University of Glasgow | School of Computing Science

Data Evolution and Quality Tracking of Scientific Data Sets

Silviya Sotirova

School of Computing Science
Sir Alwyn Williams Building
University of Glasgow
G12 8QQ

Level 4 Project — March 18, 2015

Data Quality

Abstract

The School of Geographical and Earth Sciences at Glasgow University is currently involved in a consortium based research project which aims to determine the rate and timing of the retreat of the last British-Irish ice sheet. This research involves the collection and analysis of thousands of rock, sediment and organic samples. In order to use the collected data, it needs to be managed, and that depends on the types of data involved, how data is collected and stored, and how it is used - throughout the research. A lecturer from the University of Glasgow - Dr Timothy Storer, and the leader of the consortium - Dr. Derek Fabel, proposed the idea behind this project, which is looking to achieve representation and visualization of the data in a way that will be organized and easy to manipulate. A prototype has been created in order to produce the data with help of matrices and graphs on a html page.
The dissertation demonstrate the project's idea and all of the requirements envisioned by Dr. Storer and Dr. Fabel.

Education Use Consent

I hereby give my permission for this project to be shown to other University of Glasgow students and to be distributed in an electronic format. **Please note that you are under no obligation to sign this declaration, but doing so would help future students.**

Name: _____ Signature: _____

Acknowledgements

I would like to extend my sincere gratitude to my supervisor, Dr. Timothy Storer, for his guidance and insight throughout this project.
I would also like to thank Dr. Derek Fabel, the project's client, for his patience and input which helped shape the final product.

Contents

1 Introduction **1**

1.1 Problem Statement 2

1.2 Dissertation Outline 3

2 Data Quality and Data Visualization - Background research **4**

2.1 First 4

2.2 Second 4

3 Technologies used **5**

3.1 First 5

3.2 Second 5

4 My analysis **6**

4.1 Provided Data 6

4.2 Numerical Analysis 6

4.3 Text Analysis 6

4.4 Image Analysis 6

4.5 recording data in time 6

5 Challenges **7**

5.1 First 7

5.2 Second 7

6 Evaluation **8**

6.1 First 8

6.2 Second 8

7 Future Improvements **9**

7.1 First 9

7.2 Second 9

7.3 The Lazy Dog 9

Appendices **11**

A Running the Programs **12**

B Generating Random Graphs **13**

Chapter 1

Introduction

The BRITICE-CHRONO project is a study to determine the rate and pattern of retreat of the last British-Irish ice sheet, which at one time covered much of the United Kingdom and Ireland. It is organized by the School of Geographical and Earth Sciences at Glasgow University and involves the collection and analyses of thousands of rock, sediment and organic samples in order to establish the dates at which the retreating ice had re-exposed the underlying surfaces at each collection location. The data will help with the illustration of the rate of the ice sheet's retreat. The results of the research will be used in various software applications for testing and modeling the ice sheet retreat in present days. The applications will be helpful for predicting the amount of ice melting arising from warmer temperatures, which may cause tides, relative sea-level changes, slope of the seabed, loss of buttressing ice shelves.These ice sheet models will be tested against observational data.

The team is, also, trying to assess which model implementations of iceberg-calving, grounding lines and ice streams are best suited for predicting retreat of existing ice sheets. The "Data Evolution and Quality Tracking of Scientific Data Sets" project will help them to measure how much of the collected data is correct and useful for the research.

The collection of the data is an difficult process needing lots of preparation and resources. As mentioned above, samples from various types of rocks are gathered, such as carbon dioxide (CO_2) stored within the ice sheets, because CO_2 may affect atmospheric CO_2 and global sea levels as rises in global temperatures reduce the volume of ice through melting. The research team is collecting data by going into the nature and exploring different types of rocks. The measurements made are written and saved in notebooks or by the use of cameras - creating images of a specific environment. However, some of the data might be lost, written or measured incorrectly, because of bad weather. Furthermore, manually typing the collected information into the computer might cause even more lost of data or data misinterpretation.

Dr.Timothy Storer and Dr. Derek Fabel are looking to represent and visualize the gathered data and its missing or imprecise values through the use of a software application which will be using various matrices, graphs and charts. The Computing Science department was approached with the proposal that the application could be created through a student project, with Dr. Fabel serving as client.

A web application for storing all of the data has been created successfully from a student of University of Glasgow - Fiona Steven. Her project allows the access to information of the BRITICE-CHRONO research and, also, to upload, edit, view and search sample data. The "Data Evolution and Quality Tracking of Scientific Data Sets" project" depends from the size and structure of the data set of the web-based application. As the data set grows through various contributions, so is the number of the data visualizations and matrices, too. The researchers record some of the data in different ways. This means that it faces the challenge of maintaining consistencies of data types and formats. All of these variations are affecting the format and consistency of the data set. As a result, strategies have been implemented in order to identify and track these diversity.

Other projects for collecting valuable information may benefit from the functionality of this project. Predicting the location and number of future or hidden bugs in a data is one of the challenges in scientific researches, such as the BRITICE-CHRONO project or paramedical projects. Researchers could use such predictions to identify the critical parts of a data domain, limit them by creating ground rules for manually entering information and facilitate a better planning of testing. However, identifying data errors in a specific set, might not be a bug in another. Data quality and visualization techniques and using software applications will highlight and indicate where the researchers need to focus and which data is in great importance for various scientific calculations and measurements.

1.1 Problem Statement

The idea behind the ”Data Evolution and Quality Tracking of Scientific Data Sets” project was born from the concept of being able to track code quality with various tools and software systems - e.g. Jenkins, open-source continuous integration server. Thanks to these tools, software developers are able to track the commits done during the implementation of a project. Moreover, these software systems provide plot plugins for illustrating the work progress of a specific application, they are able to show and suggest which part of the code might need improvement and they visualize statistics made on the code tests, builds and trends. Software applications are written from people which make mistakes while writing code - the same goes with data for scientific researches. A tool for tracking, managing and recording changes of data in time will be convenient and effective for reforming quality. Dr. Storer proposed a research in the field of tracking data quality, as a starting point of which is the BRITICE-CHRONO project.

The process of improving the data set of the BRITICE-CHRONO project includes a web page storing all of the data. Before the creation of the web page, the collection of the data has been done by entering it in hundreds of file spreadsheets. These files were being combined into a single master spreadsheet which contained a reduced selection of the data fields, because some of the fields were not found within the individual sheets. Managing the data and creating viewings was a daunting procedure even with the incomplete portion of the data set. As a result of it, a web-based application was created, which is replacing the master spreadsheet and allowing data uploading and management of sample records, access to the complete data set by authorized users and creation of an interactive map on which sample data is displayed. However, even in the successfully created web page, the data could still be inaccurate, entered incorrectly or in different text formats.

The prototype, created during this project, aims to explore the data set of the BRITICE-CHRONO research and to outline errors that have occurred in the data with the help of technology.

One of the biggest challenges was managing data inconsistencies. The cause of them is the fact that different individuals are completing the forms with information in vast styles depending on the data format. The variations of the forms are affected from the alteration of the sheets and from different interpretations of the data types. The BRITICE-CHRONO project comprises a huge number of researchers, which means that the density of errors in the data is extremely high.

Another challenge that is taken under consideration is developing extensive knowledge of the data that has been collected. The research is in the field of geology and with my narrow understanding of the subject, relationships between the data fields would be misinterpreted, leading to errors in the data visualizations such as graphs and calculations.

The final set of challenges to be resolved was the exploring of the data domain, the analysis that needed to be done in order to present the data in the most effective way for the researchers and the ability to design and create a practical web page where all of the data visualizations are shown.

1.2 Dissertation Outline

The next section is concerned with the data inconsistencies that software engineers have encountered and ways of handling them.

In Chapter 3 introduces the technologies used for the project - which software tools and applications are the best for data visualization and tracking data quality, why they are the most useful one and how they are used. It includes details about other familiar existing projects.

In Chapter 4 begins with describing the structure and content of the provided data. This is followed by different types of analysis done on the data - numerical analysis, text analysis and image analysis. The chapter finishes with introduction to algorithms of how the data could be recorded in time.

Chapter 5 goes into detail through the challenges faced along the implementation.

The Evaluation chapter documents both the testing and client's evaluation process.

Finally, the Conclusion chapter discusses the state of the prototype at the end of the project. It, also, explores future ideas for improvement and further development. The Conclusion chapter is followed by the Appendices and Bibliography.

Details of the submission content can be found in the first appendix.

Chapter 2

Data Quality and Data Visualization - Background research

2.1 First

2.2 Second

Chapter 3

Technologies used

3.1 First

3.2 Second

Chapter 4

My analysis

4.1 Provided Data

4.2 Numerical Analysis

4.3 Text Analysis

4.4 Image Analysis

4.5 recording data in time

Chapter 5

Challenges

5.1 First

5.2 Second

Chapter 6

Evaluation

6.1 First

6.2 Second

Chapter 7

Future Improvements

7.1 First

7.2 Second

The quick brown fox jumped over the lazy dog. The quick brown fox jumped over the lazy dog. The quick brown fox jumped over the lazy dog. The quick brown fox jumped over [**?**] the lazy dog. The quick brown fox jumped over the lazy dog. The quick brown fox jumped over the lazy dog. The quick brown fox jumped over the lazy dog. The quick brown fox jumped over the lazy dog.

7.3 The Lazy Dog

The quick brown fox jumped over the lazy dog. The quick brown fox jumped over the lazy dog. The quick brown fox jumped over the lazy dog.

The quick brown fox jumped over the lazy dog. The quick brown fox [**?**] jumped over the lazy dog. The quick brown fox jumped over the lazy dog. The quick brown fox jumped over the lazy dog. The quick brown

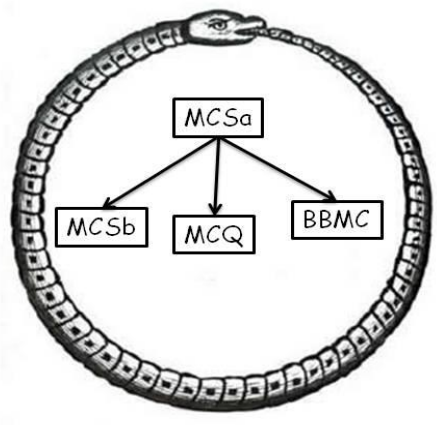

Figure 7.1: An alternative hierarchy of the algorithms.

fox jumped over the lazy dog. The quick brown fox jumped over the lazy dog.

Appendices

Appendix A

Running the Programs

An example of running from the command line is as follows:

```
> java MaxClique BBMC1 brock200_1.clq 14400
```

This will apply $BBMC$ with $style = 1$ to the first brock200 DIMACS instance allowing 14400 seconds of cpu time.

Appendix B

Generating Random Graphs

We generate Erdós-Rënyi random graphs $G(n, p)$ where n is the number of vertices and each edge is included in the graph with probability p independent from every other edge. It produces a random graph in DIMACS format with vertices numbered 1 to n inclusive. It can be run from the command line as follows to produce a clq file

```
> java RandomGraph 100 0.9 > 100-90-00.clq
```