Il futuro del motore di ricerca: "cerca e troverai". Un'intervista con Frank van Harmelen*

Frank Van Harmelen[1]

[1]Frank.van.Harmelen@cs.vu.nl http://www.cs.vu.nl/~frankh/

AI Department, Division of Mathematics and Computer Science, Faculty of Sciences, Vrije Universiteit Amsterdam, de Boelelaan 1081a, 1081HV Amsterdam , The Netherlands

Abstract. The paper presents, in form of interview, the author's view on the Semantic Web and its development in the future.

Sommario. L'articolo presenta, in forma di intervista, le opinioni dell'autore sul Semantic Web e sui suoi sviluppi futuri.

Il Web è talvolta descritto come una enorme biblioteca dove qualcuno ha gettato tutti i libri in una grande pila sul pavimento, e dove dobbiamo indossare una benda sugli occhi quando cerchiamo l'informazione giusta. I motori di ricerca come Google ci sono naturalmente di grande aiuto, ma talvolta perfino loro soccombono al caos che troviamo nell'attuale World Wide Web. Cambierà mai tutto questo? In questa intervista proviamo a dare un'occhiata al futuro dei motori di ricerca.

Il Consorzio per il World Wide Web (o W3C), l'organizzazione che promuove lo sviluppo e l'evoluzione del World Wide Web, è ben consapevole della crescita turbolenta del Web e di tutti i problemi che questa porta con sé. Già da qualche tempo, quindi, si è messo all'opera per trovare soluzioni per questo tipo di problemi. La risposta, che dovrebbe costituire la prossima generazione dell'attuale World Wide Web, si chiama Semantic Web. Abbiamo chiesto a Frank van Harmelen, professore di Computer Science alla Vrije Universiteit di Amsterdam, e attivamente impegnato nello sviluppo del Semantic Web, di spiegarci di che cosa si tratta.

D: "Cominciamo con la domanda più ovvia: che cos'è il Semantic Web?"

Van Harmelen: "Il modo migliore per spiegarlo è esaminare i difetti del World Wide Web attuale. Da un lato è un grandissimo successo – dieci anni fa nessuno avrebbe potuto predire che il Web avrebbe esercitato una influenza così enorme sulla nostra vita quotidiana, personale e professionale – ma dall'altro ha molti difetti signi-

* Traduzione di Margherita Benzi

Networks **2**: 7-13, 2003
 http://www.swif.uniba.it/lei/ai/networks/

ficativi. Il Web attuale si può usare soltanto se si parla l'inglese (o qualche altra lingua molto diffusa), o se si è in grado di riconoscere le immagini e i disegni. Gli umani lo sanno fare molto bene, ma i computer non lo sanno fare per niente. I computer non sanno affatto come affrontare il Web attuale, almeno per quanto riguarda i suoi contenuti. Di conseguenza, al momento presente noi otteniamo soltanto un aiuto molto limitato dai nostri computer quando cerchiamo informazioni sul Web: il solo lavoro che i nostri costosi PC fanno per noi consiste nello spostare informazioni da un posto all'altro, e poi mostrarcele sullo schermo. Ma tutte le attività come comprendere le informazioni, combinarle, interpretarle, selezionarle e giudicarle sono lasciate esclusivamente all'utente umano. Il computer non ci può essere d'aiuto, semplicemente perché non capisce che cosa dicono tutte quelle pagine Web. Che dire, allora, del Semantic Web? Beh, l'idea fondamentale alla base del Semantic Web è di provare a estendere il Web attuale con ulteriori informazioni che rendano possibile per il computer l'effettiva comprensione del contenuto delle pagine Web. Naturalmente, questo non significa che il Web attualmente esistente debba essere eliminato o sostituito: il Semantic Web è un'estensione, uno strato aggiuntivo posto al di sopra del Web attuale; questo significa semplicemente che dovremmo arricchire l'informazione nelle pagine correnti per renderle comprensibili ai computer. E siamo molto vicini a ottenere questo risultato."

D: "Può dirci qualcosa di più sul modo in cui ciò sarà fatto?"

Van Harmelen: "Abbiamo sviluppato diversi linguaggi che possono essere elaborati dai computer e che possono descrivere al computer qual è il contenuto di una particolare pagina Web. Per esempio, voi potreste affermare in uno di questi linguaggi che vi è qualcosa che viene chiamato la 'Vrije Univesiteit di Amsterdam', che esiste qualcuno chiamato 'Frank van Harmelen', e che tra questi due oggetti sussiste una relazione specifica, e cioè 'è un dipendente della'. Potreste anche definire il concetto di 'edificio', il fatto che un edificio specifico 'è parte della' Vrije Universiteit, e che Frank van Harmelen 'lavora in' quell'edificio. Naturalmente tutte queste relazioni devono essere definite: le relazioni tra me e l'edificio ('lavora in') è fondamentalmente diversa dalla relazione tra l'università e quell'edificio ('parte di'). Quando descrivete tutte queste cose in uno dei linguaggi che abbiamo costruito, il computer capirà che cosa state cercando semplicemente perché gliel'avrete spiegato prima. In questo modo il computer potrebbe già darvi un supporto molto più efficace nella vostra ricerca di una specifica informazione."

D: "Allora l'intero sistema dipende da come si forniscono al computer tutte queste informazioni?"

Van Harmelen: "Proprio così. Facciamo un altro esempio: immaginate di stare cercando l'indirizzo di lavoro di Frank van Harmelen. Forse con Google non lo troverete, perché sono stato troppo pigro per mettere quest'informazione sul mio sito Web. Ma se il computer sapesse che io lavoro per la Vrije Universiteit, e se potesse trovare l'indirizzo dell'università, allora potrebbe dedurre che questo è un indirizzo utile da fornire come risposta alla vostra domanda. Naturalmente, tutto ciò dipende dall'informazione di sfondo (la conoscenza) che abbiamo dato al computer. Dunque,

in effetti, tutto dipenderà dalla qualità di quella che chiamiamo l’*ontologia*: una collezione di termini e di relazioni tra questi termini.”

D: “In altre parole, la qualità dell’ontologia determina la qualità dell’aiuto che il computer può darci”.

Van Harmelen: “Precisamente. Potreste considerare un’ontologia come un modo strutturato di rappresentare il significato delle parole in un dominio dato. Per tornare all’esempio: perché le cose funzionino, dovete spiegare al computer che cosa è un’università, che cosa è un dipendente, qual è la relazione tra i due (per esempio, che l’indirizzo di lavoro del dipendente è l’indirizzo dell’università), e così via. Tutte queste informazioni insieme sono chiamate *metadati*. E questo è proprio ciò che ci serva per creare il Semantic Web. Senza questi metadati non ci sarà nessun Semantic Web.”

D: “Ma da dove verranno fuori tutti questi metadati?”

Van Harmelen: (sorride) “Questa è proprio la domanda che mi fanno più spesso quando faccio una conferenza di presentazione del Semantic Web. Naturalmente gli utenti non scriveranno questi metadati. Se guardiamo alle origini del Web attuale – diciamo le prime centomila pagine o giù di lì – allora vedremo che quelle pagine erano ancora scritte dalle persone, che usavano la tastiera per digitare pagine HTML, ma naturalmente ora è diverso: non abbiamo accumulato i più di tre miliardi di pagine del Web attuale digitandole sulla tastiera! Quello che invece succede è che la maggior parte delle pagine sono generate da database, da programmi per computer, e così via. Ebbene, in futuro quei database e quei programmi non genereranno soltanto gli HTML (destinati ad essere consumati dagli umani), ma anche i metadati (destinati ad essere consumati dai computer). Un semplice esempio è dato dal sito Web di Amazon. Com’è ovvio, quel sito Web è generato semplicemente da un database. Tutta l’informazione contenuta in quel database è trasformata in pagine HTML in maniera tale che noi, gli utenti umani, possiamo leggerla e comprenderla. Ma naturalmente Amazon potrebbe anche usare lo stesso database per generare i metadati in un linguaggio che sia accessibile a un computer, e questo consentirebbe al mio agente personale per lo shopping di assistermi nella ricerca di libri o di prodotti musicali conformi ai miei gusti personali, ancora una volta descritti sotto forma di metadati. Dunque questi database sono già una grande fonte di metadati. Un’altra fonte importante di metadati sono i programmi specializzati che possono comprendere – in maniera superficiale – i linguaggi naturali quali, per esempio, l’inglese o l’italiano, e che sono in grado di generare metadati da enunciati in linguaggio naturale. Questi programmi esistono già ed esistono già imprese che debbono loro i propri proventi. Per riassumere: i metadati saranno per lo più generati da macchine in maniera automatica o semiautomatica”.

D: “Ma affinché sia possibile lo scambio di metadati, questi non dovrebbero essere in qualche modo standardizzati?”

Van Harmelen: “Questo è un aspetto molto importante. Per esempio, se io parlo di "dipendente", e qualcun altro parla di "membro del personale", il computer deve sapere che le due espressioni designano il medesimo concetto, in maniera tale che quando cerca "dipendente" dovrebbe anche cercare "membro del personale". Questo è precisamente quanto speriamo che il Semantic Web faccia per noi. I motori di ricerca attuali si limitano nella maggior parte dei casi a confrontare delle stringhe di caratteri. D’accordo, fanno anche delle cose un po’ più intelligenti, ma fondamentalmente si basano sul reperimento della medesima sequenza di cifre all’interno delle pagine Web. L’uso di ontologie dovrebbe porre un rimedio a questo problema: non ci limiteremo più a cercare le stringhe che corrispondono a "dipendente", ma il concetto "dipendente" e le sue relazioni con altri concetti saranno definiti all’interno un’ontologia sottostante.

Naturalmente, quando qualcun altro usa la parola "unità di personale", dovrebbe fare riferimento ad un’ontologia simile, e queste due ontologie dovrebbero essere collegate l’una all’altra. Soltanto allora il computer sarà in grado di comprendere che questi due concetti sono il medesimo, e che l’uno può essere usato quando si cerca l’altro. Affinché tutto questo funzioni sono davvero necessari linguaggi standardizzati per i medatati, quali quelli sviluppati da W3C.”

D: “Il computer non può farselo da solo il collegamento? Non abbiamo la tecnologia per far sì che ciò avvenga?”

Van Harmelen: “Questo è un tema di ricerca di importanza centrale. In realtà riusciamo già a ottenere questo risultato negli esperimenti quando usiamo domini di prova scelti accuratamente, ma non riusciamo ancora a farlo nel grande mondo aperto del Web reale. Prevedo che il problema del collegamento tra ontologie darà origine a un settore di mercato completamente nuovo.

Vi sono già aziende che vendono ontologie: per esempio, un’azienda potrebbe avere un’ontologia commerciale molto grande, che descrive, tra gli altri, termini quali "datore di lavoro", "dipendente", "membro del personale", "indirizzo", "prodotto", "prezzo", insieme a una descrizione delle relazioni tra questi termini. Voi pagate, e ottenete il permesso di collegarvi a termini appartenenti a questa ontologia; successivamente, l’azienda provvede a collegare la propria ontologia con le altre ontologie. In altri termini, l’azienda vi sta fornendo una specie di servizio di indicizzazione semantica, che consente ad altre persone di trovare la vostra informazione in maniera più veloce e più facile. È verosimile che siate disposti a pagare per ottenere un servizio di questo tipo”.

D: “Chi naviga sul Web abitualmente si accorgerà di qualcosa? Da quello che mi sta dicendo, appare più che altro come un’operazione dietro le quinte. Che cosa cambierà per il comune internauta?”

Van Harmelen: “Il Semantic Web sarà un successo soprattutto se rimarrà invisibile. Tutta la tecnologia di cui abbiamo parlato fino ad ora è davvero "sotto la superficie". La sola cosa che si noterà navigando sul Web è che la qualità dei risultati che vengono restituiti dai motori di ricerca sarà molto migliore rispetto al passato. I motori di ricerca attuali funzionano molto bene per quanto riguarda il *recall*: trovano tutto

quello che c’è da trovare. Ma non sono altrettanto soddisfacenti per quanto riguarda la precisione: oltre a trovare quello che stavate cercando, vi danno un sacco di cose in più che non volevate. Naturalmente sto un po’ esagerando, ma il livello attuale di precisione potrebbe essere molto migliorato. E mi aspetterei anche un miglioramento del modo in cui l’informazione viene presentata all’utente. Per esempio, se io ora digito il mio nome ‘Van Harmelen’ in un motore di ricerca, ottengo due tipi di risultati: alcuni riguardano me e il mio lavoro scientifico, mentre altri riguardano il villaggio olandese Harmelen. Il problema è che il motore di ricerca non può distinguere tra i due, e semplicemente mette tutta l’informazione in un singolo grande elenco. Quando il Semantic Web sarà ultimato, il motore di ricerca sarà in grado di determinare che ci sono in realtà due tipi di risultati, e che questi dovrebbero essere mostrati separatamente; oppure, meglio ancora, dovrebbe chiedermi che cosa stavo cercando, se la persona Frank van Harmelen o il villaggio Harmelen”.

D: “Dunque la ricerca diventerà più facile. Ci sono altri vantaggi?”

Van Harmelen: “Un aspetto importante che non abbiamo discusso è quello della personalizzazione. Se io guardo un particolare sito Web e tu guardi lo stesso sito, vediamo entrambi la medesima informazione. Naturalmente, questa situazione è lungi dall’essere ideale, dal momento che i tuoi interessi sono molto diversi dai miei. Prendiamo ancora Amazon come esempio: non guadagnerebbero molto se potessero riuscire a mostrarti una pagina diversa da quella che mostrano a me, confezionata su misura rispetto ai tuoi interessi? Si potrebbe persino usare la personalizzazione per ridurre il sovraccarico di informazione di cui oggi tutti soffriamo: si può evitare di presentarti quanto non corrisponde al profilo dei tuoi interessi.”

D: “Quante persone stanno lavorando in questo momento al Semantic Web?”

Van Harmelen: “Al momento il W3C è un’organizzazione abbastanza ristretta. Ha molti membri, ma la quantità di personale è molto limitata: una dozzina di persone in tutto il mondo, e molte di loro si occupano anche di altre cose oltre al Semantic Web. Sono i dipendenti delle organizzazioni membri del W3C a svolgere il lavoro che conta: nei gruppi di lavoro sul Semantic Web potete trovare gente dell’IBM, della Hewlett-Packard, della Sun, ma anche della Nokia, della Philips e della Daimler-Chrysler. Alcuni di questi nomi potrebbero sorprendervi, ma molte organizzazioni differenti hanno buone probabilità di guadagnare parecchio dal Semantic Web. Per esempio, la Hewlett-Packard vede la possibilità di usare il Semantic Web per trasformare le proprie stampanti in dispositivi che si autodescrivono: ogni stampante avrà il suo specifico profilo, scritto in un linguaggio del Semantic Web, e che cosa succederà? Succederà che voi entrate in un edificio, per esempio un centro congressi, e tutte le stampanti si renderanno note al vostro laptop o al vostro PDA, e quando vorrete stampare qualcosa, il vostro laptop o il vostro PDA avrà già deciso qual è la stampante più vicina a voi e più adatta per un certo lavoro specifico. E un’impresa come Nokia punta a rendere disponibili molti nuovi servizi di telefonia mobile attraverso i propri apparecchi. Potete quindi comprendere perché tutte queste aziende partecipino allo sviluppo del Semantic Web: nessuno vuole rimanere indietro.”

D: “E per quanto riguarda Google, Alta Vista, Yahoo, ecc., sono anch’essi coinvolti nello sviluppo del Semantic Web?”

Van Harmelen: “Ho sentito di recente alcune persone di Google e sono stato sorpreso nel vedere quanto fossero … beh, direi garbatamente riluttanti. Conoscevano bene gli ultimi sviluppi, ma ci hanno detto che non stavano sfruttando attivamente la tecnologia. Ma nel contempo si può vedere che stanno sperimentando. Avete mai sentito parlare della Open Directory? È un progetto dove migliaia di volontari classificano innumerevoli pagine Web secondo decine di migliaia di categorie. Ebbene, Google sta già collegando i suoi risultati di ricerca secondo la gerarchia tematica della Open Directory. Per molti risultati delle vostre ricerche troverete già un collegamento ipertestuale alla categoria nella quale quel risultato è classificato secondo nella Open Directory; questo ci dà la possibilità di andare a vedere le altre voci che appartengono a quella categoria. Quindi, senza ammetterlo, stanno già usando la tecnologia del Semantic Web, ed è naturale che non possano permettersi di ignorarla: dopotutto la popolarità di un motore di ricerca è determinata soltanto dalla qualità dei risultati che esso produce.”

D: “Altre applicazioni?”

Van Harmelen: “Un esempio concreto è l’ontologia sviluppata dal WC3 per descrivere le capacità dei dispositivi. Essa spiega a un computer che cosa un tipo particolare di telefono può fare e che cosa non può fare e quale tipo di informazione può essere scambiato da questi dispositivi. Esistono già ontologie estese per domini specifici; ad esempio, il settore biomedico ha sviluppato un’ontologia ampia e di alta qualità che descrive i termini medici e i nomi dei farmaci. Anche l’industria automobilistica è piuttosto avanzata: la Daimler-Chrysler è persino membro del gruppo di lavoro W3C. Naturalmente queste sono applicazioni che non sono visibili per l’utente ordinario del Web; l’applicazione principale a breve termine sarà nei mercati business-to-business.”

D: “E quando possiamo aspettarci di vedere le prime applicazioni per i navigatori e i consumatori ordinari?”

Van Harmelen: “Al momento, vedo molte "isole" di Semantic Web in corso di sviluppo all’interno di settori specifici. A lungo termine, mi aspetto che queste isole si uniscano e allora avremo un vero "Web" semantico. Sarà soltanto a questo punto che le cose si faranno interessanti per il consumatore medio. Qualcosa che vedo come imminente – e la Philips è già molto attiva in quest’area – è lo sviluppo di ontologie per contenuti multimediali. Un esempio potrebbero essere quei siti Web che presentano l’elenco dei programmi televisivi giorno per giorno: al momento attuale, tali siti sono leggibili soltanto da parte degli umani; ma con i metadati, che si basano su di un’ontologia sottostante, il mio computer o il mio PDA potrebbero leggere questo tipo di pagine, confrontare gli elenchi con il profilo dei miei interessi e avvertirmi dei programmi interessanti che verranno trasmessi. Allo stesso modo potrei immaginare ontologie per gli stili musicali o per i generi cinematografici. Tutto quello che avete da fare è indicare quali stili o generi vi piacciono e il compu-

ter può fare il resto. Mi aspetto che queste applicazioni possano apparire nel giro di pochi anni.”

D: “Per finire, una domanda difficile: quando si aspetta che ci sarà la grande svolta del Semantic Web?”

Van Harmelen: (sogghigna) “Questa è davvero una domanda difficile. Predire il futuro è sempre difficile, e lo è in modo particolare per il settore della IT [*tecnologia dell’informazione* – N.d.T.]. E predire gli eventi del World Wide Web è totalmente impossibile; Tim Berners-Lee – il padre del World Wide Web – usa la metafora della slitta da bob: all’inizio dovete spingere parecchio per farlo partire, ma una volta che è partito dovete saltarci dentro in fretta prima che parta senza di voi. Beh, in questo momento il Semantic Web si trova nella fase della spinta: dobbiamo parlare alle industrie per illustrare in modo convincente i suoi benefici; ma io non dubito che il Semantic Web ci sarà. Anche Tim Berners-Lee vede il Semantic Web come l’unico sbocco possibile per il World Wide Web. Ma lasciatemi specificare meglio: sarei molto deluso se nei prossimi due o tre anni non vedessimo nessuna applicazione per i comuni navigatori, in particolare per quanto riguarda l’area del commercio elettronico – questo settore ha molto da guadagnare dalla personalizzazione. La conversione di tutto il Web in un Semantic Web richiederà molto più tempo, ma io sono convinto che avverrà.”