Nuove funzionalità degli OPAC e *relevance ranking*

di Maria Teresa Biagetti

Introduzione

Il dibattito sull'opportunità di definire nuovi modelli di OPAC e di dotarli di funzionalità analoghe a quelle offerte dai motori di ricerca nel Web, dei quali vengono apprezzate la semplicità dell'uso e la chiarezza nella presentazione dei risultati delle ricerche, ha costituito negli ultimi anni un'occasione di riflessione ed approfondimento scientifici per il settore della *Library and Information Science*, ed è stato seguito con costante attenzione anche in Italia sia da parte dell'ambiente scientifico che di quello professionale[1].

Le indagini sul comportamento degli utenti durante le sessioni di ricerca (*Transaction-log analysis),* in particolare di ricerca semantica, attraverso un OPAC, hanno evidenziato che i modelli mentali che gli utenti adottano nelle ricerche attraverso gli OPAC sono significativamente influenzati dall'uso delle funzionalità offerte dai motori di ricerca nel Web, e che la tendenza degli utenti a preferire la ricerca per parole chiave alla ricerca per soggetti è sostenuta anche dalla sequenza di presentazione delle opzioni di ricerca nel menu, che infatti frequentemente privilegia l'offerta della ricerca per parole, disponendola in prima posizione.

MARIA TERESA BIAGETTI, Università di Roma "La Sapienza", Scuola speciale per archivisti e bibliotecari, viale Regina Elena 295, 00161 Roma, e-mail mariateresa.biagetti@uniroma1.it

L'articolo costituisce un ampliamento, con modifiche, approfondimenti e rielaborazioni, dell'intervento dal titolo *Pertinence perspective and OPAC enhancement*, presentato alla 11. Conferenza internazionale di ISKO (Roma 23-26 febbraio 2010), e pubblicato in una versione abbreviata negli Atti, *Paradigms and conceptual systems in knowledge organization: proceedings of the eleventh international ISKO Conference 23-26 February 2010, Roma*, edited by Claudio Gnoli and Fulvio Mazzocchi, Würzburg: Ergon Verlag, 2010, alle p. 334-340.

Ultima consultazione dei siti web: 13 gennaio 2011.

1 Lo scenario italiano di riferimento è costituito principalmente dai lavori di Paul G. Weston – Salvatore Vassallo, "*... e il navigar m' è dolce in questo mare": linee di sviluppo e personalizzazione dei cataloghi*, in: *La biblioteca su misura: verso la personalizzazione del servizio*, a cura di Claudio Gamba e Maria Laura Trapletti, Milano: Editrice Bibliografica, 2007, p. 130-167, di Riccardo Ridi, *La biblioteca come ipertesto: verso l' integrazione dei servizi e dei documenti*, Milano: Editrice Bibliografica, 2007, p. 116-148, e dagli interventi di Paul Gabriele Weston, *Caratteristiche degli opac e strategie delle biblioteche*, di Giovanni Bergamin, *OPAC: migliorare l' esperienza degli utenti*, di Pino Buizza, *Gli opac: funzionalità e limiti nel mondo del web*, di Claudio Gnoli, *Blopac semantici*, tutti su «Bibliotime», XI (2008), n. 1, e recentemente di Antonella Iacono, *Opac, utenti, rete. Prospettive di sviluppo dei cataloghi elettronici*, «Bollettino AIB», 50 (2010), n. 1/2, p. 69-86.

L'influenza della familiarità degli utenti degli OPAC con la ricerca nel Web attraverso i motori è evidente: i ricercatori nel Web sono gli stessi che poi si rivolgono ai cataloghi elettronici e tendono a mantenere la stessa modalità di ricerca, tranne, come è stato rilevato in alcune indagini, che nell'uso degli operatori booleani, frequente nella ricerca in rete e scarso in quella negli OPAC[2]. Gli utilizzatori degli OPAC si aspettano di trovare alcune delle funzioni tipiche dei motori di ricerca e delle librerie online, come il controllo dello *spelling* delle parole, la possibilità di usare il linguaggio naturale nelle *queries* e di operare attraverso *browsing*, e soprattutto la disposizione dei risultati secondo *relevance*, ma anche di poter utilizzare il *relevance feedback*, la possibilità di riformulare le *queries* sulla base della rilevanza per l'utente di una risposta ottenuta, una tecnica nata nell'ambito dell'*information retrieval* ed applicata ai motori di ricerca. Yu e Young, nella loro indagine, hanno suggerito di operare per il miglioramento degli OPAC in particolare attraverso l'adozione della funzione di *browsing* e della disposizione dei risultati secondo la rilevanza, stabilita in base a criteri particolarmente adatti all'utente dell'OPAC, tra cui avranno un peso la data di pubblicazione del documento, le intestazioni per soggetto e soprattutto, per quanto riguarda le monografie, la frequenza dei termini nei sommari.

Negli ultimi anni, oltre alla presa di coscienza del fatto che il mantenimento del formato originale, basato su MARC, ha reso molto difficile provvedere gli OPAC di funzionalità ipertestuali che consentano, grazie all'uso dei *link*, di navigare tra i nodi semantici, sono state proposte alcune strategie di avanzamento e di miglioramento, tra le quali ha assunto un rilievo particolare la presentazione personalizzata dei servizi, realizzata attraverso l'analisi del comportamento di ricerca degli utenti e l'organizzazione dei risultati in base a *cluster* definiti, spesso utilizzando strategie di *data mining*[3], ed in particolare di *bibliomining*, applicato al miglioramento dei servizi delle biblioteche digitali[4].

Le inchieste realizzate da OCLC negli ultimi anni[5], in particolare quella dedicata all'analisi della percezione delle biblioteche da parte degli utenti svolta nel 2005 in Australia, Canada, India, Singapore, Regno Unito e Stati Uniti[6], che ha messo in evidenza i comportamenti e le preferenze degli utenti durante la ricerca, hanno contribuito a chiarire che le funzionalità offerte dalla ricerca in rete attraverso i motori raccolgono sempre maggiori consensi. L'indagine compiuta dalla California Digital Library tra il giugno 2005 ed il giugno 2006 con *The Melvyl Recommender Project*[7] ha evidenzia-

2 Holly Yu – Margo Young, *The impact of Web search engines on subject searching in OPAC*, «Information technology and libraries», December 2004, p. 168-180.

3 Gerald Benoit, *Data mining*, «Annual Review of Information Science and Technology», 36 (2002), n. 8, p. 265-310.

4 Scott Nicholson, *The basis for bibliomining: frameworks for bringing together usage-based data mining and bibliometrics through data warehousing in digital library services*, «Information processing & management», 42 (2006), n. 3, p. 785-804.

5 Le inchieste sono analizzate in modo approfondito in Paul G. Weston – Salvatore Vassallo, *"... e il navigar m' è dolce in questo mare"* cit. Andrea Marchitelli, nel suo intervento *La biblioteca nella percezione degli utenti: i risultati di tre indagini di OCLC*, «AIB Notizie», 20 (2008), n. 4, <http://www.aib.it/aib/editoria/n20/0413.htm3>, ha offerto la traduzione in italiano di tre articoli pubblicati su «American libraries» tra il 2004 ed il 2007, che contengono la presentazione dei tre Report di OCLC sull'argomento.

6 *Perceptions of libraries and information resources. A report to the OCLC membership*, Dublin (OH): OCLC, 2005 <http://www.oclc.org/reports/2005perceptions.htm>.

7 California Digital Library, *The Melvyl recommender project final report*, July 2006 <http://www.cdlib.org/services/publishing/tools/xtf/melvyl_recommender/report_docs/Mellon_final.pdf>.

to la necessità di coprire la distanza che si è instaurata tra OPAC tradizionali e principali motori di ricerca commerciali agendo in cinque settori: adozione di sistemi di ricerca che scandaglino i testi, come XTF, eXtensible Text Framework, robusto software *open source* che permette l'accesso ai contenuti digitali, consente all'utente di compiere una ricerca per parole all'interno dei testi, inserite nel loro contesto, di compiere ricerche nelle parti di una monografia, ad esempio in un capitolo, e di utilizzare gli operatori booleani; impiego di sistemi di correzione dello *spelling* delle parole, di interfacce utente rinnovate, di *recommendation systems* analoghi a quelli delle librerie on line, e soprattutto l'adozione di un sistema di *relevance ranking* dei risultati.

L'arricchimento dei *record* catalografici, attraverso l'inserimento dei TOC (*Tables of contents*) e la possibilità di accedere alle recensioni ed al testo stesso delle pubblicazioni tramite il collegamento con le versioni elettroniche gestite da *repositories* affidabili (*Web access to works in the public domain*), è stato uno degli obiettivi più significativi dei progetti avviati dalla Library of Congress nell'ambito del programma di miglioramento delle funzionalità degli OPAC e del superamento delle tradizionali potenzialità, avendo come modello i motori di ricerca e le librerie on line[8].

La Library of Congress in quegli anni ha messo in discussione l'opportunità di continuare a spendere fondi molto consistenti per mantenere complessi strumenti catalografici, dal momento che si possono ritrovare facilmente i documenti usando le parole chiave ed i motori di ricerca[9]. Nel Rapporto preparato nel 2006 per la Library of Congress, Karen Calhoun[10] insisteva sul declino del catalogo come strumento di ritrovamento dei documenti e, oltre a suggerire di abbandonare l'indicizzazione semantica di tipo tradizionale, realizzata dai bibliotecari, e sostituirla con la classificazione automatica e con la ricerca attraverso le parole chiave, ed a sperimentare nuove funzionalità incentrate su FRBR, raccomandava l'integrazione dei cataloghi con altri strumenti di ricerca e, in particolare, l'organizzazione dei risultati usando il *relevance ranking*, dal momento che lo scopo è quello di soddisfare le necessità degli utenti, definiti come *information seekers*, che preferiscono usare i motori di ricerca e sono abituati alle modalità offerte dalle librerie on line ed all'agevole ritrovamento di documenti in *full text*.

Nel suo esame critico del r*eport* di Calhoun, Thomas Mann[11], bibliotecario della Library of Congress, oltre a stigmatizzare l'uso costante di una terminologia tipicamente imprenditoriale ed il riferimento a modelli concettuali di tipo commerciale

8 Il resoconto dei progetti della Library of Congress è stato presentato da John D. Byrum Jr. alla 71. IFLA Conference (Oslo, 14-18 agosto 2005), ed è stato pubblicato nella traduzione italiana su «Biblioteche oggi», 10 (2005), p. 5-14.

9 Deanna B. Marcum, *The future of cataloging. Address to the Ebsco leadership seminar*, Boston, Massachusetts January 16, 2005, <http://www.guild2910.org/marcum.htm>.

10 Karen Calhoun, *The changing nature of the catalog and its integration with other discovery tools*, Final Report, March, 17, 2006, prepared for the Library of Congress, <http://www.loc.gov/catdir/calhoun-report-final.pdf>.

11 Thomas Mann, *The changing nature of the catalog and its integration with other discovery tools. Final Report. March 17, 2006. Prepared for the Library of Congress by Karen Calhoun. A critical review*, prepared for AFSCME 2910, April 3, 2006, <http://www.guild2910.org/AFSCMECalhounReview.pdf>. L'intervento è stato tradotto in italiano e pubblicato sul «Bollettino AIB», 46 (2006), n. 3, p. 186-205, preceduto da un articolo di Alberto Petrucciani, *La catalogazione, il mercato e la fiera dei luoghi comuni*, Ivi, p. 177-185.

applicati alle biblioteche ed ai cataloghi, metteva in evidenza la necessità di distinguere tra i "ricercatori di informazioni veloci", i quali possono trarre beneficio dall'entrare in contatto con qualsiasi tipo di informazione, e gli "studiosi", i quali affrontano le ricerche a livello scientifico, desiderano trovare su un argomento i libri più recenti classificati nel contesto della letteratura esistente sullo stesso argomento, e per il cui soddisfacimento è necessario quindi realizzare una indicizzazione tradizionale, usare precise categorie concettuali ed un vocabolario controllato. La necessità primaria per gli studiosi è mantenersi aderenti al contesto concettuale ed evitare il più possibile nei risultati delle ricerche il "rumore", frequente invece nelle modalità di ricerca attraverso le parole chiave. Mann ribadisce che la ricerca per parole chiave ed i risultati disposti utilizzando un ordinamento per rilevanza alla maniera di Google non possono affatto risolvere i problemi specifici della ricerca scientifica degli utenti.

I ricercatori, infatti, «wish to avoid having to sort through huge lists or displays – from any source – in which relevant materials are buried within inadequately-sorted mountains of chaff having the "right" keywords in the wrong conceptual contexts» (Mann, p. 8).

Strategie per il miglioramento degli OPAC: il *relevance ranking*

Tra gli strumenti che si propone di utilizzare per ampliare le funzionalità degli OPAC e renderli il più possibile simili ai motori di ricerca, *relevance ranking*, cioè la presentazione dei risultati della ricerca secondo una graduatoria basata su una rilevanza presunta, strategia ampiamente utilizzata dai motori di ricerca e quindi molto familiare ai ricercatori di documenti in rete, e *relevance feedback*, la riformulazione automatica di *query* in base all'informazione esplicita di rilevanza di un documento da parte dell'utente, sono molto citati nel dibattito internazionale che si è sviluppato negli ultimi anni intorno all'argomento.

Di quale tipo di rilevanza si tratta?

La disposizione dei risultati di una ricerca secondo *relevance* è fondamentalmente determinata dal prodotto della frequenza dei termini, TF, e della frequenza inversa nei documenti, IDF. TF riguarda la frequenza dei termini nel *full-text* di ciascun documento, o nel *record* catalografico che lo descrive; IDF, *Inverse Document Frequency*, riguarda invece la frequenza dei termini in una base di dati o in una raccolta di documenti, e prevede la determinazione di un peso maggiore nel caso in cui il termine ricorra in un numero basso di documenti, cioè sia poco frequente. Il computo della frequenza inversa è utilizzato allo scopo di bilanciare gli effetti della frequenza dei termini. IDF di un termine viene calcolato considerando il rapporto tra il numero dei documenti presenti nella collezione ed il numero dei documenti nei quali si trova il termine.

La misura conosciuta come IDF, *Inverse Document Frequency*, che prevede il computo del peso dei termini, è stata proposta nel 1972 da Karen Spärck Jones[12] come approccio innovativo nell'ambito dell'*information retrieval*. La specificità dei termini impiegati nell'indicizzazione viene valutata su base statistica, come funzione dell'uso più che come funzione del significato; i termini ricevono un peso in base alla frequenza con cui occorrono nella collezione, ed in tal modo i documenti ritrovati grazie ai termini poco frequenti e più specifici hanno più valore di quelli ritrovati in

12 Karen Spärck Jones, *A statistical interpretation of term specificity and its application in retrieval*, «Journal of documentation», 28 (1972), n. 1, p. 11-21.

base a termini frequenti. Il valore del ritrovamento di un documento attraverso un termine di indice viene così collegato con la frequenza di quel termine nella collezione. OKAPI BM25 è una versione più avanzata della misura basata su TF-IDF, nella quale, oltre alla frequenza dei termini ed alla frequenza inversa, vengono prese in considerazione la lunghezza di ciascun documento e la lunghezza media dei documenti della collezione.

IDF è la versione semplificata, applicabile quando non si dispone di alcuna informazione di rilevanza per l'utente, della *Relevance Weighting Theory*[13], elaborata alcuni anni più tardi da Stephen Robertson e Karen Spärck Jones, che utilizza invece il giudizio di rilevanza da parte dell'utente ed affida pesi diversi ai termini usati nelle ricerche, seguendo un'impostazione di tipo probabilistico. I documenti devono essere ordinati in base alla probabilità della rilevanza rispetto ai termini usati nelle richieste degli utenti. L'ordinamento secondo la probabilità di rilevanza si può realizzare assegnando pesi ai termini utilizzati nelle *query*; ciascun documento riceve quindi un punteggio che deriva dalla somma dei pesi che sono stati attribuiti ai termini usati per la *query* e presenti nel documento, o scelti per indicizzarlo. Interpretare la misura IDF in termini di probabilità che un documento contenga un certo termine, costituisce un avvicinamento all'approccio di tipo probabilistico all'*information retrieval*, di cui van Rijsbergen[14] è stato uno dei più convinti sostenitori, come ha rilevato Robertson[15].

L'altro modello applicativo largamente utilizzato nell'ambito dell'*information retrieval*, e che è stato adottato in primo luogo per l'indicizzazione automatica dei documenti *full-text*, è *Vector Space Model*, e dal momento che è un modello per rappresentare come vettori sia i documenti, trasformandoli in un insieme lineare di termini-vettori, che i termini usati nelle *query*, ha trovato applicazione anche nella misurazione della distanza semantica tra i documenti e le richieste degli utenti al fine di determinare il livello di rilevanza. Il grado di similarità tra i documenti di una collezione viene misurato in base alla distanza che si determina tra i documenti-vettori ed i termini-vettori presenti nelle *query*: se i termini usati per l'interrogazione del sistema non sono presenti nei documenti, il grado di similarità è pari a zero. Gli spazi vettoriali, vere e proprie collezioni di vettori, vengono quindi usati per rappresentare la similarità tra i documenti, la quale è sempre basata sul computo della presenza delle parole. La rilevanza dei documenti e la loro disposizione in una gerarchia, dal più rilevante al meno rilevante, si basa sempre sull'applicazione del modello TF-IDF, con cui automaticamente vengono assegnati pesi ai termini[16].

Il limite più significativo dei sistemi basati su *Vector Space Model* e sulle applicazioni di TF-IDF, è costituito dal fatto che le parole presenti nei testi sono considerate come indipendenti le une dalle altre, ed in tal modo vengono ignorati i legami, le relazioni e le associazioni semantiche anche implicite che i termini della lingua naturale possie-

13 Stephen Robertson – Karen Spärck Jones, *Relevance weighting of search terms,* «Journal of the American society for information science», 27 (1976), n. 3, p.129-146.

14 Cornelis Joost van Rijsbergen, *Information retrieval*, Second edition, London-Boston: Butterworths, 1979 (Prima edizione: 1975), in particolare p. 111-143.

15 Stephen Robertson, *Understanding inverse document frequency: on theoretical arguments for IDF* «Journal of documentation», 60 (2004), n. 5, p. 503-520.

16 Gerard Salton – Andrew Wong – Chung-Shu Yang, *A Vector space Model for automatic indexing*, «Communications of the ACM», 18 (1975), n. 11, p. 613-620.

dono. Diversi correttivi sono stati proposti per superare questo limite. Uno dei più significativi, ad esempio, è *Context Vector Model*[17], che impiega i dati provenienti dal computo della frequenza della co-occorrenza dei termini per stabilire la dipendenza delle parole dai diversi contesti in cui sono utilizzate. È stato dimostrato che il suo funzionamento è piuttosto soddisfacente nel caso di *query* che rispondono a bisogni informativi di livello generale, per le quali si avranno come risposta molti documenti rilevanti.

Un altro parametro preso in considerazione per stabilire la rilevanza, riguarda la prossimità dei termini in un testo. *Cover Density Ranking* è una misura di *relevance ranking* applicata ai risultati costituiti dalle pagine Web, ritrovate attraverso una ricerca in base a due o tre parole con un motore di ricerca. Dopo la prima creazione di raggruppamenti di documenti sulla base della semplice frequenza di uno o più termini, la procedura prevede l'applicazione di una seconda misura di *ranking* all'interno dei diversi raggruppamenti creati, e cioè la misura relativa a *Cover Density*, che non si basa sulla semplice frequenza dei termini della *query*, ma sulla frequenza, la prossimità e la densità delle occorrenze dei termini nei testi dei documenti. Rispetto ad un insieme di termini di una *query*, vengono generati i *cover* computando la distanza minima tra i termini nel testo: minore è la distanza tra i termini cercati nel testo in base alla *query*, maggiore è la rilevanza del testo corrispondente, e più *cover* risultano contenuti in un documento, maggiore è la rilevanza del documento[18].

Il modello IDF proposto da Karen Spärck Jones continua in ogni caso a manifestarsi come straordinariamente robusto, e resta al centro di quasi tutti i metodi di *ranking* usati nei motori di ricerca.

Disponendo di una informazione parziale di rilevanza da parte dell'utente, che permetta di determinare quali tra i documenti che contengono i termini della query sono stati riconosciuti dal ricercatore come rilevanti, è possibile realizzare la *query expansion*, cioè possiamo usare i termini che sono correlati con la rilevanza per espandere la *query* originale[19]. L'espansione dei termini della *query* consiste nella riformulazione dei termini impiegati nella fase della ricerca, manualmente, a cura dell'utente stesso, oppure automaticamente, determinata dal sistema, o infine in modo semiautomatico, dall'utente col supporto del sistema stesso. Ciascuna delle modalità di espansione dei termini può essere realizzata adottando come fonti in cui scegliere i nuovi termini da utilizzare, o i risultati stessi della ricerca offerti dal sistema nella prima fase, innescando così il processo del *relevance feedback*, nel quale i documenti recuperati ed identificati come rilevanti costituiscono la fonte dei termini per l'espansione della ricerca, oppure *cluster* di termini automaticamente generati, basati sui documenti presenti nella collezione, ma anche su thesauri, dizionari e lessici, ontologie o basi di dati lessicali come WordNet[20], indipendenti dalla collezione sulla quale si stanno compien-

17 Holger Billhardt – Daniel Borrajo – Victor Maojo, *A context vector model for information retrieval*, «Journal of the American society for information science and technology», 53 (2002), n. 3, p. 236-249.

18 Charles L.A. Clarke – Gordon V. Cormack – Elizabeth A. Tudhope, *Relevance ranking for one to three term queries*, «Information processing and management», 36 (2000), p. 291-311.

19 Efthimis N. Efthimiadis, *Query expansion*, «Annual review of information science and technology», 31 (1996), p. 121-187.

20 Per una presentazione di WordNet <http://wordnet.princeton.edu/> e di alcune ontologie basate sulla logica formale, rinvio al mio articolo *Le ontologie come strumenti per l' organizzazione della conoscenza in rete*, «AIDA informazioni», 28 (2010), n. 1/2, volume monografico su: *Le ontologie*, a cura di Maria Teresa Biagetti, p. 9-31.

do le ricerche. L'uso delle ontologie durante la fase di ricerca, per l'espansione dei termini presenti nella *query*, può consentire al ricercatore di condurre ricerche su tutti i sinonimi dei termini scelti in partenza, o addirittura tutti i termini che denominano i concetti sussunti sotto una classe. L'espansione dei singoli concetti in un insieme più ampio di concetti ad esso semanticamente correlato, utilizzando una ontologia, può inoltre fornire all'utente un insieme più esteso di risultati rilevanti, facendo affidamento sulle relazioni semantiche che sottostanno ai documenti[21].

Relevance feedback è entrato a far parte degli argomenti oggetto di analisi anche delle Conferenze TREC[22] nel 2008. Quell'anno la conferenza si è incentrata sulla valutazione degli algoritmi di *relevance feedback* per proporre una metodologia per valutarli e compararli che potesse essere comunemente accettata[23], e nel 2010 l'analisi è stata riservata al giudizio di rilevanza sui singoli documenti.

Nei sistemi di *information retrieval*, il *relevance feedback* esplicito si configura come una tecnica di *query expansion*, basata sulle misure dell'equivalenza rispetto ad un dato documento ritrovato, ed indicato dall'utente come rilevante. Gli algoritmi per il *feedback* possono usare la similarità, stabilita dal sistema sulla base dei termini presenti nei titoli o nelle stringhe dei soggetti. Perciò il sistema permette di ritrovare documenti simili a quelli indicati come veramente rilevanti da parte del ricercatore. Le tecniche di *relevance feedback* implicito utilizzano invece il monitoraggio del comportamento degli utenti durante le loro ricerche, senza che vi sia un giudizio sulla rilevanza da parte degli utenti.

Applicare agli OPAC le tecniche utilizzate dai motori di ricerca per realizzare il *ranking* secondo la rilevanza è obiettivamente più difficile, e fornisce risultati ancora meno attendibili rispetto a quelli offerti dall'analisi dei documenti in *full-text*, dal momento che vengono per lo più scandagliati i campi della descrizione, che accolgono testi molto brevi, e quindi con possibilità limitate di reperimento.

Un posto più alto o più basso nella graduatoria di *relevance* può essere ottenuto, rispettivamente, se i termini presenti nella *query* sono ritrovati tutti insieme in un campo di un *record*, oppure ciascuno in un campo diverso, ma assumono importanza anche altri elementi, come la presunta popolarità e l'interesse mostrato per il documento da altri utenti che hanno proposto i medesimi termini per la ricerca[24].

21 Matias Frosterus – Eero Hyvönen, *Bridging the search gap between the Web of pages and Web of data by combining ontological document expansion with text search*, in: *ICSD international conference for digital libraries and the semantic web: proceedings*, Trento, September 2009, p. 90-104.

22 TREC (Text REtrieval Conference <http://trec.nist.gov/>), col sostegno del National Institut of Standards and Technology (NIST) e del Department of Defense, ha assunto dal 1992 l'eredità del lavoro avviato nel 1966 da Cyril W. Cleverdon con i Progetti Cranfield al College of Aeronautics. Scopo delle conferenze annuali è migliorare le metodologie di *text retrieval* misurando il grado di rilevanza topica (*aboutness*), e le tecniche di valutazione dell'efficacia dei sistemi di I. R. che impiegano le misure di *recall* e *precision*. Per ciascuna conferenza viene preparato un insieme di documenti e di *query* per permettere ai partecipanti di testare i loro sistemi di I.R. presentando una lista di documenti ritrovati disposti in base alla rilevanza, ed infine i risultati vengono sottoposti alla valutazione di NIST. Negli ultimi anni l'ambito d'interesse di TREC si è ampliato, fino a coinvolgere l'analisi anche di *Video retrieval*, *Web retrieval* e, dal 2008, dei comportamenti dei ricercatori e degli utilizzatori dei BLOG (The Million Query (1MQ) Track).

23 Chris Buckley – Stephen Robertson, *Relevance feedback track overview: TREC 2008,* <trec.nist.gov/pubs/trec17/papers/REL.FDBK.OVERVIEW.pdf>.

24 Marshall Breeding, *Thinking about your next OPAC*, «Computers in libraries», 27 (2007), n. 4, p. 28-

L'OPAC della North Carolina State University Libraries (NCSU) <http://www.lib.ncsu.edu/>, uno tra i più analizzati e citati, dispone i risultati delle ricerche usando il *relevance ranking* per *default*. L'algoritmo considera la frequenza dei termini e la frequenza inversa (TF-IDF); oltre a ciò, considera più rilevanti i ritrovamenti che si verificano nel campo del titolo, rispetto a quelli che si verificano, ad esempio, nel campo delle note; l'algoritmo considera anche il numero delle volte che i termini presenti nella *query* compaiono in ciascun risultato, e la data di stampa può costituire uno dei pesi. Vengono ritenuti più rilevanti i risultati delle ricerche nei quali sono presenti esattamente i termini della *query* nella forma in cui sono stati inseriti, cioè senza la possibilità di correggere la forma, senza l'uso del troncamento e senza il controllo dei termini su di un thesaurus. In caso di *query* con più termini, sono considerati più rilevanti i risultati che contengono la frase esatta[25].

L'algoritmo di *relevance* può essere adattato alle esigenze della singola biblioteca per ciascuna tipologia di ricerca, per parola chiave, per la ricerca nei titoli, nei soggetti, nei nomi degli autori o *anywhere*, e la biblioteca può decidere in quale campo debbano essere reperiti i termini per avere un grado di rilevanza più alto. La ricerca lanciata utilizzando, ad esempio, il campo *keyword anywhere*, produce risultati disposti per rilevanza seguendo una gerarchia che vede al primo posto la ricerca nei campi dell'autore e del titolo, e poi nei campi del soggetto, della serie, del sommario, delle note ecc.[26].

Il prototipo RELVYL, realizzato da California Digital Library dopo la pubblicazione nel 2006 del rapporto finale di uno studio durato alcuni anni[27], in cui si ponevano in luce le criticità degli OPAC tradizionali e si suggerivano le possibili correzioni di strategia, trasferisce sul piano concreto i suggerimenti innovativi presenti nel rapporto. Il catalogo unico Melvyl attualmente ordina i risultati ottenuti a seguito di una richiesta semplicemente ponendo all'inizio i documenti catalogati più recentemente, ma nello studio si propone di realizzare il prototipo Relvyl con l'adozione di strumenti di *relevance ranking* basati sull'analisi del contenuto (*content-based*) ed opzionalmente con l'aggiunta di pesi attribuiti sulla base dei dati relativi alla circolazione dei documenti ed alla presenza di essi in più raccolte. L'adozione del sistema di ricerca testuale, *eXtensible Text Framework* (XTF), permette di utilizzare funzionalità come il *ranking* dei risultati in base alla rilevanza dei dati contenuti nei record catalografici rispetto ai termini delle *query*. Questo sistema viene definito *content-based* o *content-ranking*, ed il contenuto è costituito dai dati catalografici presenti nei campi dei record, ma attraverso la modalità di ricerca denominata *analysis* si prevede di fornire la possibilità di aggiungere ulteriori elementi al computo della rilevanza, utilizzando due opzioni: i dati provenienti dalla circolazione libraria, usati per stabilire un peso maggiore ai documenti più chiesti, ed i dati provenienti dalle raccolte possedute dalle diverse biblioteche universitarie, usati per attribuire maggior valore ai documenti posseduti da più biblioteche. Il punteggio di rilevanza viene cioè aumentato in base al fatto che un documento sia stato usato frequentemente, oppure che sia posseduto da molte biblioteche. Quest'ultima scelta risulta piuttosto

31, <http://www.librarytechnology.org/ltg-displaytext.pl?RC=12575>.

25 Kristin Antelman – Emily Lynema – Andrew K. Pace, *Toward a twenty-first century library catalog*, «Information technology and libraries», September 2006, p. 128-139.

26 Maria Collins – Jacquie Samples – Charley Pennell – David Goldsmith, *Magnifying the ILS with Endeca*, «The serials librarian», 51 (2007), n. 3-4, p. 75-100.

inopportuna, in realtà, dal momento che la metodologia prevalente per le acquisizioni librarie invita a distribuire gli acquisti secondo piani di coordinamento tra le biblioteche (metodologia CONSPECTUS).

Anche LIBRIS (http://libris.kb.se/), il catalogo collettivo nazionale svedese, presenta i risultati delle interrogazioni di ricerca usando *relevance ranking* per *default* ed è possibile il raffinamento per campi disciplinari, per lingua, secondo le tipologie bibliografiche ed in base alla disponibilità del *full text*.

AQUABROWSER Library Platform, usata da Queens Library <http:// aqua.queenslibrary.org/> ed in Italia dal Museo Galileo – Istituto e Museo di storia della scienza di Firenze <http://colombo.imss.fi.it/IMSS/>, organizza i risultati delle ricerche di *default* per *relevance*, utilizzando un insieme di criteri come la popolarità ed il peso dei campi di dati.

Bisogna comunque rilevare che questi OPAC, pur avendo adottato il criticabile ordinamento per rilevanza di *default*, consentono all'utente di scegliere altre tipologie di presentazione dei risultati. In realtà, essi risultano molto più apprezzabili per aver introdotto altre funzionalità innovative. L'OPAC di North Carolina State University Libraries (NCSU) permette l'accesso a basi di dati e strumenti di *reference* come enciclopedie, manuali, *directories* e biografie, ma anche a collezioni speciali di manoscritti e libri rari, nonché ai documenti conservati negli archivi dell'Università, consente la ricerca mirata alle biblioteche NCSU o all'insieme delle biblioteche costituito da NCSU, UNC (University of North Carolina), Duke, and NCCU (North Carolina Central University), ed infine di espandere la ricerca attraverso l'accesso a WorldCat. È basato sulla piattaforma Endeca ProFind™ che usa il *software* ENDECA Information Access Platform (IAP) Guided Navigation, ed il MDEX Engine, che permette ricerche avanzate e ritrovamento di dati strutturati e non strutturati. ENDECA Technologies Inc. è una ditta che serve diversi clienti nel campo dell'*e-commerce*, della finanza e del mondo bancario, ed ha adattato la sua tecnologia alle esigenze delle biblioteche. Le biblioteche NCSU hanno acquistato il *software* Information Access Platform (IAP) nel maggio 2005, e le nuove funzionalità del catalogo sono in uso dal gennaio 2006. Information Access Platform di ENDECA presenta alcune caratteristiche avanzate e sicuramente migliorative, come le nuove opportunità di *browsing*, che a partire da ampie classi disciplinarie della Library of Congress Classification, permettono all'utente di scandagliare l'intera collezione.

NCSU OPAC offre una funzionalità interessante attraverso il raffinamento che si può realizzare utilizzando varie *dimensions*, che riguardano i soggetti, gli autori, i generi, la lingua, il formato, la disponibilità in biblioteca, le classi della Library of Congress Classification, le regioni geografiche ed i periodi temporali. L'uso delle *dimensions* costituisce un raffinamento dei risultati ottenuti: si tratta di un raffinamento post-coordinato, che permette agli utenti di far aumentare la *precision* o il *recall* selezionando o deselezionando i termini proposti sullo schermo come ENDECA *dimensions*. Tuttavia, oltre all'inopportunità del raffinamento per autori, considerati come un aspetto attraverso il quale raffinare la ricerca, bisogna rilevare che anche il raffinamento attraverso la *dimension* "soggetti" presenta alcune criticità, dal momento che offre una molteplicità di termini provenienti da stringhe diverse, che in alcuni casi possono sovrapporsi o costituire inutili doppioni, mentre il raffinamento per classi della LC garantisce risultati più puliti. Anche una ricerca per parole nel campo dedicato a *subject headings* può restituire infatti risultati "rumorosi", poiché il sistema è in grado di ritrovare le singole parole all'interno di una stringa di soggetto, ma non è in grado di distinguerne il significato: una ricerca con la parola "Jaguar" nel campo dedicato ai soggetti mi fa recuperare sia documenti che trattano del mammifero, sia docu-

menti che riguardano la nota marca di automobili, e devo effettuare il raffinamento per ottenere solo i documenti che concernono il giaguaro, ad esempio.

LIBRIS, recentemente riorganizzato con l'obiettivo di offrire un servizio centrato sull'utente, permette la ricerca bibliografica nelle raccolte delle biblioteche universitarie e di ricerca, oltre che di una ventina di biblioteche pubbliche, complessivamente più di 170 biblioteche. Dal 2008 è stata sperimentata la possibilità di rendere disponibili i dati catalografici, controllati, strutturati e di buona qualità, alle applicazioni del Web semantico, dotando i *record* di Persistent URI e trasformando i dati in triple RDF, ed inserendosi nel grandioso progetto del *data linking*, per il momento un tentativo di dare vita alla rete di collegamenti tra dati contenuti nei documenti e nei *record* catalografici[28]. Si offre la possibilità all'utente di scegliere di ricercare nel campo del titolo, o del soggetto o ovunque nel record attraverso la ricerca avanzata, e di effettuare il *browsing* delle stringhe dei soggetti, dei titoli, dei nomi degli autori e degli enti, ecc. Nei record si trovano frequentemente TOC ed a*bstract*, molti documenti sono disponibili in *full text*, vi è la possibilità di visualizzare i record relativi ad altre edizioni della stessa opera e di accedere a *Google Book Search*.

AQUABROWSER Library Platform permette la ricerca per termini nei titoli, nelle stringhe di soggetto, nei TOC e negli *abstract*. Una funzionalità molto apprezzata di questo sistema è il *Discovery Layer* "Word Cloud", che consente di visualizzare automaticamente in una mappa, partendo da un termine di ricerca, lo stesso termine tradotto in altre lingue, i termini correlati ed i termini con variazione nello *spelling*, e di utilizzarli per ampliare la ricerca: ogni termine che l'utente sceglie nella *Word Cloud* verrà aggiunto al termine di partenza, arricchendo lo spazio conoscitivo e determinando un progressivo aumento della *precision*. È possibile il raffinamento dei risultati utilizzando classi disciplinari, serie, lingue, soggetti ecc. e la navigazione tra i termini nei titoli, nelle stringhe dei soggetti e nei titoli delle serie.

Rilevanza e pertinenza

In linea di massima, la disposizione dei risultati secondo la rilevanza, stabilita sulla base del computo della frequenza dei termini e della frequenza inversa, non può ritenersi soddisfacente per le necessità di ricerca dei ricercatori, i quali, come ha bene rilevato Thomas Mann, non si accontentano di reperire qualsiasi informazione la ricerca nel Web possa offrire, ma sono interessati al contesto concettuale nel quale l'informazione è inserita e desiderano risposte congruenti con le *query*, hanno bisogno di ritrovare i documenti nei quali gli argomenti cercati sono considerati secondo il senso preciso ricercato e gli aspetti voluti. Dal momento che costituisce un avvicinamento alle modalità utilizzate dai motori di ricerca, aver scelto il *relevance ranking* come strumento attraverso il quale assicurare un miglioramento degli OPAC, ha contribuito a rendere questi maggiormente appetibili agli occhi degli utenti che si fermano a considerare solo le caratteristiche più superficiali o più appariscenti.

Per ottenere un reale avanzamento delle funzionalità degli OPAC, invece, personalmente ritengo che sarebbe necessario sviluppare funzionalità che seguano linee del tutto indipendenti da quelle tracciate ed offerte dai motori di ricerca, ed impe-

27 California Digital Library, *The Melvyl Recommender Project final report*, July 2006, <http://www.cdlib.org/services/publishing/tools/xtf/melvyl_recommender/>.

28 Martin Malmsten, *Making a library catalogue part of the Semantic Web*, in: *Metadata for semantic and social applications: proceedings of the international conference on Dublin Core and metadata applications*, Berlin, 22-26 September 2008, edited by Jane Greenberg and Wolfgang Klas, published

gnarsi nella definizione di strategie in grado di restituire al ricercatore risultati quanto più possibile congruenti con le sue necessità informazionali. È quindi opportuno approfondire il concetto di *relevance* prendendo in considerazione la sua evoluzione nell'ambito della *Library and information science*.

In un fondamentale articolo del 1975, Tefko Sarácevic – attualmente professore emerito alla *School of Communication & Information* della Rutgers University (New Jersey) – definiva la rilevanza il concetto centrale, che poteva fungere da fondamento teoretico, della *Information science*. Il concetto di rilevanza implica il concetto di relazione: un dato, una informazione, una porzione di conoscenza sono *rilevanti* rispetto ad altri dati ed altre conoscenze, e generalmente questo dipende dal nostro grado di conoscenza preesistente. La comunicazione della conoscenza avviene quando si verifica un cambiamento significativo all'interno di una struttura cognitiva in grado di recepire conoscenza, e quindi aggiungere, eliminare o riorganizzare, porzioni di conoscenza o d'informazione significa modificare una conoscenza strutturata.

Per Sarácevic la rilevanza è appunto la misura di questo cambiamento, di questa modifica; è la misura, il grado o la relazione di incontro e di corrispondenza tra una citazione bibliografica o il contenuto di un articolo ed una richiesta che viene presentata da un utente di una biblioteca.

Nell'articolo Sarácevic presentava sostanzialmente due diversi orientamenti, distinguendo tra *Subject knowledge view of relevance*, considerata come la relazione tra una *query* e la conoscenza registrata, organizzata e pubblica, facente parte di un modello di pensiero, esistente su un argomento, e *Pertinence view of relevance*, la relazione tra le conoscenze personali di un singolo ricercatore e la conoscenza pubblica, esistente su di un certo argomento[29]. Questo secondo orientamento corrisponde infatti al concetto di pertinenza, che coinvolge la conoscenza privata, e riguarda i modelli di conoscenza esistenti nella mente del singolo ricercatore; riguarda la natura, la struttura e l'estensione dell'insieme di conoscenze di una persona, la sedimentazione di queste conoscenze ed i processi di selezione.

La pertinenza è una proprietà relativa, non la si può assegnare a un documento in modo assoluto, non è una caratteristica permanente di un documento, ma è dinamica e varia col tempo, è il giudizio umano di rilevanza, alla cui base sono le conoscenze di un utente. Si stabilisce una relazione di pertinenza di volta in volta, e spesso lo stesso libro può soddisfare diverse richieste, e dunque quel libro possiede pertinenze di grado diverso. Già Douglas J. Foskett[30], nel 1972, aveva considerato la rilevanza come espressione di un tipo di conoscenza pubblica, un modello di pensiero universalmente accettato, riferendosi anche alla teoria dei paradigmi scientifici emergenti e dominanti in un dato periodo storico, che lo storico della scienza Thomas Kuhn[31] ha posto a fondamento della sua teoria sulle rivoluzioni scientifi-

by the Dublin Core Metadata Initiative, Singapore and Universitätsverlag Göttingen, 2008, p. 146-150. Gli atti sono disponibili all'URL: <http://webdoc.sub.gwdg.de/univerlag/2008/DC_proceedings.pdf>.

29 Tefko Sarácevic, *Relevance: a review of and a framework for the thinking on the notion in information science,* «Journal of the American society for information science», 26 (1975), n. 6, p. 321-343.

30 Douglas J. Foskett, *A note on the concept of relevance*, «Information storage and retrieval», 8 (1972), p. 77-78.

31 Thomas S. Kuhn, *The structure of scientific revolutions*, Chicago: The University of Chicago, 1962, trad. ital. *La struttura delle rivoluzioni scientifiche*, Torino: Einaudi, 1995 (I ed. 1969). I paradigmi sono costituiti dalle «conquiste scientifiche universalmente riconosciute, le quali, per un certo periodo, for-

che, e la pertinenza invece come l'espressione del modello di conoscenza che si manifesta nella mente di un ricercatore.

Negli anni seguenti il concetto di rilevanza è stato ulteriormente articolato da Sarácevic prendendo in considerazione cinque dimensioni:

System or algorithms relevance, la relazione tra una *query* e l'informazione ritrovata o non ritrovata attraverso un algoritmo.
Topical relevance, la relazione tra il soggetto espresso in una *query* ed il soggetto coperto da un documento.
Pertinence or cognitive relevance, la relazione tra la necessità d'informazione di un utente, il suo stato di conoscenze ed il documento recuperato, connessa anche alla corrispondenza del livello cognitivo dell'utente, al potenziale di novità ed alla qualità dell'informazione del documento.
Utility or situational relevance, la relazione tra una situazione particolare ed il documento ritrovato, che coinvolge l'appropriatezza delle informazioni e la possibilità di ridurre l'incertezza nell'affrontare e risolvere un problema specifico.
Motivational relevance, la relazione tra le intenzioni e le motivazioni di un utente e i documenti recuperati dal sistema[32].

Infine, in una recente rassegna nella quale prende in esame le diverse concezioni di rilevanza[33], Sarácevic, ribadendo di considerarla una nozione appartenente alla sfera umana, come tale quindi non necessariamente del tutto razionale e certamente complessa, ha proposto la distinzione tra:

Topical relevance, *or Subject relevance*, concetto molto vicino a quello di *aboutness*, che concerne la relazione tra il topico espresso in una *query* ed il topico trattato in un testo scritto o verbale, o cui un'immagine o qualsiasi artefatto umano si riferisca.
System relevance, *or Algorithmic relevance*, anch'essa basata sulla *Topical relevance*, si riferisce alla relazione tra una *query* ed un oggetto potenzialmente informativo ritrovato o non ritrovato attraverso un sistema d'*information retrieval*, senza considerare l'utente. Si tratta di una rilevanza di livello debole, in quanto i sistemi di I. R. determinano il *match* tra una *query* con un oggetto informativo in larga misura sulla base della similarità delle parole, secondo Sarácevic, anche se altre strategie vengono adottate, sia per la musica che per le immagini.
User relevance, che presenta due diverse dimensioni: la *Situational relevance or utility*, una prospettiva pragmatica attraverso la quale considerare una rilevanza situazionale, il rapporto cioè tra gli oggetti informativi ed una particolare situazione, un compito o un problema da risolvere, una rilevanza connessa con l'utilità e l'appropriatezza delle informazioni, ad esempio per risolvere un determinato problema; la *Cognitive relevance or pertinence*, la rilevanza determinata dalla relazione tra gli oggetti informativi e lo stato cognitivo della conoscenza di un utente in un determinato momento, basata sulla corrispondenza cognitiva, il livello, la qualità dell'informazione, e la novità dell'informazione.

niscono un modello di problemi e soluzioni accettabili a coloro che praticano un certo campo di ricerca» (p. 10). La trasformazione di un paradigma costituisce una rivoluzione scientifica. La scienza evolve attraverso la continua sostituzione di paradigmi.

32 Tefko Sarácevic, *Information Science*, «Journal of the American society for information science», 50

Secondo Sarácevic, il concetto di pertinenza riguarda la conoscenza personale ed i modelli di conoscenza nella mente dell'utente, la natura, la struttura e la sedimentazione della conoscenza di ciascun utente. È una peculiarità relativa, una relazione che viene stabilita di volta in volta all'interno di una singola ricerca, non una caratteristica persistente, attribuita una volta per tutte ad un documento.

Rifiutati come inadeguati i criteri basati su algoritmi ed accettato l'orientamento cognitivista che conduce a considerare la rilevanza un concetto multidimensionale, soggettivo e dinamico, Xu e Chen[34] individuano in *topicality* e *novelty* le due dimensioni fondamentali tra quelle che caratterizzano il concetto di rilevanza, e che comprendono anche *reliability*, l'attendibilità scientifica dei documenti, *understandability*, il livello della comprensibilità dei documenti e *scope*, l'appropriatezza dei documenti rispetto alle necessità dell'utente. Oltre che sulla topicalità, l'accento è quindi posto sulla novità, cioè la percezione della differenza di un documento rispetto agli altri che l'utente già conosce. Si tratta di una novità soggettiva, non determinabile superficialmente riferendosi alla novità della pubblicazione, ovviamente, ma strettamente collegata invece alla dimensione della pertinenza, che riguarda la conoscenza personale nella mente dell'utente, così come la mancanza di conoscenza, il bisogno di informazione percepito dall'utente. Gli autori suggeriscono infine che nel futuro i sistemi di I. R. possano essere dotati di una nuova funzionalità, ad esempio siano in grado di quantificare la novità dei documenti per un utente. Alcune delle domande che vengono poste dai due autori sono molto stimolanti, ma per il momento non vengono proposte soluzioni:

> How could we capture a reader's cognitive state before document evaluation?
> How could we measure the novelty of a document against the cognitive state?
> How could we combine novelty and topicality in an overall relevance score[35]?

Nelle sue più recenti riflessioni, Birger Hjørland[36], riprendendo il concetto di rilevanza come *Subject knowledge view of relevance* definito da Sarácevic nel 1975, ne propone un'interpretazione ispirata ad un paradigma di tipo sociale, ricorrendo alla teoria della *Domain analysis*[37] da lui stesso elaborata come paradigma interpretativo e possibile fondamento teoretico dell'*Information Science*. In sostanza, una terza via tra *Subject knowledge view of relevance*, la rilevanza concepita come universalmente valida, e *User-based view of relevance*, concezione legata al punto di vista cognitivista e ad un criticabile, secondo Hjørland, mentalismo latente. Nella terza via proposta prevale il riconoscimento della soggettività della conoscenza, condivisa da gruppi di persone che seguono lo stesso paradigma. Anche la *System or Algorithmic relevance*, secondo Hjørland, è in ultima analisi riconducibile all'ambito della soggettività,

(1999), n. 12, p. 1051-1063, p. 1059.

33 Tefko Sarácevic, *Relevance: a review of the literature and a framework for thinking on the notion in information science. Part II*, «Advances in librarianship», 30 (2006), p. 3-71.

34 Yunjie Xu – Zhiwei Chen, *Relevance judgment: what do information users consider beyond topicality?*, «Journal of the American society for information science and technology», 57 (2006), n. 7, p. 961-973.

35 Ivi, p. 971.

36 Birger Hjørland, *The foundation of the concept of relevance*, «Journal of the American society for information science and technology», 61 (2010), n. 2, p. 217-237.

37 Birger Hjørland – Hanne Albrechtsen, *Toward a new horizon in Information Science: Domain-Analy-*

in quanto viene in ogni caso determinata dalle scelte sia dei programmatori, che degli indicizzatori che stabiliscono gli accessi semantici, o decidono quali pesi affidare ai termini che verranno ricercati, o definiscono la struttura dei *link*.

Il giudizio di rilevanza è determinato dagli orientamenti scientifici e teoretici seguiti dagli utenti all'interno di ciascun campo disciplinare. Il ricercatore che si occupa di schizofrenia – spiega Hjørland – e segue l'orientamento scientifico dell'eziologia di tipo ambientale, riterrà rilevanti tutti i documenti concernenti gli studi sulla famiglia, mentre il ricercatore che segue l'orientamento dell'eziologia di ordine genetico, considererà rilevanti solo i documenti che trattano dei disordini cromosomici. A questa concezione espressa da Hjørland, e che risente della sua teoria della *Domain analysis*, si può obiettare che i ricercatori che desiderano approfondire il tema della schizofrenia, ma non seguono ancora alcun orientamento, troveranno rilevanti tutti i documenti che trattano di quella patologia, a prescindere dai diversi orientamenti eziologici, e la dimensione relativa al *Subject knowledge view of relevance* teorizzata da Sarácevic può risultare quindi ancora ampiamente utilizzabile.

La terza via prospettata da Hjørland non si presenta del tutto convincente, ma non ritengo che sia sufficiente neppure la considerazione della sola dimensione della *Topical relevance*, incentrata sul concetto di *aboutness*, che prevede la strutturazione di relazioni di corrispondenza di livello generale tra i concetti espressi in una richiesta e l'argomento trattato in un documento.

Come ha chiarito Robert Fugmann[38], un sistema di *Information Retrieval* può dare risposta ad una *query*, ma può soltanto supporre quale sia l'*information need* di un utente. In sostanza, può essere in grado di determinare più o meno la rilevanza dei documenti rispetto ad una richiesta, ma non sarà mai in grado di decidere rispetto alla loro effettiva pertinenza. I documenti che un ricercatore trova pertinenti alla sua ricerca, possono essere diversi, del tutto o soltanto in parte, dall'insieme di documenti che erano stati richiesti, dal ricercatore stesso o dal bibliotecario-documentalista, sulla base della rilevanza topica.

Il sistema può ritrovare gli *items* che sono rilevanti e risulteranno poi pertinenti dal punto di vista di un particolare ricercatore, ma il sistema può anche recuperare *items* che sono rilevanti e che non saranno poi considerati pertinenti secondo il punto di vista di quel ricercatore. Al contrario, ci possono essere *items* considerati pertinenti da un ricercatore ma non recuperati tra i rilevanti dal sistema.

Dopo aver definito i parametri in base ai quali avviare una ricerca semantica, si possono ottenere diversi sottoinsiemi di documenti, libri o articoli pubblicati su riviste, nei quali, secondo Fugmann, giocano un ruolo fondamentale sia la rilevanza generica che la prospettiva della pertinenza:
1) documenti rilevanti, ritrovati e poi verificati come pertinenti alla propria ricerca;
2) documenti rilevanti, ritenuti comunque pertinenti, ma non ritrovati, per difetto o imperfezione del sistema;
3) documenti non ritenuti rilevanti, ma ritrovati casualmente, o per scarsa specificità o imperfezione del sistema, e successivamente definiti pertinenti (ritrovamento casuale e felice, *serendipity*);
4) documenti rilevanti, correttamente ritrovati, ma poi risultati non o scarsamente pertinenti;
5) documenti non rilevanti, quindi correttamente non ritrovati, ma che invece risulterebbero pertinenti alla ricerca;

sis, «Journal of the American society for information Science», 46 (1995), n. 6, p. 400-425; Birger Hjørland, *Domain analysis in information science: eleven approaches traditional as well as innovative*,

6) documenti rilevanti, ma nonostante ciò non ritrovati, e comunque ritenuti non pertinenti;
7) documenti non rilevanti e non pertinenti, ritrovati erroneamente (rumore);
8) documenti estranei ai parametri della ricerca e quindi correttamente non ritrovati;
9) documenti rilevanti e pertinenti alla ricerca, ma non ritrovati perché non presenti nella collezione[39].

In una recente indagine sui criteri adottati dagli utenti per stabilire la rilevanza dei documenti, Taylor, Zhang, e Amadio propongono di considerare la rilevanza come la percezione da parte dell'utente dell'importanza dei documenti per le proprie necessità informazionali, e la presentano come *real world view of relevance*[40], una concezione di rilevanza agganciata cioè alla realtà della ricerca scientifica, e che assume come fondamentali, quindi, proprio gli aspetti legati alla pertinenza.

Per migliorare le funzionalità della ricerca semantica attraverso gli OPAC, e renderle più vicine alle necessità della ricerca degli utenti, quindi, appare più opportuno adottare, come paradigma sul quale modellare nuove e più efficaci funzionalità, la dimensione della pertinenza.

Nuove prospettive per la ricerca semantica attraverso gli OPAC

L'adozione del *relevance ranking* offre un miglioramento molto modesto, un avanzamento più apparente che reale delle funzionalità dell'OPAC, dal momento che è basato sul computo della frequenza dei termini e sull'analisi della prossimità dei termini stessi; addirittura può costituire un elemento fuorviante per il ricercatore, che potrebbe limitarsi a considerare solo i primi risultati offerti, attratto da una disposizione secondo una presunta rilevanza, i cui meccanismi comunque ignora.

È necessario invece spostare l'attenzione dal livello dell'organizzazione dei risultati delle ricerche (*ranking*) al livello dell'indicizzazione semantica, studiare strategie di miglioramento che arricchiscano le funzioni di ricerca semantica espandendone le potenzialità intrinseche, ed elaborare interventi che rendano possibile utilizzare attraverso le interfacce gli esiti di un'indicizzazione più approfondita e sfaccettata. Un miglioramento interessante è certamente già costituito dall'arricchimento degli OPAC con i sommari delle monografie, in particolare nel settore delle scienze umane, permettendo la ricerca per parole; i sommari consentono una definizione relativamente precisa dei concetti e degli argomenti trattati nella monografia, mentre l'inserimento delle immagini delle copertine delle monografie non costituisce un grande aiuto alla ricerca scientifica.

Le relazioni semantiche che si possono mettere in evidenza sono potenzialmente infinite. I sistemi informativi e gli OPAC non possono conoscere in anticipo quali saranno i bisogni informazionali dell'utente di volta in volta, e quali documenti potranno realmente servire per una specifica ricerca. Un miglioramento veramente significativo per la ricerca potrebbe essere raggiunto adottando un'ottica che privilegia la prospettiva della pertinenza invece di limitarsi alla rilevanza topica (*aboutness*), e permettendo all'utente di conoscere la varietà degli argomenti presentati da un documento e qual'è la profondità, la ricchezza e la complessità del trattamento degli argomenti, dal momento che solo l'utente è in grado di stabilire quali documenti soddisfano pienamente, o parzialmente, le proprie necessità di ricerca. Nell'ambito delle

«Journal of documentation», 58 (2002), n. 4, p. 422-462.

38 Robert Fugmann, *Subject analysis and indexing. Theoretical foundation and practical advice*, Frankfurt M.: Indeks, 1993.

riflessioni intorno al concetto di *aboutness*, quindi del concetto che oggi avviciniamo alla rilevanza topica, nel 1978 William J. Hutchins[41] aveva proposto all'attenzione della comunità di studiosi di *Library and information science* una particolare concezione di *aboutness*, ed aveva suggerito di tenere conto del livello di conoscenza che si presuppone che i potenziali lettori di un determinato libro possiedano e di realizzare un'indicizzazione che si basi sull'*aboutness* relativo al livello di conoscenza posseduto dal lettore, dimensionata sul reale livello di conoscenze dell'utente, in modo da permettergli di procedere da ciò che egli conosce già a ciò che non conosce ancora. È un riconoscimento dell'importanza della funzione di pertinenza.

I recentissimi sviluppi nell'ambito del Web semantico, di cui sono parte lo standard SKOS[42] e la reingegnerizzazione come ontologie formali di thesauri e soggettari elaborati in ambito documentario e biblioteconomico, invitano ad alcune riflessioni. La trasposizione nel Web dei dati catalografici strutturati e le relazioni tra di essi, presenti negli OPAC, grazie agli strumenti elaborati per il Web semantico, dopo averli convertiti in triple RDF (*Resource Description Framework*), descritti sulla scorta di vocabolari condivisi (Dublin Core, SKOS:Concepts, FOAF per persone ed enti) ed identificati attraverso URI, al fine di realizzare una interconnesione di dati comprensibili dalle macchine che renda possibile la ricerca semantica dei contenuti degli OPAC direttamente sul Web, è certamente un progetto grandioso, che permetterebbe di partecipare massicciamente alla visione del *data linking* concepita da Berners-Lee[43], nel quale la Library of Congress, alcune altre biblioteche statunitensi ed il catalogo collettivo svedese LIBRIS attualmente sono già inseriti[44]. Prima di attuare questa trasposizione sarebbe opportuno però riflettere su quale intreccio di dati l'utente si troverebbe a percorrere sul Web, quale tipologia di dati semantici, ad esempio, sarebbe realmente disponibile.

I dati attualmente presenti nei cataloghi elettronici, resi disponibili come *linked data*, navigabili seguendo i percorsi istituiti dalle triple di RDF, e nel caso dei soggetti, arricchiti dalla possibilità di navigare attraverso collegamenti tra *broader*, *narrower* e *related concepts*, permetteranno ricerche comunque limitate alle possibilità oggi offerte dal livello di indicizzazione semantica già presente. Il progetto di inserimento dei dati catalografici degli OPAC nella visione del Web semantico prevede infatti la completa mappatura dei campi di MARC 21 con RDF. È necessario invece riflettere sulle modalità di indicizzazione semantica ed affrontare il problema delle strategie di indicizzazione che sarebbe preferibile adottare.

Limitandoci per il momento a utilizzare le strategie tradizionali di indicizzazione semantica, si possono presentare alcune prospettive di miglioramento che si fondano sui suggerimenti offerti molti anni fa da Alfredo Serrai[45], ma risentono anche delle

39 Ivi, p. 44-45.

40 Art Taylor – Xiangmin Zhang – William J. Amadio, *Examination of relevance criteria choices and the information search process*, «Journal of documentation», 65 (2009), n. 5, p. 719-744, in particolare p. 727.

41 William J. Hutchins, *The concept of 'aboutness' in subject indexing*, Paper presented at a Colloquium on aboutness held by the Co-ordinate Indexing Group, 18 April 1977, «Aslib proceedings», 30 (1978), n. 5, p. 172-181.

42 <http://www.w3.org/TR/2009/REC-skos-reference-20090818/>

43 Tim Berners-Lee, *Linked Data* <http://www.w3.org/DesignIssues/LinkedData.html>.

44 Sharon Yang – Yan Yi Lee – Amanda Xu, *The Semantic Web and libraries in the United States: experi-*

precisazioni e delle esemplificazioni di Benedetto Aschero[46], suggerimenti che sono qui rielaborati ed applicati alla realtà della moderna catalogazione elettronica.

Il modello di indicizzazione che si propone come ipotesi di lavoro, come esempio di possibile incremento delle funzioni di ricerca semantica negli OPAC, si potrebbe sviluppare seguendo queste linee:

1) L'analisi concettuale dei documenti dovrebbe mettere in pratica due diverse strategie di indicizzazione ogni volta che ciò sia necessario, ad esempio ogni volta che ci si trova a dover indicizzare libri dalla forte impronta ideologica:

a) mettere in evidenza gli Oggetti di cui i documenti trattano. Per una monografia che tratti delle diverse eresie (ariana, catara, ecc.) l'Oggetto della trattazione sono le "Eresie".

b) mettere in evidenza il Soggetto di cui i documenti trattano, cioè considerare il discorso che l'autore fa intorno ad un concetto o un argomento di studio. Una monografia può presentare un'analisi secondo il punto di vista cattolico intorno alle eresie che si sono storicamente sviluppate, ed il Soggetto non sarebbe allora "Eresie", ma potrebbe essere "Ortodossia cattolica".

2) Le strategie di indicizzazione rivolte all'analisi concettuale dei documenti (libri ed articoli) dovrebbero permettere di evidenziare i diversi livelli ai quali gli argomenti vengono trattati, cioè permettere di sapere chiaramente se un argomento è trattato completamente in una monografia, e quali argomenti, secondari rispetto alla trattazione principale, sono analizzati nel documento solo in funzione dell'argomento principale. Una monografia che affronti il tema dell' "Assolutismo nel 17° secolo" può anche presentare trattazioni limitate e secondarie che possono riguardare la "Finanza" o le "Condizioni sociali". Gli argomenti secondari sono limitati a considerazioni che si riferiscono all'argomento (o agli argomenti) principale, sono trattati in funzione dell'argomento principale cui la monografia è dedicata, e questa differenza di livello nella trattazione deve essere presentata con evidenza all'utente attraverso nuove funzionalità degli OPAC.

3) Si possono usare modalità differenziate per rappresentare gli argomenti principali e quelli secondari: uso di codici diversi, o di corpi diversi, o di diversi colori, permetteranno di differenziarli immediatamente. L'utente deve avere la possibilità di sapere qual è il livello di profondità dell'argomento e se si tratta dell'argomento principale trattato nella monografia o di un argomento affrontato in funzione della trattazione principale.

Questo modello può consentire di mantenere ed utilizzare la metodologia indicizzatoria tradizionale, cioè l'analisi concettuale dei documenti, e realizzare un sistema articolato: Oggetti, Soggetti, trattazione principale ed argomenti secondari, nei diversi gradi di approfondimento.

Quale potrebbe essere l'effetto di questa strategia sugli OPAC? L'interfaccia di ricerca dovrebbe consentire agli utenti di scegliere prima se ritrovare documenti che sono stati indicizzati tenendo presenti gli Oggetti del discorso oppure i Soggetti. All'utente deve essere permesso di decidere di ritrovare "Ortodossia cattolica" come Oggetto o come Soggetto.

Utilizzando le interfacce visive predisposte ad esempio dalla piattaforma di AquaBrowser Library, che permette di distinguere i termini graficamente, si potrebbero differenziare Oggetti e Soggetti ed i diversi gradi di trattamento degli argomenti potreb-

mentation and achievements, Satellite Meetings IFLA 2009, Emerging trends in technology: libraries between Web 2.0, semantic web and search technology, Florence, 19-20 August 2009. <http://www.ifla2009satelliteflorence.it/meeting3/program/assets/SharonYang.pdf>

bero essere resi immediatamente evidenti. Una presentazione grafica che sfrutti le possibilità offerte dallo standard *Topic Maps*[47], ad esempio, consentirebbe di evidenziare le relazioni tra i concetti (*topics*) precisamente definiti, stabiliti dagli indicizzatori e non estratti automaticamente, e costituirebbe un'evoluzione successiva e particolarmente apprezzabile. Oltre a ciò, dovrebbe essere possibile evidenziare i diversi livelli di trattazione degli argomenti, e consentire al ricercatore di recuperare insieme i documenti in cui l'argomento cercato sia stato trattato, ad esempio, in modo esaustivo.

ABSTRACT Bollettino **AIB**, ISSN 1121-1490, vol. 50 n. 4 (dicembre 2010), p.339-356.

MARIA TERESA BIAGETTI, Università di Roma "La Sapienza", Scuola speciale per archivisti e bibliotecari, viale Regina Elena 295, 00161 Roma, e-mail mariateresa.biagetti@uniroma1.it

Nuove funzionalità degli OPAC e *relevance ranking*

L'articolo si inserisce nell'ambito del dibattito sulle strategie da seguire per il miglioramento delle funzionalità degli OPAC, ed in particolare prende in considerazione la modalità che prevede la disposizione dei risultati delle ricerche in base ad una presunta rilevanza per l'utente, basata in larga parte sulla frequenza dei termini (TF/IDF), utilizzata dai motori di ricerca e dalle librerie on line, e ora applicata anche agli OPAC, alcuni dei quali sono descritti (NCSU Libraries, LIBRIS, AQUABROWSER Library Platform).
L'analisi critica del *relevance ranking* offre lo spunto per una indagine di tipo teoretico sui concetti di rilevanza e di pertinenza, quest'ultimo legato allo stato delle conoscenze dei singoli utenti, e per una riflessione sul miglioramento delle funzionalità della ricerca semantica negli OPAC, che si basi su strategie di indicizzazione che consentano di cogliere la varietà e il livello di approfondimento degli argomenti piuttosto che sull'adozione di discutibili modalità di organizzazione dei risultati.

New OPAC functionalities and relevance ranking

This paper aims to contribute to the debate on OPAC enhancement, and in particular it concerns the use of ranking queries results considering the supposed relevance for users, mainly based on the terms frequency and inverse document frequency (TF/IDF), until now used by search engines and on line bookstores, and recently applied to OPAC. Some notable example of new generation OPAC (NCSU Libraries, LIBRIS, AQUABROWSER Library Platform) are described.
The critical examination of the use of *relevance ranking* gives the opportunity to analyse the concepts of relevance and pertinence from a theoretical point of view. Pertinence perspective, in particular, that considers the level of knowledge of users, offers the chance to enhance semantic search functions in OPAC, following indexing strategies which allow to put in evidence topic variety and different levels of topic examination.

45 Alfredo Serrai, *Del catalogo alfabetico per soggetti*, Roma: Bulzoni, 1979.