Global Relevance Scores for DBpedia Facts

Knowledge bases with facts about real-world entities (e.g., people, locations, products, dates, etc.) have inspired much of the recent research on semantic technologies. The hope is that knowledge bases may constitute the basis for machine-to-machine interoperability and improve semantic search [1, 2, 3, 4]. However, so far, there has been no ground-breaking key application showing the indispensability of knowledge bases.

The hypothesis underlying this proposal is that in order to boost the usefulness of knowledge bases, there is a need for a general notion of relevance for facts. For example, when asking for the classes to which Albert Einstein belongs, many well-known knowledge bases return correct, but rather impractical facts, e.g., that he is a person, an academic, a humanitarian, etc. (see Fig. 1).

	Id	Subject	Property	Object
1	#2126808	Albert Einstein	type	19th-century German people
2	#2125460	Albert Einstein	type	Academics of the Charles University
3	#2123400	Albert Einstein	type	American humanitarians
4	#2125900	Albert Einstein	type	American pacifists
5	#2125652	Albert Einstein	type	American people of Swiss descent
6	#2124020	Albert Einstein	type	American philosophers
7	#2124060	Albert Einstein	type	American physicists
8	#2123828	Albert Einstein	type	American scientists of German descent
9	#2125384	Albert Einstein	type	American socialists
10	#2125608	Albert Einstein	type	American vegetarians

Fig. 1: YAGO answers to the query asking for the classes to which Albert Einstein belongs

For the above query, it would be more practical to rank facts stating that Albert Einstein was a theoretical physicist, philosopher, etc., first in the result list. Similarly, when asking about people who are physicist we would expect to have Albert Einstein, Newton, Niels Bohr, etc., ranking higher in the result list than generally less known physicists (e.g., see Fig. 2). Previous work [6, 7] has coined this notion of general relevance for facts as *informativeness* and has suggested its computation by means of web-based co-occurrence statistics for the entity pairs in the facts. For example: *Albert Einstein* and *physicist* co-occur more frequently on the web than *Albert Einstein* and *academic.* Consequently, *physicist* is a more relevant class for Albert Einstein than *academic.*

There exists no work that computes the above co-occurrence statistics at web-scale. Hence the goal of this project is twofold: (1) derive global relevance scores for all DBpedia facts by computing co-occurrence statistics for entity pairs from a web-scale corpus and (2) provide a

ranking mechanism that combines these scores to produce top-*k* results for an important fraction of SPARQL [8].

We plan to provide a web-based query interface (for the DBpedia corpus with global relevance scores) and to publish the results in a top-tier international conference.

	Id	Subject	Property	Object
1	#624248315	Aage Bohr	type	physicist
	#931072	physicist	means	physicist
2	#625180219	Aaldert Wapstra	type	physicist
	#931072	physicist	means	physicist
3	#586255639	Aarne Arvonen	type	physicist
	#931072	physicist	means	physicist
4	#625134271	Aaron Klug	type	physicist
	#931072	physicist	means	physicist
5	#649086659	Aaron Lemonick	type	physicist
	#931072	physicist	means	physicist
6	#651400215	Abd al-Rahman al-Sufi	type	physicist
	#931072	physicist	means	physicist
7	#570749595	Abd as-Salam al-Alami	type	physicist
	#931072	physicist	means	physicist

Fig. 2: YAGO answers to the query that asks about physicists

[1] http://linkeddata.org/

[2] http://de.dbpedia.org/

[3] http://www.wolframalpha.com/

[4] http://www.trueknowledge.com/

[5] https://d5gate.ag5.mpi-sb.mpg.de/webyagospotlx/WebInterface

[6] G. Kasneci, F. M. Suchanek, G. Ifrim, M. Ramanath, G. Weikum: *NAGA: Searching and Ranking Knowledge*, ICDE 2008.

[7] Gjergji Kasneci, Shady Elbassuoni, Gerhard Weikum: *MING: mining informative entity relationship subgraphs*, CIKM 2009

[8] http://www.w3.org/TR/rdf-sparql-query/