TrawlExpert: Tool for Watershed Biological Research

Trawlstars Inc. (Group 11)
Lab section: L01
Version: 1.0
SFWRENG 2XB3

Christopher W. Schankula, 400026650, schankuc
Haley Glavina, 001412343, glavinhc
Winnie Liang, 400074498, liangw15
Ray Liu, 400055250, liuc40
Lawrence Chung, 400014482, chungl1

April 5, 2018

Revision History

Revision	Date	Author(s)	Description
1.0	05.04.18	HG	created

By virtue of submitting this document we electronically sign and date that the work being submitted by all the individuals in the group is their exclusive work as a group and we consent to make available the application being developed through SE-2XB3 project, the reports, presentations, and assignments (not including my name and student number) for future teaching purposes.

Abstract

TrawlExpert is a powerful tool to enable researchers to analyze and filter large datasets from fish trawl surveys in order to perform environmental research on fish and invertebrate populations. The tool gives researchers the ability to intelligently filter and query datasets based on biological classification such as family, genus or species, or based on location or timeframe. Advanced outputs display data as a histogram or geographical map, each depending on population abundance as a function of time and spatial distribution. Additionally, *TrawlExpert* provides a tool for finding local subpopulations within a larger query. A dataset of thousands of Great Lakes trawl surveys from 1958-2016 will be used as a demonstration of *TrawlExpert*'s capability to help researchers narrow down large datasets and glean data which pertains to their research. *TrawlExpert* will be designed to be used easily and effectively as the first step in a groundbreaking climate and ecological research pipeline.

Contents

1 Project Scope

1.1 Objective

Provide a statistical and visual tool for the analysis of water ecosystems, based on scientific water trawl data. Gives researchers with tools to analyze large datasets to find patterns in fish populations, including the plotting of historical population data on a map, the analysis of population trends over time and the determination of subpopulations of a certain biological classification.

1.2 Motivation

The diminishing of fish populations in the Great Lakes became a problem in the latter half of the 20th century, with the total prey fish biomass declining in Lakes Superior, Michigan, Huron and Ontario between 1978 and 2015 (**?**). Annual bottom trawl surveys involve using specialized equipment to sweep an area and are used to determine the relative temporal variation in stock size, mortality and birth rates of different fish species (**?**). These surveys are performed annually and often have hundreds of thousands of records, making manual analysis infeasible. The ongoing protection and development of the Great Lakes water basins is considered an important topic for scientists in both Canada and the United States, as evidenced by grants such as the *Michigan Sea Grant* (**?**).

TrawlExpert will give researchers tools to filter through these large amounts of data by allowing them to search through data based on class, order, genus, family or species. This will help support scientific researchers and fishing companies as they study fish populations. These studies help inform initiatives to preserve fish populations and conduct their business in an environmentally friendly way going forward. As more data is collected on an annual basis, TrawlExpert can easily be injected with the new data and will adjust and scale accordingly, combining the new data with the old data for continued analysis.

TrawlExpert will also analyze the trawl data to find connected subpopulations within the data, giving researchers tools to analyze the portions of the water body that contain different populations and even track these specific subpopulations over time.

The focus of the project will be to develop these unique data searching and querying tools as a first step in a complete trawl survey analysis. For a complete analysis, tools like stratified statistical analysis are required by the researcher (**?**). For purposes of maintaining a manageable scope for this project, the implementation of advanced trawl survey scientific and statistical analysis tools will be relegated to future developments.

1.3 Dataset

The test dataset that will be used for purposes of this project is the *USGS Great Lakes Science Center Research Vessel Catch Information System Trawl* published by the United States Geological Survey (**?**). Compiled on yearly operations taking place from early spring to late fall from 1958 until 2016, the dataset contains over 283,000 trawl survey records in the five Great Lakes, including the latitude and longitude co-ordinates and biological classification such as family, genus and species.

2 Algorithmic Opportunities

This project provides several algorithmic challenges and opportunities, which can be broken into the categories of searching, sorting and graph algorithms. These algorithms and their use for the project are described in this section.

2.1 Searching Algorithms

A modified form of binary search will be used for quickly locating the first of a given key in the large dataset and will be a crucial building block of all three main types of output. This will allow all entries of a given query to be found.

2.2 Sorting Algorithms

Sorting will be crucial for both ordering data in chronological order as well as supporting binary search, given that it requires sorted data. The mergesort algorithm will be advantageous due to its fast and predictable runtime.

2.3 Graph Algorithms

Graph algorithms will support advanced searching features. Firstly, the biological classification of each organism forms a tree (see figure **??**) from which species in the same genus, for example, can be located.

Secondly, a graph algorithm will be used to find connected components for generating output of type 4 described in section 5.2.4. Nodes are connected together based on their distance to surrounding points (**?**). Breadth-first search will be used to determine connected components (**?**) (see figure **??**).

3 Project Plan

3.1 Milestones

The following milestones will help inform our progress towards completing the goals. Given that the team contains 5 people, the team will be divided up into two subteams of 2-3 people for maximum efficiency. Individual members' tasks can be decided by the subteams based on the progress of that team towards completing the next milestone's goal(s). Subteams should be in constant communication and in agreement about the inputs and outputs of modules to avoid the need rewriting of code.

Milestone	**Subteam A**	**Subteam B**
Milestone 1 ("Bedrock") (End of Week 1)	Finished parsing module for .csv data to create Java objects that can be used for analysis; start data cleansing	General binary search module underway
Milestone 2 ("Quartz") (End of Week 2)	Cleansed the data to remove or correct entries not containing all of the columns; start generating biological classification tree module	Finished and tested binary search; start mergesort
Milestone 3 ("Granite") (End of Week 4)	Finished and tested biological classification tree; start formatted text output	Finished and tested mergesort; start writing query module for output type 1, 2 and 3 making use of the biological classification tree, mergesort and binary search to get results from data
Milestone 4 ("Sandstone") (End of Week 6)	Continue text output tools. Google Maps API should be explored if time allows, otherwise it can be omitted to keep the project within scope	Continue query module: finished output 1, 2 & 2; start output type 4
Milestone 5 ("Diamond") (End of Week 8)	Finished data visualization or text output tools; prepare keynote presentation	Finished output type 4; prepare keynote presentation

While this schedule provides a good reference and a way to monitor progress, team members should be flexible and remain in communication to ensure the project is kept on schedule. For example, if a milestone is reached before its given date, the next milestone should start development early. Approximately 1-2 weeks have been purposely left as padding at the end in case of unforeseen circumstances.

3.2 Team Roles

Member	**Role**	**Subteam**
Christoper Schankula	Team Lead	A
Ray Liu	TA & Professor Liaison	A
Winning Liang	Project Log Admin	A
Haley Glavina	Meeting Minutes Admin	B
Lawrence Chung	Head of Booking	B

3.3 Workflow

The team will use **GitLab** as the primary way of sharing code and keeping up to date. The Git repository will be split into two branches, in addition to the *master* branch: *TeamA* and *TeamB*. Each Subteam will develop on their respective branch, then issue merge requests so the team can evaluate and approve merges into the *master* branch.

3.4 Communication

The team will use **Slack** and **Facebook Messenger** as primary and secondary means of communication. A Slack group has been created and each of the members were invited to it. **Google Drive** will be used to keep track of documentation such as the *Project Log* and meeting minutes.