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

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.

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. 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 our team 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. 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. 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. 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. 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.

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. 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 our team 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. 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. 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. 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.

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:

- Our project is like a minimalistic search engine which gives a collated information in form of a time line.
- This system can be used by people exploring the web for information, it can be used by enterprises and also by researchers analysing statistical data.
- Enterprises can use the time line view to analyse their past events and how they helped in there advancement. They can also use it to analyse their competitor's timeline to study and compare the expansion over the time.
- Our system will also be exposing a REST api for researchers interested in our event data. Researchers can use this event data and map it with other statistical information like variation of stock prices over the time, popularity index etc. to get a newer perspective on data.
- 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: People exploring the web can access the event data by searching for the name of the entity they are interested in.
- The real world scenarios: Please insert the real world scenarios in here, as follows.
  - 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.
  - Scenario1 description: A student trying researching about some famous person like "Alan Turing".
  - System Data Input for Scenario1: "Alan Turing".
  - Input Data Types for Scenario1: String (Spaces allowed).
  - System Data Output for Scenario1: All the event's data related to "Alan Turing" placed on a timeline.
  - Output Data Types for Scenario1: Date of event, text corresponding to the event and URLs of the sources of information.
- The user's interaction modes: Researchers interested in our event data.
- The real world scenarios: Please insert the real world scenarios in here, as follows.
  - 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.
  - Scenario1 description: Students at a university 'B' want to study the marketing and publicity patterns for multiple organizations.
  - System Data Input for Scenario1: Multiple REST requests, one for each organization.
  - Input Data Types for Scenario1: String (URL).
  - System Data Output for Scenario1: 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 Scenario1: JSON
- Project Time line and Divison of Labor. The project has three main components - Web-scrapper over hadoop, Web-application with mongoDB backend and Machine learning techniques used over 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.
  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. 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. 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. 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. 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.

B. *Stage2 - The Design Stage.*

Transform the project requirements into a system flow diagram, specifyng the different algorithms, data types and structures required for processing and their associated op-

erations. The deliverables for this stage include the system flow diagram containing a graphical representation and textual descriptions of the corresponding data trasnformations, high level pseudo code of the overall system operation, and overall system time and space complexity.

Please insert your deliverables for Stage2 as follows:

- Short Textual Project Description. Please insert here the flow diagram textual description here together with its overall time and space complexity.
- Flow Diagram. Please insert your system Flow Diagram here.
- High Level Pseudo Code System Description. Please insert high level pseudo-code describing the major system modules as per your flow diagram.
- Algorithms and Data Structures. Please insert a brief description of each major Algorithm and its associated data structures here.
- Flow Diagram Major Constraints. Please insert here the integrity constraints:
  - Integrity Constraint. Please insert the first integrity constraint in here together with its description and justification.

  Please repeat the pattern for each integrity constraint.

C. Stage3 - The Implementation Stage.

Specify the language and programming environemnt you used for your implementation. The deliverables for this stage include the following items:

- Sample small data snippet.
- Sample small output
- Working code
- Demo and sample findings
  - Data size: In terms of RAM size; Disk Resident?; Streaming ?;
  - List the most interestng findings in the data if it is a Data Exploration Project. For other project types consult with your project supervisor what the corresponding outcomes shall be. Concentrate on demonstrating the Usefuness and Novelty of your application.

D. Stage4 - User Interface.

Describe a User Interface (UI) to your application along with the related information that will be shown on each interface view (How users will query or navigate the data and view the query or navigation results). The emphasis should be placed on the process a user needs to follow in order to meet a particular information need in a user-friendly manner. The deliverables for this stage include the following items :

- The modes of user interaction with the data (text queries, mouse hovering, and/or mouse clicks ?).
- The error messages that will pop-up when users access and/or updates are denied
- The information messages or results that wil pop-up in response to user interface events.
- The error messages in response to data range constraints violations.
- The interface mechanisms that activate different views in order to facilitate data accesses, according to users' needs.
- Each view created must be justified. Any triggers built upon those views should be explained and justified as well. At least one project view should be created with a justification for its use.

Please insert your deliverables for Stage4 as follows:

- The initial statement to activate your application with the corresponding initial UI screenshot
- Two different sample navigation user paths through the data exemplifying the different modes of interaction and the corresponding screenshots.
- The error messages popping-up when users access and/or updates are denied (along with explanations and examples):
  - The error message:
  - The error message explanation (upon which violation it takes place): Please insert the error message explanation in here.
  - The error message example according to user(s) scenario(s): Please insert the error message example in here.
- The information messages or results that pop-up in response to user interface events.
  - The information message: Please insert the error message in here.
  - The information message explanation and the corresponding event trigger
  - The error message example in response to data range constraints and the coresponding user's scenario Please insert the error message example in here.
- The interface mechanisms that activate different views.
  - The interface mechanism: Please insert the interface mechanism here.

II. Project Highlights.

  - Only working applications will be acceptable at project completion. A running demo shoul be presented to your project advisor at a date to be specified after the second midterm. A version of your application shall be installed in a machine to be specifed later during the semester. Your final submissiom package will also include a final LaTeX report modeled after this document, as well as a Power Point Presentation.
  - The presentation (7 to 8 minutes) should include at least the following items (The order of the slides is important):
    1) Title: Project Names (authors and affiliations)
    2) Project Goal
    3) Outline of the presentation

4) Description
5) Pictures are essential. Please include Interface snapshots exemplyfing tthe different modes of users's interaction.
6) Project Stumbling Blocks
7) Data collection, Flow Diagram, Integrity Constraints
8) Sample Findings
9) Future Extensions
10) Acknowledgements
11) References and Resources used(libraries, languages, web resources)
12) Demo(3 minutes)

Please follow the sample presentation mock up that is posted on Sakai.

- By Dec 1 your group should have completed the final submission. This includes a presentation (7 to 8 minutes) to your project advisor as well as a convincing demo of your project functionalities (3 minutes): every group member should attend the demo (and presentation) indicating clearly and specifically his/her contribution to the project. This wil allow us to evaluate all students in a consistent and fair manner.
- Thank you, and best of luck!