Sportsvaganza

Do you know the score?

Author: Akash Chugh: Software Engineering department,
Arizona State University
Tempe, AZ

Author: Akshay Jain: Software Engineering department,
Arizona State University
Tempe, AZ

Author: Himani Shah: Software Engineering department,
Arizona State University
Tempe, AZ

Author: Rutuja Faldu: Software Engineering department,
Arizona State University
Tempe, AZ

Abstract**—** A semantic web application which aims to provide an amalgamation of sports news with Wikipedia, Twitter and BBC events calendar and google fusion tables. The application provides users insight into overall background information of news about a particular sport of the country which with help of dbpedia data set and the current on going trends and user reviews by the Twitter feeds of famous sports players of the country. Moreover the BBC events calendar to provide brief on events of sport extracted from news from and the country entered by the user.

To map this idea user is provided with a search keyword and then the consolidated ontology is searched which is generated from the above mentioned data sources and coloborative result is displayed.

Keywords—component; formatting; style; styling; insert (**key words**)

I. GOALS

The application developed, aims to keep the user informed about the sports news of a country. Along with the main sports news, the tweets by the sports legend (for sport identified in the news article) is displayed. The application also aims to apprise the user about the upcoming events and a wiki description about the events. The application revolves around the sports domain and educates the user about all the dimensions regarding sports.

II. RELATED WORK IN THE DOMAIN

Work in this direction is done previously by Rnews. Rnews does a seamless job if integrating the various aspects of news but that does in general. rNews is an approved standard for using semantic markup to annotate news-specific metadata in HTML documents. rNews has been developed by the IPTC, a consortium of the world's major news agencies, news publishers and news industry vendors. The Semantic markup allows publishers to attach specific *meanings* to various regions of an article page. One such semantic markup standard is called RDFa. RDFa is a framework for embedding semantic markup into HTML documents, but to apply RDFa to a specific domain it is necessary to develop terminology and data models specific to that domain. Another markup standard is called HTML5 Microdata. Microdata is another framework for embedding semantic mrkup into HTML document, it is adopted by schema.org as the preferred syntax. [4][5]
Other Apps and their functionality comparison:
Google News: **Google News** is a free news aggregator provided and operated by Google, selecting news from thousands of news websitesIt provides news specific to a country but does not provide the tweets and the upcoming events. Also, it provides you a link which further redirects on being clicked.
The website http://wiseworks.org/ provides sports news and upcoming events but does not provide the event description and the tweets by sports legends. Instead, it provides tweets about the own company Wiseworks.
The mobile application theScore does display sports news but does not filter it with a country and there is no information about the venues about the upcoming events.

III. UNIQUENESS OF PROJECT

There exists no such platform where the user can get sports news filtered based on the country. Our application provides sports and related information just based on a single filter(country). The user is provided with the trending news, the upcoming sports events along with their descriptions. Additionally, for social sports enthusiasts the feed is provided with tweets from iconic sports legends associated with the sport the trending news is talking about. All these services based on a single filter is provided on a single platform which is one of a kind and unique.

Moreover, the semantic annotation[1] of the news done can be an important future because it makes easier to understand for the browsers as it is not just a an formatted HTML page but a tag with just formatted heading taggers.

IV. DISCUSSION OF SEMANTIC DATA MODEL[1]

Semantic Web has five main components which help in accomplishing the required task and define the functioning of the web:

1. Uniform Resource Identifier :

A URI is simply a Web identifier: like the strings starting with "http:" or "ftp:" that you often find on the World Wide Web. Anyone can create a URI, and the ownership of them is clearly delegated, so they form an ideal base technology with which to build a global Web on top of. In fact, the World Wide Web is such a thing: anything that has a URI is considered to be "on the Web".

A URI may be classified as a locator (URL), or a name (URN), or both. A Uniform Resource Name (URN) functions like a person's name, while a Uniform Resource Locator (URL) resembles that person's street address . In other words: the URN defines an item's identity, while the URL provides a method for finding it.

The URI syntax consists of a URI scheme name followed by a colon character, and then by a scheme-specific part. The specifications that govern the schemes determine the syntax and semantics of the scheme-specific part, although the URI syntax does force all schemes to adhere to a certain generic syntax that, among other things, reserves certain characters for special purposes (without always identifying those purposes). The URI syntax also enforces restrictions on the scheme-specific part, in order to, for example, provide for a degree of consistency when the part has a hierarchical structure. *Percent encoding* can add extra information to a URI.

A *URI reference* is another type of string that represents a URI, and (in turn) represents the resource identified by that URI. Informal usage does not often maintain the distinction between a URI and a URI reference, but protocol documents should not allow for ambiguity.

A URI reference may take the form of a full URI, or just the scheme-specific portion of one, or even some trailing component thereof – even the empty string. An optional fragment identifier, preceded by #, may be present at the end of a URI reference. The part of the reference before the # indirectly identifies a resource, and the fragment identifier identifies some portion of that resource.

Web document markup languages frequently use URI references to point to other resources, such as external documents or specific portions of the same logical document.

2. RDF:

The Resource Description Framework (RDF) is a family of World Wide Web Consortium (W3C) specifications originally designed as a metadata data model. It has come to be used as a general method for conceptual description or modeling of information that is implemented in web resources, using a variety of syntax formats.[2]

The RDF data model is similar to classic conceptual modeling approaches such as Entity-Relationship or Class diagrams, as it is based upon the idea of making statements about resources (in particular Web resources) in the form of subject-predicate-object expressions. These expressions are known as *triples* in RDF terminology. The subject denotes the resource, and the predicate denotes traits or aspects of the resource and expresses a relationship between the subject and the object. For example, one way to represent the notion "The sky has the color blue" in RDF is as the triple: a subject denoting "the sky", a predicate denoting "has the color", and an object denoting "blue". RDF is an abstract model with several serialization formats (i.e., file formats), and so the particular way in which a resource or triple is encoded varies from format to format.

A collection of RDF statements intrinsically represents a labeled, directed multi-graph. As such, an RDF-based data model is more naturally suited to certain kinds of knowledge representation than the relational model and other ontological models. However, in practice, RDF data is often persisted in relational database or native representations also called Triplestores, or Quad stores if context (i.e. the named graph) is also persisted for each RDF triple. As RDFS and OWL demonstrate, additional ontology languages can be built upon RDF.

The subject of an RDF statement is either a Uniform Resource Identifier (URI) or a blank node, both of which denote resources. Resources indicated by blank nodes are called anonymous resources. They are not directly identifiable from the RDF statement. The predicate is a URI which also indicates a resource, representing a relationship. The object is a URI, blank node or a Unicode string literal.

In our application RDFs were created by converting CSV data fetched from various sources i.e. NYtimes, twitter, BBC events calender, Wikipedia, google tables to RDF triples with help of google refine. Which help in mapping the ontology elements to the URIs and generating triples.

3. OWL:

The Web Ontology Language (OWL) is a family of knowledge representation languages for authoring ontologies endorsed by the World Wide Web Consortium. They are characterised by formal semantics and RDF/XML-based serializations for the Semantic Web.

The data described by an ontology in the OWL family is interpreted as a set of "individuals" and a set of "property assertions" which relate these individuals to each other. An ontology consists of a set of axioms which place constraints on sets of individuals (called "classes") and the types of relationships permitted between them. These axioms provide semantics by allowing systems to infer additional information based on the data explicitly provided. [3]

The classes in OWL were generated using Protégé. The concept about the data fetched and needed for executing user queries was generated as classes or properties to relate them. And the various relations were set up which integrated the various data fetched from various sources.

4. Open Refine

OpenRefine, formerly called *Google Refine*, is a standalone open source desktop application for data cleanup and transformation to other formats, the activity known as data wrangling. It works in 3 steps of exploration of data, cleaning and transformation of data and reconciliation and matching of data.
It takes up the data in CSV format extracted from the individual python codes and converts the data to RDF format by mapping it to the classes of the ontologies. This is performed by an adding an rdf extension.
Thus the data in rdf triples is generated.

5. Apache Jena and Fuseki Server

Apache Jena is an open source Semantic Web framework for Java. It provides an API to extract data from and write to RDF graphs. The graphs are represented as an abstract "model". A model can be sourced with data from files, databases, URLs or a combination of these. A Model can also be queried through SPARQL.
Fuseki is a SPARQL server. It provides REST-style SPARQL HTTP Update, SPARQL Query, and SPARQL Update using the SPARQL protocol over HTTP.
Thus the data generated in different RDF is clubbed to a single RDF file and the triple are merged in fuseki and that data is queried using Jena+fuseki framework.

V. Challenges Faced in Integrating Data

1. Vastness: The World Wide Web contains at least 48 billion news articles and existing technology has not yet been able to eliminate all semantically duplicated terms. Any automated reasoning system will have to deal with truly huge inputs. Thus combination of data from varied wide data sources can be tedious.

2. Vagueness: Imprecise concepts arises from the vagueness of user queries, of concepts represented by content providers, of matching query terms to provider terms and of trying to combine different knowledge bases with overlapping but subtly different concepts this was faced during information integration about sports name with Wikipedia and events fetching from BBC calendar. Moreover, extraction of information of news is about which sport was a complication. In addition to it fetching the twitter handles for various sports players was cumbersome.

3. Uncertainty: These are precise concepts with uncertain values. For example, news might relate to a number of sports or there are more than one twitter accounts for the celebrity. Probabilistic reasoning techniques are generally employed to address uncertainty.

4. Inconsistency: These are logical contradictions which will inevitably arise during the development of large ontologies .Deductive reasoning fails catastrophically when faced with inconsistency, because "anything follows from a contradiction“. For example, news might be about the Olympic events which can neither relate to a single sport or single country.

5. Deceit: This is when the producer of the information is intentionally misleading the consumer of the information. There can be spoofed news which can be misleading.

VI. GENERATION OF INSTANCE DATA

Data in CSV format is then converted into RDF triple format. This is performed by the use of OpenRefine (formerly called Google Refine). OpenRefine provides easy handling of large data sets and its cleaning. It also provides many data transformation functionalities. Using the RDF Extension 0.8.0 for exporting as RDF, raw data in the form of CSV files or spreadsheets are uploaded to this offline tool to transform the data in RDF/XML format. Some data transformations are then performed to get only the desired data from the csv files in the RDF files. After obtaining the data in the desired form, we design the following skeleton using "Edit RDF Skeleton..." command available under the RDF menu.

- Set the base URI for the RDF file

- *Add Prefix* option is used to add the ontology prefix and upload the ontology from the local machine

- Match the properties of the added ontology to the columns of the tabular data

RDF data is then extracted as RDF/XML from the Export menu. This way the RDF instance data are generated.

VII. QUERYING THE LINKED DATA

The next step of the process is to query the aggregated RDF data. Like SQL queries the tables of a relational database, SPARQL query language is used in the Semantic data model that can query RDF triples. SPARQL is a W3C standard. Pattern matching is the basic idea behind the working of SPARQL. Similar to SQL, SPARQL chooses data from the RDF query data set by the use of a SELECT statement to figure out which subset of the selected data is returned. Additionally, SPARQL utilizes a WHERE clause to define graph pattern to discover a match for in the query data set. A graph pattern in a SPARQL WHERE clause consists of the subject, predicate and object triple to find a match for in the data. Following is one of the queries applied in the project.

```
PREFIX
sports:<http://www.semanticweb.org/newsanalytics/ontologies/2016/>
 ?snippet ?headline ?sport_name
WHERE {
  ?id
<http://www.semanticweb.org/newsanalytics/ontologies/2016/
/has_country> "USA".
  ?id
<http://www.semanticweb.org/newsanalytics/ontologies/2016/
/snippet> ?snippet.
  ?id
<http://www.semanticweb.org/newsanalytics/ontologies/2016/
/name> ?headline.
  ?id
<http://www.semanticweb.org/newsanalytics/ontologies/2016/
/name> ?sport_name.
}
LIMIT 1
```

The SELECT statement requests the three variables to be returned namely ?snippet, ?headline, and ?sport_name. In SPARQL query language, variable names are prefixed with the question mark ("?") symbol. In the above SPARQL query, ?snippet returns the snippet of the news, ?headline returns the headline of the news, and ?sport_name returns the name of the sports which match the four search patterns given in the query. SPARQL queries are executed using Apache Jena Fuseki server. We ran Fuseki from inside a Java program to provide SPARQL services for the application. The advantage of using Fuseki server is that it provides REST-style SPARQL Query, and SPARQL Update using the SPARQL protocol over HTTP.

VIII. KEY FUNCTIONALITIES PROVIDED BY THE APPLICATION

This app provides news for sports enthusiasts covering all its dimensions ranging from the top trending sports news to the upcoming events about that particular sport is rendered with description of the events also included. All this information is provided on a single platform. News above is coupled with tweets by the sports personalities associated with the sports news. The above service takes in country as a sole filtering criteria and provides comprehensive news.

ACKNOWLEDGMENT

We would like to show our warm thank to Dr. Srividya Bansal who supported me at every bit and without whom it was impossible to accomplish the end task.

REFERENCES

[1] Berners-Lee, Tim; James Hendler; Ora Lassila (May 17, 2001). "The Semantic Web". Scientific American Magazine. Retrieved March 26, 2008.

[2] Optimized Index Structures for Querying RDF from the Web Andreas Harth, Stefan Decker, 3rd Latin American Web Congress, Buenos Aires, Argentina, October 31 to November 2, 2005, pp. 71–80.

[3] "OWL 2 Web Ontology Language Document Overview". W3C. 2009-10-27.

[4] "The State of rNews — One Year after its Release" - presentation by Stuart Myles, Evan Sandhaus and Andreas Gebhard, 5 June 2012, Semantic Technology & Business Conference San Francisco

[5] "7 ideas for rNews" - presentation by Stuart Myles, 21 April 2011, Lotico New York Semantic Web Meetup