggetto e Descrizione

Dewey, il valore sarebbe stato molto più basso: nel caso di "Mondadori", ad esempio, sarebbero state recuperate 4.768 registrazioni a fronte delle 177.402 attinenti.
La bassa precisione del canale Titolo dipende dal fatto che sono state recuperate registrazioni che non contengono il termine "Gadda" nei campi che fanno capo al canale, ma in quelli del blocco 4XX, utilizzato per i collegamenti. È il caso, ad esempio, delle 11 registrazioni relative ad album degli Iron Butterfly, un gruppo rock statunitense, che contengono il brano *In a gadda da vida* (i dati del brano sono riportati nel campo 464, Piece-Analytic Level):

LDR			02162njm1 22003133i 4500
001			IT\ICCU\DDS\0756478
200	1	\	**$a** Iron Butterfly live **$f** Iron Butterfly
464	\	1	**$1** 001IT\ICCU\DDS\0756484 **$1** 2001 $aIn-a-gadda-da-vida$fIron Butterfly$v6 **$1** 71002$aIron Butterfly$3IT\ICCU\DDSV\018382$4590 **$1** 702 1$aIngle$b, Doug$3IT\ICCU\DDSV\170772$4070
801	3	\	**$a** IT **$b** ICCU **$c** 20180917

Tabella 5

C'è anche il caso opposto, che riguarda le 19 registrazioni relative a tracce che fanno parte delle diverse versioni dell'album *In a gadda da vida*, che deve il suo titolo al brano omonimo (i dati dell'album sono riportati nel campo 463, Piece Level):

LDR			00734nja2 22001931i 4500
001			IT\ICCU\DDS\0756665
200	1	\	**$a** Flowers and beads **$f** Iron Butterfly
463	\	1	**$1** 001IT\ICCU\DDS\0756663 $1 2001$aIn-a-gadda-da-vida$fIron Butterfly$v2 **$1** 71202$aAtlantic$c <etichetta discografica>$3 IT\ICCU\DDSV\000152$4650 **$1** 71002$aIron Butterfly$3IT\ICCU\DDSV\018382$4590
801	\	3	**$a** IT **$b** ICCU **$c** 20180917

Tabella 6

Un altro caso ancora è quello delle 36 registrazioni relative agli spartiti di Giulio Gadda, un compositore che è stato anche organista della cattedrale milanese dal 1863 al 1903. Nel campo 500, relativo al titolo uniforme, è stato inserito anche il nome dell'autore e questo ha fatto sì che le registrazioni venissero recuperate attraverso il canale Titolo:

LDR			01625ncm1 22003613i 4500
001			IT\ICCU\MUS\0104030
200	1	\	**$a** Tre divertimenti per pianoforte **$f** di Giulio Gadda
500	1	0	**$a** Tre divertimenti **$3** IT\ICCU\CMP\0027018 **$9** Gadda, Giulio
801	3	\	**$a** IT **$b** ICCU **$c** 20180917

Tabella 7

Nella tabella che segue sono elencati i 10 autori che compaiono più spesso nelle registrazioni, con i titoli e i soggetti a essi associati (nella colonna Titolo sono state conteggiate le registrazioni recuperate attraverso il canale Titolo relative a opere di o sull'autore).

Autore	Canale Autore	Canale Titolo	Canale Soggetto
Gedda, Nikolai	829	0	0
Gadda, Carlo Emilio	524	487	188
Gadda Conti, Piero	108	3	2
Gadda Conti, Giuseppe	65	0	3
Gadda, Giulio	42	0	0
Gadda, Monica	37	0	0
Gadda, Giuseppe	26	0	1
Gadda Giuseppe (1822-1901)	14	3	3
Gadda, Gianluca	13	0	0
Gadda Conti, Rosy	12	0	0

Tabella 8

Nikolai Gedda è il nome d'arte di Harry Gustaf Nikolai Gädda, un tenore svedese morto nel 2017 che ha inciso dischi con i più grandi direttori del secolo scorso. Il numero delle registrazioni che fanno capo a Gedda si spiega con il fatto che si tratta in realtà di 340 monografie (album) e 489 spogli (tracce musicali). Lo stesso, in scala minore, vale per Giulio Gadda. In questo caso abbiamo 21 monografie e 21 spogli.
Possiamo quindi dire che sul totale delle 2.395 registrazioni recuperate, 972 sono relative a manifestazioni che contengono opere di o su Carlo Emilio Gadda e 340 a manifestazioni che contengono opere di Nikolai Gedda. Il terzo autore più rappresentato è Piero Gadda Conti, cugino di secondo grado di Carlo Emilio: 110 registrazioni sulle 2.395 recuperate sono relative a manifestazioni che contengono opere di o su lui. Gli altri 41 autori che contengono il termine "Gadda" in una delle forme del nome utilizzate per indicarli hanno numeri molto più bassi.
Dato che la maggior parte delle registrazioni è stata recuperata attraverso il canale Autore e che tra gli autori quello più rappresentato è Carlo Emilio Gadda, sarebbe ragionevole supporre che l'algoritmo utilizzato nel catalogo dell'Indice SBN per stabilire la rilevanza delle registrazioni riportasse in testa all'elenco dei risultati quelle relative alle edizioni più recenti dei

capolavori dello scrittore milanese (del *Pasticciaccio* sono state recuperate 105 registrazioni, della *Cognizione del dolore* 55).
In Google, la prima pagina dei risultati di una ricerca su "Gadda" comprende nove collegamenti a pagine su Carlo Emilio Gadda e un collegamento al sito web di un istituto di istruzione superiore intitolato allo scrittore. Una ricerca analoga su Amazon mostra nelle prime dieci registrazioni nove opere di Gadda e uno studio recente su di lui di Mauro Bersani (*Gadda*, Torino, Einaudi, 2012). Lo stesso vale per IBS, dove però il saggio è quello di Paola Italia su *Come lavorava Gadda* pubblicato a Roma da Carocci nel 2017 (il confronto con i motori di ricerca e i siti di commercio elettronico è uno dei metodi per valutare l'efficacia della rilevanza dal punto di vista del sistema o dell'algoritmo).[24]
La situazione nelle prime 5 pagine dei risultati del Catalogo dell'Indice SBN è rappresentata nella tabella che segue (il titolo è riportato solo nel caso di opere anonime o quando non sia presente il termine "Gadda" nel nome dell'autore):

Autore	Titolo	Numero registrazione	Canale Autore	Canale Titolo	Canale Soggetto
Gadda Conti, Piero		1	•	•	•
Gadda Conti, Piero		2-21	•		
Gadda, Emanuele		22	•		
Gadda, Giuseppe		23	•		
Gadda, Conti Piero		24-38	•		
Gadda, Giulio		39	•		
Gadda		40	•		
Gadda, Giuseppe		41	•		
Gadda, Carlo Emilio		42	•		
Gadda, Walter		43	•	•	•
	Per l'inaugurazione del monumento a Giuseppe Gadda in Rogero	44	•	•	•
Gadda Conti, Piero		45	•		
Gadda, Giulio		46	•		
Bersani, Mauro	Gadda	47		•	•
Patrizi, Giorgio	Gadda	48		•	•
Gadda Conti, Piero		49-50	•		

Tabella 9

Per trovare la prima registrazione relativa a una manifestazione che contiene un'opera di Gadda si deve scorrere l'elenco fino alla seconda registrazione della quinta pagina (ogni pagina contiene dieci registrazioni). La registrazione 42 è relativa a una traduzione tedesca dei racconti pubblicata dall'editore Suhrkamp nel 1965. Si torna poi ad altri Gadda e solo dall'ottava pagina in poi comincia l'elenco delle registrazioni relative a manifestazioni che contengono opere di Carlo Emilio Gadda. Per Nikolai Gedda bisogna arrivare fino alla pagina 176.
Non esiste documentazione che possa essere consultata sull'algoritmo di rilevanza usato all'interno del Catalogo dell'Indice SBN. Abbiamo quindi cercato di ricostruirlo partendo dal concetto centrale per il recupero delle informazioni: l'indice inverso.
Nel caso di un catalogo di una biblioteca creare un indice inverso significa individuare tutti i termini presenti nei campi delle registrazioni bibliografiche associati ai diversi canali ed elencarli in ordine alfabetico con accanto gli identificativi delle registrazioni in cui sono presenti.
Per la registrazione relativa alle *Opere di Carlo Emilio Gadda* riportata nella Tabella 2 avremo ad esempio un indice per il canale Autore ricavato dai campi 700 e 702 che potrebbe essere rappresentato in questo modo:

Termine	Identificativo
Carlo	IT\ICCU\CFI\0155806
Dante	IT\ICCU\CFI\0155806
Emilio	IT\ICCU\CFI\0155806
Gadda	IT\ICCU\CFI\0155806
Isella	IT\ICCU\CFI\0155806

Tabella 10

La nostra ipotesi è che, a seguito di una ricerca semplice, il sistema di IR del catalogo dell'Indice SBN inserisca gli identificativi delle registrazioni presenti negli indici inversi dei canali Autore, Titolo, Soggetto e Descrizione Dewey associate al termine o ai termini della query in un unico elenco, dopo averli naturalmente deduplicati. A questo punto a ogni registrazione sarebbe assegnato un punteggio sulla base dei canali in cui è presente l'identificativo e delle occorrenze del termine all'interno dei singoli canali (una stessa registrazione può avere più autori con lo stesso nome e più di un soggetto sullo stesso argomento). È anche probabi-

le che ai vari campi riconducibili a uno stesso canale sia stato assegnato un peso diverso. Questo spiegherebbe perché le registrazioni relative a Nikolai Gedda siano state relegate nelle ultime posizioni. Il campo UNIMARC utilizzato per la forma variante del nome dell'autore potrebbe essere stato tenuto meno in considerazione rispetto a quello per la forma accettata.
Per molte registrazioni il punteggio risulterebbe però identico. Nel caso del nostro esempio, non ci sarebbero differenze tra i punteggi assegnati alle registrazioni relative a opere di Piero Gadda Conti e quelli assegnati a registrazioni relative a opere di Carlo Emilio Gadda.
Trovandosi di fronte a registrazioni che hanno lo stesso punteggio, il sistema restituisce un elenco di risultati che, dal punto di vista di chi ha formulato la query, è assolutamente casuale.

SebinaYOU (Università degli studi di Milano)

La ricerca semplice di SebinaYOU, a differenza di quella del Catalogo dell'Indice SBN, prende in considerazione tutti i canali disponibili in quella avanzata: Autore, Titolo, Soggetto, Classificazione, Abstract, ISBN/ISSN, Editore, Luogo di pubblicazione, Anno di pubblicazione, Lingua, Paese, Tecnica, Genere di pubblicazione, Tipologia, Natura, Possessore e Personaggio.
La query formulata attraverso la ricerca semplice e per cui è stato utilizzato solo il termine "Gadda" recupera 214 registrazioni, tutte attraverso i primi 3 canali:

Canale di ricerca	Registrazioni recuperate	Registrazioni che avrebbero dovuto essere recuperate	Registrazioni attinenti	Recupero	Precisione
Autore	128	128	128	1	1
Titolo	86	86	86	1	1
Soggetto	19	19	19	1	1
Totale	**214**	**214**	**214**	**1**	**1**

Tabella 11

Il recupero in questo caso è pari a 1, come anche la precisione. Sono state infatti recuperate tutte e solo le registrazioni che contengono il termine "Gadda" in uno dei campi in cui sono strutturate.
L'ordine in cui sono elencate le prime 50 registrazioni è quello riportato in Tabella 12.
La prima registrazione è relativa alle *Opere di Carlo Emilio Gadda*, ma la seconda e la terza sono relative

Autore	Titolo	Numero Registrazione	Canale Autore	Canale Titolo	Canale Soggetto
Gadda, Carlo Emilio		1	•	•	•
Gadda, Carlo Emilio		2-3	•		
Bersani, Mauro	Gadda	4		•	
Patrizi, Giorgio	Gadda	5		•	
Gadda, Leonardo		6	•		
Rinaldi, Rinaldo	Gadda	7		•	
Pecoraro, Aldo	Gadda	8		•	•
Donnarumma, Raffaele	Gadda modernista	9		•	
Italia, Paola	Come lavorava Gadda	10		•	
Terzoli, Maria Antonietta	Gadda	11		•	
Gadda, Giulio		12	•		
Gadda, Giovanni		13	•		
Gadda, Carlo Emilio		14-20	•		
Matt. Luigi	Gadda	21		•	
Gadda, Carlo Emilio		22-24	•		
	Gadda: produttività e scrittura	25		•	
Gadda Conti, Piero		26-27	•		
Gadda, Carlo Emilio		28-29	•		
Faravelli, Gian Carlo	Ritratto di Gadda	30		•	
Gadda, Carlo Emilio		31	•		
Baldi, Guido	Carlo Emilio Gadda	32		•	
Gadda, Carlo Emilio		33	•		

Tabella 12

a *Gadda in "zona Guarneri"* e *Le carte militari di Gadda*, due raccolte di lettere e appunti che non sono certo tra i capolavori dello scrittore.
I risultati sembrano migliori rispetto a quelli del Catalogo dell'Indice SBN solo per il fatto che l'insieme delle registrazioni è molto più ridotto e omogeneo (164 sulle 214 registrazioni recuperate sono relative a manifestazioni che contengono opere di o su Carlo Emilio Gadda, contro le 972 su 2.395 del Catalogo dell'Indice SBN), ma in realtà l'ordine è altrettanto casuale.
Anche in questo caso calcolare la rilevanza basandosi quasi esclusivamente sulle occorrenze del termine o dei termini utilizzati nella query invece che su un'analisi dei bisogni informativi che i diversi canali si propongono di soddisfare ha dato risultati poco soddisfacenti.

Conclusioni

Il Catalogo dell'Indice SBN e SebinaYOU utilizzano come motore di ricerca Lucene SOLR, al cui cuore c'è l'idea di un documento che contiene campi di testo: "This flexibility allows Lucene's to be independent of the file format. Text from PDFs, HTML, Microsoft Word, Mind Maps, and OpenDocument documents, as well as many others (except images), can all be indexed as long as their textual information can be extracted".
Utilizzare un'applicazione progettata per indicizzare documenti di testo con registrazioni che hanno una struttura molto più articolata e, soprattutto, sono molto più brevi (in media 700 caratteri) fa sì che, da un lato, non si sfruttino le possibilità offerte dai dati presenti nelle registrazioni stesse e, dall'altro, non sia possibile applicare i meccanismi utilizzati per calcolare la rilevanza in documenti più estesi. È il caso, ad esempio, della prossimità ("Lucene includes a feature to perform a fuzzy search based on edit distance").
Gli algoritmi utilizzati nel Catalogo dell'Indice SBN e in SebinaYOU, messi a confronto con l'ordine dei risultati dei motori di ricerca e dei siti di commercio elettronico, rivelano la loro inadeguatezza.
Sembrerebbe che la possibilità di ordinare per rilevanza le registrazioni dei cataloghi delle biblioteche sia una di quelle idee che avrebbero dovuto essere uccise dall'evidenza, ma che si rifiutano di morire.[25] In realtà, per poterlo affermare sarebbe necessario svolgere ulteriori indagini con altri campioni di dati e, soprattutto, analizzare i *log* che registrano le interazioni tra gli utenti e i cataloghi.[26]
È però quasi impossibile estrarre registrazioni dai cataloghi delle biblioteche. Attraverso il *client* Z39.50 si può ottenere dal Catalogo dell'Indice SBN solo una registrazione in formato UNIMARC per volta, mentre SebinaYOU esporta le registrazioni solo nel formato SUTRS (Simple Unstructured Text Record Syntax), poco adatto alle elaborazioni.
Per quanto riguarda i *log*, quelli riportati nelle pagine delle statistiche del Catalogo dell'Indice SBN non sono utilizzabili per questo tipo di analisi. La pagina relativa alle statistiche di SebinaYOU non è consultabile liberamente.
Uno studio più ampio, effettuato in collaborazione con i bibliotecari e gli studiosi di scienza dell'informazione, permetterebbe a chi progetta questi sistemi migliorare la presentazione dei risultati a beneficio degli utenti. Nel frattempo, sarebbe possibile sfruttare gli indici inversi impiegati per generare i filtri allo scopo di rendere più semplice la consultazione.
I filtri compaiono a fianco dell'elenco dei risultati e danno la possibilità di raffinare la ricerca: tecnicamente sono chiamati faccette. Nel Catalogo dell'Indice SBN le faccette riguardano il livello bibliografico, il tipo di documento, l'autore, il soggetto, il codice Dewey, il luogo di pubblicazione, l'editore, l'anno di pubblicazione, la collezione, la lingua, il paese, il titolo uniforme, la forma musicale e la biblioteca.
Nel catalogo dell'Università degli studi di Milano sono previste faccette per la biblioteca, l'autore principale, l'anno di pubblicazione, la lingua e il materiale. Quelle relative al soggetto, al codice Dewey, al titolo uniforme e alla forma musicale non sono state inserite perché queste informazioni non sono sempre presenti. Nessuna delle 5 registrazioni del catalogo relative al *Pasticciaccio*, ad esempio, ha un titolo uniforme. Questo vale anche per il Catalogo dell'Indice SBN, dove delle 111 registrazioni relative al *Pasticciaccio* solo 14 hanno il titolo uniforme.
Attraverso gli indici inversi impiegati per le faccette sarebbe possibile stabilire a quale entità si è voluto più probabilmente fare riferimento con "Gadda", ordinare i risultati sulla base di questo presupposto e contemporaneamente creare un raggruppamento per ogni canale, seguendo l'esempio di Apple Music e Spotify.
Se si cerca "Dead Flowers" in questi servizi di streaming musicale si ottiene una serie di registrazioni suddivise per Migliori risultati, calcolati sulla base della popolarità tra gli utenti, Brani, Album, Videoclip e Artisti. Nei termini dell'IFLA LRM si potrebbe dire che i risultati sono raggruppati insieme in base al fatto di essere manifestazioni unitarie, manifestazioni che aggregano, persone o agenti collettivi. Le manifestazioni unitarie sono a loro volta suddivise per formato (Brani, Videoclip), mentre quelle che aggregano sulla base del fatto di essere state pubblicate o di essere creazioni degli utenti (Album, Playlist).
All'interno di ogni raggruppamento è possibile scorrere un elenco dei brani, degli album, dei videoclip o degli artisti con questo titolo o con questo nome.
Se al posto "Dead Flowers" si cerca "Sticky Fingers", l'album dei Rolling Stones che contiene un brano con quel titolo, cambia l'ordine in cui sono presentati i diversi raggruppamenti, con gli album che in questo caso precedono i brani.

Lo scorrimento all'interno di raggruppamenti è considerato una valida alternativa alla ricerca per parole chiave, soprattutto quando gli utenti non sono sicuri sui termini da utilizzare per esprimere i loro bisogni informativi.[27] Si tratta infatti di un modo estremamente efficace di guidare chi consulta un catalogo verso ciò che più lo interessa. Sarebbe importante capire perché è utilizzato dai servizi di streaming musicale e non è stato ancora preso in considerazione dalle biblioteche, che pure dispongono di un modello concettuale dell'universo bibliografico ormai consolidato che avrebbe dovuto suggerire loro già da tempo questo tipo di approccio.

NOTE

[1] Christopher D. Manning, Prabhakar Raghavan, Hinrich Schütze, *Introduction to Information Retrieval*, Cambridge, Cambridge University Press, 2008, p. 1.

[2] Djoerd Hiemstra, *Information Retrieval Models*, in Ayse Göker, John Davies, *Information Retrieval: Searching in the 21st Century*, Chichester, Wiley, 2009, p. 1-19, p. 1-2.

[3] Ricardo Baeza-Yates, *Modern Information Retrieval. The Concepts and Technology behind Search*, Harlow, AddisonWesley, 2010[2], p. 1.

[4] *Ibidem*.

[5] Pia Borlund, *The Concept of Relevance in IR*, "Journal of The American Society for Information Science and Technology", 2003, August, p. 913-925, p. 914.

[6] Carlo Bianchini, *"Funziona come Google, vero?". Prima indagine sull'interazione utente-catalogo nella biblioteca del Dipartimento di musicologia e beni culturali (Cremona) dell'Università di Pavia*, "AIB Studi", 57 (2017), 1, p. 23-49, p. 38.

[7] Tifko Saracevic, *Relevance Reconsidered*, in *Proceedings CoLIS 2, Second International Conference on Conceptions of Library and Information Science. Integration in Perspective, October 13-16, 1996*, editors Peter Ingwersen, Niels Ole Pors, Copenhagen, Royal School of Librarianship, 1996, p. 201-218.

[8] David Bade, *Relevance Ranking Is Not Relevance Ranking or, When the User Is Not the User, the Search Results Are Not Search Results*, "Online Information Review", 31 (2007), 6, p. 831-844.

[9] Elaine G. Toms, Heather L. O'Brien, Rick Kopak, Luanne Freund, *Searching for Relevance in the Relevance of Search*, in *Context: Nature, Impact, and Role. 5th International Conference on Conceptions of Library and Information Sciences, CoLIS 2005, Glasgow, UK, June 4-8, 2005. Proceedings*, F. Crestani, I. Ruthven (eds), Berlin, Heidelberg, Springer, 2005, p. 59-78, p. 67.

[10] International Federation of Library Associations and Institutions, *IFLA Library Reference Model. A Conceptual Model for Bibliographic Information*, Pat Riva, Patrick Le Boeuf, Maja Žumer, Consolidation Editorial Group of the IFLA FRBR Review Group, Den Haag, IFLA, 2017, p. 97-99.

[11] Jennifer Knievel, Jina Choi Wakimoto, Sara Holladay, *Does Interface Design Influence Catalog Use? A Case Study*, Boulder, University of Colorado, 2009 ("University Libraries Faculty & Staff Contributions", 46).

[12] Elaine Svenonius, *The Intellectual Foundation of Information Organization, Cambridge*, Massachusetts, London, The MIT Press, 2000. Traduzione italiana di Maria Letizia Fabbrini, *Il fondamento intellettuale dell'organizzazione dell'informazione*, introduzione di Mauro Guerrini, Firenze, Le Lettere, 2008, p. 89.

[13] Shiyali Ramamrita Ranganathan, *Reference Service*, London, Asia Publishing House, 1961[2]. Traduzione italiana *Il servizio di reference*, a cura di Carlo Bianchini, Firenze, Le Lettere, 2009, p. 81.

[14] Lucia Sardo, *La lingua del catalogatore (parte 1). L'autore*, "Bibliothecae.it", 6 (2017), 2, p. 339-383.

[15] Christine L. Borgman, *Why are Online Catalogs Hard to Use? Lessons Learned from Information Retrieval Studies*, "Journal of The American Society For Information Science", 37 (1986), 6, p. 387-400.

[16] Christiane Behnert, *Relevance Ranking. State of the Art in Web Search and Library Catalogs*, Hamburg, Hamburg University of Applied Sciences, Department of Information, 2015, p. 6.

[17] Dirk Lewandowski, *Using Search Engine Technology to Improve Library Catalogs*, "Advances in Librarianship", 32 (2010), p. 35-54.

[18] Dan Wu, Renmin Bi, *Impact Of Device On Search Pattern Transitions: A Comparative Study Based On LargeScale Library Opac Log Data*, "The Electronic Library", 35 (2017), 4, p. 650-666.

[19] Christopher D. Manning, Prabhakar Raghavan, Hinrich Schütze, *Introduction to Information Retrieval*, cit., p. 152.

[20] International Federation of Library Associations and Institutions, *IFLA Library Reference Model. A Conceptual Model for Bibliographic Information*, cit., p. 15.

[21] Un elenco completo di campi del formato è disponibile sul sito dell'International Federation of Library Associations and Institutions nella sezione dedicata all'UNIMARC Strategic Programme.

[22] Giuliana Sapori, *Regole secondo il Nuovo soggettario*, disponibile all'indirizzo manualesapori.unimi.it. Il soggettario è quello della Biblioteca nazionale centrale di Firenze: *Nuovo soggettario. Guida al sistema italiano di indicizzazione per sog-*

getto. Prototipo del Thesaurus, Milano, Editrice Bibliografica, 2007.

[23] Anna Lucarelli, *Servizio bibliotecario nazionale e indicizzazione semantica: un decollo difficile, una rotta da condividere*, "Bibelot", 22 (2016), 3, p. 10-16.

[24] Elaine G. Toms, Heather L. O'Brien, Rick Kopak, Luanne Freund, *Searching for Relevance in the Relevance of Search*, cit., p. 68.

[25] Paul Krugman, *The Ultimate Zombie Idea*, "The New York Times", 3 novembre 2012.

[26] Eng Pwey Lau, Dion Hoe-Lian Goh, *In Search of* Query *Patterns: A Case Study of a University OPAC*, "Information Processing and Management", 42 (2006), p. 1316-1329.

[27] Christopher D. Manning, Prabhakar Raghavan, Hinrich Schütze, *Introduction to Information Retrieval*, cit., p. 352.

ABSTRACT

The article describes the use of algorithms for relevance ranking in Italian OPACs. The results show that these algorithms are not able to order the bibliographic records in order to respond to the information needs of the user who made the search. The solutions adopted by Amazon and by the streaming music services (Apple Music and Spotify), which could also be used in the library catalogs, are therefore considered, since there would be no need to add data to those already present in the records.

DOI: 10.3302/0392-8586-201901-018-1