Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence (AAAI-16)

Pragmatic Querying in Heterogeneous Knowledge Graphs

Amar Viswanathan
Tetherless World Constellation
Rensselaer Polytechnic Institute
110 Eighth Street, Troy, NY USA 12180

Summary

Knowledge Graphs with rich schemas can allow for complex querying. My thesis focuses on providing accessible Knowledge using Gricean notions of *Cooperative Answering* as a motivation. More specifically, using *Query Reformulations*, *Data Awareness*, and a *Pragmatic Context*, along with the results they can become more responsive to user requirements and user context.

Motivation

Given the rich schema definitions in Knowledge Graphs (KG), user querying is becoming quite expressive. This is also changing the search landscape. For example, user queries such *Movies directed by Martin Scorcese featuring Jose Pesci and Robert De Niro* or *Michelin star rated restaurants in New York City* are easily answered by resources like DBpedia(Auer et al. 2007). Thus, KGs are able to handle more expressive precise querying. This makes it more ideal for a variety of search tasks. However, while the research world is focused on adding increased expressiveness, an often overlooked part is the complexity of data access. With an increase in schema definitions, the user finds it harder to find information suitable to his needs in the KG space. Concepts can interact in a myriad of different ways. This means that even if the user has a precise need, translating that to a specific query becomes difficult with the increase in concept size and concept interaction space. Also given the long association with databases, querying the data from a table point of view is a strong impediment to utilize the richness of the Knowledge Graph. We contribute the following as potential reasons:

- **Ever Expanding Schema:** Heterogeneous Knowledge Graphs have all the required information, but their ever expanding schema, which is not consolidated, leads to mis-understanding of schema semantics and inhibits precise querying(Dolog et al. 2009).
- **Incomplete and Inconsistent data:** Even if the schema is complete sometimes the data is not always complete. In addition given the reliance on automatic extraction techniques, a lot of the extracted facts can be redundant, irrelevant or errors (Pujara et al. 2013). This leads to inaccurate results.
- **Complex Querying:** Querying data with SPARQL requires the user to build the exact triple patterns for his requirement. However, such requirements aren't always mapped exactly, and there is a huge gap between the *user intent* and what has been formulated as a query. Users are very often able to provide only broad queries, whereas to build a precise query they need more schema familiarity. (Dolog et al. 2009).
- **Unadaptable to User needs:** Current interfaces to the Knowledge Graph are also very unadaptable to the user and make no effort in augmenting the user knowledge. In addition current interfaces also don't provide mechanisms to suggest various interpretations of the same query.

Targeted Research Contributions

In this thesis we would like to target the problem of making Knowledge Graphs accessible for users by providing pragmatic and constrained reformulations along with results in a faceted interface. Viewing at search as a continuous iterative conversation between a user and the search system, we take the Gricean approach of *cooperative answering*(Grice 1970), where both the participants in the conversation contribute towards achieving mutual conversational ends. We demonstrate this in our system, which takes as input a user query and provides alternate possible ranked reformulations which are aligned to both the schema and to the pragmatic context. We have developed a *data aware* Pragmatic Query Reformulation system , which provides a user with a set of reformulated queries that takes into account *data availability* and a set of well-defined RDFS entailment rules for query reformulation. In brief, the contributions are: *(a)* Better data awareness in terms of context and data availability *(b)* Query reformulations constrained by the pragmatic context *(c)* Faceted discourse oriented user interface to enable a dialog, which utilizes the reformulations.

Our evaluation root queries would include *star, chain* and *complex* query types.Due to the lack of benchmarks, we are looking at evaluation on both synthetic -LUBM[1] and non-synthetic- ACE datasets.

[1] http://swat.cse.lehigh.edu/projects/lubm/

Background and State-of-the-art

The foundations of my approach lie in the application of Grice's principles(Grice 1970). Grice proposed that, *talk exchanges do not normally consist of a succession of disconnected remarks, but they are, to some degree, cooperative efforts and the participants recognize in them, a common set of goals*. To be able to achieve this in the current context a system should be able to talk to the user and provide contextually relevant information or additional similar relevant queries. So we rely on *Query Relaxations* and *Query Reformulations*, which are a part of *Cooperative Answering*. Generally Reformulations for RDF graphs are focused on Relaxations or Generalizations(Hurtado, Poulovassilis, and Wood 2008) aimed at pushing more relevant content to the users. Such relaxations are either deductive relaxations or use RDFS semantics i.e. type hierarchy or property hierarchy(Poulovassilis and Wood 2010) to relax triple patterns to generate more data. While such systems work on the concept and property level, they do not consider the implications of *data availability* and user query context. In addition they do not seek to always restrict the amount of new reformulations. This work also aims at using pragmatic context, which would include data availability, user preferences, domain expertise and geographical coordinates as a key factor in constraining the generated reformulations.

Current Progress

Our first approach was to devise a novel *Pragmatics and Data Aware* Query Reformulation Algorithm. This is a work in progress and we are targeting the 25th International World Wide Web Conference [2]. We summarize the results of this reformulation with an example query $q_1 \in \{$*Find all nations who are involved in attacks* $\}$, which looks like :

$$\boxed{\begin{array}{r} q_1\,\{entity\ \ event\ \ role\} := \\ entity\ \ role\ \ event \\ entity\ \texttt{rdf:type}\ individual \\ event\ \texttt{rdf:type}\ attack \end{array}}$$

Table 1 shows the results of the reformulation for query q_1 using techniques developed by our Algorithm. This example is queried on a sample Knowledge Graph KG that is extracted from 75,000 documents, which are in the ACE'05 [3] schema. In q_1, the given query is matched against {ENTITY,ROLE,EVENT} from the schema. The Schema built from the documents has a total of 132 classes and 233 Logical axioms, along with 37 Object Properties and 10 Data Properties. In addition to these results we have built a discourse enabled faceted user interface that eases the interaction with the KG.

TECHNIQUE	#REFORMULATIONS
Query Relaxation	74,620
Entity Aware Restriction	11480
Event Restriction	840
Domain&Range Restriction	48
Similarity Restriction	24
Data Aware Restriction	24

Table 1: Reformulation Steps for q_1 applying **Algorithm 1**

Future Directions

In the immediate future I would like to investigate how *Query Reformulation* fits into the larger scheme of *Relevance* and *Pragmatic Context*. Defining the parameters that would form a Pragmatic Context for KGs would be the basis for evaluating relevance. In the next phase, our work will be focused on extending this system to cater to more expressive Knowledge Graphs i.e. graphs with a denser representation of concepts and their interactions. A denser schema along with a more complete Knowledge Graph can provide for more relevant contextual clues. This would significantly improve the quality of results and reformulations that become available for ranking. In addition these reformulations also form a reformulation graph. Adapting the faceted interface to use the reformulations, thus providing a capability for intelligent dialog and constrained navigation of the underlying data would be the final phase of this work.

Acknowledgments

The Research was sponsored by the Army Research Laboratory and was accomplished under Cooperative Agreement Number W911NF-09-2-0053 (the ARL Network Science CTA). The principal project investigators are myself, my advisor, James A.Hendler and Geeth De Mel from IBM Research. I would like to acknowledge their help and generous suggestions for shaping this work.

References

Auer, S.; Bizer, C.; Kobilarov, G.; Lehmann, J.; Cyganiak, R.; and Ives, Z. 2007. *Dbpedia: A nucleus for a web of open data*. Springer.

Dolog, P.; Stuckenschmidt, H.; Wache, H.; and Diederich, J. 2009. Relaxing rdf queries based on user and domain preferences. *Journal of Intelligent Information Systems* 33(3):239–260.

Grice, H. P. 1970. *Logic and Conversation*.

Hurtado, C. A.; Poulovassilis, A.; and Wood, P. T. 2008. Query relaxation in rdf. In *Journal on data semantics X*. Springer. 31–61.

Poulovassilis, A., and Wood, P. T. 2010. Combining approximation and relaxation in semantic web path queries. In *The Semantic Web–ISWC 2010*. Springer. 631–646.

Pujara, J.; Miao, H.; Getoor, L.; and Cohen, W. 2013. Knowledge graph identification. In *International Semantic Web Conference*.

[2] http://www2016.ca/

[3] http://www.itl.nist.gov/iad/mig//tests/ace/ace05/doc/