Treo: Best-Effort Natural Language Queries over Linked Data

André Freitas[1], João Gabriel Oliveira[1,2], Seán O'Riain[1], Edward Curry[1], and João Carlos Pereira da Silva[2]

[1]Digital Enterprise Research Institute (DERI)
National University of Ireland, Galway
[2]Computer Science Department
Universidade Federal do Rio de Janeiro

Abstract. Linked Data promises an unprecedented availability of data on the Web. However, this vision comes together with the associated challenges of querying highly heterogeneous and distributed data. In order to query Linked Data on the Web today, end-users need to be aware of which datasets potentially contain the data and the data model behind these datasets. This query paradigm, deeply attached to the traditional perspective of structured queries over databases, does not suit the heterogeneity and scale of the Web, where it is impractical for data consumers to have an a priori understanding of the structure and location of available datasets. This work describes *Treo*, a best-effort natural language query mechanism for Linked Data, which focuses on the problem of bridging the semantic gap between end-user natural language queries and Linked Datasets.

Keywords: Natural Language Queries, Linked Data

1 Introduction

End-users querying Linked Data on the Web should be able to query data spread over potentially a large number of heterogeneous, complex and distributed datasets. However, Linked Data consumers today still need to have a previous understanding of the available datasets and vocabularies in order to execute expressive queries over Linked Datasets. In addition, in order to query Linked Data, end-users need to cope with the syntax of a structured query language. These constraints represent a concrete barrier between data consumers and datasets, strongly limiting the visibility and value of existing Linked Data.

In this scenario, natural language queries emerge as a simple and intuitive way for users to query Linked Data [1]. However, unrealistic expectations of achieving the precision of structured query approaches in a natural language query scenario and in the Web scale, brings the risk of overshadowing short-term opportunities of addressing fundamental challenges for Linked Data queries. With the objective of reaching the balance between precision, flexibility and usability, this work focuses on the construction of a *best-effort natural language query*

approach for Linked Data. Search engines for unstructured text are an example of the success of best-effort approaches on the Web. One assumption behind best-effort approaches is the fact that part of the search/query process can be delegated to the cognitive analysis of end-users, reducing the set of challenges that need to be addressed in the design of the solution. This assumption, applied to natural language queries over Linked Data, has the potential to enable a flexible approach for end-users to consume Linked Data.

This work presents *Treo*, a query mechanism which focuses on addressing the challenge of bridging the semantic gap between user and Linked datasets, providing a precise and flexible best-effort semantic matching approach between natural language queries and distributed heterogeneous Linked Datasets.

2 Description of the Approach

In order to address the problem of building a best-effort natural language query mechanism for Linked Data, an approach based on the combination of *entity search*, a *Wikipedia-based semantic relatedness measure* and *spreading activation* is proposed. The center of the approach relies on the use of a Wikipedia-based semantic relatedness measure as a key element for matching query terms to dataset terms. Wikipedia-based relatedness measures address limitations of existing works which are based on similarity measures/term expansion based on WordNet [2].

The query processing starts with the determination of key entities present in the user natural language query, using named entity recognition. Key entities are entities which can be potentially mapped to instances or classes in the Linked Data Web. After detected, key entities are sent to the entity search engine which determines a list of pivot entities in the Linked Data Web. A pivot entity is an URI which represents an entry point in the Linked Data Web for the spreading activation search. After the entities present in the user natural language query are determined, the query is analyzed in the query parsing module. The output of this module is a *partial ordered dependency structure* (PODS), which is a reduced representation of the query, targeted towards maximizing the matching probability between the structure of the terms present in the query and the subject, predicate, object structure of RDF.

The PODS and the list of pivot entities are used as the input for the spreading activation search algorithm. Starting from the URI representing the first pivot entity in the list, Treo dereferences the pivot entity URI, fetching its associated RDF description. The RDF description contains a set of properties and objects associated with the entity URI. For each associated pair (*property, object*), *Treo* calculates the semantic relatedness measure between the pair and the next query term. The Wikipedia-based relatedness measure allows the computation of the semantic proximity between the query terms and the terms present in the datasets (in both terminological and instance level). This step represents the core of the flexibility introduced in the semantic matching process. Objects associated with pairs with high relatedness discrimination are further explored

(i.e. their URIs are dereferenced) and the semantic relatedness of the properties and objects in relation to the next query term is computed. The process continues until the end of the query is reached. The process of node search using semantic relatedness defines a spreading activation search where the activation function is defined by the semantic relatedness measure. The spreading activation search returns a set of ranked triple paths, i.e. connected triples starting from the pivot entities which are the result of the search process. The set of triple paths are further post-processed and merged into a graph which is the output displayed to users. The characteristic of the approach of using semantic relatedness to determine the direction of navigation on the Linked Data Web and the set of returned triple paths, defined the name Treo for the prototype, the Irish word for path and direction.

The components of the prototype's architecture are depicted in figure 1. In order to minimize the query execution time, three caches are implemented. The relatedness cache stores the values of previously computed semantic relatedness measures between pairs of terms. The RDF cache stores local copies of the RDF descriptions for the dereferenced URIs, in order to minimize the number of HTTP requests. The URI-term cache stores labels for each URI which are used in the semantic relatedness computation process. An important characteristic of the proposed approach, which was not implemented in the current version of Treo, is the natural level of parallelization that can be achieved in the relatedness-based spreading activation process.

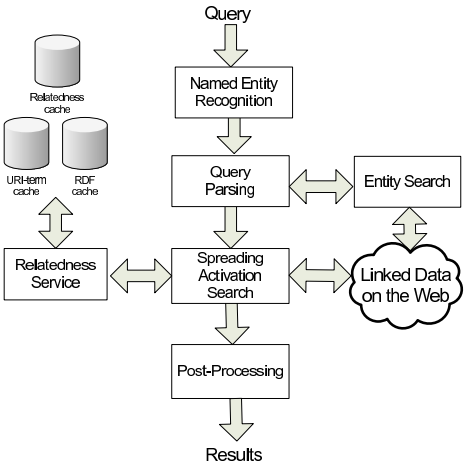

Fig. 1: Components of Treo's architecture.

3 Query Example

Figure 2 shows the output for the query *'From which university did the wife of Barack Obama graduate?'*. The output shows an example of merged triple

paths containing the correct answer for the user query (the URIs for *Princeton University* and *Harvard Law School*) together with triples which are not part of the desired answer set. The best-effort nature of the approach delegates to the user the final cognitive validation of the answer set. In order to support users in the validation process, the relationships from the pivot entity (in this query, *Barack Obama*) to the final answer resources/values are displayed. This example emphasizes the flexibility introduced by the semantic relatedness measure in the query-dataset matching, where the query term *'wife'* is matched with the dataset term *'spouse'* and the query terms *'graduate'* is matched with the *'university'* type in the dataset. Additional results for different queries can be found in [3].

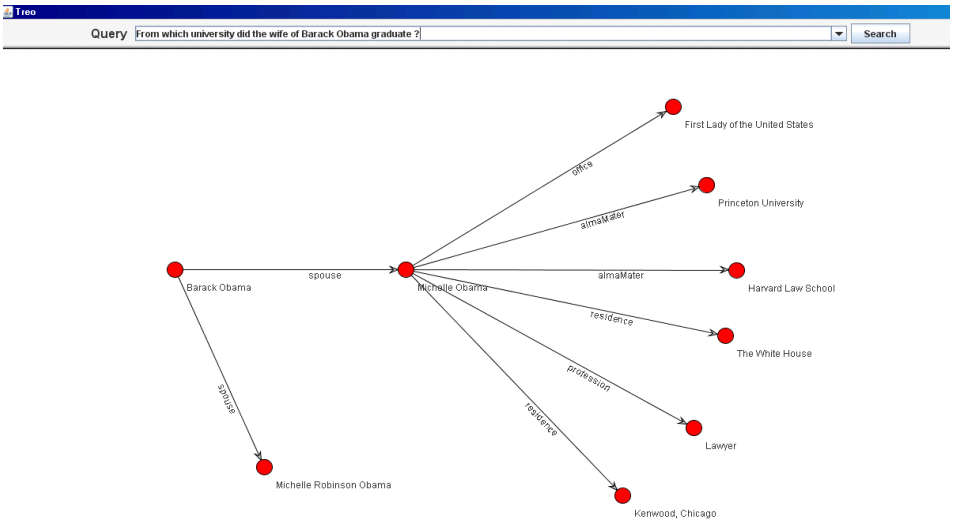

Fig. 2: Treo result set for the query *'From which university did the wife of Barack Obama graduate?'*.

Acknowledgments. The work presented in this paper has been funded by Science Foundation Ireland under Grant No. SFI/08/CE/I1380 (Lion-2).

References

1. Kaufmann, E., Bernstein, A.: Evaluating the usability of natural language query languages and interfaces to Semantic Web knowledge bases. Web Semantics: Science, Services and Agents on the World Wide Web 8 393-377, 2010.
2. Freitas, A., Oliveira, J.G., O'Riain, S., Curry, E., Pereira da Silva, Querying Linked Data using Semantic Relatedness: A Vocabulary Independent Approach. In Proceedings of the 16th International Conference on Applications of Natural Language to Information Systems (NLDB), 2011.
3. Treo Query Examples, `http://treo.deri.ie/gallery/nldb2011.htm` (2011).