Karlsruher Institut für Technologie
Institut für Angewandte Informatik und Formale Beschreibungsverfahren (AIFB)
Dr.-Ing. Michael Färber
michael.faerber@kit.edu

15.07.2019

Bachelor-/Masterarbeit "Entwicklung einer Suchmaschine für Datensätze" zu vergeben!

Um was geht es?

Die Anzahl an Datensätzen ist in den letzten Jahren dramatisch gestiegen. WissenschaftlicherInnen und EntwicklerInnen stehen daher häufig vor der Herausvorderung, passende Datensätze im Web zu finden. Das Ziel der Abschluss ist es dhaer, Metadaten über Datensätze zu crawlen, semantisch zu repräsentieren und durchsuchbar zu machen. Im Unterschied zum Google Dataset Search [1,2] sollen die Daten öffentlich verfügbar sein, semantisch modellieren sein (sofern möglich mit Links zu anderen Dingen im Web) und letztendlich mittels keywords o.ä. durchsuchbar sein.

Das Vorgehen:

1. Wir extrahieren die Metadaten über Datensätze aus Webseiten. Als Korpus verwenden wir Common Crawl oder einen vergleichbaren Korpus (siehe http://commoncrawl.org/the-data/get-started/). Die Metadaten sind im HTML und anderen verlinkten Dateien der Webseiten beschrieben. Konkret wäre die Idee, – analog zum Paper "Google Dataset Search" – http://schema.org/Dataset, http://schema.org/DataCatalog und http://www.w3.org/ns/dcat#Dataset als zu erkennende "tags" in HTML und im markup von RDFa, microdata und JSON-LD zu verwenden.
2. Die extrahierten Fakten modellieren wir dann in RDF und erstellen damit einen RDF knowledge graph. Dieser ist über das Web erreichbar und jeder kann mittels keywords nach Datensätzen suchen.

Neben diesen Kernelementen kann das entwickelte System erweitert werden. Beispielsweise kann ein User-Interface gebaut werden, um die Datensätze zu durchsuchen. Zudem können Daten in dem erstellten Wissensgraphen zu anderen Wissensgraphen in der Linked Open Data cloud mittels graph-basierter und String-basierter Algorithmen verlinkt werden (um z.B. Datensätze mit Publikationen zu verbinden).

Die Herausforderung besteht in der Extraktion von Fakten aus Milliarden von Webseiten. Es kann geeignete Infrastruktur am SCC oder bei AWS verwendet werden.

[1] https://toolbox.google.com/datasetsearch

[2] https://storage.googleapis.com/pub-tools-public-publication-data/pdf/77547c8d2a7fba472e76c774028cf2b3c0afdb8a.pdf

Welche Voraussetzungen sind notwendig?

- Erfahrung im Umgang mit großen Datenmengen sehr vorteilhaft
- Erfahrung im Text Mining/Natural Language Processing von Vorteil (z.B. MapReduce), aber nicht notwendig
- Mindestens mittlere Kenntnisse in Python oder einer ähnlichen Programmier-/Skriptsprache.

Im Falle guter Resultate können die Ergebnisse der Arbeit gerne in Zusammenarbeit mit den Betreuern wissenschaftlich veröffentlicht werden.

Kontakt:
Dr. Michael Färber | Anna Nguyen
michael.faerber@kit.edu | anna.nguyen@kit.edu