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

Il termine è coniato da Google per indicare uno strumento usato dal suo motore di ricerca. Il grafo permette di essere processato facilmente da una macchina e allo stesso tempo permette un livello di astrazione soddisfacente; si possono anche effettuare query come in un database a grafo. Il modello a grafo permette inoltre una facile integrazione di sorgenti diverse. La costruzione di grafi di conoscenza è spesso effettuata 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.

Esistono grafi di conoscenza aperti, quali DBpedia, Yago o Wikidata; anche alcune aziende private sviluppano il proprio per facilitare la propria attività imprenditoriale.

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

È impossibile effettuare 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.

1 *Linking Data.*

I dati sono spesso collegati in modo automatico grazie a programmi di apprendimento automatico. Una ricerca su internet non è fatta per documenti ma per contenuti: un motore di ricerca in passato offriva una serie di documenti senza offrire informazioni aggiuntive; oggi invece un motore di ricerca tenta direttamente di rispondere con *factual information* rilevanti nella ricerca. Per *fatto* si intende un'informazione interpretabile come vera o falsa (al contrario, alcuni dati quali immagini o suoni non sono interpretabili per veridicità). I fatti costituiscono un elemento centrale per l'analisi.

Le risposte fattuali sono personalizzate in base alla natura della ricerca, secondo criteri *data driven* (ovvero statistico). Informazioni di tipo diverso hanno caratteristiche diverse: si produce un grafo di conoscenza per gestire meglio entità di tipo diverso con caratteristiche peculiari diverse. Le *preview* dei contenuti sono generate chiedendo agli sviluppatori di inserire dei contenuti nel codice HTML che sono poi interpretati dal motore di ricerca.

I chatbot sono costruiti con l'ausilio di reti neurali e rappresentazioni del mondo reale.

Internet può essere interrogato

2 Spazio vettoriale.

Le query e i documenti sono interpretabili come vettori; per calcolare la similarità tra due vettori, si usa la distanza coseno.

$$\begin{aligned} sim(r,u) &= cos(\theta) \\ &= \frac{uq}{|u||q|} \\ &= \end{aligned}$$

La rappresentazione vettoriale è alla base dell'analisi testuale. Si usa un sistema di pesi più sofisticato per valutare diversamente l'importanza di una parola all'interno di un documento. Una parola è tanto più importante nel documento quante più ricorrenze ci sono nel documento (in percentuale).

$$\begin{aligned} tf_{i,j} &= \frac{n_{i,j}}{|d_j|} \\ idf_i &= \log \frac{|D|}{|\{d : i \in d\}|} \\ tfidf_{i,j} &= tf_{i,j} \times idf_i \end{aligned}$$

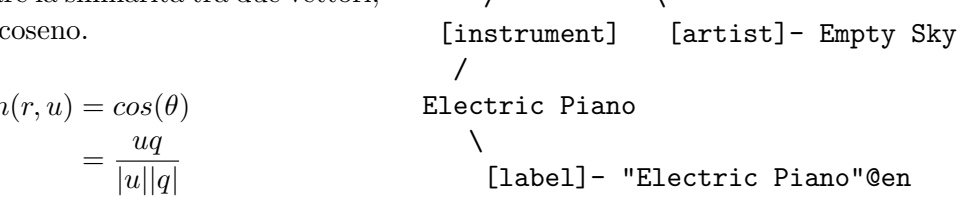
Le entità, le loro proprietà e i collegamenti tra di loro sono iscritti nel grafo di conoscenza. Esistono linguaggi formali, definiti a inizio '900 che permettono di dichiarare relazioni tra entità, ma che non sono leggibili da una macchina. Inoltre si possono usare assiomi logici per definire le ricorrenze nel testo.

3 Standard.

RDF (*Resource Des...*) è lo standard per il web semantico, ovvero per l'internet elaborabile dalle macchine. L'unità base per rappresentare l'informazione è rappresentata da triple (affermazioni), ovvero grafi etichettati, identificati da URI (*unique resource identifier*). Esempi di triple sono:

```
<Electric Piano, label,
                    "Electric Piano"@en>
<Elton John, instrument, Electric Piano>
<Sails, artist, Elton John>
```

Le triple sono rappresentabili come un grafo diretto, aciclico ed etichettato:



Alle triple si applicano degli identificativi globali (una sorta di chiave primaria SQL): si usano generalmente identificativi web (abbreviati), che specificano come recuperare l'informazione (qualsiasi cosa a cui si può attribuire un valore costante è definibile come URI). Tutti gli URI sono risorse, e rappresentano l'ontologia del dominio.

Le triple sono strutturate con un soggetto, un predicato e un oggetto. Il soggetto è composto da un URI o da un *black node* (ovvero costanti), il predicato da un URI, un *black node* o da un letterale, a cui può essere attribuito un tipo (stringhe, stringhe, date o booleani) per permettere operazioni. I *black node* sono nodi anonimi per consentire una buona costruzione del grafo.

I *linked data* riassumono delle pratiche per pubblicare e unire dati provenienti dal web:

- usare URI come nomi delle cose;
- usare indirizzi HTTP come URI;
- usare informazioni utili su un URI cercato;
- includere link ad altri URI.