CSCI 586

DATABASE INTEROPERABILITY

PROJECT MID TERM REPORT

Group 1: Integrated Ontology for Sports

Team:

- Abhishek Agrawal
- George Sam
- Hari Haran Venugopal
- Noopur Joshi

Project Aim:

To create a Federated Ontology on sports primarily focusing on the following sports: Tennis, Football (Soccer), and Cricket. Each domain having diverse classes ranging from history, rules, leagues, player/team ranking and news depending on respective sports.

Our proposed project (application) aim at providing a brief information on player background, tournament details (schedule, location, prize money etc.)

Currently there are no specific data sets available for the different domains in sports. The data is scattered and consists of different fields. For example, tennis is an individual based sport and football and cricket are team based sports. The data sets will have different attributes, which need to be federated. We intend to collect data by scrapping websites and extracting information about some specific leagues and tournaments. For football we want to focus on the following leagues to get the data.

- English Premier League
- Spanish La Liga

For tennis there is no data sets available with all the current players and their information. We will scrape the ATP website to fetch information about the players and the tournaments. We intend to collect the data for all the grand slams

We also wish to scrape data for cricket players and cricket teams. This will give us player information for the cricket teams and the match statistics for the sports.

Approach Used to solve the problem:

1. Scrape websites to generate JSON and CSV files of datasets.
2. Generate Ontologies for the individual sports.
3. Create a federated ontology for tennis, football and cricket, which will cover information for all the sports using Protégé.

4. The ontology will then be loaded into a tool like Karma to model the data into the desired format.
5. Using Google OpenRefine, we will be performing data cleaning on the data sets. Data cleaning will involve filtering unwanted data attributes from the data sets, initializing the blank data values to null values, and creating filters to extract only relevant rows of data.
6. We will create RDF graphs after modelling the data sets according to properties and predicates in the ontology and using dcterms and schema.org properties.
7. Once a meta-ontology of the data sets is created we will convert it into RDF triples.
8. The rdf triples will then be loaded into an RDF triple store. We intend to create our own triple store. The triple store will be created using Apache Jena and Sesame OpenRDF libraries. We intend to create a federated persistent triple store and host it so that it will be available online.
9. Once the triple store is created, we will then use SPARQL queries to query the triple store to extract the required information.
10. The SPARQL queries will consist of queries of the following kind:
    a. To find the ranks of the winners of the major tennis grand slams since 2004
    b. To find the nationality, age, etc of the top 10 ranked players in football, tennis and cricket.
    c. Find which teams/ player won specific tournaments with the location and score of the match
    d. Find which football stadium hosted the maximum matches for a given year
    e. Find coaches' information for cricket and football

Tools Used:

- Libraries:
  JSoup, JSON, Request, BeautifulSoup, Sesame OpenRDF, Apache Jena, Scrapy, WebScraper.
- Karma
- Google OpenRefine
- Protege

Individual Contribution:

Abhishek Agrawal:

Scrapped websites (http://www.espncricinfo.com/) and (http://www.icc-cricket.com/) and (http://cricsheet.org/) using a combination of Scrapy python library and Chrome Web scrapper. Dataset includes information about Players, Teams and Tournament (T20,Test,ODI).

George Sam:

Responsible for creating web-scrapers to scrape websites and fetch player and team details. Libraries used for scraping are JSoup. Scraped the following websites to get the data sets. Scraper created using Java.

http://www.footballsquads.co.uk/

http://www.soccerbase.com/

http://openfootball.github.io/

Also responsible for creating the ontology for football players and for football teams. Worked towards merging different data sets into single data sets using Google OpenRefine, and also converting different CSV and JSON files into uniform JSON format.

Noopur Joshi:

Responsible for creating ontologies for each sport and the federated ontology for all sports combined on Protégé. The approach being used is the Hybrid approach for creating the ontologies. Initially an ontology for each individual sport namely Football, Tennis and Cricket is created. These ontologies will then be mapped to the federated ontology for all sports. Will be handling Data cleaning for the datasets using Google OpenRefine. The process will include removing attributes that are not required based on the ontology, removing null values. The paper referred for creating ontologies:

http://130.88.198.11/tutorials/protegeowltutorial/resources/ProtegeOWLTutorialP4_v1_3.pdf

Hari Haran Venugopal:

Built a python script to scrap websites, collected relevant dataset in JSON format for domain tennis, primarily focusing on players details: Name, Bio-Data, Ranking, Personal History, Coach, Age, Nationality and tournament details considering all grand slams: year, winner, scores.

Website: http://www.atpworldtour.com.
Python Library: requests, beautiful soap 4 and json.
Language: Python.

Project Repository: https://github.com/hari316/web_semantics

Data Modelling in Karma:
This task will be performed by all members.
Creation of Triple Store:
This will be handled by Noopur Joshi and Hariharan Venugopal.
Queries:
Query creation will be handled by Abhishek Agrawal and George Sam.