CS512 (Fall 2015) - Project Playback

Ashish Jindal
Rutgers University
Piscataway, NJ, USA
Email: ashish.jindal@rutgers.edu

Yikun Xian
Rutgers University
Piscataway, NJ, USA
Email: siriusxyk@gmail.com

Sanjivi Muttena
Rutgers University
Piscataway, NJ, USA
Email: sm1727@scarletmail.rutgers.edu

***Abstract*—Events in the past play a major role in shaping our future. World Wide Web has abundant information related to such events, but it is scattered all over the web. There is no automated and efficient way to organize this information and to study such events for a specific entity of interest, where entity here refers to a generic term to represent an organization, country and person etc. This project aims to build a web-based automate system that can collect such data from heterogeneous online sources, extract condensed description of events, and integrate them into a timeline pertaining to an entity. This largely benefits those who would like to study history events of a certain company or organization in chronological order.**

***Index Terms*—Heterogeneous Sources, Event Extraction, Topic Model, Text Mining**

I. Project Description

Project Playback is about gathering data from web using various web scrapping methods and then extracting information about events of interest from the collected data using several machine learning algorithms. We will use Java to develop most of our system components along with a back-end built using Hadoop and MongoDB. Our project falls in two categories *Massive Algorithmics* and *Scalable Algorithms Infrastructure*.

We believe that a timeline view of events and activities for an entity of interest will give a new perspective for exploring data on Internet. Instead of viewing information organized in form of sections we will be showing it organized in form of a timeline. The intent is to give a bird's eye view of evolution of the entity over the time and study that evolution process based on data we call as the events of interest.

It is feasible to complete the system in 6 weeks if we consider only a limited set of entities to analyse. For example, if we limit the system's input to only consider fortune 500 companies or 200 most significant people in computer science field, then we should be able to deliver a working system within 6 weeks. The main stumbling blocks for this project are identifying seeding sources for data, integrating a real-time web-crawler into the system, cleaning the retrieved data, extracting events of condensed description, reorganizing events of companies and finally visualizing timeline data in a web-based system.

The project has four stages: requirement gathering, design, infrastructure implementation, and user interface.

A. Stage1 - The Requirement Gathering Stage

Following are the deliverables for this stage:

a) General Description: This project is like a minimalistic search engine which gives collated information and events about the query company or person in form of a timeline.

b) User Type 1: Job-hunters who want to know background of a company or a person.

- User Interaction Modes: Keyword based search for people exploring the web.
- Real World Scenarios:
  - Scenario 1 Description: A Person trying to get insight into a company like ”Microsoft”.
  - System Data Input for Scenario 1: ”Microsoft”.
  - Input Data Types for Scenario 1: String (Spaces allowed).
  - System Data Output for Scenario 1: All data related to ”Microsoft” placed on a timeline.
  - Output Data Types for Scenario 1: Date of event, text corresponding to the event and URLs of the sources of information.
  - Scenario 2 Description: A student trying to know about some famous person like ”Alan Turing”.
  - System Data Input for Scenario 2: ”Alan Turing”.
  - Input Data Types for Scenario 2: String (Spaces allowed).
  - System Data Output for Scenario 2: All the event's data related to ”Alan Turing” placed on a timeline.
  - Output Data Types for Scenario 2: Date of event, text corresponding to the event and URLs of the sources of information.

c) User Type 2: Researchers or students who want to look for techniques or patents created by a person.

- User Interaction Modes: Keyword based search for people exploring the web.
- Real World Scenarios:
  - Scenario 1 Description: A researcher who collects technique inventions by a company like ”Google”.
  - System Data Input for Scenario 1: ”Google”.
  - Input Data Types for Scenario 1: String (Spaces allowed).
  - System Data Output for Scenario 1: All research data related to ”Google” placed on a timeline filtered by data category.
  - Output Data Types for Scenario 1: Date of event, text corresponding to the event and URLs of the sources of information.

- Scenario 2 Description: A student trying to collect papers by some researchers like ”James Abello”.
- System Data Input for Scenario 2: ”James Abello”.
- Input Data Types for Scenario 2: String (Spaces allowed).
- System Data Output for Scenario 2: List of papers published by ”James Abello” placed on a timeline.
- Output Data Types for Scenario 2: Date of event, text corresponding to the event and URLs of the sources of information.

d) User Type 3: Analysts who would like to fetch financial statistics data of a company

- User Interaction Modes: RESTful API access for researchers interested in our event data.
- Real World Scenarios:
  - Scenario 1 Description: Researchers at a company ’Bloomberg’ trying to study their stock fluctuations over the past months and want to map the stock data to the event’s data.
  - System Data Input for Scenario 1: A RESTful request of GET type.
  - Input Data Types for Scenario 1: String (URL).
  - System Data Output for Scenario 1: Date of event, text corresponding to the event and URLs of the sources of information.
  - Output Data Types for Scenario 1: JSON.
  - Scenario 2 Description: Students in Rutgers University want to study the marketing and publicity patterns for multiple organizations.
  - System Data Input for Scenario 2: Multiple REST requests, one for each organization.
  - Input Data Types for Scenario 2: String (URL).
  - System Data Output for Scenario 2: All the event’s data consisting of date of event, text corresponding to the event and URLs of the sources of information.
  - Output Data Types for Scenario 2: JSON

e) Project Timeline and Division of Labor: The project has three main components - Automated web scrapper/crawler, web application with MongoDB backend and timeline visualization, and Machine learning techniques used on Hadoop to extract events of interest. Each of the team member will be responsible for completion of one of the above tasks, including work related to development, testing and documentation. We estimate the project completion in about 6 weeks.

- Week 1: By the end of week 1 we expect to have identified all the seeding sources for our application and built a working multi-threaded web-scraper.
- Week 2: After we have tested it’s working on localhost, 1 person in team will be responsible for deploying Hadoop in fully-distributed mode which will be integrated into web application. In parallel other members will start building the web application and exploring the machine learning on event extraction and topic models. At the end of week 2 we expect to have bunch of data crawled by web-scraper and a functional web application front end with Hadoop platform.
- Week 3: In week 3 we will link our web application with a MongoDB backend which will be loaded with some dummy data to make the web application fully operational to test its functionality. Meanwhile, we will also be working on cleaning the data and testing some machine learning algorithm to extract event data from it.
- Week 4: We will implement some effective models and integrate them into web application. At the end of week 4 we should be able to extract useful event information and save it into MongoDB to be available for query from web application.
- Week 5: In week 5 we will work on testing and making the application stable.
- Week 6: In week 6 we will work on making the application stable. At the end of week 6 we should have a stable working system with final project report and presentation.

References

[1] M. Naughton, N. Kushmerick, and J. Carthy, “Event extraction from heterogeneous news sources,” in *proceedings of the AAAI workshop event extraction and synthesis*, 2006, pp. 1–6.

[2] J. Dean and S. Ghemawat, “Mapreduce: simplified data processing on large clusters,” *Communications of the ACM*, vol. 51, no. 1, pp. 107–113, 2008.

[3] D. M. Blei, A. Y. Ng, and M. I. Jordan, “Latent dirichlet allocation,” *the Journal of machine Learning research*, vol. 3, pp. 993–1022, 2003.

[4] C. C. Aggarwal and C. Zhai, *Mining text data*. Springer Science & Business Media, 2012.

[5] O. Etzioni, M. Cafarella, D. Downey, A.-M. Popescu, T. Shaked, S. Soderland, D. S. Weld, and A. Yates, “Unsupervised named-entity extraction from the web: An experimental study,” *Artificial intelligence*, vol. 165, no. 1, pp. 91–134, 2005.