Answering Natural Language Queries over Linked Data Graphs: A Distributional Semantics Approach

André Freitas, Fabrício F. de Faria, Seán O'Riain, Edward Curry
Digital Enterprise Research Institute (DERI)
National University of Ireland, Galway
IDA Business Park, Lower Dangan, Galway
firstname.lastname@deri.org

ABSTRACT

This paper demonstrates *Treo*, a natural language query mechanism for Linked Data graphs. The approach uses a distributional semantic vector space model to semantically match user query terms with data, supporting *vocabulary-independent (or schema-agnostic) queries* over structured data.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval-Search Process

Keywords

Semantic Search, Knowledge Graphs, Schema-Agnostic Queries, Distributional-Compositional Semantics, Natural Language Queries

1. MOTIVATION

Structured data is becoming a central element in the evolution of the Web search experience. Applications such as Google Knowledge Graph[1] are using structured data to provide direct answers to users' informational queries. The use of knowledge graphs in search engines is synchronized with the growth of structured and open data available on the Web, in particular, Linked Open Data.

Within the realm of the Web and Big Data, databases following the Entity-Attribute-Value (EAV) are becoming progressively more popular, where data is more sparse and the schema is more complex and heterogeneous. Consuming this data demands search/query mechanisms with the semantic flexibility necessary to cope with the semantic gap that exists between user queries and the representation of the data. Traditional structured query mechanisms for databases allow expressive queries at the expense of usability: the semantic matching process is delegated to data consumers (in the construction of SQL or SPARQL queries). On the other side of the usability spectrum, information retrieval (IR) approaches allow users to search using intuitive keyword-based interfaces. High usability comes at the expense of query expressivity and effectiveness. Traditional IR approaches do not provide expressive queries over structured data. At the core of this usability-expressivity trade-off is the *semantic gap* between the way users express their information needs and the way structured data is represented. By addressing this semantic gap, *vocabulary-independent/schema-agnostic queries* provide users with greater freedom and efficiency for querying and searching large heterogeneous data sources.

2. TREO: VOCABULARY INDEPENDENT QUERIES OVER LINKED DATA GRAPHS

This paper presents the demonstration for *Treo*, a vocabulary independent query mechanism for Linked Data graphs and EAV databases. Treo allows users to search over EAV Databases using expressive natural language queries. To enable semantic flexibility and vocabulary independency, a principled distributional-compositional semantic model [3] is used to build a distributional structured vector space model (T-Space). The distributional semantics component of the model, based on the Explicit Semantic Analysis [2], supports a semantic approximation between query and dataset terms: operations in the T-Space are mapped to *semantic relatedness* operations using the distributional background knowledge of Wikipedia. The robustness of distributional semantics, compared to WordNet based or ontology-based approaches, lies in the scope of its semantic matching, which allows matching between terms from different grammatical classes and it is not dependent on manually created structured resources (distributional semantic models are automatically built from texts).

The elements of the query processing approach are depicted in Figure 1. In the *query pre-processing* phase the natural language query is analyzed by the *Interpreter* component, where a set of *query patterns and features* are detected in the user query. The second phase consists of the *vocabulary independent query processing approach* which defines a sequence of search and data transformation operations over the data graph embedded in the T-Space, targeting a maximum semantic matching with the query. The *Query Planner* generates the sequence of *graph semantic search and navigation operations* which defines the *query processing plan*, based on a set of *query features* which are determined in the pre-processing phase. Each query feature maps to a set of graph search and navigation operations. The third phase

[1] http://googleblog.blogspot.ie/2012/05/introducing-knowledge-graph-things-not.htm, 2012

SIGIR'13, July 28–August 1, 2013, Dublin, Ireland.
ACM 978-1-4503-2034-4/13/07.

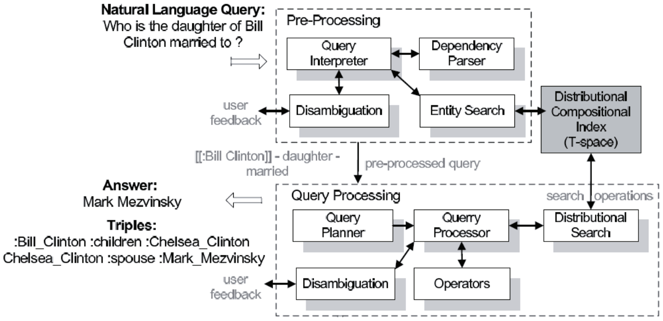


Figure 1: Query processing architecture.

consists in the execution of the query processing plan operations over the T-Space index. The T-Space VSM model index is implemented over *Lucene* 3.5 IR framework.

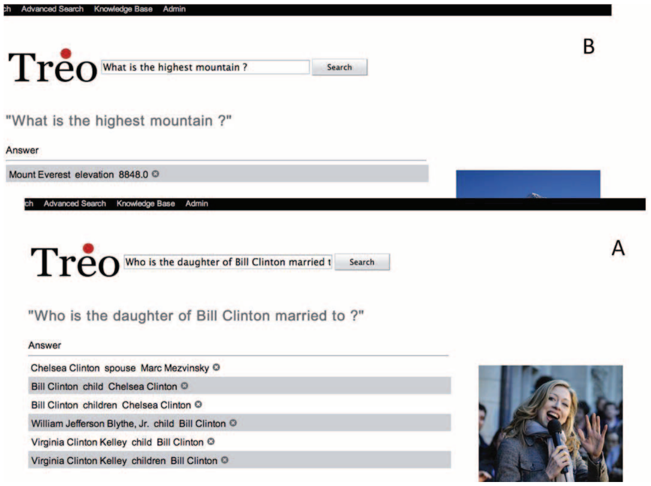
3. SYSTEM DEMONSTRATION

The system is demonstrated over the open-domain DBpedia 3.7 and YAGO Linked datasets[2]. The full dataset consists of 128,071,259 triples (17GB) loaded into the Treo index. A set of natural language queries from the Question Answering over Linked Data Challenge [1], extracted from real user queries, are used to demonstrate the system. Users input free natural language queries and the system returns two types of results: (i) a list of highly related triples or (ii) post-processed/aggregate results, depending on the query type.

For the example query (A) *'Who is the daughter of Bill Clinton married to ?'* (Figure 2), the system returns a list of triples which have the answer for the user query, including the final answer *'Bill Clinton's child is Chelsea Clinton'* and *Chelsea Clinton's spouse is Marc Mezvinsky*. This query also returns related triples which are not the answer for the query but which can be quickly filtered by users. This *semantic best-effort* behavior places Treo in a hybrid scenario between a *question answering* and a *semantic search* system. In the example, the distributional semantic model was used to match *daughter* to *child* and *married* to *spouse*. For the second query (B), *'What is the highest mountain ?'* (Figure 2), the system returns a precise answer. In answering this query, the system first determined the class *Mountain* in the dataset, expanding all instances associated with this class. The system used the distributional semantic information to match *highest* with all properties available for the instances of *Mountain*, finding *elevation* to be the most related. The next step sorts all instances by elevation and retrieves the top most triple. Further examples can be found at the Treo website[3].

4. RELATED WORK

PowerAqua [4] is a question answering system which uses PowerMap, a hybrid matching algorithm comprising terminology level and structural schema matching techniques with the assistance of large scale ontological or lexical resources. In addition to the ontology structure, PowerMap uses WordNet-based similarity approaches as a semantic approximation strategy. Exploring user interaction techniques, FREyA [5] is a QA system which employs feedback and clarification dialogs to resolve ambiguities and improve the domain lexicon with the help of users. User feedback is also used to enrich the semantic matching process by allowing manual input of query-vocabulary mappings. Compared to existing approaches, Treo focuses on the use of distributional semantic models to support a more comprehensive query-data matching mechanism, with a lower adaptation effort for new datasets. Additionally, Treo uses a *distributional structured vector space model* (T-Space) [3] to support semantically approximate queries over labelled data graphs (EAV data model), providing a principled associated IR framework over structured data.


ch Advanced Search Knowledge Base Admin
B
Treo
Search
ch Advanced Search Knowledge Base Admin
A
Treo
Search

What is the highest mountain ?

"What is the highest mountain ?"

Answer

Mount Everest elevation 8848.0

Who is the daughter of Bill Clinton married t

"Who is the daughter of Bill Clinton married to ?"

Answer

Chelsea Clinton spouse Marc Mezvinsky
Bill Clinton child Chelsea Clinton
Bill Clinton children Chelsea Clinton
William Jefferson Blythe, Jr. child Bill Clinton
Virginia Clinton Kelley child Bill Clinton
Virginia Clinton Kelley children Bill Clinton

Figure 2: Example queries: (A) semantic best-effort answer (B) exact answer.

5. ACKNOWLEDGMENTS

This work has been funded by Science Foundation Ireland under Grant No. SFI/08/CE/I1380 (Lion-2).

6. REFERENCES

[1] 1st Workshop on Question Answering over Linked Data, http://www.sc.cit-ec.uni-bielefeld.de/qald-1, 2011.

[2] E. Gabrilovich and S. Markovitch, Computing semantic relatedness using Wikipedia-based explicit semantic analysis, in *Proc. International Joint Conference On Artificial Intelligence*, 2007.

[3] A. Freitas, E. Curry, J. G. Oliveira, S. O'Riain, A Distributional Structured Semantic Space for Querying RDF Graph Data. International Journal of Semantic Computing (IJSC), 2012.

[4] V. Lopez, E. Motta, V. Uren, PowerAqua: Fishing the Semantic Web, Proc. 3rd European Semantic Web Conference ESWC, Vol. 4011. p. 393-410, 2004.

[5] D. Damljanovic, M. Agatonovic, and H. Cunningham, FREyA: An Interactive Way of Querying Linked Data Using Natural Language, Proc. 1st Workshop on Question Answering over Linked Data (QALD), (ESWC 11), 2011.

[2] http://dbpedia.org

[3] http://treo.deri.ie/SIGIRDemo