Project Playback - Fall 2015

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

Yikun Xian
Rutgers University
Piscataway, NJ, USA
Email: alan1936@cs.rutgers.edu

Sanjivi Muttena
Rutgers University
Piscataway, NJ, USA
Email: alan1936@cs.rutgers.edu

***Abstract*—** Events in the past, play a major role in shaping our future. World Wide Web has all the information related to such events, but it is scattered all over the web and there is no automated way to unite this information and study such events for entities of interest. Our aim is to build a system which can collect such data from web and use it to build a time-line of major events pertaining to an entity. Here entity is a generic term we are using to represent an organization, country, person etc.

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 various machine learning techniques. We will be using Java to develop most of our system components along with a backend built using hadoop and mongoDB. Our project falls in two categories ”Massive Algorithmics” and ”Scalable Algorithms Infrastructure”.

We believe that a time-line 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 time-line. 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 time if we consider only a limited set of entities to analyse. For e.g. if we limit the system’s input to only consider fortune 500 companies or 200 most significant people in 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, deploying a web-scraper with hadoop backend, cleaning the retrieved data, analysing the data using machine learning techniques and finally creating a web application to query for time-line data.

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 various machine learning techniques. We will be using Java to develop most of our system components along with a backend built using hadoop and mongoDB. Our project falls in two categories ”Massive Algorithmics” and ”Scalable Algorithms Infrastructure”.

We believe that a time-line 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 time-line. 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 time if we consider only a limited set of entities to analyse. For e.g. if we limit the system’s input to only consider fortune 500 companies or 200 most significant people in 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, deploying a web-scraper with hadoop backend, cleaning the retrieved data, analysing the data using machine learning techniques and finally creating a web application to query for time-line data.

The project has four stages: Gathering, Design, Infrastructure Implementation, and User Interface.

A. Stage1 - The Requirement Gathering Stage.

Following are the deliverables for this stage:

- The general system description: Our project is like a minimalistic search engine which gives collated information about the searched item in form of a time line.
- The types of users (grouped by their data access/update rights):
- The user’s interaction modes: Keyword based search for people exploring the web.
- The real world scenarios:
  - Scenario1 description: A Person trying to get insight into a company like ”Microsoft”.
  - System Data Input for Scenario1: ”Microsoft”.
  - Input Data Types for Scenario1: String (Spaces allowed).
  - System Data Output for Scenario1: All the data related to ”Microsoft” placed on a timeline.
  - Output Data Types for Scenario1: Date of event, text corresponding to the event and URLs of the sources of information.
  - Scenario2 description: A student trying researching about some famous person like ”Alan Turing”.
  - System Data Input for Scenario2: ”Alan Turing”.
  - Input Data Types for Scenario2: String (Spaces allowed).
  - System Data Output for Scenario2: All the event’s data related to ”Alan Turing” placed on a timeline.

- Output Data Types for Scenario2: Date of event, text corresponding to the event and URLs of the sources of information.

- The user's interaction modes: REST api access for researchers interested in our event data.
- The real world scenarios:
  - Scenario1 description: Researchers at a company 'A' 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 Scenario1: A REST request.
  - Input Data Types for Scenario1: String (URL).
  - System Data Output for Scenario1: Date of event, text corresponding to the event and URLs of the sources of information.
  - Output Data Types for Scenario1: JSON.
  - Scenario2 description: Students at a university 'B' want to study the marketing and publicity patterns for multiple organizations.
  - System Data Input for Scenario2: Multiple REST requests, one for each organization.
  - Input Data Types for Scenario2: String (URL).
  - System Data Output for Scenario2: 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 Scenario2: JSON
- Project Time line and Divison of Labor. The project has three main components - Web-scrapper running over hadoop, Web-application with mongoDB backend 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 web-scraper with a hadoop backend running in pseudo-distributed mode.
  - 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 and running the web-scrapper on top of it. In parallel other members will start building the web application and exploring the machine learning techniques. At the end of week 2 we expect to have a web-scraper running on top of a fully-distributed hadoop and a functional web application front end.
  - 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. In parallel we will also be working on cleaning the data and exploring machine learning techniques to extract event data from it.
  - Week 4: At the end of week 4 we should be able to extract useful event information from hadoop and pass it onto 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 5 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.

s