Semantic Search and Browsing nell'ambito dei Beni Culturali

Claudio Cortese

CILEA, Servizi applicativi e assistenza

Abstract

Negli ultimi anni, il mondo dei beni culturali si sta rivelando una delle aree più promettenti e stimolanti per quanto concerne la sperimentazione e la diffusione degli standard e delle tecnologie che rientrano nell'ambito del "Semantic Web". All'organizzazione dei dati secondo gli standard semantici si accompagna la necessità di fornire agli utenti funzionalità innovative per reperire le informazioni e navigare tra di esse, soprattutto in presenza di dati caratterizzati da un elevato grado di interconnessione e che possono essere decisamente eterogenei, come quelli riconducibili all'ambito culturale. Il contributo, dopo una breve introduzione generale sul Semantic Web, analizza le principiali problematiche legate al "semantic search and browsing", anche alla luce delle esperienze pratiche condotte da CILEA che, dal 2008, è impegnato nelle creazione di *Digital Library* semantiche.

Today, cultural heritage represents one of the most promising and challenging areas for the application of the standards and technologies that fall within the domain of the "Semantic Web". Organization of data according to the Semantic Web standards is accompanied by the need to provide users with innovative tools for finding and browsing the information, particularly in the presence of data with a high degree of interconnection and which can be very heterogeneous, like those belonging to the cultural heritage domain. The paper, after a short introduction, analyzes the major issues related to semantic search and browsing, also in the light of the practical experiences carried out by CILEA, which since 2008 is engaged in the creation of Semantic Digital Libraries.

Keywords: Semantic Web, Semantic Search and Browsing, Reti Semantiche, Digital Libraries.

Il Semantic Web

Negli ultimi anni, il mondo dei beni culturali si sta rivelando una delle aree più promettenti e stimolanti per quanto concerne la sperimentazione e la diffusione degli standard e delle tecnologie che rientrano nell'ambito del "Semantic Web" [1], [2]. In particolare diversi enti si stanno rivolgendo con attenzione sempre crescente a questi strumenti per arrivare ad offrire all'utenza avanzate funzionalità di ricerca e di navigazione basate sulle relazioni semantiche tra dati anche di diversa origine (archivistici, bibliografici, museali etc.) e a creare di set di dati altamente interoperabili.

Il Semantic Web viene solitamente definito come un'estensione del web attuale, in cui l'informazione può essere espressa in un formato comprensibile dalla macchina e può essere processata automaticamente da parte di agenti software.

Il Semantic Web permette l'interoperabilità, la condivisione e il riutilizzo dei dati da parte di applicazioni e comunità eterogenee. Esso si basa principalmente sul *Resource Description Framework* (RDF), per mezzo del quale è possibile definire le relazioni tra i dati e quindi arrivare a creare reti semantiche che possono essere rappresentate mediante grafi etichettati orientati.

Nell'ambito del Semantic Web l'informazione viene organizzata sulla base di ontologie che descrivono i concetti pertinenti ad un determinato dominio. Un'ontologia è essenzialmente un vocabolario che definisce una serie di concetti e le relazioni esistenti tra essi.

Tali vocabolari possono essere sviluppati utilizzando linguaggi specifici come l'RDF Schema Language (RDFS) e l'Ontology Web Language (OWL).

L'utilizzo delle tecnologie Semantic Web, in particolare nell'ambito dei beni culturali, è ancora agli inizi e mancano ancora applicazioni in grado di implementare vasti *dataset* utilizzabili da parte di un vasto pubblico.

Tuttavia, i risultati di recenti ricerche hanno provato che l'approccio del Semantic Web può rivelarsi molto uti le per rendere l'informazione

relativa al patrimonio culturale maggiormente interconnessa e dunque utilizzabile.

In questo senso, un esempio di come potrebbe configurarsi il "web del futuro" grazie a collegamenti tra i dati basati sulla loro valenza semantica è fornito dall'approccio denominato dei "Linked Data". È al di fuori degli scopi di questo contributo discutere nel dettaglio questo tipo di approccio. Basti dire che si tratta di una serie di "*best practises*" finalizzate a creare dei link tra differenti basi di conoscenza codificate in formato RDF, in modo da permettere di navigare ed effettuare ricerche in maniera integrata all'interno delle basi di conoscenza collegate, favorendo quindi una più completa interoperabilità. Una volta collegati gli uni agli altri, infatti, i diversi set di dati possono essere uniti e riutilizzati dando origine a quello che viene chiamato *Web of Data* (termine spesso utilizzato come sinonimo di Semantic Web).

Semantic Search and Browsing

L'organizzazione dei dati secondo gli standard del Semantic Web può costituire il punto di partenza per fornire agli utenti funzionalità innovative per reperire e mettere in relazione le informazioni, soprattutto in presenza di dati caratterizzati da un elevato grado di interconnessione e che possono essere decisamente eterogenei, come quelli che vengono gestiti nell'ambito dei beni culturali.

La sfida è quella di realizzare portali ed interfacce che permettano agli utenti di effettuare ricerche e navigare tra i dati, sfruttando appieno la ricchezza informativa delle reti semantiche, condizione indispensabile per far comprendere il valore aggiunto che queste tecnologie possono apportare.

In quest'ottica è il caso di sottolineare come le caratteristiche delle ricerca in ambito semantico si differenziano in certa misura da quelle che caratterizzano i motori di ricerca "tradizionali".

Innanzitutto quello che cambia nella navigazione di una rete semantica (e che spesso risulta spiazzante per l'utente che si avvicina per la prima volta a queste tecnologie) è il fatto che, in quest'ultimo caso, l'obiettivo non è quello di restringere sempre più la ricerca, ma quello di "allargarla", includendo nel risultato qualunque istanza che sia in relazione, anche alla lontana, con il *focus* della ricerca.

Nell'ambito del Semantic Web quindi viene utilizzato il concetto della "*query expansion*" [2], che mira appunto a valorizzare all'interno dei risultati della ricerca tutti i concetti che hanno una qualsiasi relazione con l'*input*. In questo modo, per fare un esempio, cercando "Parigi" verrebbe restituito tra i risultati anche il concetto "Montmartre", anche se la parola "Parigi" non compare tra i metadati di "Montmartre". Il rischio è ovviamente quello di esagerare e di perdere dunque di vista la precisione nella risposta all'interrogazione. Per questo i motori di ricerca forniscono la possibilità di parametrizzare la distanza (in termini di nodi del grafo) tra il concetto di partenza e quelli ad esso legati entro la quale effettuare la ricerca.

Un altro elemento da considerare è che generalmente l'utente non conosce le classi e le proprietà che caratterizzano l'ontologia secondo la quale sono organizzati i dati, fatto che potrebbe portare spesso ad effettuare interrogazioni che restituiscono un risultato nullo.
Per risolvere questo problema spesso i motori di ricerca semantici utilizzano strumenti come il *faceted browsing* e l'autocompletamento.

Mediante il *faceted browsing,* infatti, l'utente può esplorare i dati, procedendo per filtri successivi legati alle sfaccettature (*facets*) dei dati stessi (ad esempio, luogo, data, autore etc.), e combinando più filtri. In questo modo è possibile da un lato arrivare a comporre *query* molto complesse, anche senza conoscere la struttura dei dati, dall'altro eliminare il rischio di interrogazioni che diano un risultato nullo.

Mediante l'autocompletamento invece il motore di ricerca cerca di "indovinare", basandosi sull'ontologia e a volte utilizzando anche tecniche di ragionamento automatico, il termine da ricercare man mano che viene digitato ogni singolo carattere (Fig. 1). In questo modo l'utente è guidato a digitare solo termini che abbiano un senso all'interno dell'ontologia.

Un ulteriore elemento da prendere in considerazione quando si trattano strumenti di *Information Retrieval* nell'ambito del Semantic Web è quello delle modalità di presentazione dei risultati della ricerca.

In questo senso molto utile si rivela la possibilità di raggruppare i risultati della ricerca a seconda del tipo di relazione che li lega al concetto di partenza, secondo una tecnica che è stata definita "*post-query disambiguation*" [3].

A completamento di queste funzionalità, poi nell'ambito dei beni culturali, risultano particolarmente efficaci quegli strumenti come mappe, *timeline* e visualizzazioni del grafo RDF che possono migliorare la fruizione delle reti

semantiche, tenendo conto delle differenti esigenze e del diverso livello di esperienza degli utenti.

Fig. 1 – Esempio di autocompletamento nell'ambito di una ricerca effettuata utilizzando il motore di ricerca semantico sviluppato dal CILEA per gli "Archivi Storici della Psicologia Italiana".

Esperienze CILEA

A partire dal 2008 CILEA ha iniziato ad approfondire il tema dell'applicazione degli standard e delle tecnologie Semantic Web al mondo dei beni culturali. In particolare due progetti hanno permesso di testare nuove tecnologie e paradigmi per strutturare l'informazione culturale e di sperimentare metodologie innovative per la ricerca delle informazioni e la presentazione di esse all'utenza.

Il primo "Biblioteca Aperta di Milano" (BAMI) [4], [5] mirava a permettere la consultazione via web di documenti di vario genere (manoscritti, documenti a stampa, documenti di archivio) appartenenti ad alcune delle principali istituzioni milanesi. Il secondo sugli "Archivi storici della psicologia italiana" (ASPI) [6], ancora in corso, mirava invece a permettere agli studiosi di effettuare ricerche e consultare i dati riguardanti le figure chiave della storia della psicologia italiana e i rispettivi archivi.

Nell'affrontare entrambi i progetti molta importanza è stata attribuita all'analisi delle diverse caratteristiche dell'utenza cui erano destinate le soluzioni tecnologiche: BAMI doveva infatti rivolgersi sì a studiosi ed esperti del dominio, ma anche al pubblico più generale dei non specialisti interessati alla storia della musica.

L'applicazione sviluppata per il progetto legato gli archivi storici della psicologia italiana (Fig. 1), invece, è stata concepita appositamente per archivisti ed esperti nel campo della psicologia e della storia della scienza.
In quest'ottica fondamentale importanza ha rivestito la scelta dell'applicazione da utilizzare per costruire l'interfaccia web.

Per il progetto BAMI è stato scelto Longwell, un *faceted browser* appositamente sviluppato da MIT (*Massachusetts Institute of Technology*) per navigare dati in formato RDF. Longwell offre molteplici funzionalità di ricerca e navigazione, e quindi è sembrato ben adattarsi ad un progetto che voleva raggiungere utenti con esigenze diverse e con un differente grado di esperienza. In particolare, alla modalità di navigazione secondo la tecnica del "*faceted browsing*" è stata affiancata la navigazione basata sulle relazioni che legano le diverse istanze, la possibilità di visualizzare ed esplorare una rappresentazione del grafo RDF, e quella di prospettare le occorrenze disponendole lungo una *timeline*.

Nell'ambito del progetto legato agli archivi storici della psicologia italiana, invece, il fine era soprattutto quello di riuscire a prospettare agli studiosi "in un colpo solo" tutte le informazioni disponibili all'interno della base di conoscenza basata sull'ontologia CIDOC-CRM, cioè tutte le relazioni esistenti tra il *focus* della ricerca e le altre istanze, indipendentemente dalla quantità di nodi del grafo RDF che fosse necessario attraversare. Per raggiungere questo obiettivo si è fatto ricorso a ClioPatria, una piattaforma basata su SWI-Prolog che offre funzionalità avanzate per la ricerca semantica che facilitano il recupero di tutti i concetti e le relazioni espressi nel grafo RDF. In particolare, l'algoritmo di ricerca controlla tutti i *literal* corrispondenti all'interno del grafo e raggruppa le risorse trovate sulla base del percorso che le lega al focus della ricerca. Grazie a questo approccio, il set di dati semantici può essere esposto in tutta la sua ricchezza, secondo le esigenze dei ricercatori cui il progetto si rivolge.

I due progetti illustrati hanno sicuramente raggiunto molti risultati positivi, tra cui spiccano la creazione di due tra le prime *Semantic Digital Library* italiane, e il popolamento di ricche basi di conoscenza su cui è possibile effettuare interrogazioni anche molto complesse, che possono contribuire a gettare luce su aspetti poco valorizzati dei dati. Tuttavia i *feedback* ricevuti hanno mostrato che, nonostante gli sforzi fatti per raggiungere un buon livello di fruibilità, a volte, gli utenti hanno trovato comunque poco comprensibili le modalità di recupero e presentazione dei dati. Ciò spesso è dovuto alla complessità delle ontologie utilizzate, FRBR per BAMI e CIDOC-CRM per ASPI la cui struttura può rivelarsi di difficile comprensione per utenti non specialisti.

Anche le funzionalità di navigazione, tuttavia, talvolta si sono rivelate spiazzanti: è emerso infatti che lo stesso *faceted browsing* non risulta sempre di facile comprensione per gli utenti.

Le difficoltà legate alla fruzione delle interfaccia riguardano comunque l'intero ambito del Semantic Web e sono uno degli argomenti maggiormente dibattuti tra gli esperti del settore nell'ambito dei convegni internazionali (si vedano in questo senso anche le osservazioni contenute in [7]).

Viene sottolineato da più parti, infatti, che lo sviluppo di strumenti efficaci per navigare basi di conoscenza ricche e complesse rappresenta un compito non facile e che in futuro i maggiori sforzi di chi realizza applicazioni di questo tipo dovranno essere indirizzati al miglioramento della fruibilità delle interfacce utente, che, all'occorrenza, devono anche essere in grado di nascondere la complessità delle ontologie.

In questo senso è stato detto [8] che ad oggi il Semantic Web è come un "brutto anatroccolo", un qualcosa di ancora incompiuto. Se infatti le tecnologie descritte in questa sede hanno permesso la creazione di basi di conoscenza molto ricche e strutturate, non si può negare che da esse è ancora troppo difficile ricavare l'informazione desiderata, a causa del fatto che spesso sono caratterizzate da interfacce ostiche da utilizzare per i non specialisti.

Sicuramente questo tipo di affermazione è da considerasi in parte provocatoria, in quanto il Semantic Web non è solo "un'interfaccia", ma è innegabile che è necessario arrivare a garantire un'esperienze di utilizzo maggiormente soddisfacenti in quanto ciò rappresenta la condizione

necessaria per radicare le tecnologie Semantic Web nell'ambito dei beni culturali.

Bibliografia

[1] Hyvönen, E., "Semantic Portals for Cultural Heritage", in Staab, S., Rudi Studer, D. (eds.) "Handbook on Ontologies", International Handbooks on Information Systems, Springer Berlin Heidelberg, 2009, pp. 757-778.

[2] Nixon, L., Dasiopoulou, S., Evain, J., Hyvönen, E., Kompatsiaris, I., Troncy, R., "Multimedia, Broadcasting and eCulture", in "Handbook of Semantic Web Technologies", Springer, 2011, pp. 901-965.

[3] Wielemaker, J., Hildebrand, M., Van Ossenbruggen, J., Schreiber, G., "Thesaurus-based search in large heterogeneous collections", in "ISWC '08", Proceedings of the 7th International Conference on The Semantic Web, 2008.

[4] Barbera, M., Cortese, C., Zitarosa, R., Groppo, E., "Building a Semantic Web Digital Library for the Municipality of Milan", Rethinking Electronic Publishing: Innovation in Communication Paradigms and Technologies – Proc. 13th International Conference on Electronic Publishing/Edited by: Susanna Mornati and Turid Hedlund, 2009, pp. 133-154.

[5] Barbera, M., Cortese C., Groppo, E., Zitarosa, R., "Una Biblioteca Digitale Semantica per il Comune di Milano", Bollettino del CILEA, n. 113, giugno 2009.

[6] Cortese, C., Mantegari, G., "Extending the Digital Archives of the Italian Psychology with Semantic Data", Semantic Digital Archives, 2011.

[7] Hildebrand, M., Van Ossenbruggen, J., "Configuring Semantic Web Interfaces by Data Mapping", Workshop on Visual Interfaces to the Social and the Semantic Web (VISSW2009), 2009.

[8] Andraz, T., "Semantic Web Interfaces: Do They Have to Be Ugly?", 2010, URL: http://www.slideshare.net/andraz/-semtech2010-do-semanticwebuserinterfa-ceshavetobeugly