Il servizio di Emeroteca Virtuale al CASPUR ed il suo nuovo motore di ricerca

GINO FARINELLI, RICCARDO FAZIO,
ILARIA DE MARINIS, STEFANO DE LUCA

Sintesi
L'Emeroteca Virtuale (EV) consente l'accesso a più di cinquemila periodici accademico-scientifici a testo completo per un totale di otto milioni di articoli. L'uso del software *Science Server è ormai datato, e sta causando un aumento eccessivo nei tempi di risposta nelle operazioni di ricerca di un singolo articolo. Tutto ciò ha spinto verso lo sviluppo di un nuovo motore di ricerca più in linea con i tempi di risposta attesi dai destinatari del servizio, anche in relazione ad una mole di articoli ricercabili che si sta avviando a superare il traguardo dei dieci milioni.*

Il servizio di Emeroteca Virtuale <periodici.caspur.it>, gestito dal CASPUR sin dal 1999, consente l'accesso permanente a più di cinquemila periodici accademico-scientifici di cinque tra i più grandi esponenti mondiali di editoria scientifica o di società professionali. Attraverso un'interfaccia web, gli utenti di circa 30 università italiane del centro-sud, che hanno sottoscritto contratti di accesso alle riviste con gli specifici editori, possono arrivare a visualizzare il testo completo del singolo articolo scientifico (si parla in tali casi di *accesso al full-text*).

A fine 2008, l'Emeroteca Virtuale ha raggiunto il traguardo degli otto milioni di articoli e Science Server, il *software* gestionale su cui si basa il servizio, ha cominciato a mostrare sempre più evidenti limiti, a partire dagli elevati tempi di risposta nelle funzioni di ricerca di un articolo.

Relativamente al modo con il quale gli utenti dell'Emeroteca Virtuale hanno la possibilità di reperire gli articoli, esistono due modalità operative: attraverso elenchi da cui selezionare la rivista desiderata e successivamente "sfogliarla" come fosse una rivista cartacea, oppure tramite articolate maschere di ricerca all'interno delle quali è possibile indicare titolo della rivista, titolo dell'articolo, *l'autore (gli autori), l'anno di pubblicazione, e molte altre chiavi di ricerca.*

Elevati tempi di risposta alle ricerche effettuate dagli utenti possono costituire un problema rilevante per l'Emeroteca. Come diversi studi di web *usability* dimostrano, infatti, gli utenti che effettuano ricerche su Internet si aspettano di ottenere risposte dai *server* web in tempi brevi, nell'ordine di una manciata di secondi al massimo (comparabili con quelli del suo più famoso motore di ricerca, Google); in caso contrario sono indotti a ritenere che il servizio non funzioni o che la ricerca non sia andata a buon fine o, ancor peggio, pensano di dover nuovamente eseguire la stessa avviandone di nuove (fenomeno dei doppi *click*), appesantendo così ulteriormente il suo *carico di lavoro.*

Al fine di non peggiorare significativamente la qualità del servizio offerto dall'Emeroteca, in ragione soprattutto della sua diffusione in àmbito universitario, si è avviata, nel 2008, una fase di studio e valutazione delle soluzioni *open source* 56 di *information retrieval* (IR) 57 disponibili in rete che potessero meglio rispondere ai due seguenti criteri: di *scalabilità*, ovvero di poter garantire la loro efficienza e funzionalità al crescere della base dati degli articoli ricercabili; di robustezza, ovvero in grado di poter essere applicate anche in condizioni di intenso utilizzo.
Questa ricerca ha portato alla scelta di Lucene [58] , un potente sofwtare open source di *information retrieval* (IR), interamente sviluppato nel linguaggio di programmazione Java dalla Apache Software Foundation [59].

56 *Open-source* è un termine utilizzato per indicare una tipologia di *software* non soggetto a *policy* d'uso di tipo commerciale e, soprattutto, del quale è distribuito in chiaro il codice sorgente.
57 Con il termine *Information Retrieval* (IR) si intende l'insieme delle tecniche utilizzate per il recupero mirato di informazioni in formato elettronico; nell'Emeroteca Virtuale le "informazioni" sono rappresentate dagli articoli scientifici in *full-text*.
58 <lucene.apache.org>
59 L'ASF <www.apache.org> è una fondazione *non-profit* formata da una comunità distribuita di sviluppatori che lavorano su progetti di *software open source* per applicazioni web.

Lucene è tanto potente e versatile quanto complesso da configurare; per tale motivo la stessa comunità dell'ASF ha sviluppato un "enterprise search server", denominato SOLR, anch'esso *open source*, che fornisce allo sviluppatore e all'amministratore del sistema un'agile interfaccia web per la gestione e la configurazione di Lucene. SOLR, per il suo funzionamento, necessita di un "Java servlet container" [60], e, allo scopo, si è deciso di utilizzare il *server* Tomcat [61], (distribuito sempre dalla ASF).

Uno dei motivi che ci hanno spinto a privilegiare Lucene come nuovo motore di ricerca dell'Emeroteca è legata proprio al sistema SOLR: quest'ultimo riunisce in due soli *file*, in formato XML, le configurazioni utilizzate da Lucene e fornisce una versatile interfaccia web per il *debugging* analitico del motore di ricerca. L'indicizzazione dei documenti all'interno del sistema di IR avviene su file XML all'interno di sessioni HTTP e nel medesimo formato sono i risultati delle ricerche effettuate.

Ha certamente contribuito alla scelta di Lucene anche il fatto che nello specifico contesto degli organismi che offrono un servizio simile a quello offerto dall'EV, Lucene venga utilizzato in àmbito internazionale dal consorzio delle Università dell'Ohio (OHIO-Link), ovvero dalla *Research Library* dei laboratori nazionali di Los Alamos: in questi casi i cataloghi contengono più di dieci milioni di *oggetti digitali,* rappresentati da articoli o libri elettronici. A livello italiano, inoltre, la Biblioteca Nazionale Centrale di Firenze utilizza Lucene come strumento di ricerca del proprio posseduto.

Uno degli aspetti nei quali si è intervenuti per la personalizzazione di Lucene all'ambiente di EV è stato quello relativo alle *procedure di stemming*: queste procedure permettono di indicizzare un termine all'interno del DB utilizzato per la ricerca, memorizzando non solo la parola completa che si vuole rendere ricercabile, ma anche la sua radice. In questo modo la ricerca del termine *cell* può restituire non solo l'articolo che contiene questa parola, ma anche quelli che contengono *cells*, *cellular*, etc. Dal momento che nessuno degli *algoritmi di stemming* adottati da Lucene si comporta come quelli utilizzati attualmente dal *software* dell'EV ("Science Direct"), si è deciso di riscrivere un algoritmo di *stemming ad hoc*, sfruttando in tal senso la caratteristica di questo *software* di essere di tipo *open-source*.

Un'altra area nella quale si sono apportate modifiche è quella relativa alle maschere di ricerca: normalmente Lucene, e gli *information retrieval* in genere, vengono utilizzati per realizzare motori di ricerca "google-like"; la maschera di ricerca è composta da una sola casella di testo e i termini da ricercare vengono posti in OR fra loro; nei risultati vengono mostrati prima i documenti che contengono tutti i termini ricercati e, a seguire, i documenti in cui sono presenti solo alcuni di loro. La ricerca degli articoli in EV è caratterizzata invece da una diversa impostazione: infatti l'utente deve avere la possibilità di effettuare ricerche su più campi contemporaneamente in AND fra loro e, se si inseriscono più termini in un dato campo di ricerca, anch'essi devono essere posti in AND fra loro (AND *implicito*). Anche in questo caso sono state apportate personalizzazioni dell'ambiente offerto da Lucene, in modo da garantirne una transizione *senza discontinuità* verso quello dell'Emeroteca.

Buona parte dell'attività di *integrazione* di Lucene nell'EV ha riguardato l'identificazione dei dati soggetti alle operazioni di indicizzazione: nell'Emeroteca Virtuale ad ogni articolo è associato un file PDF, che contiene il *full-text* dell'articolo (ovvero l'articolo completo), e un *file* XML con la descrizione di tutte le informazioni legate all'articolo stesso (i cosiddetti *metadati*).

Struttura e contenuto dei metadati nei file XML sono stati, quindi, attentamente analizzati, per definire quali informazioni degli articoli dovessero essere presenti nelle maschere di ricerca e, quindi, ricercabili dall'utente e quali dovessero essere mostrati nella scheda con i risultati.

---

60 Lo Java Servlet Container è un *server* Java utilizzato per lo sviluppo di siti web con contenuto dinamico

61 <tomcat.apache.org>

Lucene prevede che ad ogni informazione da trattare sia associato un *tipo* di dati (testuale, numerico, ecc.) il cui *comportamento* può essere definito dall'amministratore (ad esempio, se e quale algoritmo di *stemming* utilizzare). Prevede inoltre che venga indicato se l'informazione sia da indicizzare, o se sia da salvare.

Ad ogni informazione da trattare è stata, quindi, associata una "tipologia". Ad esempio, il "titolo dell'articolo" è un *tipo* di dato che deve essere soggetto allo *stemming,* e deve essere indicizzato e salvato. Questo perché tale informazione deve essere ricercabile anche nelle sue varianti e deve essere mostrata nei risultati della ricerca. Viceversa, l'*abstract* di un articolo viene indicizzato in quanto ricercabile, ma non viene salvato, perché non compare nei risultati; o ancora, il *coverdate* (la data di pubblicazione dell'articolo) è un'informazione che non va indicizzata in quanto non ricercabile, ma deve essere salvata perché viene visualizzata tra i risultati.

Per quanto riguarda l'EV, le informazioni indicizzate in Lucene per ciascun articolo comprendono: l'editore, il titolo della rivista, il codice ISSN[62], il titolo dell'articolo, l'autore(i), l'*abstract*, le *keyword* dell'articolo, l'anno di pubblicazione ed il DOI[63].

L'attività di personalizzazione di Lucene nell'ambiente dell'EV si è conclusa con lo sviluppo di un programma (*parser*) che effettua l'analisi dei file XML con i metadati degli articoli ed estrae soltanto le informazioni necessarie all'indicizzazione in un formato adatto. Da ultimo, sono state sviluppate tutte le procedure per l'interrogazione attraverso sessioni HTTP del DB di Lucene e per la visualizzazione dei documenti trovati (si veda lo schema a blocchi della figura 1).

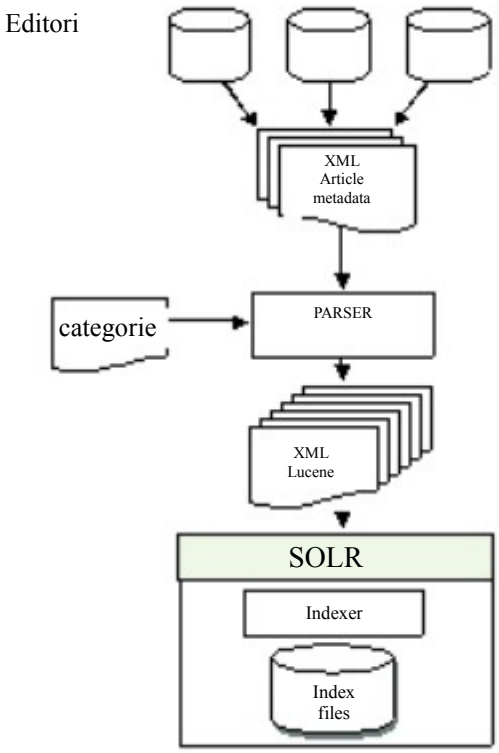

Fig. 1 Schema a blocchi dell'integrazione di Lucene con l'Emeroteca Virtuale

Nell'ultimo trimestre del 2008 sono stati effettuati dei *test* di indicizzazione degli articoli nel DB di Lucene, sempre più massivi, fino ad arrivare all'indicizzazione di tutti gli articoli presenti in Emeroteca Virtuale (otto milioni circa).

---

62 Il codice ISSN è una stringa alfanumerica di 8 caratteri, identificativa, nel panorama delle pubblicazioni cartacee o elettroniche, dello specifico prodotto editoriale (periodico o monografia)

63 DOI: *Digital Object Identifier*; è un identificatore univoco degli oggetti digitali (articolo, presentazione, immagine, video, ecc.) che ne permette non solo l'identificazione, ma anche il recupero indipendentemente dalla sua specifica collocazione all'interno della Rete.

Parallelamente, sono state effettuate sul DB diversi *test* di ricerca, differenziate per tipologia (ricerca semplice, avanzata, esperta) per valutare le prestazioni ed i tempi di risposta di Lucene all'aumentare del numero degli articoli indicizzati.

In tutti i casi i risultati sono stati estremamente incoraggianti, dal momento che i tempi di risposta si sono attestati ben al di sotto del secondo, anche nel caso di *query* relativamente complesse dal punto di vista del numero degli articoli trovati (ad esempio, la ricerca effettuata con la chiave *cell* ne fornisce più di 250.000).

Come nota conclusiva, si vuole osservare come l'attività di integrazione iniziata con il *software* Lucene (che andrà in linea entro la fine del 2009) costituisca un primo passo verso una riprogettazione completa del servizio di Emeroteca Virtuale sia in termini di contenuti sia di *layout*. Il tutto al fine di fornire agli utenti un servizio più fruibile, efficiente ed in linea con le nuove tendenze (di *format* e contenuti) che stanno emergendo nel campo degli strumenti di presentazione *online* delle pubblicazioni scientifiche.