Question Answering on Statistical Linked Data

AKSW Colloquium paper presentation

Konrad Höffner

Universität Leipzig, AKSW/MOLE, PhD Student

2015-2-16

1. Introduction

2. Corpus & Benchmark

3. Algorithm

Motivation

Statistical Linked Data

- Increasing amounts available
- Highly relevant for decision making

Problems

- Multidimension data opaque to the end user
- Reliance on predefined visualizations: problems of bias, coverage, adequacy

Motivation

Question Answering

- Intuitive and expressive way of accessing Linked Data
- Generic algorithms cannot process statistical Linked Data
- Specific algorithms for statistical Linked Data do not exist

Contributions

- Corpus of natural language questions with statistical information needs
- Benchmark based on the corpus
- First QA algorithm for statistical RDF data
- First results and discussion of challenges to open up statistical QA as new research field

1. Introduction
2. Corpus & Benchmark
3. Algorithm

Corpus

- High precision requires analysis of typical user questions
- Open survey where participants were asked to provide question with statistical information needs
- 50 questions to no particular existing dataset

Corpus

Excerpt of Questions

How much money, does Leipzig and Dresden spend on child care in relation to the birth rate in comparison to the average in Saxony.
What is the average monthly income of a German citizen?
How much money was invested to fight bicycle thefts in Leipzig?
How many citizens live in a certain area?
How much does Germany spend on research a year?

Call for Participation

- Everyone (who hasn't yet), please fill out `http://tinyurl.com/statisticalqa`
- More entries → more effective benchmark, more accurate evaluation, better basis for algorithm development

Corpus Properties

restriction	dimension value	29
	dimension value range	5
	measure value range	2
	top k measure	5
	top k dimension	1
expected answer type	measure value	14
	measure value aggregate	10
	dimension count	2
	dimension value	7
referenced	measure name	30
	measure unit	2
	dimension name	3

Benchmark

- Subset of the corpus that has an identifiable correct interpretation
- Rewrite to one specific dataset: LinkedSpending Finland foreign aid[1]

[1] http://linkedspending.aksw.org

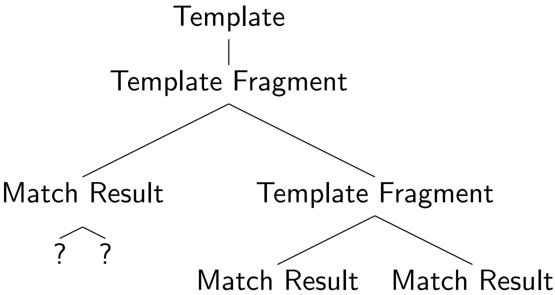
1. Introduction
2. Corpus & Benchmark
3. Algorithm

Observations and Assumptions

- Questions are structurally complex but semantically simple
- All answerable questions ask for a subset of a data cube (optionally + an aggregate)
- Query model as conjunction, empty question selects everything, phrases are restrictions on dimension values → doesn't model all questions but many and leads to efficient implementation

TCQA—Tree Based Question Answering

- Recursive visit of parse tree
- Stanford statistical english parser resulting in phrase structure
- adaptable to other languages
- Top-down matching, bottom-up combining



Template

Template Fragment

Match Result

Template Fragment

? ?

Match Result

Match Result

Match Result

- $m = (N, V)$
- N—scored component property (name) references
  $R \subseteq P \times [0, 1]$
- V—scored comp. property value references
  $V \subseteq P \times L \cup U \times [0, 1]$
- P—component properties
- L—literals
- U—uris

Combining Match Results to Template Fragment

- Fragment $c(m_1, m_2) = (N, V, R)$
- $R \subseteq P \times V \times V$ (restriction is property with value range)
- N and V are unions of N and V from m_1 and m_2 minus property references and values in the restriction
- R at most one element: combination of property reference and fitting property value between both match results with highest score product
- Combining template fragments works in the same way but existing restrictions are integrated as well

Converting Template Fragment to Template, Execution

- Leftover property value references for unmatched properties over a threshold are converted to restrictions
- All other references are discarded, the set of restrictions ($+$ aggregate) is the template
- Restrictions in template are transformed to a SPARQL query
- Query execution → answer

Evaluation

Criterion	$\varnothing p$	$\varnothing r$	$\#p = 1$	$\#r = 1$
all component properties	68.12%	38.04%	10	0
dimensions	70.37%	64.81%	8	9
attributes	0%	0%	0	0
measures	72.72%	63.63%	8	6
restrictions	46.15%	30.77%	6	2