Università di Salerno

Dottorato di ricerca in
Ingegneria dell'informazione

Probabilistic Lightweight Ontology per l'estrazione e rappresentazione della Semantica

Abstract

Autore:
Fabio Clarizia

Tutor:
Prof. Massimo De Santo

Coordinatore:
Prof. Angelo Marcelli

Anno Accademico 2009-2010, IX ciclo - Nuova Serie

L'estrazione e la rappresentazione della semantica contenuta nel linguaggio sono tra i principali argomenti che da sempre animano le discussioni in psicologia cognitiva e intelligenza artificiale. Nell'ambito di queste comunità scientifiche il dibattito è riferito principalmente al problema di individuare il modo migliore per rivelare il significato che risiede in un qualsiasi atto comunicativo: *scrivere, leggere, parlare, ecc.* In alcuni rami applicativi dell'intelligenza artificiale così come nella Knowledge Engineering, i risultati di questi dibattiti sono stati utilizzati per introdurre nuovi linguaggi formali grazie ai quali è stato possibile sia rappresentare la semantica su un calcolatore che manipolarla per l'esecuzione di ragionamenti automatici. Questi linguaggi, tra cui XML (eXtensible Markup Language), RDF (Resource Description Framework), OWL (Ontology Web Language), sono stati la risposta tecnologica alla nascente visione del Web Semantico introdotta da Tim Berners-Lee, formalmente l'inventore del World Wide Web. Secondo questa nuova visione, il Web dovrebbe essere una rete altamente interconnessa di dati facilmente accessibile e comprensibile ad un qualsiasi calcolatore, un desktop o palmare, dove agenti software intelligenti sono in grado di risolvere richieste complesse dell'utente.

A tutt'oggi alcune importanti questioni legate alla nascita del Web Semantico non sono state risolte: questioni che si riferiscono principalmente al modo per rivelare e rappresentare la semantica stessa. Il problema fondamentale è che scoprire le intenzioni dell'autore, ad esempio contenute all'interno di un testo, può essere un processo molto complesso e soprattutto ambiguo per via del fatto che il *significato* stesso esiste indipendentemente dal linguaggio utilizzato e dal processo d'interpretazione. E' ben noto, infatti, che a causa delle imperfezioni del linguaggio umano i processi d'interpretazione ovvero codifica e decodifica del messaggio-significato potrebbero essere imperfetti e dunque corrotti da rumore. La rivelazione della semantica del testo, dunque, non può essere fatta semplicemente associando un significato al testo attraverso linguaggi formali ma piuttosto deve essere realizzata utilizzando metodi in grado di considerare il fattore rumore intrinseco ai processi stessi e, di conseguenza nasce così l'esigenza di utilizzare ulteriori linguaggi - per forza di cose probabilistici - in grado di gestire al meglio tali processi rumorosi.

La Knowledge Engineering, dal canto suo, dovrebbe poi fornire semplicemente gli strumenti per facilitare l'interazione dell'utente con i

dati, visto che i processi in gioco possono essere considerati situati e temporali in cui il testo stesso agisce come condizione a contorno e nel quale l'utente è, ex necessitate, il protagonista. Un processo d'interpretazione del significato dovrebbe così portare in conto i diversi livelli di rappresentazione della semantica, che partono dal testo - livello più basso - fino all'utente - livello più alto - usando un formalismo tale da consentire il trattamento dell'incertezza. Progettare modelli per la semantica in grado di non tralasciare tutti questi aspetti sembrerebbe l'unico modo per assegnare il significato ad un testo: significato che può essere rappresentato attraverso un linguaggio formale oltre che dal linguaggio naturale e che si presta facilmente ad essere manipolato da un agente artificiale anche e, come nel nostro caso, attraverso l'utilizzo di metodi automatici basati su ontologie.

In questo lavoro di tesi è stato così introdotto un sistema a livelli, come accennato in precedenza, per la manipolazione del significato secondo il quale l'estrazione della semantica può avvenire attraverso l'analisi delle relazioni tra i diversi tipi di elementi tra cui *parole*, *concetti* e *percetti*. Lo sforzo impiegato nella formalizzazione di questi aspetti ha dato luogo a due differenti tendenze. Una prima che si concentra principalmente sulla struttura delle relazioni associative tra parole del linguaggio naturale e sulle relazioni tra concetti e parole, che si definisce come la parte più superficiale della semantica (che possiamo estrarre direttamente dai testi), o *light semantics*. Una seconda che esalta strutture concettuali astratte, concentrandosi su relazioni tra concetti, relazioni tra concetti e percetti, relazioni tra percetti e azioni, e che si definisce come la componente più profonda della semantica (che si può estrarre ad esempio studiando l'utente ed i suoi comportamenti), o *deep semantics*. Una volta introdotto il modello generale, si è posta attenzione solo ad alcuni degli aspetti discussi precedentemente (si veda la figura 1) cosicché il cuore della nostra proposta è quindi la definizione di un tipo di conoscenza informale, indicata da noi come *informal Lightweight Ontology* (*iLO*)e che può essere desunta automaticamente da documenti.

Tale rappresentazione della conoscenza può essere ancora vista come un *Grafo di Termini* o *Grafo di Concetti*, composto cioè da nodi (i concetti stessi) e da collegamenti pesati tra essi (in grado di conservare e rappresentare le relazioni semantiche tra concetti) e dove ogni nodo-concetto può essere specificato attraverso un ulteriore grafo (in questo

Figura 1 Rappresentazione a livelli della semantica: il modello computazionale proposto in questo lavoro si riferisce alle relazioni incluse nella cornice rossa.

caso si parlerà di parole come nodi e legami/archi, ancora una volta pesati, tra parole). A questo punto, se per entrambi i grafi, i pesi dei diversi legami (tra concetti e tra parole) sono intesi attraverso una probabilità, è dunque possibile compiere inferenza e quindi apprendere una rappresentazione di un concetto e/o di un grafo di concetti attraverso tecniche probabilistiche[1].

Nell'ambito di questo lavoro è stato poi realizzato un ambiente sperimentale in grado di testare l'approccio proposto e verificarne la sua effettiva bontà quando impiegato in casi reali per l'interpretazione dell'intenzione utente. Specificatamente abbiamo realizzato così due scenari di interesse:

I) il primo sfrutta le potenzialità di categorizzazione concettuale delle *iLO* su grandi collezioni di dati testuali, per esempio repository di pagine web;

II) il secondo più focalizzato in ambito di *User Satisfaction* sfrutta la stessa tecnica per recuperare da un repository di pagine web, contenuti più vicini alle intenzioni degli utenti quando questi effettuano queries di tipo informazionali.

Nel primo ambiente di testing, proprio per sollevarci dalla soggettività dei dati sperimentali, abbiamo previsto il confronto con un motore di ricerca text-based puramente sintattico molto diffuso in ambiente open source, *Lucene*, attraverso l'uso di indici di prestazione specifici

[1] All'uopo si è pensato di utilizzare una versione smoothed della Latent Dirichlet Allocation conosciuta in letteratura anche come Topic Model così come sarà ben descritto nel seguito di questo tesi.

(*Precision* e *Recall* tra gli altri) dei motori di ricerca su web. Allo stesso modo e per avere una ulteriore conferma sulla bontà del nostro sistema, all'interno della seconda fase di sperimentazione sono stati così condotti esperimenti in diversi contesti e, per ogni uno di essi, è stato richiesto ad alcuni gruppi di esseri umani di assegnare dei giudizi di rilevanza per il set di pagine web restituite sia da un motore di ricerca classico (nella fattispecie una versione personalizzata di Google - *Google Custom Engine*) e dal nostro motore di ricerca implementato con l'ausilio di *Lightweight Ontology*.

In entrambi i casi, i risultati ottenuti hanno confermato che la tecnica proposta aumenta sicuramente le prestazioni in termini di rilevanza e rispondenza alle vere intenzioni dell'utente e, poiché un'ontologia, almeno per come l'abbiamo intesa noi in questo lavoro, consiste di concetti e di collegamenti tra essi, una maggiore specializzazione di intenti, ovvero una rappresentazione coerente del significato, risulta molto utile per ridurre i problemi legati l'ambiguità intrinseca del linguaggio.