CONFIGURAZIONE SERVER SOLR

INTRODUZIONE

Il server su cui si basa l’applicazione “Dblp-Solr Search” contiene all’interno tutti i dati relativi alle pubblicazioni scientifiche presenti sul sito “dblp.org” ed è gestito dal software di Apache Lucene/Solr.

Tale software permette il salvataggio dei documenti o delle parti di essi più importanti, e con un sofisticato algoritmo crea un indice di ricerca, struttura che consente la risposta pressoché immediata del server stesso alle richieste fatte dagli utenti dell’applicazione. È infatti grazie all’indice che è possibile reperire le corrispondenze con titolo, autore, parole nei documenti, eseguire raggruppamenti e filtrare i risultati.

Affinché il sistema di Solr funzioni perfettamente con i nostri dati, è necessario eseguire alcune configurazioni iniziali. Così facendo, sarà anche possibile ottimizzare il server specificatamente per i nostri input.

Nel file S.R.S. di “Dblp-Solr Search” è possibile trovare il link al manuale online di Solr, dove sono spiegate con cura le tecniche di configurazione del server di ricerca.

FILE CONFIGURAZIONE

I due file principali da creare/modificare per un corretto funzionamento del sistema sono “schema.xml” e “solrconfig.xml”. Entrambi devono essere inseriti all’interno della cartella di configurazione di Solr, e saranno utilizzati per l’inizializzazione del server e durante l’indicizzazione dei documenti.

* SOLRCONFIG

Questo file contiene la maggior parte delle informazioni necessarie per l’inizializzazione di Solr.

Nella nostra situazione, il file fornito da Solr di default è un’ottima base di partenza, sono solo necessarie alcune aggiunte e alcune modifiche per ottenere l’ambiente di lavoro di cui abbiamo bisogno.

1. Inserire alcune proprietà della JVM su cui il server si appoggerà. (a seconda delle esigenze particolari)
2. Rimuovere il riferimento alla funzione di auto completamento dello schema.

* SCHEMA

Questo file contiene tutte le informazioni relative alla tipologia dei documenti che andremo ad inserire nell’indice. Vi sono definiti i vari campi dei documenti, i tipi di dato e il modo in cui vengono trattati da Solr una volta incontrati tali campi. (divisione in token, indicizzazione, storage del valore ecc.)

Nel nostro caso, è sufficiente fare riferimento al file DTD scaricabile dal sito dblp.org, dove sono elencati i vari attributi delle pubblicazioni e il tipo dei dati.

Questo schema viene modificato e aggiornato ogni volta che alcuni campi vengono tolti, aggiunti o cambiati.

* SOLR.IN.CMD

Questo file è utilizzato dallo script di inizializzazione del server su Windows, e permette di definire all’interno variabili di Solr, quali la memoria dell’Heap riservato, proprietà della JVM e molto altro. L’equivalente Linux è “solr.in.sh”.

TRASFORMAZIONE XSLT

Il file XML scaricato da dblp.org deve essere indicizzato attraverso il sistema standard di Solr. Per fare ciò, è necessario che esso diventi conforme alla tipologia di documento XML riconosciuto da Solr. (si veda il manuale online per i dettagli)

Bisogna perciò creare un file XSLT che trasformi l’XML di partenza in un file riconoscibile dal server. Tale file deve essere inserito in un’apposita cartella nominata “xslt” all’interno della cartella di configurazione creata in precedenza.

Per indicizzare i dati ora, basterà richiamare da linea di comando la richiesta specificando nella stringa di connessione il nome del file XSLT da utilizzare.

È naturalmente possibile creare un apposito programma Java per gestire tramite richieste HTTP l’indicizzazione dei dati.