Initial Plan: Project 133 Information Extraction From Webpages.

Author: Ashley Sean James

Project Supervisor: Andrew Jones

Module Number: CM3203

Module Title: Final Year Project.

Number Of Credits: 40 credits

Table of Contents

Project Description

Information extraction from webpages at its most basic is retrieving useful information from suitable webpages for the purpose of analysis. Additionally for my project, I will also be finding the inner meaning of the information which is stored on such webpages. The problem to be overcome is finding this meaning programmatically and it is something which is hard to do. Particularly, making use of retrieved information to infer new information correctly is a challenging task and one which I will hope to overcome.

Extracting information depending on different styles of webpage design is going to be an interesting part of this project. Furthermore, the nature of the internet and associated webpages is one which is always evolving and the content within is always changing. Adapting to different page designs and mark-up languages and being able to gather information where no formal structure exists will be difficult.

I will be retrieving information from a variety of web page sources to get accurate and relevant information. The reason for this approach will be to check that information is actually correct. The wrong information will produce the wrong inferred information and so it is very important to eliminate this problem through checking for commonality between the pages used for information build up.

The project will be programmed in a modular way using the Java programming language. Java offers portability and also modularity which I believe are key features for this project. The information extraction process will be performed by a Java function which returns the necessary information about classes. This will be achieved by examining the mark-up (e.g. HTML, XML) of the page in question to construct a function which extracts necessary information from the different pages.

I propose to work with the theme of animal classifications based on their attributes to determine which family (“Birds”, “Reptiles”, “Amphibians”, “Mammals” etc.) a specific animal belongs to. I chose Animal classification because it is an area which still requires additional research to understand why Animals do belong to their respective classification families.

For example, given a set of properties to define a Reptile class and a “Lizard” class found through information extraction from “Reptile” and “Lizard” webpages. The program would be able to infer that a “Lizard” is part of the “Reptile” class. Perhaps I can also look for relationships between Dinosaurs and how they differ to 20th century Animals if time allows.

Once I have created these relationships between attributes I will be able to test this further by querying the model to find out how accurate my inferences are. This may even gleam new facts such as how a “Bird” differs from a “Lizard” in terms of classification. Furthermore, I would hope to be able to create a hierarchical “food chain” model of which animal eats which other animal.

Project Aims and Objectives

I plan to achieve the following objectives at the end of this project:

- Extract information from a variety of webpages
    - This will ensure accurate and relevant information retrieval.

- Retrieve the information efficiently from inside a well-documented programming language.
    - This will produce modular, well-structured and maintainable code.

- Create an ontology of classes and relationships between class objects.
    - Building this model should give an accurate model for classifying objects which belong in the hierarchy.
    - I think creating an ontology for the hierarchy of the animal kingdom is a good choice for this. It offers interesting relationships but also can be well tested.

- Infer whether one class belongs to another.
    - Applying a reasoner such as “Pellet” should provide the inferencing capabilities.
    - E.g I would hope to be able to Infer that a “Lizard” is a “Reptile”.

- Query the inferred model to find new relationships about different classes.
    - I think that SPARQL is the best choice for this because it is designed for querying Semantic Web data.

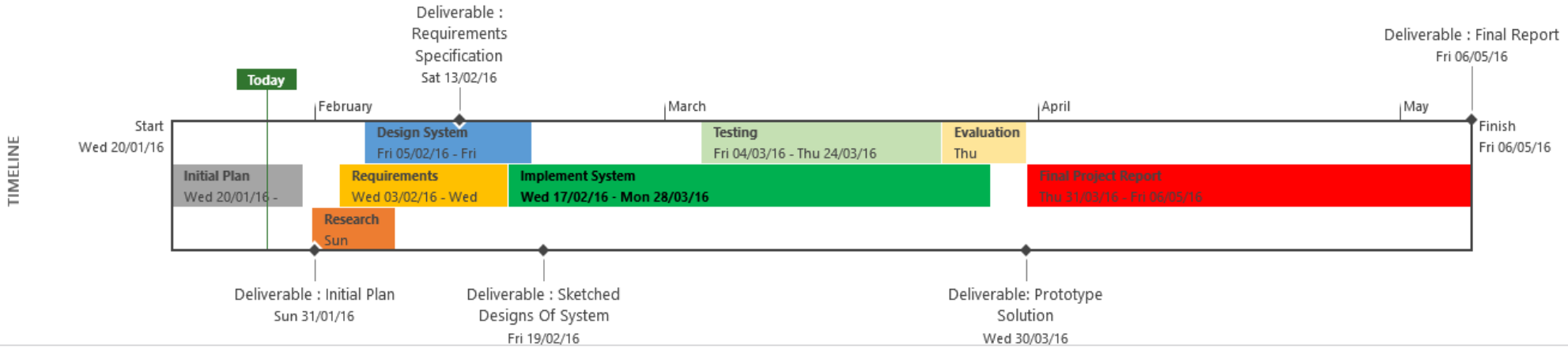
    - E.g A “Bird” classification would typically be any animal which flies and lives in a nest. Although with all things there are some exceptions which will need to be accounted for.

If I have time:

- Build up a hierarchy of related animal relationships to infer which animal eats which other animal based on the one animal “eats” another animal relationship

Overall, my aim is to develop a system that can work out the relationships between species from online information gathered from inside a well-structured programming language.

Work Plan



GANTT CHART

	Task Mode	Task Name	Duration	Start	Finish	Add New Column
0		**Gantt Chart**	**2579 hrs**	**Wed 20/01/16**	**Fri 06/05/16**	
1		**Initial Plan**	**11 days**	**Wed 20/01/16**	**Sun 31/01/16**	
2		**Deliverable : Initial Plan**	**11.46 days**	**Wed 20/01/16**	**Sun 31/01/16**	
7		**Research Into Project**	**7 days**	**Sun 31/01/16**	**Sun 07/02/16**	
8		2nd Meeting with Dr Jones : discussing research.	1 hr	Sun 31/01/16	Sun 31/01/16	
9		Investigate Information Extraction Techniques	4 days	Sun 31/01/16	Thu 04/02/16	
10		Investigate Semantic Web Programming and JENA	2 days	Tue 02/02/16	Thu 04/02/16	
11		**Requirements Specification**	**14 days**	**Wed 03/02/16**	**Wed 17/02/16**	
17		**Design System**	**14 days**	**Fri 05/02/16**	**Fri 19/02/16**	
22		**Implement System**	**40 days**	**Wed 17/02/16**	**Mon 28/03/16**	
28		**Testing**	**20 days**	**Fri 04/03/16**	**Thu 24/03/16**	
30		**Evaluation**	**7 days**	**Thu 24/03/16**	**Thu 31/03/16**	
32		**Final Project Report**	**36.46 days**	**Thu 31/03/16**	**Fri 06/05/16**	

Further work plan timeline explanations can be found below.

Risk mitigation decisions.

The following decisions were made to mitigate the risks of this project due to the short timescales available for the project:

- Conducting research alongside requirements specification because it is important to understand what can actually be achieved in the project.

- Working on the requirements and design stages together, I believe gives me the best ability to ensure the system I have designed will be both suitable but also achievable.

- I will design the solution and begin implementation as soon as possible. To ensure that I have a prototype solution to add more complex functions to, this will hopefully mean the final solution matches the main key requirements.

- Testing as I go will hopefully prevent errors being introduced, but will also identify areas for improvement.

- Evaluating the system will provide me with the ability of reflection on the experience and the problems which occurred. Mitigating risk for similar projects in the future.

- I have given myself a lot of time to complete the final report because this is the most important deliverable overall.

Milestones to limit risks further.

Further risks at each stage include:

- Having undefined requirements
    - Requirements clearly specifying what is possible, what could be implemented and what is not likely will help to overcome this.

- Having a solution which does not fit what the supervisor wants to achieve.
    - Showing the designed solution regularly to the supervisor and getting feedback which can be acted on will produce a solution which is more suitables.

- Having a good implementation model.
    - This stems from good research, design of algorithms and the programming approach used, if all of these are documented then the solution will be less risky.

- Having a robust solution taking into account different inputs and handling these appropriately.
    - Performing effective testing will limit this risk.

All of the mentioned risks will be mitigated by being deliverables for the project.

Appendix

Detailed Gantt Chart Items

Task Name	Duration	Start	Finish
Gantt Chart	**2579 hrs**	**Wed 20/01/16**	**Fri 06/05/16**
Initial Plan	**11 days**	**Wed 20/01/16**	**Sun 31/01/16**
Deliverable : Initial Plan	**11.46 days**	**Wed 20/01/16**	**Sun 31/01/16**
Research Into Project	**7 days**	**Sun 31/01/16**	**Sun 07/02/16**
2nd Meeting with Dr Jones : discussing research.	1 hr	Sun 31/01/16	Sun 31/01/16
Investigate Information Extraction Techniques	4 days	Sun 31/01/16	Thu 04/02/16
Investigate Semantic Web Programming and JENA	2 days	Tue 02/02/16	Thu 04/02/16
Requirements Specification	**14 days**	**Wed 03/02/16**	**Wed 17/02/16**
Deliverable : Requirements Specification	**11 days**	**Tue 02/02/16**	**Sat 13/02/16**
3rd Meeting With Dr.Jones : Discuss What He Wants From System	1 hr	Thu 04/02/16	Thu 04/02/16
Initial Requirements Specification	3 days	Thu 04/02/16	Sun 07/02/16
Feedback on Requirements	2 hrs	Sun 07/02/16	Sun 07/02/16
Final Requirements	6 days	Sun 07/02/16	Wed 13/02/16
Design System	**14 days**	**Fri 05/02/16**	**Fri 19/02/16**
Implement System	**40 days**	**Wed 17/02/16**	**Mon 28/03/16**
Testing	**20 days**	**Fri 04/03/16**	**Thu 24/03/16**
Evaluation	**7 days**	**Thu 24/03/16**	**Thu 31/03/16**
Final Project Report	**36.46 days**	**Thu 31/03/16**	**Fri 06/05/16**