Data Semantics

Sommario

Data Semantcs si occupa di comprendere il significato dei dati, nella pratica della scrittura del programma. È necessario prestare attenzione al significato dei dati nell'integrazione di più dataset; inoltre la semantica del dato è necessaria per la condivisione di dataset (cioè renderli fruibili da chi non ha prodotto il dataset). Altro problema centrale è la capacità di usare dati non strutturati, in modo tale da facilitare query.

Scopo del corso è strutturare dei modelli per la semantica dei dati in modo tale da facilitarne l'uso; inoltre si stabiliscono strategie per attribuire semantica ai dati. Bisogna inoltre capire il ruolo della semantica nell'integrazione dei dati. Il corso tratta della semantica dei dati nei *big data*, dell'estrazione dello *knowledge graphs* (ovvero le relazioni tra gli elementi di un database, *data linkage*) o la costruzione di sistemi di raccomandazione. Inoltre saranno analizzate alcune tecniche di *natural language processing* e la costruzione di rappresentazioni a partire dai dati.

Le esercitazioni si occuperanno di interrogare *knowledge graphs*, modellare e costruire grafi di conoscenza e integrare fonti di dati.

L'esame orale sarà accompagnato da un progetto software (effettuato in gruppo di, circa, 3 persone), di cui sarà fatta una presentazione orale; in alternativa al progetto è possibile scrivere un articolo di approfondimento su una tematica. La preparazione sarà "ragionevolmente" dettagliata su tutti gli argomenti, e sarà approfondito l'argomento del progetto.

Parte I
Grafi di conoscenza

La costruzione di grafi di conoscenza è spesso effettauta a mano da una moltitudine di utenti. Il modello *Semantic Web* ha costruito linguaggi e strumenti, approvati dal W3C, per definire, interrogare e fare inferenza su grafi di conoscenza. Nel mondo reale tuttavia non sono usati questi linguaggi.

Internet produce enormi quantità di dati diversi tra di loro, usati spesso per altri fini: la semantica dei dati si occupa di integrare grandi quantità (*data volume*) da diverse fonti di dati (*data variety*). Questo permette la costruzione di intelligenze artificiali, ovvero di programmi che eseguono task tipicamente umani con risultati simili.

I dati possono essere *strutturati* (tabelle ordinate), *semi-strutturati* (tabelle annidate) o *non strutturati* (testi).

È impossibile effettaure a mano certi compiti particolarmente ardui, come l'integrazione di serie temporali con altri documenti riguardanti lo stesso tema, soprattutto con una scarsa conoscenza del dominio.