12/11/2014

CIS 4911 Senior Project

Intelligence Inference Engine
Feasibility Study and Project Plan

Group Members
Jose Acosta
Lazaro Herrera

Mentor
Eric Kobrin

Instructor
Masoud Sadjadi

Intelligence Inference Engine : iie-dev.cs.fiu.edu

Abstract

Most web searches currently depend on matching keyword or phrases in order to return results. Matching keyword and phrases works as long as you know what your are searching for and you only want information to that specific thing. In the modern world, when it comes to cyber security new exploits and attacks are constantly coming out. It is not feasible to keep track and know everything about these attacks because there are so many. Fortunately, most new attacks and exploits are usually related to old attacks and exploits. This is where the Semantic Web comes in. Data in Semantic Web is represented by relations, this makes it easy to search for things even without knowing exactly what you are looking for. The only problem with the Semantic Web is that there is a barrier of entry, knowing a specific querying language. This project aims to lower that barrier of entry.

Table of Contents

Introduction

Problem definition

Humans are capable of using the Web to carry out tasks such as finding the German translation for "eight days", reserving a library book, and searching for the lowest price for a DVD. However, machines cannot accomplish all of these tasks without human direction because web pages are designed to be read by people, not machines. The semantic web is a vision of information that can be readily interpreted by machines, so machines can perform more of the tedious work involved in finding, combining, and acting upon information on the web.

Unfortunately the Semantic Web has a problem, it is designed so that it easily read by machines meaning that it is hard for a human user to directly read the data that is on the Semantic Web. This means that someone must query the data to find what they are looking for. This is a problem because it requires the person to know one of the Resource Description Framework (RDF) querying languages.

Background

The Semantic Web is about linked data. Linked Data is resource-based linking of information. The Semantic web is built of by large of linked data which is defined by the Resource Description Framework (RDF). Each RDF data point consist of three parts: a subject, a predicate, and an object. In essence, RDF give you little building blocks of data that can be connected to other building blocks of data in both directions. When you build complicated webs of connected information, you end up with really specific detailed structures of conceptual knowledge over which you can answer complex question programmatically.

Right now, if you went to Google and put a search query like “Everything related to exploits on SSL that affects Linux and similar systems ”, the results would mostly be articles that cover SSL, some that may cover Linux, and maybe some that cover exploits on SSL. That search result would be useless because it does not give you what you asked for, instead it just matched the keywords you put in, Google does not know what systems are similar to Linux by keywords alone, it does not know what how to put all that together to find specifically what your are looking for . This means that you have to know what you are looking for before you look for it. With Semantic Web you can use relations between different data points and tries to answer the query using relations.

Definitions, Acronyms, and Abbreviations

OSINT - Open-source intelligence

Cyber-attack - Any type of offensive maneuver employed by individuals or whole organizations that targets computer information systems, infrastructures, computer networks, and/or personal computer devices by various means of malicious acts usually originating from an anonymous source that either steals, alters, or destroys a specified target by hacking into a susceptible system [1]

Triple store - A triple-store is a purpose-built database for the storage and retrieval of triples through semantic queries. A triple is a data entity composed of subject-predicate-object. [2]

RDF - 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 notations and data serialization formats. [3]

Semantic Web - collaborative movement led by international standards body the World Wide Web Consortium (W3C). The standard promotes common data formats on the World Wide Web. By encouraging the inclusion of semantic content in web pages, the Semantic Web aims at converting the current web, dominated by unstructured and semi-structured documents into a "web of data". The Semantic Web stack builds on the W3C's Resource Description Framework (RDF). [4]

SPARQL - an RDF query language, that is, a semantic query language for databases, able to retrieve and manipulate data stored in Resource Description Framework format. [5]

Overview of document

This document introduces the problem that the system will be tyring to solve and gives background information on why this problem is something that should be tackled. This document also covers the basic definitions and abbreviations in field relating to the problem. Within this document, The current way that the problem is tackled is described and also there will also be a description of the new system that will be implemented in order to combat the problem. The high level user requirements that specify what the system must be able to do are stated. There is also some investigation in alternative solutions and implementations to this problem and those alternatives are are compared to the suggested solution and analyzed. Finally, this document provides what the necessary hardware and software is and describes the team structure. A project plan is also included which describes that has been accomplished and what work still needs to be done.

Feasibility Study

Description of Current System

Currently, there exist many websites that catalog cyber-attacks, data breaches and vulnerabilities. Although all of these websites are critical to today's security researchers, missing any of them can leave important data behind as well as not knowing the correct term to search for can cause some of the important information to be left behind. Consider the possibility that an attack may be filed as targeting "banks" yet the security researchers might be searching for an attack that is targeting "financial institutions". Similar terms being used but a standard search engine does not take these into account.

Purpose of New System

The purpose behind the new system that will be created is to allow users to be able to easily share information with the public while also being able to access said information about OSINT data as well as internal cyber-attack data using semantic web techniques. The front end of the system will be a web interface to facilitate the access and use of the system. From the front end, the user will be able to easily submit data about recent cyber-attacks or possible cyber-attacks into the system through an easy to us web form. The data that users provide will be stored in a triple store. Once the data is being stored, users will be able to query the data that has been collected to see any relevant data to that user. Also, the system will be able to make inferences about future data from current data that is being stored. The system should provide storing for intelligence-critical metadata such as assertion provenance and confidence of assertions.

High-level Definition of User Requirements

- A user must be able to submit data to be stored through some type of web form
- The data should be query-able directly using Sparql or Datalog
- A user must be able to set up predefined queries which are accessible by other users
- The system must be able to graphically display all the data that it currently stores.
- The system must be able to import data given a file with a specific format.
- The system must be able to export data to a file in a specific format.
- The data must be stored in one of the existing semantic web triple- or quad-stores such as Mulgara or Jena
- The system must be able to set a confidence interval for the data that is being collected.

Alternative Solutions

Description of Alternatives

An alternative to the proposed solution would be to change the entry and query of data from a web form to an independent client that a user would install. The system could also use a relational database rather than a triple store to store all the data the system will receive.

Selection Criteria

1) Ease of use
2) Ease of development
3) Extensibility (support newer technology)
4) Reliability
5) Continued Support

Analysis of Alternatives

For our project we will be using a bootstrap-based web form built on Jena. A web form is being used for easy user access from both desktops and mobile devices instead of providing a native client. Jena's triple-store along with it's REST interface allows us to perform queries directly from the web form in multiple languages (Sparql/Tql).

Jena was chosen over sqlLite for the inference engine mainly because of its ease of development. Jena is significantly more efficient for what this system needs to do as opposed to using sqlLite. A bootstrap web form was chosen because it is much easier to develop a single web client that anyone with a web browser can access as opposed to native clients for each operating system. Also with a web client there is no installation so its a little easier to use than a web client.

Recommendations

Our final recommendation is that we utilize Jena with a jQuery/Bootstrap powered UI to allow for both desktop and mobile use of our system through an easy-to-use web form.

Project Plan

Project Organization

Project Personnel Organization

Lazaro Herrera - Developer / UI Design / Testing

Jose Acosta - Developer / Database / Query Library Development

Eric Kobrin - Project Manager / Client

Hardware and Software Resources

Hardware - Single core CPU / 2 GB Ram / 30 GB hard drive

Software - Apache Jena / Easy Dev PHP / Bootstrap / jQuery / MySQL / Selenium / D3JS

Identification of Tasks, Milestones and Deliverables

For our Trello board, we have decided to use color tagging to both identify the assignments of tasks and the completion status of tasks

Colors and Definitions for Priority

Blue - Card created by students

Purple - Card created by client

Colors and Definitions for Completions (only one active at a time)

Green - This task has a completion of 25% or more

Yellow - This task has a completion of 50% or more

Orange - This task has a completion of 75% or more

Red - This task has a completion of 100%

The following epics currently exist on our board.

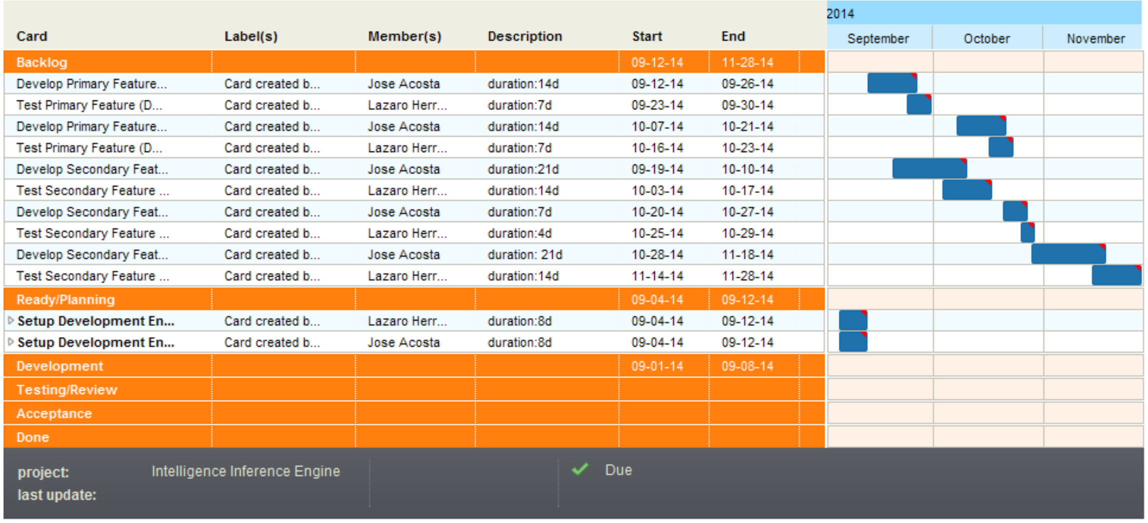
Some of these are receiving checklists to be converted into stories.

- Setup Development Environment (Jena)
- Setup Development Environment (Web Server + Bootstrap)

- Develop Primary Feature (Data Entry Web Form)
- Test Primary Feature (Data Entry Web Form)
- Develop Primary Feature (Data Retrieval Web Form)
- Test Primary Feature (Data Retrieval Web Form)
- Develop Secondary Feature (Database Explorer)
- Test Secondary Feature (Database Explorer)
- Develop Secondary Feature (Custom Reusable Queries)
- Test Secondary Feature (Custom Reusable Queries)
- Develop Secondary Feature (Confidence Ranking)

Appendix

Appendix A - Project schedule



Card	Label(s)	Member(s)	Description	Start	End
Backlog				09-12-14	11-28-14
Develop Primary Feature...	Card created b...	Jose Acosta	duration:14d	09-12-14	09-26-14
Test Primary Feature (D...	Card created b...	Lazaro Herr...	duration:7d	09-23-14	09-30-14
Develop Primary Feature...	Card created b...	Jose Acosta	duration:14d	10-07-14	10-21-14
Test Primary Feature (D...	Card created b...	Lazaro Herr...	duration:7d	10-16-14	10-23-14
Develop Secondary Feat...	Card created b...	Jose Acosta	duration:21d	09-19-14	10-10-14
Test Secondary Feature ...	Card created b...	Lazaro Herr...	duration:14d	10-03-14	10-17-14
Develop Secondary Feat...	Card created b...	Jose Acosta	duration:7d	10-20-14	10-27-14
Test Secondary Feature ...	Card created b...	Lazaro Herr...	duration:4d	10-25-14	10-29-14
Develop Secondary Feat...	Card created b...	Jose Acosta	duration: 21d	10-28-14	11-18-14
Test Secondary Feature ...	Card created b...	Lazaro Herr...	duration:14d	11-14-14	11-28-14
Ready/Planning				09-04-14	09-12-14
▷ **Setup Development En...**	Card created b...	Lazaro Herr...	duration:8d	09-04-14	09-12-14
▷ **Setup Development En...**	Card created b...	Jose Acosta	duration:8d	09-04-14	09-12-14
Development				09-01-14	09-08-14
Testing/Review					
Acceptance					
Done					

project: Intelligence Inference Engine

last update:

✓ Due

Appendix B – Feasibility Matrix

	Jena	SqlLite
Ease of Use	n/a	n/a
Ease of development	10	1
Extensibility	6	8
Reliability	9	9
Continued Support	3	9
	Bootstrap web form	Native Client
Ease of Use	9	8
Ease of development	10	5
Extensibility	5	5

Reliability	7	7
Continued Support	9	9
	Easy Dev PHP	Custom Web server
Ease of Use	n/a	n/a
Ease of development	10	1
Extensibility	4	9
Reliability	9	9
Continued Support	9	10
	Selenium	Manual Testing
Ease of Use	n/a	n/a
Ease of development	8	1
Extensibility	5	9
Reliability	8	4
Continued Support	8	9

Appendix C – Cost Matrix

Cost Matrix	Weeks Required	Jose - $30 hour / 12hr week	Lazaro - $30 hour / 12hr week	Total Cost
Data Entry Form Implementation	2	$720.00	$720.00	$12,672.00
Data Entry Form Testing	1	$360.00	$360.00	
Data Retrieval Form Implementation	2	$720.00	$720.00	
Data Retrieval Form Testing	1	$360.00	$360.00	

Database Explorer Implementation	3	$1,080.00	$1,080.00	
Database Explorer Testing	2	$720.00	$720.00	
Custom Queries Implementation	1	$360.00	$360.00	
Custom Queries Testing	0.6	$216.00	$216.00	
Confidence Ranking Implementation	3	$1,080.00	$1,080.00	
Confidence Ranking Testing	2	$720.00	$720.00	

Appendix D - Diary of Meetings

Date: 9/3/2014

Meeting Begins: 6:03PM

Meeting Ends: 6:45PM

Medium of Communication: Skype

Present Members: Eric, Jose, Lazaro

6:03PM - 6:15PM Eric (Formal Introductions, Intro to Inference Engine)

6:15PM - 6:30PM Use Case Elicitation

6:30PM - 6:33PM Architecture concerns and discussion

6:33PM - 6:33PM Meeting with client formally disbands

6:33PM - 6:45PM Jose and Lazaro discuss required documentation, recap about requirements.

Assignments:

Lazaro and Jose will install Mulgara and attempt to execute basic queries, we'll also be looking at bootstrap-based approach to generating simple queries from a web form.

Lazaro and Jose will also begin writing the required initial drafts for our Feasibility Study, Requirements Document and Project Plan.

Date: 9/6/2014

Meeting Begins: 11:30AM

Meeting Ends: 6:00PM

Medium of Communication: Physical Meeting

Present Members: Jose, Lazaro

11:30AM - 1:30PM: Trello Board Upgrades and Discussion

1:30PM - 4:30PM: Feasibility Study Documentation

4:30PM - 5:30PM: Requirements Documentation

Date: 9/21/2014

Meeting Begin: 10:00 AM

Meeting Ends: 3:00 PM

Medium: Physical Meeting

Present members: Jose, Lazaro

In this meeting Lazaro started work on implementing the UI of the website. The initial UIs were done. He also installed Mulgara and testing it to make sure it worked. Jose begin looking for a library to do the REST calls that the UI needed to make in order to make queries. He found a library and started testing it to make sure that it had everything that was needed for the project

Date: 10/4/2014

Begin : 11:00 AM

End : 4:00 PM

Medium : Physical Meeting

Present Members : Jose, Lazaro

In this meeting Lazaro worked mainly on getting the confidence ranking to show up on the forms when evidence was added. Apart from that, he also included a lot of bug fixes for code that was previously submitted. Jose worked on the PHP scripts that are going to generate SPARQL queries based on what users entered in the forms. He also worked on some bug fixes for the scripts.

Date: 10/19/2014

Begin : 11:00 AM

End : 4:00 PM

Medium : Physical Meeting

Present Members : Jose, Lazaro

In this meeting Lazaro began working on the custom query subsystems of the system. He also continued to work on the previous forms. Jose continued testing and developing the PHP scripts so that they generated valid SPARQL queries and correctly connected to the server.

Date: 10-28-2014

Meeting Begins: 4:30PM

Meeting Ends: 5:15PM

Medium of Communication: Skype

Present Members: Jose, Eric

4:30 - 4:40 Discussed what parts of the system I was showing today

4:40 - 4:50 Talked about the data entry form and how predicates worked

4:50 - 4:55 Discussed how predicates currently worked and changes to be made

4:55 - 5:05 Talked about the data search form and some specific queries

5:10 - 5:15 Talked about some good ontologies to use

Date: 10-31-2014

Meeting Begins: 11:30

Meeting Ends: 1:30

Medium of Communication: Physical Meeting

Present Members: Jose, Eric, Lazaro

11:30 - 11:50 Tour of Eric's workplace

11:50 - 12:20 Lazaro showed Eric some of the features of the system and asked for comments

12:20 - 12:30 Eric talked to us about how he would like to be implemented

12:30 - 12:50 We discussed about some good ontologies to use and which default ontologies we should use

12:50 - 12:55 Jose asked about the web crawler and if Eric has any recommendations

12:55 - 1:10 Eric explained what he wanted the crawler to do as we thought it had a different function

1:10 - 1:30 We got some minor features that Eric wanted us to implement and also got comments on the current system

Date:11/1/2014

Begin : 11:00 AM

End : 4:00 PM

Medium : Physical Meeting

Present Members : Jose, Lazaro

In this meeting, we discussed the need to change our core technology. We decided it was necessary to replace Mulgara with something that supported everything that was necessary for the project. Lazaro continued work on the front end forms. Jose continued work on the scripts and modified them for the new back end triple store that was going to be used.

Date: 11/17/2014

Begin : 11:00 AM

End : 4:00 PM

Medium : Physical Meeting

Present Members : Jose, Lazaro

Jose modified the PHP scripts so that they returned XML and prepared the PHP scripts to be retired. He also began work on the javascript library that were going to replace the PHP script for query generation. Lazaro worked on getting auto-complete for the different text fields working. He also added the base code import and export.

Date: 12/1/2014

Begin : 11:00 AM

End : 4:00 PM

Medium : Physical Meeting

Present Members : Jose, Lazaro

Lazaro began the testing of the front end part of the system. Jose added some of the javascript files for query generation.

Date: 12/6/2014

Begin : 11:00 AM

End : 6:00 PM

Medium : Physical Meeting

Both Jose and Lazaro finalized their code for their respective parts and began integrating the parts together. Both also fixed any bugs that came up from integration in their respective parts. Both continued to test their parts and tested the integration.

References

[1] - S. Karnouskos: *Stuxnet Worm Impact on Industrial Cyber-Physical System Security*. In:*37th Annual Conference of the IEEE Industrial Electronics Society (IECON 2011), Melbourne, Australia*, 7-10 Nov 2011. Retrieved 20 Apr 2014.

[2] - Jack Rusher, Semantic Web Advanced Development for Europe (SWAD-Europe), Workshop on Semantic Web Storage and Retrieval - Position Papers

[3] - http://www.w3.org/TR/PR-rdf-syntax/ "Resource Description Framework (RDF) Model and Syntax Specification

[4] - Berners-Lee, Tim; James Hendler; Ora Lassila (May 17, 2001). "The Semantic Web". *Scientific American Magazine*

[5] - Hebeler, John; Fisher, Matthew; Blace, Ryan; Perez-Lopez, Andrew (2009). *Semantic Web Programming*. Indianapolis, Indiana: John Wiley & Sons. p. 406. ISBN 978-0-470-41801-7