****

**Assignment Part 2**

**Intelligent Web (COM6504)**

**Course: MSc (ENG) Advanced Software Engineering**

**Prepared by:**

|  |
| --- |
| **Jianyue Ni** |
| **Iman Rastkhadivmasouleh** |
| **Amirah Syazana Binti Ibrahim** |

**Submitted to:**

**Professor Fabio Ciravegna**

Table of Contents

[1.0 Introduction 1](#_Toc388785338)

[2.0 Task Specification 1](#_Toc388785339)

[2.1 Querying the social Web 1](#_Toc388785342)

[2.2 Design an Ontology and Triple Store 4](#_Toc388785343)

[2.3 Web Interface Implementation 5](#_Toc388785344)

[2.4 Quality of the Solution 5](#_Toc388785345)

[2.5 Additional Features 5](#_Toc388785346)

[3.0 Conclusion 6](#_Toc388785347)

[4.0 Extra Information 7](#_Toc388785348)

[5.0 Bibliography 7](#_Toc388785349)

**List of Abbreviations**

|  |  |
| --- | --- |
| **API** | [Application programming interface](http://en.wikipedia.org/wiki/Application_programming_interface) |
| **CSS** | Cascading Style Sheets |
| **HTML** | Hyper Text Markup Language |
| **JSP** | Java Servlet Pages |
| **JSTL** | JavaServer Pages Standard Tab Library |
| **MVC** | Model View Controller |
| **RDF** | Resource Description Framework |
| **URL** | Uniform Resource Locator |
| **URI** | Uniform Resource Identifier |

# Introduction

The objectives of this assignment is to query and accumulate the information of others people from public domain. There are two kinds of social network that using in this assignment which are Twitter and Foursquare. Twitter is a social networking that enable user to send and read text message where there are limited character number of text to display. While Foursquare is a social network that manages on location based where user can check-in in specific place that they are visited. The API for twitter and foursquare used that can help in writing the programming code for that application. Twitter (APIs) allow real-time access huge amounts of content. Furthermore, Twitter provides data streamed through the APIs include metadata about the authors, including the text location from user profile.

In the second stage of this assignment requirement is design and implement the ontology and triple store. Ontology can describe as concepts of the domain that has a relationship or related in a hierarchy with classes of object. The purpose of ontology in the semantic web is to organize information using structural frameworks and use RDF schema to describe ontology. RDF is a standard data model for machine-processable semantics. The RDF also is a language that describes the statement about objects (resources).

Most of the tasks on the assignment are performed according to specification by each category. All scenario questions for section 1.1 are completed perform with the required output. The additional scenarios parts for section 1.5 are completely performed based on the requirement provide. In addition, each scenario is able to be browse and perform queries by user in a web interface. The design of ontology and quality of solution also has been implementing in this assignment.

# Task Specification

This section describes the task for each section includes the issue and design choice of each requirement. The design choice of the project solution also describe in this section. In order to justify the designing choices, the requirements of the projects are required to be described in details in this part.



## Querying the social Web

**1. Tracking discussions:**

* **Issues**

This part refers to people discussion on a specific topic on twitter. User should be able to insert any keyword or hashtag topic and give the output from everyone that mentions the specific topic.

The challenge that faces out for this task is regarding in obtaining retweet message from up to 10 users.

* **Design Choices**

MVC design pattern is used to develop most part of this application:

● Controller: Contain the servlet file to link to the web page.

● View: Contain JSP file as the web site view page.

● Model: Data model represents an object or and carrying data.

The default location for searching keywords or hashtags is set to Sheffield. User is able to unchecked the checkbox and remove the default location. For separating retweeted text from the original tweet text, two different list of class Person is created. Person represent tweeter user and contains their properties.

The two lists of people are created when a user queries a keyword. One is refer to retweeted people. This list is set as a session attribute and is shown on the retweet.jsp page. Each tweet messages show the number of retweet. If it is greater than 0 then the user is able to click on the number and view the list of retweeted people. The Twitter API and tweet stream are used to query this part.

* **Requirement**

The design of this part is meet the assignment requirement where user required entering the input keyword, hashtag, longitude, latitude, and radius. The display outputs in the list of table are image, name, username, tweet text of the each tweeter user and the retweet message.

**2. Queries about specific users:**

a) What specific users discuss about:

* **Issue**

Twitter is the platform that used in order to queries about frequent keyword that used by users in specific days. The challenge is that to find most frequent keywords used by up to 10 users. We save all the keywords and their occurrence into hash map. Since we had to use java 6, sorting the map by its value was challenging. Iterating throw the map in the view page for each user was also challenging

* **Design Choices**

Each user has a map which stores keywords as a key and the number of each keyword as a value. And there is a map for sorting all the keywords and their number which different users has tweeted. In our design services are used in order to follow single responsibility design patterns. For example stopword service is only responsible to remove stop words from a given tweet. To design this past required the Twitter API.

* **Requirement**

This query part is meeting the assignment requirement where when user enters the User ID, Number of keyword and Number of date the output are display in the list of table contain keywords and total counter of keyword for each twitter user.

* **Limitations**

The provided code to meet this requirement is completely extensible. All this classes are used for other parts of the assignment. For example for Mixed query we reused this classes to get the most frequent keywords in a specific geographic point.

b) What points of interest (venues) a specific user has visited in the last *X* days:

* **Issue**

In this part User has to input User Id and the days to display output the place that user have been visited on that particular days (and start streaming if the day is 0). How to get the venue from tweets is one of the major issues. One challenge is to retrieve the output from streaming API and stop it properly. Another challenge is that the twitter api only provides information for less than 7 days. Things like how to get page refreshed partly in certain frequency is also an issue.

* **Design Choice**

The design is by using filter the result tweets and return only the tweets with expandedURL from Foursquare. Click on the short URL link, it would redirect to another page to display the user information of foursquare and the information of that specific venue. Once click the search button (with or without the key word of nearby interests) there would be a marker on google map and include some information about the foursquare venue on it. If user make a search use the proper key word (e.g. hotel), the map would also show the nearby intents around the position of the same foursquare venue including the star symbol. Once the star is clicked it would show the information of that interest from google map with the format of info window. JSON has been choosing as the response from the servlet. The other requirement is, use jQuery to check the validation of input and Ajax to display the dynamic output from servlet. The page would start a timmer to constantly request the servlet for result when it comes to streaming. In this part, the streaming result is stored in a list of venue models in the servlet. Therefore, the result list would be available when the page automatically request for it. In addition, the valid range days are design between the ranges of 0 to 7 day and if the day is 0 the system has been set to start streaming. We save all of the result into the RDF file, and use it to limit the search of twitter. (we search for the tweet in RDF first so that we don’t need to query twitter again for the same tweets)

* **Requirement**

After input the username and click on the shortURL link, it would go another page and display the information about venues both on list and google map. With the shortURL and the key word of interest, there would be a list of venues from google shown on the map with the marker of star. Besides, the information of the venue inside the map info window is included as well. We can display the venue through url and a map with a marker to show the point of that venue with a list of details.

* **Limitations**

The result from twitter is limited to tweets with foursquare short urls. The ideal way to do that is get all tweets with geolocation and foursquare short urls and send both to google map. Issues like twitter recently change the query method which would cause queries like “from:user” filter part of the result from foursquare( check my account njy0612 you would only get one result but actually there are a few should be available). Another limitation is that twitter only gives back records within 7 days. Ideally, we should go to twitter authorized website to get the former data, but considering the time limit, we just put a validation for that.

c) Who is visiting venues in a specific geographic area in the last *X* days:

* **Issue**

The main issue in this part is in order to search from twitter with location. This is because the twitter4j API only support search with geolocation and not near. Another issue is to filter the same user with multi tweets from the same area. Moreover, the streaming also becomes a problem.

* **Design Choice**

In order to get the most correct result of a venue name input, thus google map API through Javascript to get the accurate coordinates has been decided to be. To deal with the same user result, the design of add filter has been created. We also save all of the result into the RDF file, and use it to limit the search of twitter.(we search for the tweet in RDF first so that we don’t need to query twitter again for the same tweets)

* **Requirement & limitaion**

By using google map javascript api, we are able to get the most accurate coordinates of that place. And with that coordinates, we are able to fetch all the tweet with that geolocation. We can get all the correct result and because we save results to RDF, it saves time and band width to get the existing results directly from RDF file.

Again, the twitter api has a time limitation of 7 days, and we don’t have enough time to deal with that.

## Design an Ontology and Triple Store

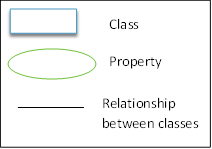
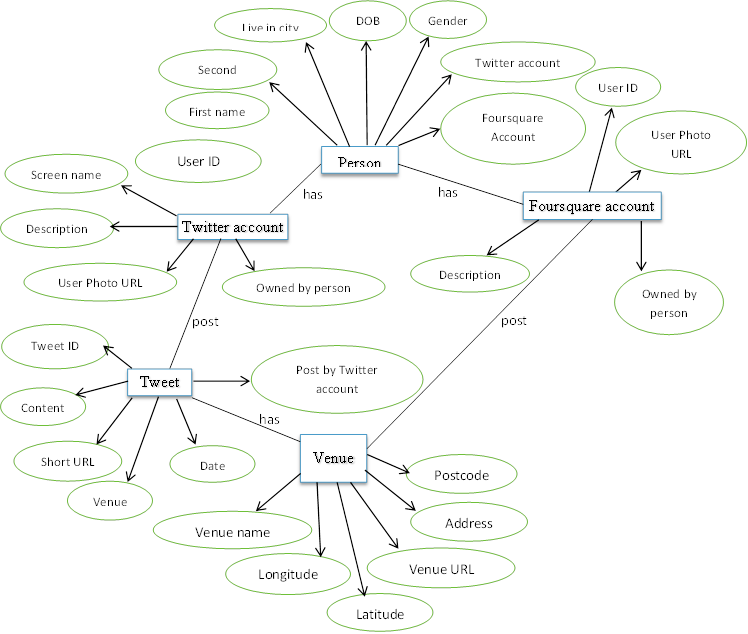
This section is replacing the first part of assignment which is storing information in the database. The second stage of assignment, there will be no longer use MySQL database and make an amendment to store information in a Jena RDF triple store.

* **Issue**

We don’t understand what is RDF and RDFs at first. And how to make RDF more useful and reusable is very difficult.

* **Design Choices**

The design of ontology contains 5 resources or class and each resource has a Uniform Resource Identifier (URI). The RDF also has properties which is part of special kind of resources and can describe the relation between resources. The following diagram is the design of the ontology. We put person to have two accounts so that the ontology would actually mean something( person has these two accounts). We thought about using account and make twitter account and foursquare account inherit from account, but it might make the insert too complicated and it is not necessary. And we also move venue out of tweet so that we don’t have to save the same venue again because it was tweeted by another tweet or check in. We have a RDfs file to save structure of our ontology.



**Figure 1.0 : The Design of Ontology**

The information is now stored in the RDF file. Use Apache Jena as the Java framework to build the semantic web and make a link data between applications. The other design choice that required in this assignment use SPARQL to query language for the data in RDF.

## Web Interface Implementation

The interface have been organise by diving in three main parts which is body file, header file and footer file. The header and footer file part will invoke in the content of interface. This task involves the implementation of the interface as a servlet and also the design using CSS, HTML, Javascript and JSP and JSTL technologies. The program is fully accessible through any web browser.

## Quality of the Solution

**i. Javascript**

Javascript and Jquery are used in all of the forms in order to validate user inputs. If user inserted an invalid data a message will be pop up to inform the user.

All the results are displayed below the forms. This is achieved by using javascript and jquery in different views. We separated header page from the main page and the link to javascritp files are placed just in the header file. In order to make the most of jquery hidden fields are used for getting data in the javascript files.

**ii. Ajax and JSON**

For retrieving data as efficient as possible from servlets and our models we make use of ajax and json. This enables the user to query as much as it is possible without worring about data loading

## Additional Features

The additional features that have been add is by adding another two queries about venues and mixed queries.

**3. Queries about venues:**

a) What points of interests (i.e. venues) are in a specific geographic area (input: name of location or coordinates)?

b) What other venues are around a specified venue (input: id of venue).

**• Issue**

How to get the information of venues around venues are a big issue

**• Design Choice**

We combine this along with get venue from short url. We use the short url to get the central marker and the search point. We ask user for an input of interest (like hotels) and get all the information (pic, description, link etc.) and display it with a star on the map which user can click and see the venue information. In some way, it fufills both of the query and display it very nicely in google map.

**• Limitations**

If there is any change on google map, the entire thing would collapse, but it is the risk of all public APIs.

**4. Mixed queries:**

Popular discussions in a geographical area can also be viewed. The Twitter search API was used in this situation. Like section 2 all the keywords are stored as a key in a map and the number of that keyword is stored as value in the map.

**● Limitations**

The mix queries is largely depend on twitter query solutions, so once twitter decided to change it , the result might not be correct.

# Conclusion

A clear description has been explained on the task specification and this assignment exposed a lot of new technology that are currently used. In this assignment are not simply queries to display output but also emphasis and require the quality of the solution that used Javascript, Ajax and Json.

The second part of assignment change the database and store information in the RDF which is give knowledge in linked data. The understanding of linked data which is about creating global database of linked things. The URI is required to build a linked data technologies because URI provide a simple and extensible to identify a resource as well as no only use in web pages.

RDF is flexible and schema-less because it is not depend only to a single schema. RDF allows using different properties and classes from different ontologies. Database schema and ontology can be distinguish based on the way of data is organised. The data in the database is organised by tables while ontology used to represent information and purpose about data in order to gather new information and check consistency. The use of semantic web can improve the current state of the World Wide Web. It is because ontologies are useful for improving the accuracy of web searches. The failure to find or query any relevant document, can be suggested to the more general query where the search engine can look for other pages that has a relation between concept in an ontology.

Furthermore, this assignment gives an understanding on the importance of technology that able to influence on ethical behaviour. Nowadays, as the Internet grows to meet the needs of billions of users and the increasing of sophisticated technology, the right and policy to keep the information of user should be taken into consideration which not disclose arbitrary of user information.

# Extra Information

There are no extra configuration has been used in the code.

# Bibliography

1. Fabio Ciravegna, Lecture Notes Intelligent Web Week 2 – 9, at The University of Sheffield.
2. Boostrap

*http://getbootstrap.com/*

1. Grigoris Antoniou and Frank van Harmelen (2008) A Semantic Web Primer, 2nd Edition, England: The Massachusetts Institute of Technology (MIT) Press
2. Apache Jena

[*http://jena.apache.org/documentation/javadoc/*](http://jena.apache.org/documentation/javadoc/)