Visual Search and Interaction with the DARPA XDATA Employment Data

Neha Ahuja
Poushali Banerjee
Renu Kanakamedala
Aldrin Rodrigues

CSCI 572 HW 3
Prof. Chris Mattmann

December 6, 2014
University of Southern California

I. Introduction

The objective of this assignment is to us the Apache Solr index constructed from upstream JSON files representing job postings collected and crawled from the upstream http://www.computrabajo.com affiliate sites that primarily serve Mexico and South American countries and visualize them using the D3.js. The task is to leverage the Apache Solr index and the D3.js data visualization technology (http://d3js.org/) to interact and visualize the search engine data. We have leveraged D3.js and the geolocation information from the data to recreate the map. We have also leveraged the Extract, Transform and Load (ETL) work, and the work in ranking and deduplication previously performed on the XDATA employment dataset, to construct a novel data visualization capability using D3.js.

II. Methodology

1. Task 1

   Constructing a web front end.

   a. Creating the map.

      i. We used an existing map.json file from D3 website with to create a json file. We then used the d3.json function to load map.json file and render the map. Once the map was loaded we could now add data points to illustrate the answers to the questions.

      ii. All the json data was present in solr, indexed and we ran several function queries through d3 to access and retrieve the necessary information. Once the results of the query was loaded into D3, we used several d3 functions to control how the data would be displayed on the screen.

      iii. For challenge queries, we used Apache Solr's Function Queries and also implemented a drop down menu such that the data could be visualized based on what user can choose to represent. The data was plotted using latitude, longitude, job type fields from the json files.

      iv. We also generated a scatter plot for one of the challenge questions.

Figure 1 : Geographical distribution of the jobs based on longitude and latitude

The Figure 1 above gives a clear distribution of locations where job postings have been recorded. The color coded scheme is used to identify the location.

b. Web-based REST access to Solr data with D3
    i. We developed our own custom method to connect D3 to Solr
    ii. Asynchronous javascript with XML (AJAX) was used to call the Solr's URL and get data using JSONP. JSONP is a communication technique used in Javascript programs running in web browsers to request data from a server in a different domain, something prohibited by typical web browsers because of the same-origin policy.

2. Task 2

Solr is a standalone enterprise search server with a REST-like API. REST is an abstraction of the architecture of the world wide web. REST is basically used for accessing the search index in Solr. In our project, we used the REST based API call to query the JSON data that was indexed in Solr. The data points were then mapped on the screen. We also used the REST calls for getting query specific data. For this, the calls were issued using the Solr function query and the URL's modified to give us the desired data as output.

3. Task 3

a.

Figure 2 : Geographical Data Distribution using Company Name

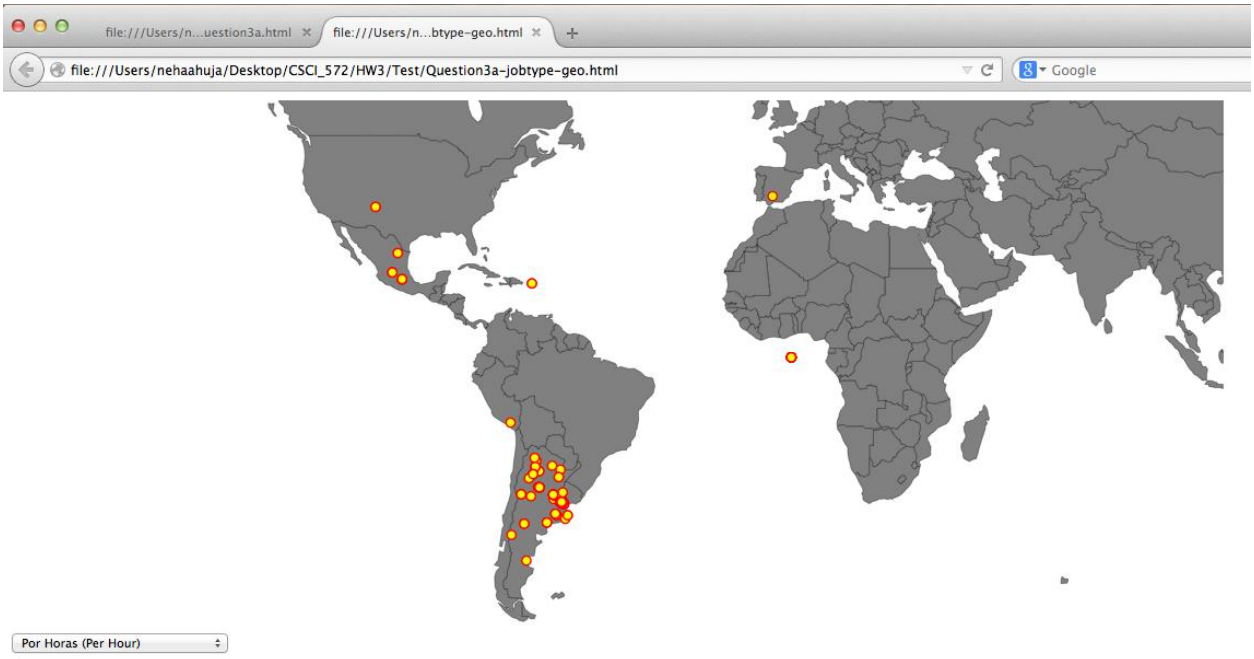

Figure 3: Geographical Data Distribution using Job Type

The Figure 2 shows the distribution of company names. The user can select the company name from the

drop down menu to get a distribution of a particular company.

The Figure 3 shows the distribution of job postings. The user can select the job type like part-time, full-time, etc from the drop down menu to get a distribution of jobs based on the job types in the areas.

The data uploaded on our Solr had only json 12051 files and it has data only from August to November and thus, we just plotted geographically and temporally, it uses the data from September to November.

For challenge question 3 a., we could not use salary field to show the distribution as we don't have appropriate values for the salary field. It is either numeric or blank (--) or 'A Convenir' as its value. Its data type is not consistent.

b.

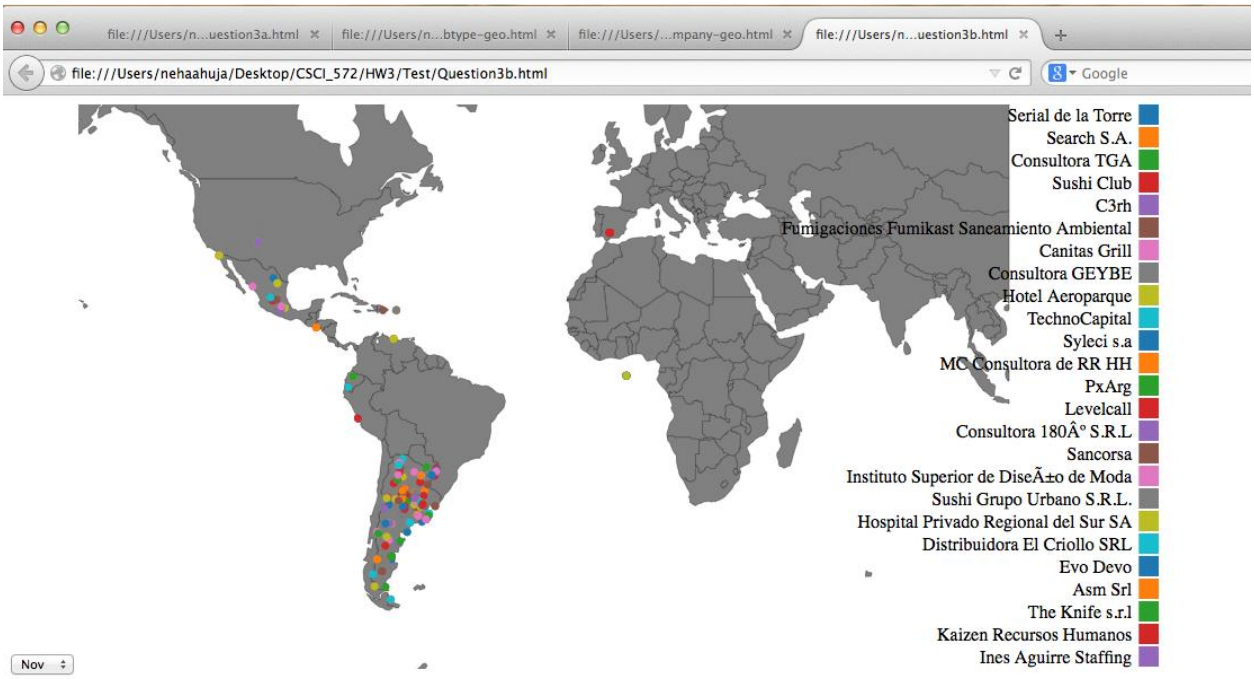

Figure 4 : Data Distribution using time

All of the our uploaded data on Solr was from 2012 and therefore, we have shown the distribution of the based on the progression of months. There is a dropdown menu from which the months of the year 2012 can be chosen and this will show the changing trends in how companies have been growing. Each map is based on the option selected from shows the count of by company, data is thus grouped by company and different colors represent the different companies. Thus as the user browses by time the growth and spread of the companies geographically becomes apparent. Basically, we have used the months for display the data based on time. We observed that the job postings have gradually increased from the month of August to December. It is difficult to say that whether the companies are entering into new domain to make money.

c.

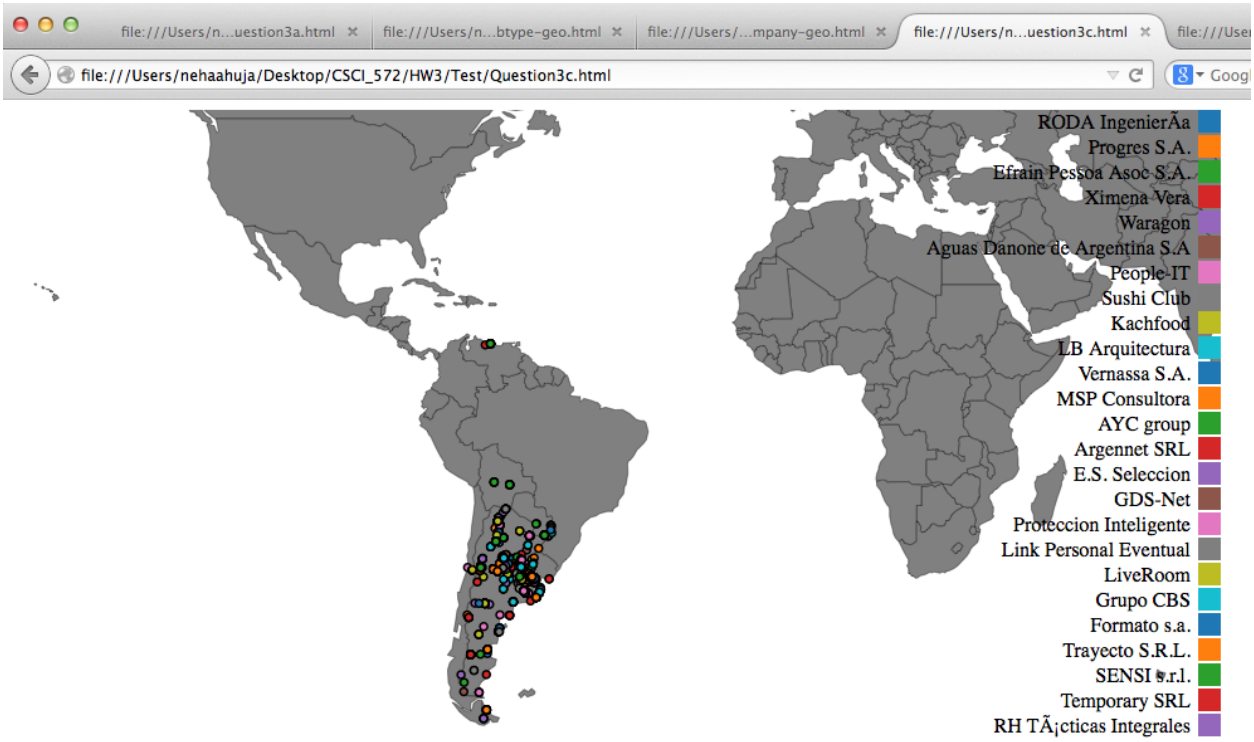

Figure 5 : Corporate Presence across South America

The Figure 5 shows the corporate presence across South America and displays the points based on the Company name. From the our dataset, we observe that there are more companies in the south part of the South America than in the north part of the South America. Hence, we can conclude that for our dataset, the southern part of South America is commercialized and north part is not.

d.

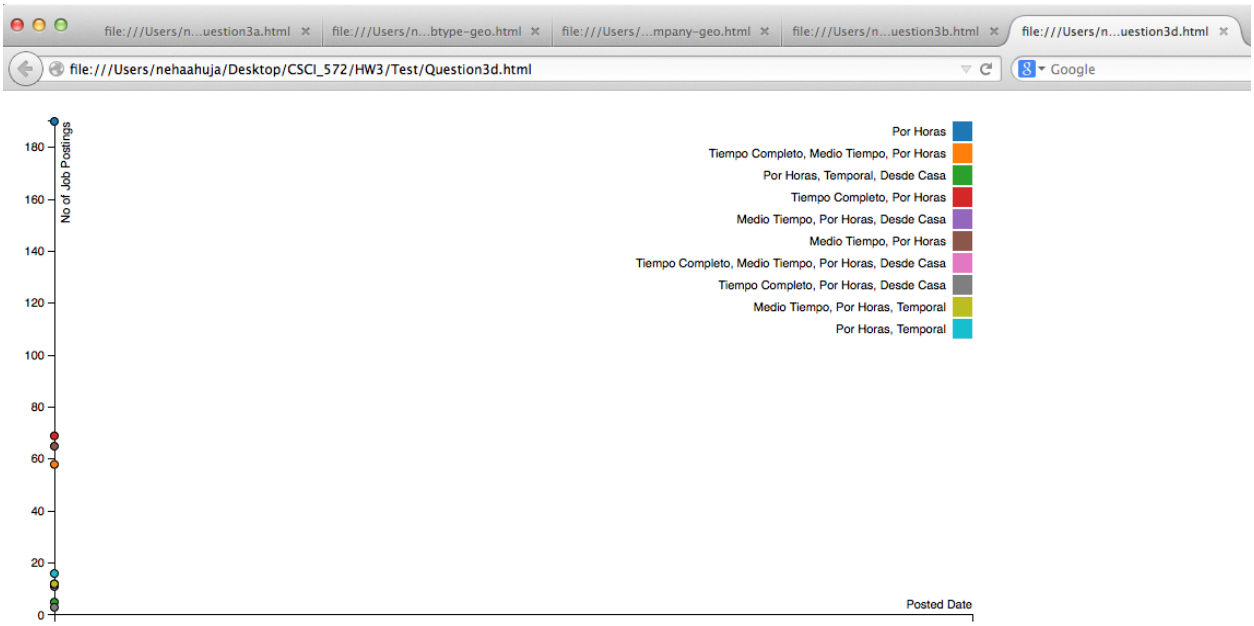

Figure 6 : Seasonal Trends of different job categories

We used the Scatter Plot graph to represent the seasonal trends of the job categories. Here, we have used the job type and the number of job postings for particular time to display the data.

III. Report Analysis

1. How easy to use was D3?

   Ans: D3 is a Javascript library that uses digital data to drive the creation and control of dynamic and interactive graphical forms which run in web browsers. Compared to other softwares of its kind, it is the first widely successful visualization platform and there are many good reasons for it. In our opinion, D3 was easy to use due to the following reasons:

   1) It is easily available to download (d3.js and d3.min.js files).
   2) There are lot of resources available as tutorials and examples.
   3) It is easy to understand.

   Along with the our opinion, there are some general reasons to use D3:

   - First, it is web based. Meaning, it is easily accessible and available to a wide audience over the web. Anyone can create their visualizations and host it on a website thus making it available to a huge audience and this is important as visualizations are largely meaningless unless they can be easily used across a wide audience.
   - It is extremely flexible and offers a wide variety of visualizations that can be created with a variety of data formats. It works well with existing web technologies and works well with the DOM which makes it easily plug-able into

many languages and technologies with which websites are rendered these days.

- The visualizations are extremely beautiful. No other tool has provided this kind of control of css before, and D3 adheres to all the modern design standards and formats making the visualizations appealing to its consumers.

2. What was the hardest part in D3?
   Ans: Because it so dynamic in its possibilities and is bundled with features, there is a steep learning curve. For us, loading, i.e, connecting Apache Solr and D3 was time consuming compared to visualizing it as it took us a while to understand how to make it work with JSON as there is not much support for JSON data format sets. Also, we found that D3 tends to run a bit slow as dataset gets larger and more complex operations are run. But once we got the hang of it, it was fun. In addition, using min version of the d3.js was not preferred over original d3.js version.

We have provided color coded visualizations along with the legend on the right to explain the significance of the data distributions. We have provided drop down menus, which users can use to get different representations of the same map by certain categories as asked in the challenge questions. For example, the distribution of companies posting job data for a given month, by selecting it from the dropdown and the distribution for that month can be viewed, this representation can also be used to understand the progress of the companies or trends in the data. Screenshots for the challenge questions are provided above which demonstrate some of the capabilities of our maps. A video is also provided for further understanding.