Group 6 - Phase 1

Kevin Dombrosky
Rochester Institute of Technology Student
610 Park Point Dr
Rochester, New York 14623
kfd6490@rit.edu

Brittany Purcell
Rochester Institute of Technology Student
107 Weldon Street
Rochester, New York 14611
blp6903@rit.edu

ABSTRACT

This paper discusses the progression of the group project for Group 6 in the Intro to Big Data class at Rochester Institute of Technology, CSCI-620. The paper will further explain the work that has been done so far, what will be done in the future, and what the expectations are for the finished project.

Keywords

Phase 2, Group 6, Plants

1. INTRODUCTION

The group project for CSCI-620, Intro to Big Data at Rochester Institute of Technology has been underway for 9 weeks now. Is has been about about 2 weeks since the last phase update, and since then a lot of work has gone into the database and visualization.

As a recap there are two requirements for the project. The first is a database management system (DBMS), which for this project will be a relational database system in MSQL which contains a dataset from the UCI Machine Learning Repository(http://archive.ics.uci.edu/ml/datasets/Plants). The second requirement is the visualization, which is an application that shows the geographical location of various types of plants in the United States and Canada. The information for the visualization will be contained in that dataset from the UCI website.

There has been steady progression toward this goal, and so far there have been no changes to the visualization design, or the dataset.

2. BODY

2.1 Database

As a reminder, the dataset is be set up in a MySQL database. In order to move the data from the .text files into the dataset, a script was written. Originally the script was going to be written in Python, however Java was chosen later due to its familiarity with the group. The script needed to break down the files in an easy to read way so the MySQL Workbench could be used to load the data into the tables. Below is, first the ER-diagram which represents how the database will interact, the second is the original file contents that need to be interpreted to work in the database.

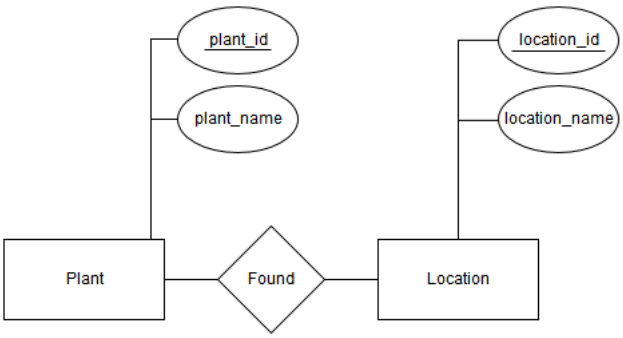

Figure 1: The ER diagram which represents the organization of the data in the database.

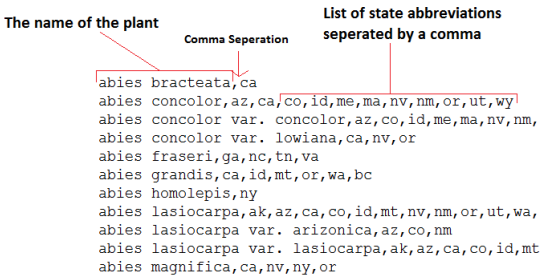

Figure 2: The organization of the original database file.

As a reminder, for the database there were going to be three tables, each with two columns.

Table Layouts:

- Plant Table
  - IDplant
  - plantname
- State Table
  - IDstate
  - statename
- PlanttoState Table
  - IDplant
  - IDstate

In order for MySQL to load the information into the tables, the .text files need to be converted so that the files contained the information in the order that it would appear in the columns, seperated by a tab character. So, for example in the Plant table, the plantdata.text file would need to contain an ID, tab character, plant name. This is the same for each of the tables that need to be loaded. If the .text files are in the correct order it is a simple LOAD call in MySQL to load the tables.

Steps to Database Loading:

1. Creating a plant.text file that contains the plant name and plant id. The plant ID was just generated, starting at 1, and going until there were no more plants in the original file (which a snippet of can be seen in the figure above). The name then, was just the first element in every line of the file. Each line in the file was a different plant, so each would have a unique ID.
2. Next, the states needed to be organized into the state table, which also contained an ID and a name. The dataset provided the following file, which contained the states and their abbreviations that are used. (http://archive.ics.uci.edu/ml /machine-learning-databases/plants/stateabbr.txt) This, for now, was used to access the full names of the locations, since that is what will be used by the visualization. Again, the ID's were created based on the amount of locations that were read in, and the state names were the second item in every line, only this time separated by a space and not a comma.
3. The last part was the trickest. Since the PlanttoState table is looking only at the ID's of the states and plants, those needed to first be determined for each plant, and added into a text file. For this the original file (seen again in Figure 2) has the abbreviations for the states. The script kept track of the abbreviation to name connection for states, and the name to ID connection. So, in order to find the state ID associated with a specific abbreviation there had to be a lookup. The plant ID was simple, since the name and ID were connected, but the state lookup was a little more in depth. In the end, there was a text file that contained a plant ID and a state ID, each line representing either another state the plant was found in, or the next plant on the list. (This is so that in the PlanttoState table there are only the two columns)
4. Once all three of the text files were created, the following LOAD call needed to be made to load the information. This call will be made three times (one for each table that needed to be loaded).

   LOAD DATA LOCAL INFILE 'filelocation' INTO TABLE tablename COLUMNS TERMINATED BY '';

2.2 Application

Group 6 decided to create a visualization for this class, which was discussed in Phase-0. As a recap, the visualization will consist of a map(functional for only the United States and Canada), as well as a search box for the user to search for a certain plant. This has been created, and the screen shot of the application is in Figure 2 below.

This application is a web page that will be written in Jade Markup Language with CSS. This is the language that is most familiar to the group, after one of the members worked with the language all summer on an independent project. The basic blueprint for the web page has already been created(again, see Figure 2), but it has not yet been hooked up to the database.

Ultimately, there will be (on the left hand side of the website right below the search bar) a list of plants that are found in the database. The user can search for a specific plant (using the search bar) which will highlight the regions that plant is located. Otherwise, the user can choose a plant from the list, and it will do the same thing. Another option for users would be to select a region they are interested in, and the list on the left would change to be only plants that can be found in that area.

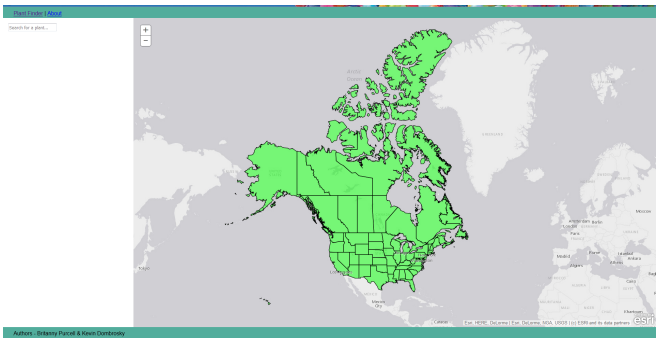

Figure 3: A screen shot of the website being used for the application.

2.3 Communication

In order for the website and database to communicate, the database will be setup on a Node.js server. This will allow the website to request information from the local database, which is set up using MySQL Server. In order for the user interface(UI) to change based on users requests, this is how the information will be transmitted from the database to the website.

2.4 Data Analyses

The data analyses will all happen through the website visualization. A user will be able to either select a state or

a plant in order to find information from the database. If a state is selected, then, on the left hand side of the website, a list of plants that can be found in this location will appear. The other option is using the search bar to find a plant. This will actually use substring searching, so as you are searching a list of possible plants will appear, and the location will be highlighted. This works the same for searching a specific plant name. The locations in which the plant can be located will become highlighted on the screen. This is a simple way to take the information found in the data set and portray it in an easy to read and easy to see way for any user. Now, a user can find a list of plants that grow in their location, or find out if a plant in a specific location can be found.

2.5 Resources

There are many different resources being used in order to develop the project. These include certain files, softwares, and online databases all assisting in the creation of the website and database. As it was already covered, the database is being set up as a MySQL database running on a local server set up using MySQL Server, the application will be a website visualization that is coded using Jade and CSS, and the database and website will communicate with a Node.js server.

The code is being stored using GitHub, and is shared among all members of the group. This is a repository to store, download, and change the existing code. The code is being written in different platforms, some members are using Notepad++ to do the coding, while others are using Visual Studio Code. The scripts to load the database were written in Eclipse, using Java coding language. Once the code is written in each of these mediums, the code is added to the GitHub repository.

LaTeX is being used to write the ACM style templates, which is the document you are currently reading, and Visual Studio Code is used for the README(which will walk users through how to run the program and access the website).

The README will be provided to establish step-by-step instructions on downloading and installing necessary software, and (once the software is installed), how to run the project and access the website. There is also README's on how to get the data set information in the .text files into a local database using MySQL Workbench.

3. CONCLUSIONS

The group has already made some progress towards finishing the project. This includes work on both the application as well as the database. There is still much work that needs to be done, but most of the thinking and setting up has been finished. Languages and applications have been chosen and are in the process of being set up. Below is a quick list of what was done, what needs to be done, and deliverables that have been set.

Accomplished:

- Developed ER Diagram
- Created Website Wire Frame
- Setup Website
- Conceptualized Application Utilization
- Wrote a Java Script to change dataset text files
- Loaded the dataset into MySQL database
- Set up the database on a local server using MySQL Server

To Be Done:

- Establish Communication Between Website and Database
- Display Information on Website
- Change Display based on User Input

Resources

- GitHub
- LaTeX
- Node.js
- Notepad++
- Visual Studio Code
- Eclipse
- MySQL WorkBench
- MySQL Server

Files

- README
- Group6-Phase2
- CSCI620-master (Folder)
- plant_data.text
- state_data.text
- plant_to_state_data.text

4. CITATIONS

- Relevant Papers
  - HÃd'mÃd'lÃd'inen, W. and NykÃd'nen, M.: Efficient discovery of statistically significant association rules. Proceedings of the 8th IEEE International Conference on Data Mining (ICDM 2008), pp. 203-212. IEEE Computer Society 2008.
- Dataset
  - USDA, NRCS. 2008. The PLANTS Database ([Web Link], 31 December 2008). National Plant Data Center, Baton Rouge, LA 70874-4490 USA.