CM0343 (40 Credits)

Autonomic approach to information discovery in crowd sourced data.

Initial Plan

Author: Liam Turner

Supervised by Dr Stuart Allen, Moderated by Dr Jianhua Shao

October 14, 2012

1 Project Description

1.1 Problem

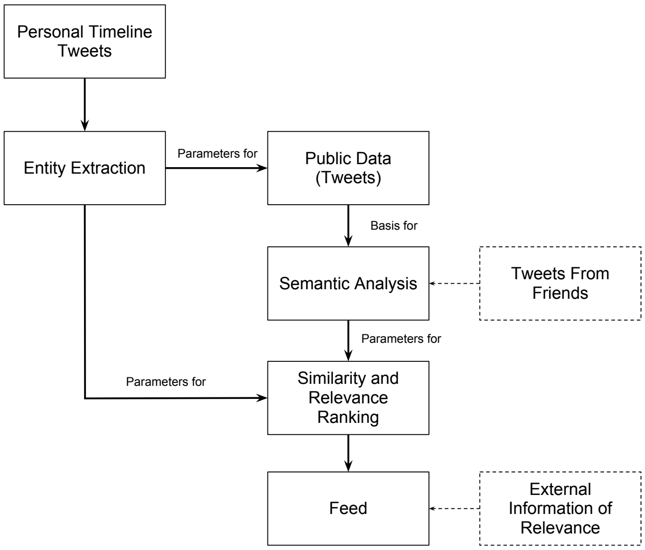
This project aims to provide a solution to the problem of efficiently managing vast amounts of community driven data posted on online micro-blogging services (e.g. Twitter) to find interesting, useful and legitimate information. Currently, factors that hinder information discovery using Twitter include: the speed of the data flow and the amount of data to process in real-time; which often result in infeasible or inefficient information processing. This is particularly the case in finding information posted outside of the user's personal following list (or local network). Twitter does offer users a 'Discover' option, compromising of: a handful of media stories relating to the users time line tweets over the past twenty-four hours and lists of individual users to follow under a category. This feature offers a limited and potentially outdated source of information and overall is predominately people based rather than content based.

1.2 Planned Solution

This project intends to create a solution to the problem of managing real-time crowd-sourced data collated from repositories such as online social networks in a manner that is both quick and useful to consumer users and 3rd party developers. Context will be derived from the crowd-sourced data, resulting in clusters of information relevant to the users personal interests; which will also be ranked in terms of calculated authenticity. Expansions on the information gathered will incorporate pre-emptively gathered semantics for sustainable information discovery; minimising manual searching and exploration of other web services for information.

In terms of scope for the project, the main considerations in terms of limitations are: available time relative to workload and the amount of data harvested against the computational complexity. With that in mind, the project will compromise of two parts: an application layer interface focusing on gathering and processing data in real-time based upon personalised topics and one or more front-end clients that will be powered by the functionality of the API.

Fig. 1. Initial planned process from data retrieval to information output.



2 Aims and Objectives

2.1 Project Aims and Objectives

- At a conceptual level, the primary aim and objective of the project is to create a system capable of providing users with a personalised and sustainable means to quickly gain knowledge and discover information based upon real-time collated data.

- Analyse psychological aspects relating to efficient information discovery, particularly exploring to what extent cognition can be optimised through implementable system mechanics.

- Develop a web service platform as an API for developers that would produce pages of semantic information from a data API backbone(s); with a similar concept to the Google Maps API, a platform on top of the raw data. Be flexible in terms of API endpoints and standards to be compatible for use with as many technologies as possible in client applications. Be sustainable for deprecation or introduction of new data backbones.

- Assess the suitability of building an API centric system revolving around real-time use of external services with text processing; including any notable advantages and disadvantages.

- Give indication as to the effectiveness of using text categorisation and ranking mechanisms to extract semantics from micro-blogging posts (e.g. Tweets); particularly with relevance to authenticity and suitability with formal and informal language expressions.

- Provide functionality for a measurable means of sustainable or seamless information discovery such as: attempting to predict desired information by integrating pre-emptive search aspects with related insights. In addition, use the authenticated user's social graph to produce recommendations of similar interests; based upon what is popular with friends.

2.2 Personal Aims and Objectives

- Exercise knowledge gained over the past two years of the degree programme such as: building upon the foundations of mainstream programming languages (such as Python) to gain experience of a project scale larger than previous academic projects; practice development paradigms (such as Object-Oriented) and project management techniques.

- Gain understanding of processes used in industry and academia which I have little to no experience of; such as version control systems (e.g Git) and project life cycles (e.g. Agile). Use the opportunity to explore emerging technologies such as Rich Internet Applications by means of frameworks (such as Node.js or Backbone.js) and standards (such as HTML5).

- Complete a project that has the objectives and complexity to be an attainable stepping stone into the area of research (Psychology in Computer Science), that I wish to taking forward into postgraduate study.

2.3 Content for Interim Report

The interim report will document the research into the background of the problem and develop ideas and prototypes for the solution at a deeper level than the conceptual level defined in the initial plan. More specifically, documentation regarding:

- Any existing solutions or partial solutions that are relevant to the aims and objectives of this project; particularly web application hybrids (such as 'mashups').

- Research into the applicability of using external APIs as the data backbone(s) for this project's API (such as Twitter or App.net). As well as the use of external APIs for elements of the semantic extraction process.

- Documentation of results from prototypes to represent the inner components of the API regarding extracting semantics and using semantics for expanding knowledge under a particular context. Document details of the chosen measures to implement fully to satisfy core API functionality aims.

- Summarise the details of system architecture, components and designs; as well as brief justifications for using an API as the core of a system, over a stand-alone application.

- Overall plan of technologies intended to be used such as: capability of any programming languages; hardware requirements, testing strategies (and mechanisms such as 'Mechanical Turk') and evaluation of version control systems. Including whether the use of particular native packages or frameworks can be applied as well as what aspects would require bespoke implementations.

- Initial review of the feasibility of any intended clients to be made using the API, such as: the amount of clients; platforms and scale of each client.

2.4 Content for Final Report

The primary objectives of the final report is to document the outcomes of the project at post-implementation stage; including evaluating and reflecting on: system functionality and the processes used; my individual performance and to what extent did the project achieve what was envisaged and why, including further steps that could be taken.

- Document specifics of the final implementation achieved as well as any design changes from the interim report that were necessary.

- Evaluation of overall project solution in terms of whether everything was achieved that was planned. As well as how well the final implementation performs against the envisaged solution including evaluations of the fully implemented prototypes that provide the core functionality.

- Reflections on whether an API over a stand-alone system was the more suitable solution including any unforeseen discoveries from concept to final implementation. As well as on the technologies used, such as: programming languages; bespoke against native implementations; data-models used and database systems and schemas used.

- Evaluation of project management strategy used including: suitability of the Agile methodology and any alterations invoked. Additionally, discuss personal performance against work plan and document any mistakes or oversights that may have been made; reflect on learning outcomes and applicable points for the future.

- Discussion regarding whether the limitations of the final implementation could be resolved with more time and/or resources. Explore any further capabilities of the API including additional or better functionality and potential clients or other applications of use.

- Reflect on the effectiveness of attempting to create an efficient means of optimising human cognition towards information discovery through the implemented API mechanisms and functionality.

A Appendix: Work Plan

This project intends to adopt an Agile approach to a project management strategy; particularly with principles such as: reviews of progress, adaptations to changing requirements, self-organisation, frequent updates and sustainable development. Therefore, weekly reviews of progress and allocation of time and resources will be carried out. Subsequently, the weekly details in the following plan become increasingly flexible and tasks become milestones and targets; with specifics intended to be detailed and agreed at least one week in advance with the project supervisor. The number of items vary each week due to complexity of the items and predicted time and resources available due to other commitments (such as other coursework deadlines in the final third of each semester).

In addition to the items below, weekly documentation of the previous weeks tasks (where necessary) will be carried out, in order to promote sustainable development of all aspects of the project.

A.1 Autumn Semester

Week Commencing 08/10/2012

- Research into API standards, including authentication ideologies (e.g. OAuth).
- Explore applicability of external API data backbones (e.g. Twitter, App.net, Google+ etc)
- Plan a process on a conceptual level for determining topic recommendations from a user's social connections.

Week Commencing 15/10/2012

- Explore API structures and draft initial structure and endpoint plan, establish requirements for any database storage.
- Explore viability of external APIs and language specific packages regarding text classification.
- Research ranking algorithms and credibility calculation algorithms.

Week Commencing 22/10/2012

- Research into suitable version control systems to put in place.
- Experiment with chosen API programming language capabilities and version control mechanisms put in place.
- Set up a local host for development of system prototypes.

Week Commencing 29/10/2012

- *MILESTONE* System structure research complete and documented.
- Establish experimental connections with chosen data backbone(s).
- Research and implement a chosen amount of text classification algorithms with using a subset of experimental data and analyse effectiveness in the context of the project's intent.

Week Commencing 05/11/2012

- ... continuation of previous week's research and prototyping.
- Research and implement a chosen amount of text authenticity detecting algorithms and analyse effectiveness with chosen data backbone(s).

Week Commencing 12/11/2012

- ... continuation of previous week's research and prototyping.
- Research into the complexity of the semantic insight functionality, investigating existing product implementations (e.g. Google search, Twitter 'Discover') and implement a prototype using a chosen amount with a subset of data.

Week Commencing 19/11/2012

- Time reserved for unexpected events that have hindered previous progress. In the event that tasks will not be allocated then the items from successive weeks will be carried out.

Week Commencing 26/11/2012

- *MILESTONE* Prototypes complete.
- Review of prototypes developed for inner components of the API and decide on chosen prototypes to integrate into the full system; as well as any back-up choices.
- Carry out research in any remaining topics for justification where neccessary for the interim report, such as testing strategies.

Week Commencing 03/12/2012

- *MILESTONE* Background research complete.
- Finalising contents of interim report.
- Review front-end client possibilities such as amount, type and size. Develop functionality and UI designs for chosen clients.

Week Commencing 10/12/2012

- Time reserved for unexpected events that have hindered previous progress. In the event that tasks will not be allocated then the items scheduled for successive weeks will be carried out.

Weeks Commencing 17/12/2012 - 21/01/2013 (inclusive)

- Implement basic framework structure of the back-end system, implement database schema ready for full development.

A.2 Spring Semester

Week Commencing 28/01/2013

- *MILESTONE* Project preparation complete for full system implementation.
- Create object-orientated class wrappers for: external API connections and calls; user sessions and database connections.
- Perform initial testing of wrappers.

Week Commencing 04/02/2013

- Integrate high level endpoints from design of system API structure using previously developed prototypes; such as authentication and user account control.

Week Commencing 11/02/2013

- ... continuation of previous week's integration of basic API endpoints.

Week Commencing 18/02/2013

- Develop basic structure of primary client and test integration with system API.

Week Commencing 25/02/2013

- *MILESTONE* Initial integration of API with primary client complete.
- Integrate topic based information cluster API endpoint using previously developed prototypes.
- Perform initial testing of information cluster API endpoint with primary client.

Week Commencing 04/03/2013

- Integrate semantic insights API endpoint using previously developed prototypes.
- Perform initial testing of semantic insights API endpoint with primary client.

Week Commencing 11/03/2013

- Integrate social graph recommendations API endpoint using previously developed prototypes.
- Perform initial testing of social graph recommendations API endpoint with primary client.

Week Commencing 18/03/2013

- *MILESTONE* Pre-final testing system build complete.
- Perform final rigorous testing of API.
- Integrate outstanding functionality of primary client.
- Perform final rigorous testing of primary client.

Weeks Commencing 25/03/2013 to 15/04/2013 (inclusive)

- Resolve any issues arisen after testing.
- Reviewing current system implementation and work schedule to determine focus of extensions to the system API or additional smaller clients.
- Implement any chosen additions to the project following review.
- Finalising contents of final report.

Week Commencing 22/04/2013

- Time reserved for unexpected events that have hindered previous progress. In the event that tasks will not be allocated then the items scheduled for successive weeks will be carried out.
- *MILESTONE* Project complete by end of academic week.

Week Commencing 29/04/2013

- Preparation for Viva examination.

Weeks Commencing 06/05/2013 to 17/06/2013 (inclusive)

- ... continuation of preparation for Viva examination.