Qualità nella ricerca dell' informazione: il ruolo delle ontologie

W3C OFFICE IN ITALY
Oreste Signore

KM CROSS

QUALITY MANAGEMENT

Sommario
La descrizione delle risorse web con metadati migliora nettamente la qualità dell' informazione. Nel Semantic Web è possibile rappresentare, condividere ed esportare la conoscenza, utilizzando un linguaggio per esprimere *dati* e *regole*. Le macchine possono quindi accedere ad un *insieme strutturato di informazioni* e a un insieme di *regole di inferenza* da utilizzare per il ragionamento automatico. Le tecnologie del Semantic Web consentono di realizzare sistemi informativi che utilizzano ontologie per la ricerca, l' integrazione e l' analisi, con evidenti miglioramenti della qualità dell' informazione e della conoscenza.

1 Introduzione

La classica interfaccia verso i sistemi di reperimento dell' informazione, basata sul confronto di sequenze di caratteri, non è in grado di identificare i concetti rilevanti che dovrebbero comparire nelle risorse reperite dal sistema. I termini specificati per formulare la *query* sono sostanzialmente utilizzati ancora come descrittori sintattici del contenuto dei documenti. L' attenzione di molti ricercatori si è focalizzata sul "*semantic search*", per consentire una ricerca dell' informazione basata sui concetti di interesse. Un aspetto da considerare nella realizzazione di una efficace ricerca dell' informazione è anche quello degli interessi specifici dell'utente, che possono essere modellati con un profilo di interesse. Sulla base delle informazioni semantiche che descrivono il contenuto dei documenti e gli interessi dell' utente è possibile migliorare l' efficacia della ricerca.

Le ontologie giocano un ruolo di primo piano nel processo di rappresentazione e utilizzo della conoscenza, essenziale per supportare un approccio al reperimento dell' informazione che garantisca risultati di qualità.

In questo lavoro viene prima presentato brevemente il modello concettuale dell' Information Retrieval, e successivamente viene data una breve descrizione del concetto di ontologia, delle tecnologie di base del Semantic Web, e del ruolo che le ontologie possono svolgere per migliorare la qualità dell' informazione.

2 Il modello concettuale dell' Information Retrieval

I sistemi di Information Retrieval sono basati su un modello funzionale che individua la *corrispondenza* tra l' insieme delle domande e l' insieme dei documenti, mediante un linguaggio intermedio di rappresentazione, detto *linguaggio di indicizzazione* o *linguaggio di classificazione*. Più in dettaglio, i passi sono i seguenti:

- l' insieme dei documenti viene analizzato e indicizzato in maniera opportuna;
- le domande degli utenti vengono analizzate e formalizzate;
- un processo determina la somiglianza tra la rappresentazione del documento e la rappresentazione della domanda;
- il sistema ritrova e restituisce i documenti ritenuti simili alla domanda.

Il processo di indicizzazione normalmente adottato, cioè l' estrazione dal testo di tutte le parole, ad esclusione delle più comuni, è di tipo essenzialmente sintattico, e ben noti sono i problemi derivanti da sinonimie, polisemie e iperonimie e iponimie.

Un approccio ben noto, e ampiamente discusso in letteratura, per identificare il grado di similitudine tra una domanda e un documento è quello dello spazio vettoriale dei documenti, in cui sia i documenti che le domande vengono rappresentate come vettori in uno spazio *n*-dimensionale, dove *n* sono le parole chiave (o termini indice). Il grado di similitudine tra la domanda e il documento è quindi calcolato come il coseno dell' angolo tra i due vettori.

Per misurare l' efficacia di un sistema di Information Retrieval, e quindi la qualità dell' informazione, si usano due parametri: precisione (*precision*) e richiamo (*recall*). La precisione (P) è il rapporto tra il numero di documenti reperiti e pertinenti e il numero di documenti reperiti. Il richiamo (R) è il rapporto tra il numero di documenti reperiti e pertinenti e il numero di documenti pertinenti esistenti nell' intera collezione. Evidentemente sia R che P sono due numeri compresi tra zero e uno, e vanno considerati entrambi nel valutare l' efficacia del sistema (il sistema "perfetto" avrebbe il valore uno per entrambi i parametri).

È esperienza comune quella della scarsa precisione, tutti gli utenti dei sistemi di ricerca dell' informazione sperimentano come un gran numero di documenti restituiti a fronte di una interrogazione siano poco o per nulla pertinenti (fenomeno noto anche come "*effetto rumore*"). Molto meno evidenti sono invece gli effetti di un basso valore di R, cioè il mancato reperimento di documenti pertinenti, noto anche come "*effetto silenzio*", in quanto è abbastanza difficile riuscire a individuare i documenti pertinenti che non sono stati restituiti dal sistema, a meno che non si conosca perfettamente il contenuto della collezione di documenti.

Si noti che la scarsa efficacia del sistema di Information Retrieval, ovvero un basso valore di P e di R, non è da ascrivere a difetti intrinseci del sistema, ma alle caratteristiche del processo di indicizzazione e alla metodologia adottata. Non ci dilunghiamo in questa sede su come individuare i termini indice, limitandoci a ricordare che un buon termine indice deve essere in grado di caratterizzare il singolo documento e di distinguerlo nell' ambito dell' intera collezione.

In effetti, i problemi sono legati alla scelta dei termini indice e all' esistenza di idonei supporti per poter formulare l' interrogazione. In primo luogo, i termini da utilizzare come termini indice non devono essere né troppo specifici (quindi con frequenza bassa) né troppo generici (quindi con frequenza alta), in quanto i primi individuano solo una piccola frazione dei documenti disponibili, e i secondi una porzione troppo ampia. Tradizionalmente si ovvia operando una *trasformazione di*

frase, che accorpa più termini con alta frequenza in una frase (che è di frequenza più bassa) o una *trasformazione di thesaurus* (che accorpa termini più specifici come elementi di una classe di termini) consentendo così di utilizzare termini più generici, quindi con frequenza più alta. In secondo luogo, una ricerca può essere efficace solo se l' indicizzatore e l' utente *condividono la stessa base di conoscenza*, e quindi se sono disponibili strumenti per rappresentare la conoscenza e renderla disponibile.

La qualità dell' informazione deriva dalla possibilità di operare indicizzazioni e ricerche di tipo semantico. Le trasformazioni di frase e di thesaurus sono state un primo passo in questa direzione.

3 La ricerca di informazioni sul Web

I moderni motori di ricerca indicizzano l' intero contenuto delle pagine web, ma in pratica la ricerca viene sempre eseguita verificando la presenza o meno, nel testo della pagina, delle sequenze di caratteri (parole o frasi) specificate dall' utente (*term matching*). In molti casi vengono utilizzati algoritmi e tecniche di Information Retrievali anche sofisticati per ordinare i risultati in ordine di pertinenza, come il *tf/idf ranking model* o il Google PageRank. Si tratta comunque sempre di un approccio di tipo sintattico, mentre si potrebbero ottenere notevoli miglioramenti con un approccio semanticamente più ricco, che permetta di comprendere meglio il significato della *query* formulata dall' utente e di individuare il contenuto delle pagine web. In questo modo i sistemi potrebbero eseguire delle ricerche più mirate e restituire i risultati effettivamente interessanti per l' utente. Probabilmente i motori di ricerca di nuova generazione non sostituiranno i vecchi, ma gli si affiancheranno, con una migliore integrazione tra l' aspetto sintattico e quello semantico.

Il Semantic Search può essere definito come un' applicazione del Semantic Web alla ricerca di informazioni, una delle attività più comuni sul web. Va osservato che molti utenti preferiscono formulare query utilizzando concetti semantici ad alto livello, più coerenti con la nomenclatura standard e la conoscenza tacita, piuttosto che specificando semplicemente delle parole da utilizzare a livello meramente sintattico. Il Semantic Search può migliorare i risultati ottenibili mediante la ricerca tradizionale, basata sulle tecnologie di Information Retrieval e quindi sulla presenza di parole nel testo, grazie alla disponibilità di informazione strutturata e comprensibile alle macchine per un ampio numero di oggetti presenti nel Semantic Web.

Possiamo distinguere tra la ricerca navigazionale (*navigational search*), in cui l' utente specifica dei termini e chiede al sistema di restituire i documenti che li contengono, e ricerca di informazioni (*research search*), in cui l' utente specifica una frase che denota un oggetto relativamente al quale desidera ottenere informazione. In questo caso, non esiste un particolare documento che contiene l' informazione richiesta, ma l' utente sta cercando di ritrovare un insieme di documenti che nel loro complesso sono in grado di fornirgli l' informazione desiderata, e le tecnologie del Semantic Web possono essere di grande utilità.

Il Semantic Search può migliorare i risultati della ricerca permettendo una migliore definizione dei termini o concetti da ricercare, o arricchendo la lista dei risultati, o aiutando a comprendere il testo dei documenti. In tutti questi processi potrebbe utilizzare ontologie di supporto e una definizione ontologica degli interessi dell' utente.

4 Le ontologie

Il termine ontologia deriva dalla filosofia, dove viene inteso come una spiegazione sistematica dell' essere. Negli anni recenti il termine si è ampiamente diffuso nella comunità del Knowledge Engineering. Esistono diverse definizioni di ontologia, ognuna dele quali mette in evidenza qualche aspetto. Secondo la definizione di Neches et al. (1991):

> *An ontology defines the basic terms and relations comprising the vocabulary of a topic area as well as the rules for combining terms and relations to define extensions to the vocabulary.*

Da notare che un' ontologia include non solo i termini che sono esplicitamente definiti in essa, ma anche la conoscenza che ne può essere derivata mediante un processo di inferenza.
Secondo la definizione data in Studer et al., che riprende e chiarisce quelle date precedentemente da Gruber (1993) e Borst (1997)

> *An ontology is a formal, explicit specification of a shared conceptualisation. A 'conceptualisation' refers to an abstract model of some phenomenon in the world by having identified the relevant concepts of that phenomenon. 'Explicit' means that the type of concepts used, and the constraints on their use are explicitly defined. For example, in medical domains, the concepts are diseases and symptoms, the relations between them are causal and a constraint is that a disease cannot cause itself. 'Formal' refers to the fact that the ontology should be machine readable, which excludes natural language. 'Shared' reflects the notion that an ontology captures consensual knowledge, that is, it is not private to some individual, but accepted by a group.*

Guarino (1998) definisce a sua volta un' ontologia come:

> *A set of logical axioms designed to account for the intended meaning of a vocabulary.*

Talvolta il concetto di ontologia viene inteso in modo meno rigido, per cui anche le tassonomie vengono considerate ontologie, perché offrono una concettualizzazione di un dominio su cui c'è il consenso di una vasta comunità. Si distingue spesso tra ontologie *lightweight* (che sono sostanzialmente delle tassonomie) e le ontologie *heavyweight* (che forniscono un modello del dominio più accurato e includono delle restrizioni sulla semantica del dominio). Le prime includono i concetti e le tassonomie di concetti, le relazioni tra concetti, e le proprietà che descrivono i concetti. Le ontologie *heavyweight* aggiungono a questo assiomi e vincoli, che chiariscono il senso inteso *(intended meaning*) dei termini presenti nell' ontologia.

Abbiamo riportato solo alcune delle tante definizioni di ontologia, ma l' esistenza di un gran numero di definizioni non deve far ritenere che possa sorgere confusione sul significato che la comunità scientifica che opera nel settore attribuisce a questo termine. Le varie definizioni enfatizzano di volta involta alcuni aspetti, ma in realtà forniscono una serie di punti di vista complementari. In alcuni casi la definizione è indipendente dal processo seguito per la costruzione di un' ontologia e il suo utilizzo nelle applicazioni, in altri casi viene invece data rilevanza al processo seguito per svilupparla. Va piuttosto posto l'accento su come le ontologie mirino a catturare la conoscenza *consensuale*, e possano essere condivise e riutilizzate tra applicazioni e gruppi di persone diversi.

Le ontologie vengono costruite in genere mediante un processo cooperativo e distribuito, e utilizzano varie tecniche di modellazione della conoscenza e diversi tipi di linguaggi. Possono perciò essere *molto informali*, *semi-informali*, *semi-formali* o *rigorosamente formali* a seconda che siano espresse in linguaggio naturale, in linguaggio naturale ristretto, in un linguaggio artificiale e definito formalmente, o fornendo una descrizione meticolosa dei termini, utilizzando una semantica formale, teoremi e dimostrazioni di proprietà. Nella classificazione delle ontologie basata sulla ricchezza della loro struttura interna, i vocabolari controllati e i thesauri si collocano nella parte bassa delle ontologie informali, mentre le ontologie in cui vengono espressi dei vincoli sui possibili valori si collocano nella parte alta delle ontologie formali.

È stato osservato come le ontologie semi-formali si siano dimostrate in pratica molto utili per raggiungere diversi obiettivi importanti, in particolare l' *information integration*. Rispetto alle ontologie rigorosamente formali, le ontologie semi-formali sono più diffuse e spesso più utili, perché possono essere realizzate a una scala adeguata per le applicazioni reali e richiedono uno sforzo di sviluppo certamente minore. La loro diffusione è legata alla necessità di dover rappresentare informazione parziale (quindi incompleta) e talvolta non completamente coerente, in particolare per quanto concerne le asserzioni. Per esempio, la Gene Ontology, che è più una nomenclatura e una tassonomia che una vera e propria ontologia, e presenta alcune incoerenze, è largamente usata con molto successo per annotare grosse moli di documenti, supportando così l' interoperabilità e l' integrazione di fonti informative diverse (http://www.geneontology.org/), dimostrando che possono essere sviluppate applicazioni reali aggiungendo un po' di semantica, o venendo a un compromesso con la completezza e il rigore richiesti da una rappresentazione più formale e dalle tecniche di inferenza ("*little semantics goes a long way*" – Jim Hendler).

5 Il Semantic Web: alcune tecnologie

5.1 I metadati e RDF

Ricordiamo che la filosofia di base del web è quella di uno spazio informativo universale, navigabile, con un *mapping* da URI alle risorse. Il Semantic Web si basa sull' ipotesi che le macchine possano accedere ad un *insieme strutturato di informazioni* e a un *insieme di regole di inferenza* da utilizzare per il ragionamento automatico. La sfida del Semantic Web è fornire un linguaggio per esprimere *dati* e *regole* per ragionare sui dati, che consenta l' *esportazione* sul web delle regole da qualunque sistema di rappresentazione della conoscenza.

Nel navigare sul web i link portano a quella che formalmente viene detta *risorsa* (resource) identificata univocamente da un URI. Le informazioni sulla risorsa vengono generalmente dette "*metadati*", definiti come *"informazioni, comprensibili dalla macchina, relative a una risorsa web o a qualche altra cosa*".

L'uso efficace dei metadati richiede che vengano stabilite delle convenzioni per la *semantica*, la *sintassi* e la *struttura*. Le *singole comunità* interessate alla descrizione delle loro risorse specifiche definiscono la *semantica* dei metadati pertinenti alle loro esigenze. Resource Description Framework (RDF) è lo strumento base per la codifica, lo scambio e il riutilizzo di metadati strutturati, e consente l'interoperabilità tra applicazioni che si scambiano sul web informazioni *machine-understandable*. Il *data model* RDF consente di rappresentare statement RDF in

modo sintatticamente neutro ed è basato su tre tipi di oggetti: *Resources* (sempre individuate da un URI), *Properties* (un aspetto specifico, identificato da un "nome", che assume un "valore"), *Statements* (triple "*soggetto-predicato-oggetto*", ovvero "*risorsa-proprietà-valore*").

Grazie ai metadati, e ad un formalismo per esprimerli, il Semantic Web ha già una struttura semantica *esplicita* distribuita e a grande scala, costruita in maniera indipendente dal testo che si intende ricercare.

5.2 Web Ontology Language

Applicazioni sofisticate richiedono di poter "ragionare" sui dati. Il Semantic Web deve quindi essere supportato da ontologie, e disporre di un linguaggio che consenta di definire la terminologia usata, le caratteristiche logiche e i vincoli delle proprietà, l' equivalenza dei termini, le cardinalità delle associazioni, etc. Un' ulteriore complessità deriva dal fatto che il web è intrinsecamente distribuito, e di conseguenza applicazioni diverse possono usare ontologie diverse, o le stesse ontologie, ma espresse in lingue diverse. Il W3C, sfruttando anche i risultati di altri progetti, quali DAML e OIL, ha definito un linguaggio, denominato OWL, che permette di esportare le ontologie in modo interoperabile. OWL offre tre sottolinguaggi, di crescente potere espressivo: *OWL Lite* (per rappresentare classificazioni gerarchiche e vincoli semplici), *OW DL* (per una maggiore potenza espressiva, garantendo comunque che tutte le conclusioni siano computabili e concluse in un tempo finito), *OWL Full* (che offre la massima potenza espressiva, ma non fornisce garanzie sui tempi di computazione, e difficilmente sarà supportato nella sua interezza da software che implementano il ragionamento). Ognuno di questi linguaggi è un' estensione del precedente, sia in termini di ciò che può essere espresso che in termini della validità delle conclusioni.

6 Supporto delle ontologie nella ricerca dell' informazione

Sono molti i modi in cui le ontologie possono essere utilizzate per migliorare la ricerca dell' informazione, sia come efficacia che come interfaccia utente. D' altra parte, questo aspetto è stato sempre considerato nell' ambito dell' Information Retrieval, dove per anni si è discusso di indicizzazione semantica e di strumenti di rappresentazione della conoscenza come thesauri e dizionari controllati. Le ontologie, più ricche e formali di un thesaurus, possono portare ulteriori vantaggi. Ricordiamo, a titolo di esempio dell' efficacia delle *ontologie di dominio* per migliorare la ricerca di informazioni, il sito GoPubMed (http://www.gopubmed.org/), fonte di citazioni nel settore biomedico, in cui la ricerca, la presentazione dei risultati e la navigazione vengono migliorate grazie all' utilizzo della Gene Ontology. I risultati delle query vengono ordinati in base all' ontologia, e vengono generati termini addizionali per la ricerca, evidenziati nella presentazione degli abstract.

Un' ontologia è una rappresentazione formale di conoscenza condivisa. Dalla sua forma più semplice (tassonomia, vocabolario controllato, thesaurus) a quella più complessa e formale può quindi egregiamente assolvere il compito di *mettere a comune la base di conoscenza* di chi indicizza i documenti e di chi li ricerca, indipendentemente dal fatto che essi siano esseri umani o macchine.

I risultati di una ricerca tradizionale sono una lista di documenti o pagine Web. È possibile *arricchire* questa lista con dati estratti dal Semantic Web, utilizzando le

proprietà definite nell' ontologia. Per esempio, a seguito della ricerca per un gruppo musicale, i risultati potrebbero essere arricchiti aggiungendo il calendario degli eventi, gli album pubblicati, etc. Analogamente, se si ricerca un aeroporto, il sistema potrebbe presentare un ventaglio di possibilità di collegamenti con le città servite da quell' aeroporto, e indicare quelli preferibili sulla base dell' ora di arrivo.

Il tipo di utente può essere preso in considerazione per selezionare, tra i documenti reperiti, quelli che maggiormente possono essere interessanti. Questo *profilo utente* può essere esplicito o derivato da un' analisi del suo comportamento (per esempio i link seguiti, i documenti scaricati, etc.).

È pensabile che non tutti i documenti siano corredati di metadati, e che sia necessario estrarli dall' analisi del contenuto. Le ontologie possono essere utilizzate nell' elaborazione del linguaggio naturale (NLP – Natural Language Processing) per estrarre i metadati, mediante una analisi del contesto e la sua *comprensione*.

Abbiamo già accenato alla *data integration* come ad un problema di grande rilevanza. Le ontologie, se espresse in OWL, possono essere distribuite sul Web, per condividere la conoscenza e consentire il reperimento di informazioni espresse in lingue diverse o con riferimenti ad ontologie diverse.

È possibile implementare uno *spazio vettoriale dei concetti*, simile allo spazio vettoriale dei documenti, per implementare un Intelligent Semantic Search basato slle relazioni semantiche tra concetti. In questa prospettiva, degli intelligent software agent potrebbero valutare se i documenti restituiti sono rilevanti rispetto alla query, e decidere autonomamente se renderla più specifica o più generica, per trovare dei risultati significativi per l' utente.

7 Conclusioni

L' approccio tradizionale al reperimento di informazioni è essenzialmente sintattico, e non è in grado di soddisfare adeguatamente le esigenze degli utenti, che sono interessati all' aspetto semantico dell' informazione. Le ontologie permettono di rappresentare la conoscenza, e di renderla disponibile in modo interoperabile grazie alle tecnologie del Semantic Web, favorendo la *data integration*.

La qualità dell' informazione può essere molto migliorata se i sistemi utilizzano le *ontologie* per mettere a comune una base di conoscenza distribuita di supporto alla formulazione delle query, per comprendere il significato dei documenti e corredarli di metadati, e per arricchire la lista dei documenti restituiti, sfruttando le relazioni definite a livello ontologico. Un miglioramento della precisione e del richiamo è anche ottenibile da un' opportuna modellazione ontologica degli interessi dell' utente e da approcci come lo spazio vettoriale dei concetti.

Le tecnologie del Semantic Web, su cui lavora attivamente il World Wide Web Consortium, sono la chiave per sfruttare adeguatamente queste possibilità e sviluppare applicazioni tecnologicamente avanzate.

Riferimenti bibliografici

Neches, R., Fikes, R., Finin, T., Gruber, T., Patil, R., Senator, T., and Swartout, W. R. (1991) **in** *Enabling Technology for Knowledge Sharing*. AI Magazine, 12(3) pp. 16-36

Rudi Studer and V. Richard Benjamins and Dieter Fensel **in** *Knowledge Engineering: Principles and Methods*, Data Knowl. Eng. 25(1-2): 161-197 (1998)

Dave Beckett's Resource Description Framework (RDF) Resource Guide, http://planetrdf.com/guide/

Web Ontology Language (OWL), http://www.w3.org/2004/OWL/

Resource Description Framework (RDF), http://www.w3.org/RDF/

Signore, Oreste: *Strutturare la conoscenza: XML, RDF, Semantic Web* - Clinical Knowledge 2003 (1st edition) - Udine, 20-21 September 2003 http://www.w3c.it/papers/ck2003.pdf, http://www.w3c.it/talks/ck2003/

T. Berners-Lee, J. Hendler, O. Lassila: *The Semantic Web*, **in** Scientific American, May 2001

Asunción Gómez-Pérez, Mariano Fernández-López, Oscar Corcho, *Ontological Engineering*, Springer-Verlag (2004), ISBN 1-85233-551-3

Salton G., McGill M.J., *Introduction to modern Information Retrieval*, McGraw-Hill, New York, 1983

Dario Bonino, Fulvio Corno, Laura Farinetti, Alessio Bosca, *Ontology Driven Semantic Search*, WSEAS Conference ICAI 2004, Venice, Italy, 2004

R. Guha, Rob McCool, Eric Miller, Semantic Search, **in** Proceedings of the 12th international conference on World Wide Web, Budapest, Hungary (2003) ISBN:1-58113-680-3

Amit Sheth, Cartic Ramakrishnan, *Semantic (Web) Technology In Action: Ontology Driven Information Systems for Search, Integration and Analysis*, **in** IEEE Data Engineering Bulletin, Special issue on Making the Semantic Web Real, U. Dayal, H. Kuno, and K. Wilkinson, Eds. December 2003.

Signore, Oreste: *Ontology Driven Access to Museum Information* **in** CIDOC 2005 Annual Conference of the International Committee for Documentation of the International Council of Museums ICOM-CIDOC - May 24 -27, 2005 Zagreb, Croatia - ISBN 953-6942-15-1

Signore, Oreste, *Semantic Web: il futuro è già qui?* - 10th Knowledge Management Forum - Siena, 24-25 Novembre 2005