Raccolta accurata di fatti da testo in linguaggio naturale di Wikipedia

M. Faretra, G. Marini, A. Martinelli

17 febbraio 2017

RIASSUNTO

Molti approcci sono stati utilizzati per estrarre informazione da Wikipedia sotto forma di fatti (entità, relazione, entità) per il popolamento di Knowledge Graphs, in particolare sfruttando le informazione contenute nelle sue infoboxes. Tuttavia queste strutture dati riportano solo una piccola parte delle informazioni contenute negli articoli. Infatti nel testo libero si concentra la maggior parte delle relazioni estraibili, tuttavia la rilevazione di essi risulta più problematica trovandosi all'interno di testo libero in linguaggio naturale. In questo lavoro si cerca di quantificare il numero di relazioni estratte da questi articoli con il supporto di un KG già popolato per aumentarne la conoscenza. In particolare, la nostra valutazione è stata effettuata utilizzando DBPedia, un KG gratuito, che ci ha portato a... (to be implemented, forse spiegare come è rappresentata la conoscenza con le triple)

1 INTRODUZIONE

L'incremento dei Knowledge Graph in questi ultimi tempi è stata di particolare interesse scientifico e hanno evidenziato la limitatezza e la mancanza di informazioni in essi presente, dovuta in alcuni casi a entità inserite automaticamente senza particolari relazioni, in altri alla limitata conoscenza reperibile dagli infoboxes.

Ad esempio il progetto DBPedia estrae informazioni da più di 125 edizioni in linguaggi differenti di Wikipedia. La più grande base di conoscenza è estratta dalla versione inglese e consiste in più di 580 milioni di fatti che descrivono 38 milioni di cose. Il progetto DBpedia mappa le infoboxes di Wikipedia da 28 edizioni in linguaggi differenti in una singola ontologia condivisa consistente di circa 685 classi e 2795 proprietà.

Considerando questi numeri e la loro fonte (infoboxes), la conoscenza estraibile dal testo naturale potrebbe essere decisamente più consistente e aumentare di molto il Knowledge Graph. L'approccio utilizzato va a scalare sfruttando i fatti contenuti nel KG stesso, è quindi dipendente anche da essi, ed in particolare dalla loro qualità: le triple sono etichettate come fidate e non fidate e con esse abbiamo effettuato misurazioni differenti, considerando le relazioni candidate estratte dalle triple appartenenti all'una o all'altra categoria.

L'utilizzo di Wikipedia è ampiamente diffuso tra i vari KG implementati data la grande affidabilità che ormai garantisce, tuttavia le infoboxes fino a pochi anni fa erano ancora molto poco diffuse e solo nell'ultimo decennio esse si trovano in più della metà degli articoli.

Se invece si va a considerare la conoscenza presente nel testo libero, si può immaginare che da esso (ovviamente presente in ogni articolo) si possa estrarre informazione non presente nel KG poichè esso riporta una serie di relazioni che probabilmente non sono presenti nell'infobox, ad esempio riguardanti entità che non sono il soggetto dell'articolo.

Il nostro approccio al problema considera pattern del tipo [entità] frase [entità], ad esempio: "Antonio è poco propenso alla Scrittura" mette in relazione le due entità "Antonio" e "Scrittura" utilizzando la frase "è poco propenso alla" che descrive un'istanza della relazione. In questo articolo descriviamo l'approccio di estrazione di conoscenza da testo libero di Wikipedia, e i risultati in termini di aumento dei fatti presenti in DBPedia, che usiamo anche come garante dell'effettiva utilità della frase per esprimere una determinata relazione.

(sommaria spiegazione del processo e qualche numero di risultati) Il resto dell'articolo è organizzato in vari capitoli: nel Capitolo 2 vengono analizzate in maggior dettaglio le risorse utilizzate; nel Capitolo 3 si presenta il nostro approccio; nel Capitolo 4 vengono mostrati i nostri risultati; nel Capitolo 5 si trovano cenni a lavori correlati e su cui ci siamo basati (se vogliamo scriverci qualche cagata su Lector da cui siamo partiti); infine, nel Capitolo 6 vengono presentate le conclusioni sul lavoro effettuato.

2 RISORSE