Group 6 - Phase 2

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 group project for Group 6 in the Intro to Big Data class at Rochester Institute of Technology, CSCI-620. This paper will discuss the progress, ideals, and implementation on creating a database with visual representation.

Keywords

Phase 1, Group 6, Plants

1. INTRODUCTION

The group project for CSCI-620, Intro to Big Data at Rochester Institute of Technology was a semester long project where members of the group had to take a dataset, and create a way to organize and understand the dataset. The project was meant to expand the members knowledge and understanding of handling large datasets, as well as create a way for others to analyze the data.

There were two requirements that the team needed to meet for the project. These were creating a database management system (DBMS) and having some form of database analytic for the system. The members used the the Plants dataset from the UCI Machine Learning Repository which contained information on the growth location and scientific names of plants in North America. The data was then put into a relational database using MySQL.

For the analytic the team decided on a visual representation using a webpage that displayed a map of the United States and Canada. The user can then either search for a plant name, or select a region on the map, and the results of the dataset would become available. There will be more specifics on the visualization later in the paper.

2. BODY

2.1 Database

As a reminder, the dataset being used is the Plants dataset from the UCI Machine Learning Repository (http://archive.ics.uci.edu/datasets/Plants), and was set up in a MySQL database. In order to add all of the data into the database, a Java script was created. The script ran through the dataset and added the elements into MySQL.

The information found in the database consists of a plant and its location in the United States/Canada. Below is the Entity-relationship(ER) diagram which was developed by Group 6. This diagram is the basis for how the information from the dataset was stored in the database.

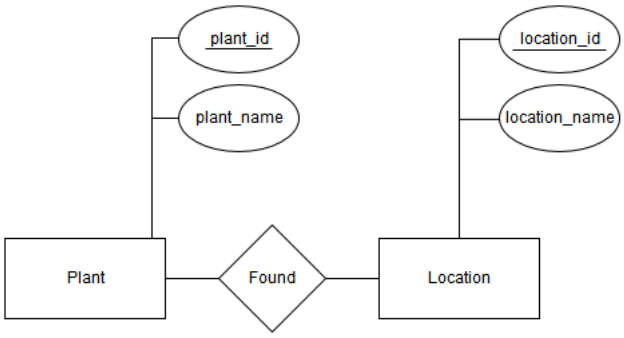

Figure 1: The ER diagram which represents the organization of the data in the database.

Later, when the visual needs to access the information in the dataset, this allows for proper organization so the appropriate calls can be made. There are three tables within the database. The first holds the information on the plants. This consists of the plant ID, as well as the name of the plant associated with that ID. The locations table is very similar. The location table has the location ID and the associated location. Lastly, there is a table which maps the plant-id to a location-id. This is how the location and plant information can be accessed using the appropriate calls, and will be simple to display in the visual.

There are a few specific sets of data which were excluded from the dataset. These were Greenland/Denmark, St. Pierre and Miquelon, and Prince Edward Island. The reason the members excluded these locations is because the map only looked at The United States and Canada. These regions were outside the scope of the map, and therefore were outside the scope of the project.

2.2 Application

The group decided that for the analytic portion of the project they would work with a map visualization on a webpage. The map is functional only for the United States and Canada, and each region is selectable. Selecting a region would bring up the list of plants that can be found in that area.

The visualization also has the ability to search by plant name. The search does a sub search, that will find all plant names containing what is typed in. When you search either using the full plant name, or a sub plant name, the locations where the plant is available will show up green. When searching or selecting a plant name, any locations where that plant is will show in the map.

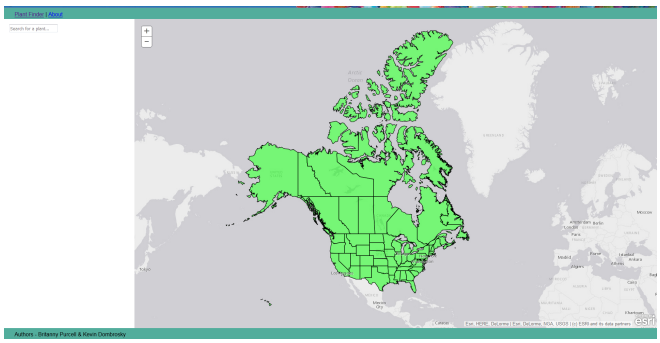

Figure 2: The blueprint for the website.

This application is a web page that was 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 can be seen in Figure 2, and in Figure 3 is the final result of the webpage.

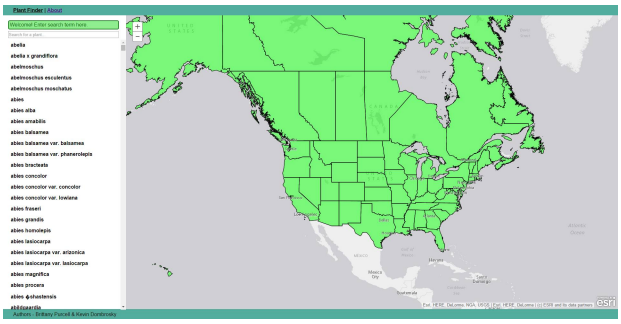

Figure 3: The final website.

2.3 Communication

In order to allow communication between the webpage and the database, the data needed to be uploaded to a server. The members used a Node.js server, and using Javascript the group could make calls to the server requesting data. The data is then transmitted to the webpage, and the results are displayed on the webpage.

In order to make the communication work the group used a MySQL package that would allow for communication. In order to secure and protect the database against a SQL injection attack, the group used a function from the MySQL package that escaped important characters in any user input.

There are a few areas where the group had to create slight workarounds in order to get the information to display nicely. There was a schema name change part of the way through the project in order to make the information easier to access. Also, there was a slight problem with Newfoundland and Labrador. These locations were separate in the database, but are the same place on the map. In order to compensate for this and display the information correctly, at runtime if the database requests information for the location, it will pull the information from both Newfoundland and Labrador and combine the results. These results will then be displayed at one place. This is the same if a plant's location is either Labrador or Newfoundland. Either way, if one or both locations are listed, it will mark that spot on the map.

2.4 Resources

There were many resources that were used in order to develop and complete the project. These include certain files, softwares, and online databases. The database was set up using a MySQL database, and the visualization is coded using Jade and CSS. THe database and webpage communicate using a Node.js Server. The communication is protected and available because of MySQL packages and protected SQL injections.

The code and files are all being stored using GitHub which is accessible by all of the members of the group. GitHub is a repository to store, download, and change the existing code. The code itself was written using many different platforms. Some members used Notepad++, while others are using Visual Studio Code. Any changes to files or code needed to be pushed to the repository.

LaTeX is being used to write the ACM style templates, which is the document you are currently reading, and Visual Studio Code/Notepad++ 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.

3. CONCLUSIONS

The group has successfully created a working webpage visualization which will allow analysis on a large dataset of plant information. They designed and implemented a worker database set up on a usable and accessible server. Information has been moved, copied, and rearranged to create the database, and the visualization has been adapted to work with the data.

Below is a list of all of the different elements that needed to be accomplished in order to create the final project. There is also a simple list of the resources that were used by the group to make the project possible. Many of these recourses were mentioned previously in the paper, but this is a full list of them. Lastly, is all of the necessary files and folders that should be available in the zip folder. These are important for the explanation of the project, the README describing the process needed to set everything up, and this document.

Work Done:

- Developed ER Diagram
- Created Website Wire Frame
- Setup Website
- Conceptualized Application Utilization
- Write Python Script for Database Input
- Set Up Database
- 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
- Javascript

Files

- README
- Group6-Phase2
- CSCI620-master (Folder)