

INTERNET

SARA' UN WEB INTELLIGENTE

Maggiordomi elettronici che fanno la spesa per noi, ricerche mirate, viaggi più facili e imprese virtuali. Questo è quello che ci offrirà il web di domani, una rete dal QI più elevato e dalle virtù "semantiche" che si materializzerà anche grazie al contributo italiano

di Raffaele Mastrodonato

Ottobre 2011. Paola, professoressa di scienze antropologiche, potrà dedicarsi molto di più alla ricerca. Per esempio, risparmiando tempo nell'ordinare i libri della biblioteca di istituto: non dovrà fare altro che inserire i titoli dei volumi su una schermata internet e un *web agent*, una sorta di maggiordomo elettronico intelligente, si occuperà di contattare al posto suo tutti i rivenditori online cercando le offerte migliori per poi procedere, autonomamente, all'acquisto.

Sempre in quell'anno Cinzia, grafica web e blogger a tempo perso, troverà in pochi secondi la foto giusta per il diario online dedicato a suo figlio di tre anni. Quando effettuerà la ricerca su internet potrà specificare tutti i parametri desiderati senza ricevere

in risposta, come avviene oggi, un diluvio di immagini disparate, ma solo quelle che corrispondono a questa precisa richiesta. Agosto 2011. Giovanni, oculato padre di famiglia, finalmente non avrà più problemi a individuare la soluzione più economica per le vacanze: chiederà semplicemente a un motore di ricerca di trovare tutte le offerte inferiori alla cifra che aveva in mente di spendere. Anche in questo caso sarà la tecnologia a individuare le soluzioni. Questi sono solo banali esempi di attività che la rete e i motori di ricerca attuali, per quanto potenti, non sono in grado di svolgere. Perché diventino realtà c'è infatti bisogno non di algoritmi più potenti ma di una internet con un quoziente intellettuale decisamente più elevato. Di un web in grado, per esempio, di capire il significato dei termini superando le barriere linguistiche. Di una rete in cui le macchine siano capaci di comunicare tra loro e svolgere da sole la maggior parte delle attività. In una parola, di un *web semantic*, un'idea in circolazione ormai da qualche tempo e che, dopo un periodo di appannamento, sembra avere finalmente ritrovato smalto e *appeal*. Anche grazie, bella sorpresa, a un robusto apporto della ricerca di base del nostro Paese.

Il futuro è (già) domani

Secondo gli esperti più ottimisti, un lustro o poco più potrebbe essere sufficiente per toccare con mano i primi esempi della rivoluzione semantica. «Se il progresso dell'interoperabilità a cui abbiamo assistito dal 2001 al 2005 continuerà allo stesso ritmo, credo che il 2011



potrebbe essere l'anno buono», racconta Maurizio Dècina, professore al Politecnico di Milano e guru delle tecnologie della comunicazione. Secondo Dècina – che si vanta di avere sempre sbagliato le proprie previsioni “per difetto” – fra cinque anni i nuovi protagonisti del web non saranno più gli uomini in carne e ossa ma creature artificiali. «Domani in rete non ci saranno solo 4 miliardi di persone che interrogano Pc, cellulari, palmari. Ma milioni di miliardi di macchine che parlano tra loro. L'uomo avrà ai suoi ordini migliaia di agenti elettronici personali». Ma per arrivare all'universo futuristico delineato da Dècina il lavoro da fare è ancora molto. Il vero nodo sta in un'operazione fondamentale e assai costosa in termini di tempo e risorse: la ridefinizione di tutta la

conoscenza del web in modo che l'intelligenza non umana possa comprenderla e interpretarla. «Qualcuno pensa che l'uso di motori semantici migliorerebbe la ricerca sulle pagine attuali. Ma non è così. Queste pagine bisogna ristrutturare aggiungendo

Sonia Bergamaschi,
coordinatrice del
progetto Network
peer for business

dei metadati scritti in un linguaggio che i computer possano leggere direttamente», spiega Marco Colombetti, professore di Ingegneria della conoscenza al Politecnico di Milano. Insomma, perché Paola, Cinzia e Giovanni possano avvertire i benefici del *semantic web*, c'è bisogno di un'autentica rivoluzione a partire dalle fondamenta: il cambiamento della rete deve precedere quello delle tecnologie che usiamo per orientarci al suo interno. Per usare una metafora “reale”, è un po' come dire che prima dobbiamo riscrivere tutti i libri per poi poter sviluppare lenti più sofisticate per leggerli e comprenderli meglio. «Solo così potremo ovviare alla mancanza di selettività dei motori attuali», spiega Colombetti.

La strada verso l'intelligenza

Dopo tutto, questo è quello che aveva in mente Tim Berners-Lee, già padre del web tradizionale, quando, nel 2001 in un articolo su *Scientific American*, delineava le caratteristiche dell'impresa semantica: intervenire sulle informazioni nascoste (*tag*) delle pagine web e renderle più esaurienti; associare a ogni “oggetto” una lista di significati e di relazioni; costruire quelle che, con un vocabolo preso a prestito dalla filosofia, sono comunemente definite “ontologie”. Per rendere possibile questa rivoluzione il W3C, il consorzio che presiede allo sviluppo del web, ha emesso in questi anni una serie di raccomandazioni sugli standard da utilizzare per l'annotazione. A cominciare dai *framework*

GLOSSARIO ESSENZIALE

GLI INGREDIENTI

XML - eXtensible Markup Language è un linguaggio per la conservazione di una struttura di dati e il loro scambio. L'Xml è adottato per la costruzione del web semantico, dal momento che permette di descrivere semanticamente (e con il dettaglio desiderato) le diverse parti di un documento. Tuttavia, la sua sintassi non definisce meccanismi espliciti per qualificare le relazioni tra documenti. Due documenti Xml possono descrivere in modo adeguato il termine “abitazione”, ma non è poi possibile stabilire se entrambi si riferiscono allo stesso oggetto.

Rdf - Il W3C ha definito anche il Resource

per la descrizione delle risorse web. La scelta è caduta su Rdf (Resource description framework), un'applicazione Xml che, sfruttando la logica dei predicati, permette di definire e mettere in relazione i dati sulla base di uno schema fisso: *soggetto* (per esempio: “Tom Cruise”), *predicato* (“è padre di”), *valore* (“Suri”). Per ciascuno di questi elementi si possono reperire in rete altrettante risorse da associare al dato attraverso un indirizzo web. Quando c'è bisogno di più raffinatezza, il W3C suggerisce un'estensione di Rdf chiamata Owl (web ontology language) che permette la costruzione di vocabolari e di gerarchie che offrono una maggiore

MOTORI DI RICERCA/1

PROGETTO DART: WEB SEMANTICO, MULTIMEDIA E P2P

Web semantico per tutti, in salsa multimediale. In Italia c'è chi ci crede e scommette sull'apporto degli utenti per farlo diventare realtà. È questa infatti l'utopia molto concreta di Dart, ovvero la sfida italiana ai colossi della ricerca web. Il nuovo search engine – che sta nascendo in Sardegna grazie alla collaborazione di Tiscali, CRS24, il centro di ricerca che vanta nel 1993 la creazione del primo sito web italiano, e l'Università di Cagliari – punta infatti su due caratteristiche innovative. La prima è un indice distribuito nei Pc di tutti gli utenti che vorranno donare un po' di Cpu e di spazio disco per migliorare la qualità

del recupero delle informazioni su internet. Il secondo è lo sforzo collaborativo degli utenti per l'annotazione semantica. Con la prima mossa, Dart spera di aggirare i costi di entrata nel mercato evitando l'acquisto di server farm e data center: se gli utenti mettono a disposizione i loro Pc come nei progetti di calcolo distribuito non c'è bisogno di investire in infrastrutture. Con il secondo tassello, si punta sull'intelligenza collettiva per rendere più intelligente la rete. Il tutto, come spiega a *Vision* Domenico Dato di Tiscali, ottimizzato specificamente per i contenuti multimediali. «Noi costruiremo un'ontologia

per i contenuti multimediali. L'approccio semantico ci consentirà di fornire contenuti video e audio personalizzati: potremo costruire palinsesti a misura di utente, impacchettarli e spedirli direttamente sul Personal video recorder del cliente». Dall'internet al set top box insomma, passando per il cuore di Dart, il *Personal media delivery framework* che, secondo il *business plan* del progetto, dovrebbe arrivare sul mercato sotto forma di prototipo entro il 2008. Un'impresa non da poco, in un terreno già popolato di giganti. «Siamo convinti che ci sia ancora spazio per nuovi entranti», conclude Dato.

DI BASE DEL WEB SEMANTICO

description framework, che standardizza la definizione di relazioni tra informazioni. Rdf è dunque uno schema per la descrizione della conoscenza nel web formato da: risorse, proprietà e valori. L'unità base per rappresentare un'informazione in Rdf è detta *statement*, un'informazione strutturata secondo lo schema: soggetto-predicato-oggetto. Il soggetto è una risorsa, il predicato è una proprietà, l'oggetto un valore. **Uri** - Ogni risorsa è identificata da un Universal resource identifier, un identificatore univoco di risorse. Qualunque cosa descritta da Rdf è infatti detta risorsa. Principalmente le risorse sono reperibili sul web, ma Rdf può

espressività attraverso la definizione di sofisticate relazioni tra le risorse (equivalenza, simmetria, eccetera). È solo questo complesso reticolo di informazioni che permetterà a un motore di ricerca di capire, per esempio, che "parrot" e "papagallo" denotano lo stesso animale. Quello che gli standard definiti dal W3C non possono fare è diminuire i costi di questo lavoro (*re-tagging*) dai benefici non immediatamente visibili e che moltiplicato per svariati miliardi di pagine web assume proporzioni titaniche. Un'impresa che, per giunta, richiede le competenze di ricercatori che provengano sia dal settore dell'intelligenza artificiale (si veda a questo proposito l'articolo a pagina 82), sia da quello della ricerca sulle basi di dati. Due approcci (visionario il primo, più pragmatico il secondo) che non sempre vanno d'accordo. È così che la risposta di Colombetti alla domanda cruciale (*chi si accolla i costi del re-tagging?*) lascia intravedere uno sviluppo del web semantico più simile alla macchia del leopardo che a quella d'olio: «Il semantic web si incomincerà a diffondere in nicchie e in contesti specializzati, là dove i costi di ridefinizione

Maurizio Dècina, professore al Politecnico di Milano



descrivere anche risorse che non si trovano in rete. Anche le proprietà – relazioni che legano tra loro risorse e valori – sono identificate da Uri. I valori, invece, possono essere considerati risorse o dati primitivi.

Owl - Web ontology language è un'estensione di Rdf che permette una maggiore espressività nella descrizione di oggetti e relazioni. Tra queste *l'equivalenza* (la possibilità di affermare che due o più Uri rappresentano lo stesso elemento) o *l'inversa* (che permette di specificare che se Romolo è fratello di Remo anche Remo, a sua volta, sarà fratello di Romolo).
[Fonte: Wikipedia]

sono contenuti e i ritorni immediatamente verificabili». Il nuovo Eldorado intelligente, almeno all'inizio, non sarà per tutti. Prepariamoci dunque a vederne i primi esempi in contesti più limitati. Come le intranet delle grandi multinazionali, dove l'esigenza di superare le barriere linguistiche tra varie sedi è un incentivo forte e, soprattutto, dove esiste una gerarchia che può porre fine alle dispute sui significati. Altro terreno fertile di sperimentazione sono poi le reti di imprese. In simili network mettersi d'accordo sui nomi delle cose e dei servizi può voler dire risparmi, efficienze e maggiore competitività. È con questa finalità, per esempio, che è nato Sewasie, prototipo di motore di ricerca semantico sviluppato dall'Università di Modena e Reggio Emilia e integrato con un sistema di business intelligence realizzato da Thinking Networks.

Sewasie permette a un'azienda di cercare informazioni sul fornitore di uno specifico prodotto e di effettuare una negoziazione sul prezzo.

L'azienda virtuale

Sulla stessa strada si colloca Network peer for business (NeP4B), progetto per un'infrastruttura di commercio elettronico che permetta alle pmi di scambiarsi dati e servizi indipendente-

MOTORI DI RICERCA/2

ASK.COM: VIVA LA SEMANTICA IMPLICITA!

C'è chi vede nel web semantico troppi rischi e difficoltà. Tra queste la (non) competenza dei webmaster nell'inserire i tag giusti nelle pagine e quello degli spammer. «Come possiamo fidarci delle descrizioni affidate agli utenti?», ha chiesto Peter Norvig di Google a Tim Berners-Lee nel luglio scorso durante una riunione della American Association of Artificial Intelligence (si veda anche l'intervista esclusiva a Norvig nella sezione LAB di questo numero). Le stesse critiche riecheggiano oggi nelle parole di Antonio Gulli, direttore prodotti di ricerca avanzata di Ask.com, il quarto *search engine* al mondo che ha da poco aperto a Pisa il suo centro R&D europeo. «Sul web ci sono 20 miliardi di pagine. Di queste neanche 5 milioni hanno informazioni strutturate. Non vedo come il resto dei dati possa essere convertito in organizzazione semantica». E allora? La soluzione è, secondo Gulli, in due parole: la semantica implicita. Ovvero, derivare indicazioni di significato dal comportamento degli utenti. ExpertRank, l'algoritmo su cui si regge Ask.com, è in grado di "leggere" all'interno delle ricerche e di effettuare automaticamente quella che in gergo si chiama *disambiguazione*. La risposta a una ricerca sul termine "Apache" suggerirà così anche una serie di comunità che definiscono gli ambiti denotati dalla parola: gli elicotteri da guerra, il popolare software open source per server, la tribù indiana. A questo punto, l'utente può scegliere all'interno del contesto di significato prescelto.

mente dalla collocazione geografica. Finanziato dal ministero dell'Università e della Ricerca con 2 milioni e 910 mila euro, il progetto conta sul contributo di alcuni centri di punta della ricerca italiana. NeP4B integrerà infatti Momis, tecnologia per l'estrazione automatica di ontologie realizzata dall'Università di Modena, il sistema di notazione semantica dei servizi web realizzato dal Cefriel di Milano, mentre il contributo dell'Isti-Cnr di Pisa permetterà di estendere la condivisione (e la ricerca) semantiche anche ai contenuti multimediali. Il tutto all'interno di una rete p2p regolata da sistemi che permetto-

