CSC 742 (Advanced Topics in Data Management)
Project Proposal

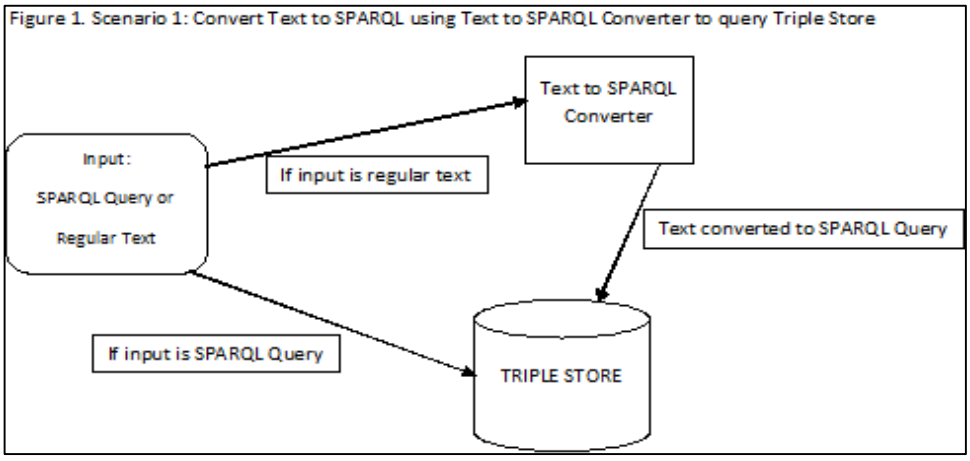
Hybrid Structured and Unstructured Search

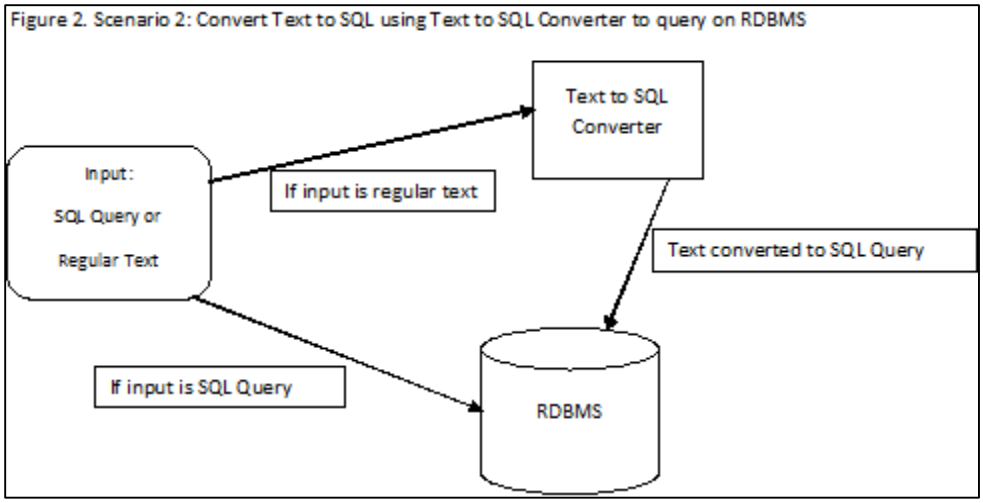
Aniket Sonmale (amsonmal)
Neelesh Salian (nsalian)
Nitin Rao (nrao2)
Omkar Dalvi (okdalvi)

Semantic web search is a sought after technology in the industry today with endless possibilities in the future. Data as we see it and understand has evolved over a period of time. We are entering a new era where slowly RDBMS' might well be replaced by the more optimized triple store databases. Triple store as we know houses information in the form of triples <subject, predicate, object>. Resource Development Framework (RDF) is a popular choice for querying data stored in triples. While RDF has been around for a while with SPARQL being the preferred choice for writing RDF based queries, not many developers across the globe are well versed with SPARQL. Our idea is to promote the use of triple stores and RDF by allowing users to query for data by means of natural language. We want to create an interface for a hybrid searching mechanism that takes in natural text, parses it into a SPARQL query which can then be used to retrieve data from the triple store. While the conversion might make the process slower, it gives more options to users who actually want to migrate to a triple store but are not familiar with the SPARQL framework.

The idea to convert plain text to SPARQL queries has been discussed for some time now and attempts have been made by several people including AutoSPARQL which appears to be discontinued. We feel that there is still some work to be done in this field and given that this might well be the future of data storage we decided to address the need to make RDF queries more user friendly. We intend to implement Natural Language Processing (NLP) to help us perform the conversion from text to SPARQL.
The approach that we intend to take for this project can be summarized by means of scenario based diagrams.

Proposed Architecture:





Here the Scenario 2 is something that we want to attempt if time permits. As mentioned earlier, our goal is to create a complete hybrid searching mechanism that unifies RDBMS and Triple Stores.

Milestones:

1) Gain understanding of Triple Store databases, RDF and using SPARQL to query.
2) Understanding NLP and how to parse plain text.
3) Parsing plain text to SPARQL.

Extra credit:

Time permitting we would like to extend this project to unify RDBMS and Triple store by allowing users to enter natural language text to query the aforementioned databases. This would provide more flexibility for users to query data on any form of database without having to resort to a particular form of programming based on the database type. It could be a unification of structured and unstructured data which would be the ultimate goal of our vision.
At the end of the semester, we aim to formulate a research paper on this topic.

References:

1. Triple-Store-Databases: http://en.wikipedia.org/wiki/Triplestore
2. LingPipe: http://alias-i.com/lingpipe/
3. Apache OpenNLP https://opennlp.apache.org/
4. Natural Language Processing: http://surface.syr.edu/cgi/viewcontent.cgi?article=1043&context=istpub
5. Natural Language Processing: http://onlinelibrary.wiley.com/doi/10.1002/asi.4630350507/pdf
6. SPARQL: http://www.w3.org/TR/rdf-sparql-query/
7. Difference between a triple store and a relational database: http://krisalexander.com/innovation/2013/07/16/the-difference-between-a-triplestore-and-a-relational-database/
8. AutoSPARQL: http://jens-lehmann.org/files/2011/autosparql_eswc.pdf
9. Processing SPARQL queries in RDF using Regular expressions

http://www.biomedcentral.com/content/pdf/1471-2105-12-s2-s6.pdf
10. http://gate.ac.uk/sale/dd/related-work/Kaufmann_nlp+reduce_ESWC2007.pdf
11. Semantic Content Access Using Domain-Independent NLP Ontologies
http://link.springer.com/chapter/10.1007/978-3-642-13881-2_4