SPARQL Query Construction from keywords

1. Objective:
With emergence of semantic web technologies there would be lots of structured information available which can logically queries for accurate search, but due to complexity involved in current semantic query languages (e.g RQL, SPARQL, RDQL) there has been a critical gap between semantic search and end users. To reduce this complexity we suggest a hybrid approach to retrieve information from semantic datasets using traditional keyword based queries by transforming them to semantic query for SPARQL. In this approach traditional IR techniques can be applied to discover URIs of respective keyword query terms. Using this URIs one can formulate SPARQL queries to semantically retrieve queried information by combining keyword based search with semantic search. To achieve this functionality we need to develop a distributed knowledge base which maps set of keywords to URIs from different datasets. The algorithm to populate this knowledge base is a challenging work given variety and ambiguity in current datasets. Our objective is to develop a prototype system which given a keyword based formal logic query, outputs SPARQL query using above suggested knowledge base. Which in order can be executed against datasets.

2. Motivation
In the last few years, there has been increasing interest in applying keyword query to structured data such as XML and Relational Database. The attractions are that users can keep the habit in traditional web search and do not need to know about the data schema. Also Results have shown that Semantic search results are accurate than keyword search given its Precision-oriented Search.¹ But current semantic information retrieval technology like SPARQL needs URI based queries of respective dataset. In order to provide more convenient keyword based searching on semantic datasets one has to resolve keywords to URIs. To experiment feasibility of this keyword to URI mapping approach we primarily replacing keywords from formal keyword based query by URIs from

¹ A Comparative Study between Keyword and Semantic Based Search Engines 
http://rgpv.ac.in/iccbdt/Papers/CL-112.pdf
In recent years many search system for web of data has been proposed. R.Delbru has examines the shift from the traditional web document model to a web data object (entity) model and studies the challenges faced in implementing a scalable and high performance system for searching semi-structured data objects over a large heterogeneous and decentralised infrastructure in 2. Towards this goal, they defined an entity retrieval model, and developed methodologies for supporting this model and show how to achieve a high-performance entity retrieval system introducing an indexing methodology for semi-structured data which offers a good compromise between query expressiveness, query processing and index maintenance compared to other approaches. They address high-performance by optimisation of the index data structure using appropriate compression techniques. Finally, they demonstrate that the resulting system can index billions of data objects and provides keyword-based as well as more advanced search interfaces for retrieving relevant data objects in sub-second time. In 3 klara has developed KWQL which is keyword based querying for the social semantic search. In 4, Researchers explored a novel approach of adapting keywords to querying the semantic web: the approach automatically translates keyword queries into formal logic queries so that end users can use familiar keywords to perform semantic search. A prototype system named ‘SPARK’ has been implemented in light of this approach. Given a keyword query, SPARK outputs a ranked list of SPARQL queries as the translation result. In 5 the methods for automatically transforming keyword-based queries into SPARQL has been suggested. Also work has been done in improving those methods in order to apply them on (a large subset of) the Linked Data Web. A heuristic method has been proposed for generating SPARQL queries out of arbitrary number of keywords.

4. Expected Outcomes
At the end of this we are expecting a prototype system which take inputs a keyword based formal query for a dataset and outputs a SPARQL query by replacing its keywords by appropriate URIs from that dataset. The system comprise of a continuously updating


3 KEYWORD-BASED QUERYING FOR THE SOCIAL, SEMANTIC WEB – THE KWQL LANGUAGE: CONCEPT,. ALGORITHM AND SYSTEM

4 SPARK: Adapting Keyword Query to Semantic Search

5 DC Proposal: Automatically transforming keyword queries to SPARQL on large-scale knowledge bases
4. Action Plan

We would first set up a scalable big table which is a sparse, distributed, persistent multidimensional sorted map consisting of keyword, URIs, Type of Resource indicated by the URI (Object, subject or Predicate). The bigtable would be implemented on Hbase which is storage system for Hadoop distributed file system. The URIs in table would be populated using crawling algorithms which take datasets as input and insert URIs and Other information about visited URI in table. For this we have chosen Dbpedia dataset. Requirements and challenges for crawling the Linked Data web are different from regular web crawling. There exists many tool to crawl rdf dump like [6] and We have chosen LDSpider which is an open-source crawling framework for the Web of Linked Data. For each Inserted URI, depending upon the type of resource it is indicating different keyword extraction algorithm would map appropriate keyword from wordnet to that URI and update the table persistently. Primarily we are experimenting this approach with algorithms to extract keyword by parsing Basic Resource structure like URI itself and tags(rdf:label). (In case of DBpedia dataset each URI contains keyword of corresponding wikipedia page title). Further we can extend the algorithms to extract keyword from a variety of Web documents using tools that extracts structured data in RDF format. One example of such tool is any23[7]. In case of user supplied keyword don’t match with given mapping, vocabulary set like wordnet can help to find alternative term for that keyword for which there exist mapping. To test accuracy of mapping and performance of translation we have formulated test queries in keywords which would be converted into URI based SPARQL queries using above knowledge base.

---

6. LDSpider: An open-source crawling framework for the Web of Linked Data

7. Apache Any23
http://any23.apache.org/index.html