Project 1: Semantic Web Research - CECS 571

Liam Nguyen, Loc Hinh, Suchitra Chinnamail, Varun Lingabathini, Phuc Dang

1 On the Use of Cloud and Semantic WebTechnologies for Generative Design

Generative Design is a new programming paradigm often used in AutoCAD. Data sources that are different and similar, validation processes and optimizing knowledge base, and storing and managing Generative Design artifacts. Utilize Cloud for data sourcing for certain Autocad simulations, processing of simple and complex validation of Generative Design artifacts, and storing and managing Generative Design artifacts. Three points to address: data sources, validating artifacts and storing, managing, indexing artifacts. Data sources are hosted on the Cloud and similarities between sources are grouped without disturbing the original source format. Artifact verification is done in two steps: first they perform simple validation locally and complex validation using Cloud processing; second they indexed their validation sources similarly to their data sources to enable reuse. Storing and managing artifacts are in progress, but suggests that using metadata to locate files stored on the Cloud, artifacts can be stored anywhere compared to the classic method of maintaining artifact relationships.

2 Business Driven Insight via a Schema-Centric DataFabric

Analytics and Business Intelligence (BI) tools are not successful due to difficulty understanding both where the data came from and what the data means. Using 'schema centric data fabric' to model business structures allows the businesses to easily choose and send their data to analytic tools. Any generated data from analytic tools will retain context as the data conforms to the schema. AWS' Neptune offers high data availability, guaranteed transactions performing parallel updates, and light analytics processing. This service allows all data from the 'schema centric data fabric' to be stored on Neptune, and manipulated all within that service for light analytics. Creating and managing more complex analytics will require some custom components and user interfaces to be written, while utilizing SHACL to communicate between cloud applications. Modeling the business is done using RDF/OWL.

3 FootballWhispers: Transfer Rumour Detection

Football whispers provides a service of showcasing the rumours of transfers of players between teams. The problem that needs to be addressed to showcase the rumours, is the estimation of the veracity of the rumours, by providing relative likelihood of a player moving into a specific team. The likelihood is predicted using the Twitter conversations related to transfers. This process has two processes, Named-Entity Linking (NEL); and the assessment and combination of rumors to determine their relative likelihood. As rumor detection and the use of knowledge graph data was evolving, this led to significantly increased player NEL performance, and identifying of these rumors concerned actual 2016 football transfers.The primary reason for this is the availability of multilingual data in the KGs and the use of language agnostic statistical disambiguation techniques in NEL.

4 Semantic Concept Discovery Over Event Data

Preparing a report on a topic or a question which is comprehensive, accurate and unbiased is a challenge by itself. The first step as a part of the process is daunting discovery task which requires searching a number of information sources without introducing bias from analyst's current knowledge or limitations of these information sources. For most of the analysis reports a common requirement is a deep understanding of different kinds of historical and ongoing reported events in the media. The goal is to provide a unified solution allowing deeper understanding of the same data which is used to perform other analysis tasks.This is done by creating a framework to discover event databases using semantic web technologies. The system takes a set of event databases and RDF knowledge bases as input and provides a set of APIs as output. These APIs provide a unified retrieval mechanism on input data and knowledge bases, and an interface to various concept discovery algorithms.

5 Federated Semantic Data Management for Business Intelligence and Healthcare: Two Case Studies

In a large e-commerce company, the inconsistency between departments like IT, fulfillment, and accountant departments when they process data. In addition, in a hospital, the system misses to identify patients who need mechanical device (Left Ventricular Assist Device – LVAD) Using the structure of Ontology Based Data Access (OBDA): OWL (Web Ontology Language) as uniform conceptual model describing the interest and stored data; R2RML (RDB to RDF Mapping Language) to map ontology with the source database; SPARQL is an RDF Query Language for managing the heterogeneous source database. When applying the structure to real-world situations, an e-commerce company shows that the time and cost have been reduced to generate reports. Meanwhile, the hospital shows that more patients are identified.

6 The Qualification Data Repository (QDR): Enhancing Interoperability of Qualifications in Europe

In the area of employment across Europe, the stakeholders (jobseekers, learners, education institutes, employment services, Member States authorities, etc) need to share information to support recruitment and career management. QDR as a software component allows providers of data to upload datasets on European web portals, online service, and in semantic assets. During the operation, QMS (Qualification Metadata Schema) ensures consistency and guarantees the translation of dataset. Developing QDR is not an easy task. 2 problems: size of dataset is big, like versioning and validation. Secondly, the inconvenience of working with RDF directly from the front end.

7 Semantic Technologies for Data Analysis in Health Care

The HMOs in US need to demonstrate satisfactory performance W.r.t NCQA measures if they wish to participate in government funded healthcare schemes. The quality measures proposed are regularly revised and updated and this makes computation of relevant quality measures very complex using existing systems like SQL and SAS tools. In order to integrate relevant data from different heterogeneous sources, RDF is used along with declarative rules and adaptable schemas. The first step is to create a data model to transform the relevant patient data to a human readable format and the second step is to generate declarative rules and execute these rules using SPARQL queries. The above approach is implemented in one of the health care company in United states known as Kaiser permanente in Georgia region. The data is translated according to the above approach and the results were compared with the existing systems.The results were extremely encouraging as only 174 rules were needed when compared to 3000 lines of complex SQL code.

8 Extracting Semantic Information for e-Commerce

E-commerce websites in general uses a large legacy taxonomy of classes to organize the items and provide the relevant searches. In order to improve the user search experience, there is a need to extend the taxonomy and this takes a massive amount of time and human effort. The author tries to provide a solution to automate the process which would eventually increase the profitability of the business. The solution is to extend the taxonomy for the classes which are usually difficult to explore and this is achieved by aggregating small set of properties which are most popular among the users with the automatically selected taxonomy subtree and the outcomes are new RDF triples. Comparing the results of the automated approach to the manual work, the results were not consistent enough but using this automated approach along with manual process will help the accuracy get to above 80%.

9 Drug Encyclopedia – Linked Data Application for Physicians

Information about drugs is scattered among various sources and changes rapidly. The need to aggregate all information into a single source is critical for physicians to update their knowledge quickly and accurately. Drug Encyclopedia is a web application for physicians to search and explore the clinically relevant information about the medical product and drugs in general. The underlying technology is RDF and Linked Data (LD) Principles. Semantic technologies make the process of integrating various publicly available data sources much more flexible and easier.

10 Building and Using a Knowledge Graph to Combat Human Trafficking

This paper utilizes the huge amount of data on the web to combat human traffickers and help victims by building a knowledge graph from heterogeneous sources. There are 4 main steps to build a knowledge graph: data acquisition, mapping data to ontology, resolving entities, generating graph.In data acquisition, web crawler by Apache Nutch perform at scale and “landmark extractor” to identify elements using landmarks defined with regular expressions. Then, the extracted data is mapped to JSON-LD, a Linked Data representation by Karma project by USC. The next step is to form potential links between data items. This is done by Minhash/LSH algorithms in O(n * log(n)) time. When the association between entities are established, the graph is generated and contains a query interface. The system is initially deployed in six law agencies with the potential to roll out to 200 agencies. It has successfully identify several victims of human trafficking. With constant stream of 162,000 ads per day, this technology rebuilds the knowledge graph under 24 hours. Future application includes research trends, autonomous system, etc....