Università degli Studi di Roma "Tor Vergata"

Metodi di costruzione dei Knowledge Graph

Manuel Fiorelli

fiorelli@info.uniroma2.it

Metodi di costruzione

- KG curati (*Cyc* e *OpenCyc*)
- Crowd-sourcing (*FreeBase* e *WikiData*)
- Estrazione da basi di conoscenza semi-strutturate a larga scala (*DBpedia*,*YAGO*)
- Metodi di *Information Extraction*
  - *Information Extraction* con un'ontologia/schema prestabiliti (*NELL*, *Google Knowledge Vault*) dall'intero Web
  - *Open Information* (schema-less) *Extraction* (*Reverb*, *OLLIE*) dall'intero Web
  - Approcci ibridi (non discussi ulteriormente)

KG curati (1/2)

Cyc è un esempio di KG curato (da *CyCorp*) ed anche uno dei più vecchi KG (risalente agli anni '80)

> Una base di conoscenza di senso comune (common sense knowledge) molto grande: fatti, regole empiriche (*rules-of-thumb*) ed euristiche per ragionare degli eventi e degli oggetti della vita di tutti i giorni
>
> La sua costruzione è partita inserendo tutta la conoscenza – esplicita ed implicita – contenuta in un centinaio di articoli dell'Encyclopaedia Britannica scelti in modo casuale.

Cyc è accessibile su licenza di ricerca o commerciale.

OpenCyc era un frammento pubblico di *Cyc* rilasciato inizialmente nel 2001, e successivamente ritirato all'inizio del 2017.

KG curati (2/2)

- Più di 900 anni uomo sono stati invesisti nella creazione di *Cyc* (Sarjant et al, 2009), eppure ci sono ancora dei buchi (Paulheim, 2016)
- Chiaramente la costruzione di un KG di uso generale è uno sforzo che supera le possibilità di qualsiasi individuo o organizzazione

Crowd-sourcing

Una possibile soluzione al problema della scalabilità del processo di costruzione dei KG curati consiste nell'affidarsi ad una comunità di utenti, tramite crowd-sourcing.

Due esempi notevoli sono:

- *Freebase* (interrotto il 31 Marzo del 2015 dopo essere stato acquisito da Google; tuttavia il suo contenuto è stato migrato all'interno di Wikidata)
- *Wikidata*

Estrazione da basi di conoscenza semi-strutturate a larga scala

- Molti siti web contengono già informazione semi-strutturata (es. tabelle, liste, metadati, etc.):
- Tuttavia, essa è spesso sepolta all'interno di pagine destinate per lo più agli essere umani, e non può essere acceduta ed interrogata direttamente dalle macchine
- DBpedia e YAGO sono due esempi di KG costruiti estraendo informazione sfruttata da Wikipedia:
  - DBpedia estrae dalle diversi versioni di Wikipedia (nelle diverse lingue) tanti KG interlinkati
  - YAGO estrae un solo KG, fondendo i contribuiti delle diverse edizioni

Estrazione da basi di conoscenza semi-strutturate a larga scala

- Molti siti web contengono già informazione semi-strutturata (es. tabelle, liste, metadati, etc.):
- Tuttavia, essa è spesso sepolta all'interno di pagine destinate per lo più agli essere umani, e non può essere acceduta ed interrogata direttamente dalle macchine
- DBpedia e YAGO sono due esempi di KG costruiti estraendo informazione sfruttata da Wikipedia:
  - DBpedia estrae dalle diversi versioni di Wikipedia (nelle diverse lingue) tanti KG interconnessi
  - YAGO estrae un solo KG, fondendo i contribuiti delle diverse edizioni

Metodi di Information Extraction (1/2)

- DBpedia e YAGO sono costruiti *estraendo informazione semi-strutturata* dalle singole pagine di Wikipedia, trasformandola e consolidandola in una KB.
- Un'alternativa è tentare di estrarre informazione strutturata dal contenuto testuale e, in generale, non strutturato delle pagine Web
- Inoltre, questa estrazione può essere operata su una scala molta vasta… persino dell'intero Web

Metodi di Information Extraction (2/2)

In particolare, distinguiamo due grandi categorie:

- *Information Extraction* con un'ontologia/schema prestabiliti (*NELL*, *Google Knowledge Vault*) dall'intero Web
- *Open Information* (schema-less) *Extraction* (*Reverb*, *OLLIE*) dall'intero Web

Riferimenti

- Paulheim, H. (2017). Knowledge graph refinement: A survey of approaches and evaluation methods. *Semantic web*, *8*(3), 489-508.
- (1997) An Introduction to the CYC Knowledge Base. Available at: https://www.cise.ufl.edu/~ddd/cap6635/Fall-97/Short-papers/36.htm