I-State

A Semantic Web Application

Malvika Guliani
Pujitha Kara
Richa Srivatsava

Arizona State University
CIDSE, Software Engineering

***Abstract*— This document gives a description about I-State which is a semantic web application. The document also summarizes the related work that is going on in the domains of our application, an overview of the application and a description about the semantic web concepts that we used to develop the project like integrating the data, generating instance data and querying the linked data.**

Keywords— semantic web, linked data, integration.

I. INTRODUCTION

I-State is a semantic web project that combines information from four different domains and gives an interesting and useful application. The datasets that are considered are the crime dataset provided by the US government, dataset of universities and static data from YELP app and TWITTER. The application integrates data from these datasets and displays data pertaining to a particular state. It is a simple web application developed using semantic web concepts like XML, XSLT, RDF, RDF-S, OWL, SPARQL, Protege, Apache Jena Semantic Web Framework. This application provides a useful functionality to the users. There are several apps that give details about crime rate and apps that give information about the universities. YELP app provides details of different events that are taking place in a particular area. Twitter is a social networking site. The major advantage of this application is that, it provides the user an integrated information instead of searching through different sites. This application can further be enhanced to integrate large data from several other domains.

II. GOALS

The major goals of the I-State application are as follows

A. To provide an interesting and useful semantic web application.

I-State is a simple, easy to use application. The only input that the user has to enter is the state. Depending on the input given by the user, the application displays crime rate in that state, the list of universities, the top trending news in the state and exciting events in the state. As the datasets we considered are from versatile domains, the user gets information that is useful like the crime and university information and also interesting data like the events and top tweets in that state.

B. To implement all the semantic web concepts we learnt in the course work.

We initially started choosing the datasets from different domains. The datasets we chose can be categorized under safety domain, education domain, social networking domain and entertainment domain. We collected the data in CSV and JSON formats. We then created a data model graph that represents the connection between the classes and properties of each domain. A detailed description about the data model is discussed in the following sections. The next phase was to generate the instance data. For the datasets in JSON format we used a Python program to generate the instance data. For the datasets in CSV format we used online tools to generate the instance data. Using the RDFs we generated, we tried to integrate the data and performed querying using SPARQL to get the desired results. We developed the web app using open source framework Apache Jena. The different APIs interact together to process the RDF data. We used Jena to query the RDF via SPARQL. For the front end we used HTML, CSS, JS, JAVA.

C. To provide a base for the development of useful applications.

The current application is a simple one that integrates data from four different domains and displays information related to a state. It can be enhanced to integrate data from more domains. This gives detailed statistics related to a state which can be used to estimate the development in the area, security, the ongoing events etc. Such integrated data can be used by normal citizens for information purpose and can be used by organizations for research purposes.

D. To develop team work and collaboration.

Apart from understanding and implementing the concepts, this group project helped us to work collaboratively by sharing the responsibilities. As the entire project was submitted in three deliverables, it helped us to do proper planning. Our preparation for the presentations and experience gained while developing the project aided us to think innovatively. So this project helped us to gain a hands-on for developing semantic web applications and also to work in teams.

III. Related work in the domains

The crime dataset majorly falls under security domain. The dataset we considered was extracted from the federal bureau of investigation site. The FBI regularly publishes the crime statistics that helps people to understand the threats both nationally and locally. The particular dataset we used is the Crime in US dataset which gives an analysis of the violent crime and property crime that is taking place in different states of the nation. There is a considerable amount of work related to semantic web technologies that is going on in this domain. Some of the references are "How can the semantic web help law enforcement?"[1],"Leveraging Semantic Web Technologies for Analysis of Crime in Social Science" [2], "Using Future Internet Technologies to Strengthen Criminal Justice",[3] .

The second dataset that we considered is the University data set which we collected from the inventory, data.gov site. This dataset gives information about universities in different states. This information include the name of the university, a web URL to the university and its address. This dataset can be considered under education domain. The user can access the university page using the web URL that is displayed by the application. By exploring the site, the user can know all the details regarding the university, the courses they offer and about the faculty. Semantic web technologies serve a great purpose for development in education domain. A few examples for work that is done in this domain include "Education and the Semantic Web",[4], "Semantic web technologies for education – time for a 'turn to practice'?",[5], "The New Challenges for E-learning: The Educational Semantic Web",[6], "Ranking Universities Using Linked Open Data",[7] .

The third dataset that we considered is tweets data from the social networking site Twitter. One profound example of related work being done in this domain is the linked data hash tag provided by the twitter application There are thousands of tweets that are posted and all are related to work on linked data , SPARQL, discussions about conferences on linked data etc. There is also a twitter page Linked data Platform which contains similar tweets.

The fourth dataset is the event dataset taken from the YELP app data. Our dataset contains information about the events taking place in a particular area. The semantic web concepts can be used in this domain to develop recommendations for the events based on popularity, user reviews and other related information collected from diverse sources. One example of such work is "Hybrid event recommendation using linked data and user diversity",[8].

IV. Already Existing Apps

In each of the domain we considered there are several apps that provide different functionalities.

The apps related to crime majorly were based on the concept of displaying crime rate in a particular area. Some apps also provided the functionality of showing crime alerts based on dynamic data generated. Some example apps are as follows

- CrimeMapping iOS app which is developed by Omega group
- CrimeReports which is an android app
- SpotCrime Crime Map, an android app

A majority of the apps related to university data tried to display information about each university. Based on the ranks of each university, they provided the feature to compare between universities. Some of the apps are as follows

- Startclass.com
- universityCompare, an iOS app
- A simple application provided by university of Texas at Austin (https://www.utexas.edu/world/univ/state/)

An ample number of applications are available for Twitter.

A large part of the applications were on providing better ways to manage and use the Twitter social networking site. A few apps are also available which try to provide some useful functionality based on the tweets.

Lastly, the apps related to events data. The major functionality provided by them is to display all the events that take place in a particular region.

V. Uniqueness of I-state Application

All the applications that provide information related to a state only concentrate on some fixed attributes like the capitols, population, area, GDP and elevation. Our application is unique as it provides a versatile information related to a state and not just the general statistics. It embeds all the functionalities provided by apps in individual domains (listed above) .It not only displays the information, it also facilitates to navigate to that particular university site or a particular venue of the event. As the user gets integrated information from completely different domains, they can use it for both entertainment purpose and information purpose. The users can also consider a subset of the information which may help to take decisions. For example, from the data that is displayed, the user may concentrate on the crime and university data, which help them to conclude about the safety in that particular university. The idea on which our application was developed is robust and can be enhanced on a larger scale to integrate data from more versatile domains. A large number of applications can be developed based on this integrated data.

VI. Key functinality provided by i-state

The major functionality of the I-State semantic web application is to provide an integrated information about a state. This information include the crime rate in the state, the universities available in the state, the top tweets made particular to that state and the popular events in that state. The uniqueness of this application from the existing apps is explained in the previous section. The user initially enters the state name and click on search. Then a page is displayed with all the information related to that state. We have considered Web URL as one of the element in the university instance

data, so the user can click on the link and navigate to the university page. The user can explore the university site to get more information. Based on the information that is displayed the user can make inferences about the safety of that particular state, the rankings of the universities available in that state, what famous events are conducted in that state and what is trending in that state. One other major functionality that we provided is compare button. Depending on the data displayed related to each state, a comparison is made and the user is shown with which state is safer, which state has better educational opportunities and which state has more events. The following images provide an overview of the I-state application

VII. The Semantic Data Model

Figure 1 and figure 2 illustrate the semantic data model for the I-State application. As every dataset information can be related to a State, we have considered State as the main class. Institutes is a sub class of the state class and it represents the university data. The Institute class has several properties out of which we chose to display only the address, name of the institute and the web URL of the university. All these properties have their domain as the Institute class and range the Literal class. Another subclass of the State class is the events class which represents the events data collected from YELP. The Events class has properties like name of the event, rating given for the event and address of the event. All of the data properties have their domain as the event class and range as the Literal class.

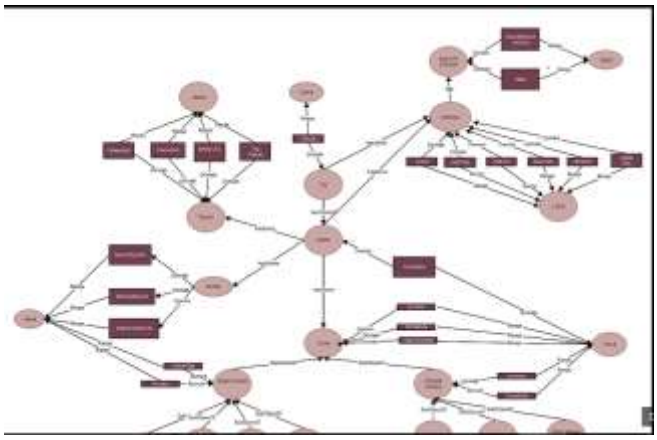


Fig1

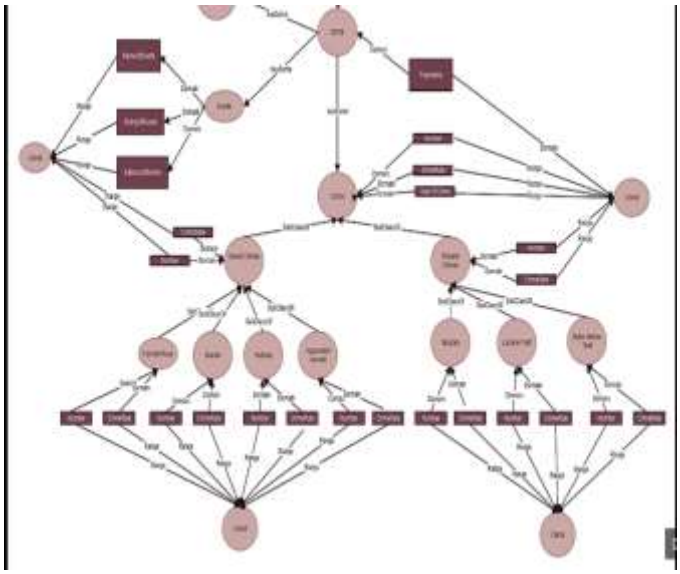


Fig 2

The third subclass of the State class is Tweets, which represent the dataset collected from twitter. This class has properties like retweeted, favorited and Media URL. All these properties have their domain as Tweets class and range as the Literal class.

The crime data is represented as properties of the State class. Out of the many properties of crime data, we considered the violent crime rate property which aggregates all the violent type crime rates and the property crime rate property which aggregates all the property type crime rates. Apart from these the state class has a city property. Its domain is the state class and range is the Literal class.

VIII. Challenges Faced in Integrating the Data

As we chose our datasets from versatile domains, we spent a considerable amount of time in figuring out the basis on which we should integrate the data. We had to first get an overview of the application and what we want the application to do. After identifying the datasets we had to integrate them. Some of the factors and challenges we had to consider are as follows

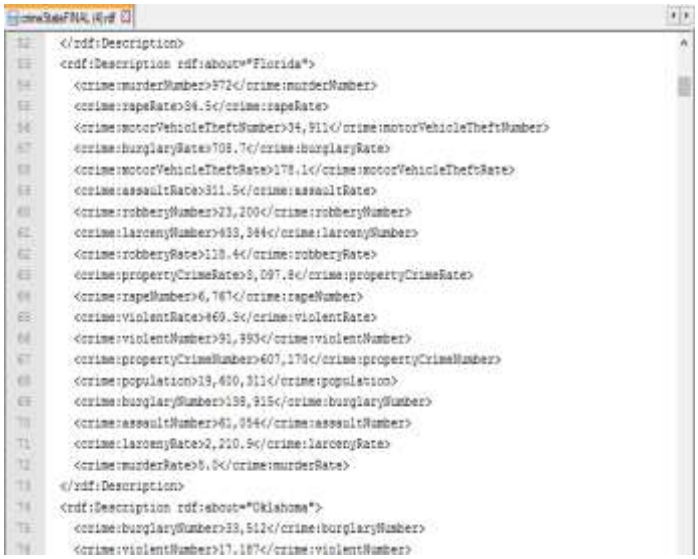
- Since the data is heterogeneous data, we had to perform a detailed analysis of characteristics and uses of each dataset.
- The next challenge was to identify the information that is needed for our application from the huge data that is present in the datasets.
- One major challenge was to extract datasets from YELP and Twitter data using respective APIs.
- We also had to consider the scalability issue while integrating the data.

IX. Generation of Instance Data

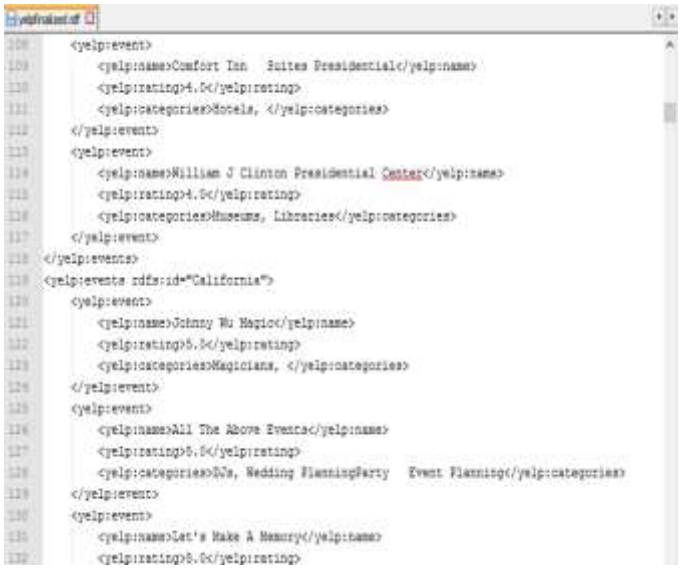
Two of the datasets that we considered are in CSV format and two are in JASON format. We used online tools like RDF123 and google refine to create instances of the datasets which are in CSV format. While generating the instance data, the datasets are validated against the owl file of our application. For the datasets in JASON format we used a python code to generate the instance data. A sample of the instance data for university dataset is given below

```
<rdf:Description rdf:about="#Fort Totten">
  <uni:address>214 1st Ave</uni:address>
  <uni:url>www.littlehoop.edu</uni:url>
  <uni:institute>Cankdeska Cikana Community College</uni:institute>
  <uni:state>North Dakota</uni:state>
</rdf:Description>
<rdf:Description rdf:about="#McKenzie">
  <uni:address>325 Cherry Ave</uni:address>
  <uni:url>www.bethelu.edu/</uni:url>
  <uni:institute>Bethel University</uni:institute>
  <uni:state>Tennessee</uni:state>
  <uni:address>16940 Highland Dr</uni:address>
  <uni:url>www.toatmckenzie.edu</uni:url>
  <uni:institute>Tennessee College of Applied Technology-McKenzie</uni:institute>
</rdf:Description>
```

A sample of instance data for crime dataset is given below



A sample of the event instance data is given below



A sample of twitter instance data is given below

X. QUERYING THE LINKED DATA

The main objective of our application is to display the data related to a state, where the user inputs the state. For this purpose we run SPARQL queries on the RDF instance data we generated. We used Jena to query the RDF via SPARQL. The SPARQL queries we used to get the desired are as follows

```
"PREFIX rdf: <http://w3.org/1999/02/22-rdf-syntax-ns#> " +
"PREFIX crime:
<http://www.semanticweb.org/dell/ontologies/2015/10/untitled-ontology-13#>" +
"PREFIX st: <file:///>" +
"SELECT * WHERE {" +
//"file:///C:/Users/DELL/workspace/Hello%20Semantic%20Web-Copy/src/org/semweb/assign6/" +
"st:" +s+ " crime:violentNumber ?vnumber ;" +
"crime:violentRate ?vrate ;" +
"crime:propertyCrimeNumber ?pnumber ;" +
"crime:propertyCrimeRate ?prate ;" +
"crime:picture ?pic." +
"}";


"PREFIX rdf: <http://w3.org/1999/02/22-rdf-syntax-ns#> " +
"PREFIX uni:
<http://www.semanticweb.org/dell/ontologies/2015/10/untitled-ontology-13#>" +
"PREFIX st: <file:///> +
"SELECT * WHERE {?univ uni:state '"+ s +"';"
+ "uni:city ?city;"
+ "uni:address ?address;"
+ "uni:url ?url}";
//"uni:institute ?inst" +
//"}";


"PREFIX rdf: <http://w3.org/1999/02/22-rdf-syntax-ns#> " +
"PREFIX twitter:
<http://www.semanticweb.org/dell/ontologies/2015/10/untitled-ontology-13>" +
//<http://www.semanticweb.org/dell/ontologies/2015/10/untitled-ontology-13#>" +
"PREFIX st: <file:///> +
"select * where {st:" +s+ " a ?type; twitter:mentionedin ?tweet. ?tweet twitter:text ?status. ?tweet twitter:retweeted ?ret. ?tweet twitter:favorited ?fav.}";
//"SELECT * WHERE {?state twitter:mentionedin ?tweet }";
//+ "?tweet twitter:text ?text}";


"PREFIX rdf: <http://w3.org/1999/02/22-rdf-syntax-ns#> " +
"PREFIX yelp:
<http://www.semanticweb.org/dell/ontologies/2015/10/untitled-ontology-13#>" +
"PREFIX st: <file:///> +
"select * where {st:" +s+ " a ?type; yelp:happening ?event. ?event yelp:name ?name. ?event yelp:rating ?rate. ?event yelp:categories ?cat.}";
//"uni:institute ?inst" +
//"}";
```

XI. CONSLUSION AND FUTURE WORK

I-State is a simple web application we developed using the semantic web technologies. We considered datasets from four domains. This application can further be enhanced on a larger scale, datasets from more versatile domains can be considered. The data displayed can be used manipulated, interpreted to develop more helpful applications.

REFERENCES

[1] Lieutenant John Liu, "How can the Semantic Web Help Law Enforcement ? , Fremont Police Department, March 19, 2012

[2] POLCOMING, "Leveraging Semantic Web Technologies for Analysis of Crime in Social Science" , Dibris, Universit`a di Genova,Italy

[3] John S.Hollywood, "Using future Internet Technologies to strengthen criminal justice", Dulani Woods, Richard Silbergilt, Brian A.Jackson, project

[4] Vladan Devedzic, "Education and the Semantic Web", Department of Information Systems and Technologies, FON – School of Business Administration, University of Belgrade, 2004

[5] Patrick Carmichael, Katy Jordan, "Semantic Web Technologies for education – time for a turn to practice", Published online, July 2012

[6] "The New Challenges for E-learning: The Educational Semantic Web", Lora Aroyo, Technische Universteit Eindhoven, Department of Mathematics and computer Science , Darina Dicheva, Winston-Salem State University, Department of Computer Science, 2004

[7] Rouzbeh Meymandpour, Ranking Universities Using Linked Open Data, Joseph G. Davis

[8] Houda Khrouf, Hybrid event recommendation using linked data and user diversity, Raphaël Troncy