A partire dagli indici: la ricerca secondo dandelon.com

L'arricchimento collaborativo del catalogo nei paesi di lingua tedesca

Mara Persello
mpersello@hotmail.com

Anche ai tempi di Internet, i libri scientifici continuano ad essere fonti primarie di informazione e formazione per milioni di studenti e ricercatori. Nell'approccio a un testo scientifico una delle prime cose che un utente fa, una volta che si trova il libro fra le mani, è leggerne l'indice, e valutarne quindi la coerenza con i propri interessi di ricerca. Questa operazione è sostanzialmente semplice se l'utente si trova nella biblioteca giusta e il libro è disponibile per la consultazione, ma l'esperienza ci dice che molto più spesso il testo si trova in un magazzino oppure in altre biblioteche e forse in altre città.
D'altronde, uno studente o un ricercatore passano oggi la maggior parte del loro tempo davanti ad un computer, lavorando ai loro testi con Internet pronto per la consultazione. Sempre più comune la ricerca diretta in Google o Wikipedia: spesso l'utente ha necessità di svolgere ricerche su temi ed espressioni tecniche specifiche, che raramente gli restituirebbero risposte nelle ricerche per soggetti dei cataloghi di biblioteche. Ogni bibliotecario addetto al reference si è trovato di fronte alla difficoltà di dover tradurre la richiesta di un utente nel linguaggio astratto della soggettazione, strumento affascinante, ma troppo spesso spuntato: perfino superfluo quando la ricerca per soggetto viene tentata direttamente da utenti completamente privi di competenze specifiche. Così, la messa in rete di informazioni, che si rivelano in certa parte inaccessibili a coloro che non ne possiedono i codici di interpretazione, diventa una contraddizione in termini. E la situazione non migliora se si passa alla ricerca libera nell'OPAC: un'applicazione più comprensiva, certo, ma per questo più generica, che crea rumore e costringe ad una ricerca nella ricerca, nei casi più fortunati, e nei casi peggiori (e frequenti) può portare a dei veri e propri abbagli. A partire da questi bisogni e da queste difficoltà, diventa naturale conseguenza pensare ad una ricerca on line che venga incontro ad utenti abituati all'uso di motori di ricerca in Internet e che permetta di valutare la pertinenza dei documenti in modo immediato, senza che si sia costretti a spostarsi di biblioteca in biblioteca, perdendo ore per ricerche che possono rivelarsi inutili. Google ha da tempo colto questa tendenza e si è inserito in questo campo offrendo soluzioni come Scholar e Print, e con motori di ricerca *inhouse* come Appliance o servizi con punti d'accesso esterni in università e biblioteche scientifiche. Google è ormai un gigante, con fatturati milionari e potere quasi assoluto: il nuovo prodotto sviluppato, il Billing-System, basandosi sul sistema *pay-per-view* e fornendo direttamente la possibilità di acquistare i libri online, non farà altro che aumentare ulteriormente gli accessi e gli introiti del gruppo.

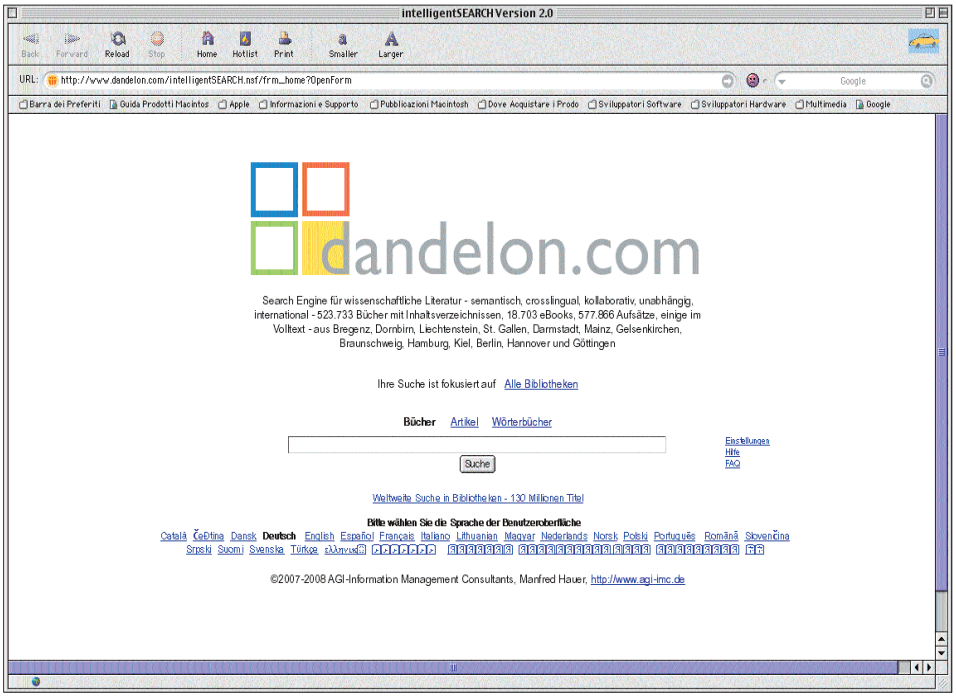
Perché dandelon.com

In ambito tedesco – Austria, Liechtenstein, Svizzera e Germania – si sta sviluppando da alcuni anni un sistema nuovo, che sta cercando la propria strada e si propone come alternativa all'attuale stato di cose, un modello che sta crescendo e si sta irrobustendo: si tratta del motore di ricerca dandelon.com.
I due fondatori di dandelon.com hanno messo in campo grande preparazione ed esperienza: Manfred Hauer, tedesco, è titolare di AGI Information Management Consultants, e Karl Rädler è responsabile della soggettazione della Biblioteca statale del Voralberg a Bregenz in Austria. L'obiettivo che i due si sono dati è di sfruttare il posseduto delle biblioteche così da far funzionare in modo nettamente più efficiente le ricerche tematiche specifiche, non costringendo per questo gli utenti a nuove interfacce o a cambiamenti. Entrambi i fondatori di dandelon.com si occupavano già da tempo di thesauri e classificazione, e oggi la ricerca, possibile in 25 lingue, comprende

1.700.000 espressioni semantiche.
L'idea è semplice, ed è nata osservando gli schemi di comportamento degli utenti. Già sopra si è notato che per valutare la pertinenza o meno di un testo scientifico si usa scorrerne il sommario; l'indice è infatti una parte densa di significato nell'economia di un testo: è qui che vengono presentati in modo sintetico gli argomenti che verranno affrontati. Proprio a partire dalla scansione dei TOC (*table of contents*), che non essendo coperti da copyright non ci portano nell'accidentato territorio dei diritti d'autore, e dalla loro lettura in OCR, vengono generati termini di ricerca, parole chiave e descrittori. Fin qui tutto chiaro: il problema che però si pone a questo punto è l'eccesso di rumore che una lista di termini acriticamente e meccanicamente estratti da un sommario restituisce. Di fronte a questo ostacolo si sono fermati la maggior parte dei progetti diretti in questo senso. Gli ideatori di dandelon.com, però, hanno fatto un passo in più, portando nella macchina l'esperienza di chi conosce e lavora con i soggetti. La lista di termini che la macchina ricava, infatti, viene messa a confronto e filtrata da un sistema di thesauri semantici, che graduano la pertinenza dei descrittori e permettono una ricerca anche per sinonimi: il termine cercato può, al limite, non apparire neanche una volta nel sommario restituito dalla ricerca, molto più rilevante la graduatoria di pertinenza con cui vengono presentati i risultati. Ma dandelon.com è andato ancora oltre, implementando anche un sistema di thesauri linguistici, perché soprattutto nella letteratura scientifica, per sua natura in continua e veloce evoluzione, spesso gli interventi vengono prodotti in inglese o nella lingua dell'autore, senza che gli editori ritengano necessario divulgarli in traduzione.

I descrittori generati vengono inseriti in dandelon.com, e la ricerca per termini restituisce non, come un normale motore di ricerca, tutte le occorrenze di una parola, ma una lista di testi e la loro collocazione nelle biblioteche associate: la ricerca da generica si fa scientifica, e rientra in biblioteca. Tutto questo avviene a partire da un'interfaccia utente scarna e semplice come quella di Google, che eventualmente può essere resa più specifica con la scelta, fra le "preferenze" (l'interfaccia è disponibile in 25 lingue), di filtri più familiari ai fruitori di OPAC, con ricerche per biblioteca, autore, anno e così via, ma anche attraverso filtri come quello della tolleranza ortografica o delle sottocategorie semantiche. Ancora, e anche in questo differenziandosi dai motori di ricerca tradizionali, dandelon.com presenta non solo una lista ragionata di testi e la loro collocazione a partire dalla ricerca, ma può mostrare i contenuti in pdf o in html, insieme con dei metadati, di ciascun testo selezionato dalla lista. Si tratta di circa 30.000 copertine, acquisite direttamente dagli editori, dati bibliografici originati dal sistema della biblioteca, abstract di case editrici e infine risultati delle indicizzazioni e rielaborazioni meccaniche.
Una buona struttura ha bisogno di contenuti, buone idee con macchinosi procedimenti diventano idee meno buone. L'obiettivo è ovviamente una massiccia collaborazione a questo progetto da parte delle biblioteche, ed è stato quindi pensato un sistema di acquisizione dei testi quanto più semplice e snello possibile. Entrando nello specifico, nelle biblioteche associate viene scannerizzato l'indice dei testi che si vogliono rendere disponibili, utilizzando l'hardware: la tecnologia alla base di dandelon.com è IBM Lotus Notes & Domino, Adobe Acrobat e Abbyy Finereader Engine, i computer Dell e gli scanner Jujitsu, il tutto predisposto in un unico pratico carrello che si muove agilmente anche fra gli scaffali compatti (l'operatore non ha quindi necessità di cercare e al termine della scansione ricollocare il testo, ma può direttamente portare la macchina sul posto), ma soprattutto usando il client-software Intelligentcapture, sviluppato da AGI, in cui lavorano contemporaneamente diversi server e moduli, in commercio o *open source*, che si integrano in un unico sistema, così da generare un'indicizzazione dei contenuti in modo automatico e altamente efficiente. Lo stesso processo avviene nell'indicizzazione delle risorse digitali, che provengono da case editrici, librerie, cataloghi di biblioteche e agenzie di stampa che partecipano insieme all'arricchimento del catalogo. In sostanza la biblioteca in fase di produzione non ottiene solo i link, ma anche i rispettivi dati (pdf, txt, tif) estratti dal sommario, così che questi siano anche in futuro utilizzabili da nuove versioni o da altri programmi, o per diversi utilizzi. La produzione continua a crescere in modo esponenziale, sempre di più dall'inizio del 2007, grazie alle stazioni mobili che sono state perfezionate, e che si possono muovere direttamente fra gli scaffali, così che viene ridotto al minimo il tempo di consultazione e ricollocazione. Il gruppo di lavoro della AGI riesce, con tre postazioni mobili, a produrre fino a mille sommari al giorno, il più veloce modo di acquisire dati finora sperimentato, ma ovviamente la biblioteca può scegliere di servirsi di personale interno: l'utilizzo del software è molto semplice ed è attivo un servizio di assistenza remota per tutto il tempo della produzione. Con poco personale e in poco tempo, quindi, si può immaginare un rafforzamento delle possibilità di ricerca enorme: i documenti richiesti dai magazzini delle biblioteche

vengono restituiti in molti casi senza esser letti, e questo costa lavoro e denaro inutilmente, costi che possono essere abbassati con l'uso di una più precisa ricerca, di una migliore reperibilità e con la possibilità di mostrare all'utente il pdf del sommario sullo schermo. Il vantaggio è ovviamente anche dell'utente, che risparmia tempo e inutili richieste di prestito, e raggiunge più efficacemente gli obiettivi della sua ricerca.
Tutti i dati raccolti vengono analizzati e valutati automaticamente, attraverso i thesauri, sviluppati in ambito biblioteconomico e collegati inoltre a livello linguistico. Partner ospite è il centro GBV per servizi bibliotecari di Gottinga. I costi di un tale sistema erano, fino ad alcuni anni fa, troppo alti, e alcune biblioteche tedesche hanno sviluppato metodi parzialmente assimilabili con finanziamenti statali. Ma dandelon.com è cresciuto con molta più velocità. Un po' di numeri possono rendere conto di quanto la cosa si stia espandendo nei paesi di lingua tedesca. Finora sono stati scannerizzati, indicizzati automaticamente e resi disponibili alla ricerca in più lingue 450.000 sommari. Questi TOC sono presenti sia in dandelon.com che negli OPAC delle biblioteche rispettive. Parallelamente sono stati acquisiti 500.000 articoli.
Con 70.000 consultazioni al mese, e una crescita, nel 2007, di 230.000 sommari, quest'anno si può considerare quello della svolta. Diventa un fatto normale la implementazione di questo genere di contenuti nei sistemi bibliotecari. Dal 2008 ne fa parte anche la Biblioteca nazionale tedesca. E il modello collaborativo fa sì che a sommari di contenuti prodotti in Italia si colleghino ricerche che rimandano alla Library of Congress o alla British Library.
Il modello collaborativo di dandelon.com, inoltre, permette di minimizzare spese ed energie, dato



La home page di dandelon.com

che alcune biblioteche trovano già presenti nel catalogo fino al 30% del loro posseduto. Va chiarito, comunque, che questo modello è accessibile solo agli utenti di intelligent CAPTURE, il sistema stesso con cui i sommari vengono elaborati e prodotti, lo strumento che AGI ha sviluppato come base di questo progetto.
In Germania, la Biblioteca nazionale, varie biblioteche regionali, biblioteche specialistiche e biblioteche universitarie (Kiel, Amburgo, Berlino, Braunschweig, Gottinga, Gelsenkirchen, Mainz, Darmstadt, Bregenz, Dornbirn, Vaduz, St. Gallen) raccolgono i sommari, li indicizzano e li pubblicano in dandelon.com, e d'altro lato, partendo dal catalogo, li rendono disponibili in pdf come collegamento, così che è possibile, durante la consultazione dell'OPAC, leggere direttamente sullo schermo, senza la necessità di avere il testo in mano, il sommario del volume d'interesse. Alcune di queste biblioteche salvano i risultati dell'indicizzazione automatica in nuovi campi dei loro sistemi di catalogazione (Aleph, Pica, Sisis, Libero): così come i descrittori, il sistema include nomi, nomi geografici, e soprattutto espressioni specifiche. dandelon.com può essere integrato nell'interfaccia utente di ciascuna biblioteca, così da poter offrire la tecnica di ricerca di dandelon.com senza bisogno di rinvii esterni e in un unico contesto di ricerca, rimandando con un click al catalogo. In questo senso AGI ha elaborato una chiave indipendente dai sistemi bibliotecari e che non crea sovrapposizioni se diverse biblioteche hanno uguali numeri di sistema. Questa chiave viene generata dalla prima biblioteca che acquisisce il documento, e trasmessa a tutte le altre biblioteche che hanno lo stesso documento. E se si può pensare che l'indicizzazione meccanica possa creare rumore, va detto che le informazioni avanzate vengono prodotte in una blackbox da sistemi di retrieval come Fastsearch e Ibm Omnifind Intern. Il grado di utilizzabilità e i formati di trasferimento sono configurati in modo flessibile e leggero.
La maggior parte dei sistemi di catalogazione possono salvare, in

base al proprio sottostante sistema di management dei dati, solo brevi testi. La ricerca avviene attraverso l'aiuto di operatori booleani, e non può restituire liste per livelli di rilevanza. Ma è proprio questa graduabilità di rilevanza che si attendono coloro che operano la ricerca, perché così li ha abituati la ricerca Internet. In uno studio condotto su 99 studenti e 300 titoli, si è potuto notare che il sistema booleano funziona bene nel caso in cui i termini di ricerca dati siano molto specifici e lavorino in collaborazione con la indicizzazione meccanica. Queste ricerche hanno successo in piccole biblioteche, fino a circa 5.000 titoli, ma sono inefficienti per le grosse collezioni. Decisamente superiore si è dimostrata, in questo esperimento, la performance di *relevance ranking* e di espansioni di ricerca semantiche in dandelon.com.

Dopo l'enorme implementazione di 230.000 titoli nel 2007, il progetto intende crescere ancora. L'intenzione è quella di coinvolgere grandi biblioteche o poli di biblioteche, di ampliare il respiro internazionale del progetto presentandosi in altri paesi e in diversi tipi di biblioteche, piccole o speciali. L'offerta di per sé molto semplice, priva di spese di hardware e senza bisogno di formazione del personale. La visione è quella di una collaborazione, proficua per tutti, che permetta a dandelon.com di diventare uno strumento sempre più ricco, nuovo, superiore ai classici OPAC, che non richiede una particolare preparazione a chi si accinge alla ricerca, che si ricollega con il proprio interfaccia ad un sistema di ricerca snello e semplice, che si dimostra comprensivo ben oltre i poli bibliotecari, includendo biblioteche di diverse nazioni. Un passo avanti nello scambio di conoscenze, un sogno di divulgazione sorretto da una tecnologia di cui si vogliono sfruttare le enormi potenzialità.

Abstract

dandelon.com, a search website meant to exploit library catalogues, already in use in German, Austrian and Swiss libraries, offers a solution to the problem that a search machine working on word recurrence on a text is often useless for a scientific research. dandelon.com analyses through a thesaurus string the relevance of searched words in previously scanned indexes, and proposes a list of suitable texts and their location in libraries. Goal of dandelon.com is a more efficient words search and an easier approach to library catalogues, avoiding the problems of subject cataloguing.