Relational Data Science & Application Development

**The George University Washginton**

**12/14/13**

Seung Kwon An Shichiji Ishiyama

**Abstract**

In almost any industry, the use of data has increased as technology began to develop to aid in data collecting. The term “big data” can be seen as a major role in a business such as the IT world and the Healthcare world. Data in business refers to any material or knowledge in a particular field that will benefit the business with improved efficiency or an increase in revenue. With so much data to process and analyze, the final product of researched data will prove to be a very valuable resource for a variety of tasks in a business. The goal of this project is to create and combine open source software projects that intend to collect or “mine” the data, analyze the data through computer coded algorithms, and create an export of the data as a readable report that will contain useful information for businesses to utilize. The project will carry on with using Java as the fundamental backbone of the software and will integrate other successful software to create a very user friendly data analyzer.

Table of Contents

[Brief Overview 3](#_Toc374897180)

[Title: 3](#_Toc374897181)

[Concept: 3](#_Toc374897182)

[Objectives and Goals 3](#_Toc374897183)

[Target Users 3](#_Toc374897184)

[Review of Related Literature 4](#_Toc374897185)

[Methodology 6](#_Toc374897186)

[Conceptual Brief Workflow 7](#_Toc374897187)

[Algorithm Map and Components 8](#_Toc374897188)

[GUI (Graphical User Interface) 9](#_Toc374897189)

[EXE. 9](#_Toc374897190)

[Start Menu 9](#_Toc374897191)

[Data Selection Menu 9](#_Toc374897192)

[Report Menu 1 10](#_Toc374897193)

[Query Menu 10](#_Toc374897194)

[Report Menu 2 10](#_Toc374897195)

[Weka 11](#_Toc374897196)

[Process 11](#_Toc374897197)

[Types of Process 12](#_Toc374897198)

[Transporting the data into Weka 12](#_Toc374897199)

[Data & Queries Matching 12](#_Toc374897200)

[Obtaining related news and website 12](#_Toc374897201)

[Obtaining similar and opposite type data within the user-uploaded data 12](#_Toc374897202)

[Calculating Data Matching Percentages 12](#_Toc374897203)

[Decisions 13](#_Toc374897204)

[EXIT (Terminator) 13](#_Toc374897205)

[Conclusion 13](#_Toc374897206)

[Resources 14](#_Toc374897207)

# Brief Overview

## Title:

　 Relational Data Science & Application Development　(The title of the application has not been decided.)

## Concept:

　　 Our project is to develop a user-friendly data science application. Most data mining and science tools are difficult to use and highly focused on the professional use. Today, the amount of data is exponentially increasing every year but the increase in the population of the people who can explore and discover something out of big data are not adequately proportionate to the growth speed of data amount. We considered this problem and came up with the idea about making user-friendly and easy to use but certainly functions as a data mining and science tool.

## Objectives and Goals

The objectives and goals in our project are:

* To make graphically and methodologically user-friendly interface and process in data science
* To change the preoccupation among people that data science is way beyond average users’ mean
* To design adequate algorithm of this application development and make it into practice

## Target Users

People who are interested in data mining and science but do not know what to do or how to use the tools for them.

# Review of Related Literature

The software selected for review is data mining software called WEKA. The Waikato Environment for Knowledge Analysis (WEKA) was first drafted in 1992 by the University of Waikato located in New Zealand. This software primarily focuses on analyzing data sets using different algorithms pre-programmed into WEKA. The algorithms are compiled to be available in WEKA at any time and the ability to add or use the algorithms without restrictions is one of its key features. Another reason for selecting this software to study and review is due to the fact WEKA was written in Java. From algorithms to the interface of the software were all developed in Java.

WEKA is an open source project that allows for other Java code to be integrated into the software or reversely, implement the algorithms in WEKA to another project or software that is compatible with Java. WEKA contains unique features that attempt to allow the users to have a better understanding of what data analysis process is going on. WEKA consists of a section called the “workbench”, which contains numerous algorithms for users to try and study the results all in one area. This feature proves to be very useful for scientists or data analysists that are attempting to create multiple results.

The workbench feature is one of the interesting options that WEKA allows in order to experiment with data. Another feature is called “experiment mode”, which allows the user to modify a series of conditions that might bring very different results when compared to the traditional data analysis methods. This is a useful feature if a particular project requires varying algorithms or to simply test how new modified algorithms might be beneficial. From altering a single variable to implementing different conditions to an existing algorithm, the experiment feature is a great tool to have when discovering new alternate ways of analyzing data, even if it seems to be an unconventional way.

In general, WEKA is used primarily for data analysis that involves large amounts of numerical data. It is widely used by scientists and data researchers. In regards to our current project proposed, our project will be focused on a smaller scale of numbers and data to analyze. Our goal through this project is to analyze business data, which might be financial reports, sales invoice, monthly inventory, etc. Also, our project will be catered to small local businesses, which generally do not contain extensive amounts of data. We chose to take this route on the project, in order to create a reachable goal. Without extensive knowledge of the data analysis field, it will be impossible to create complex algorithms at this time.

Our hope through the project is that more business related data mining and analysis algorithms will be available for people to use as an open source. We will also add the option of making our algorithms easily accessible for use with the WEKA software to show our support of the open source mission WEKA is striving towards. When it comes to business data analysis there are software and companies that already offer the service, but at a very high cost. We want our project to be free of charge and create software that everyone will be able to understand and use. Currently, even for WEKA, the interface is not very straightforward, and it is difficult for a regular user to make use of the program even if he/she desires to do so. The goal of the capstone project is to create a small business data analyzer, but at the same time create software anyone can use at any experience level.

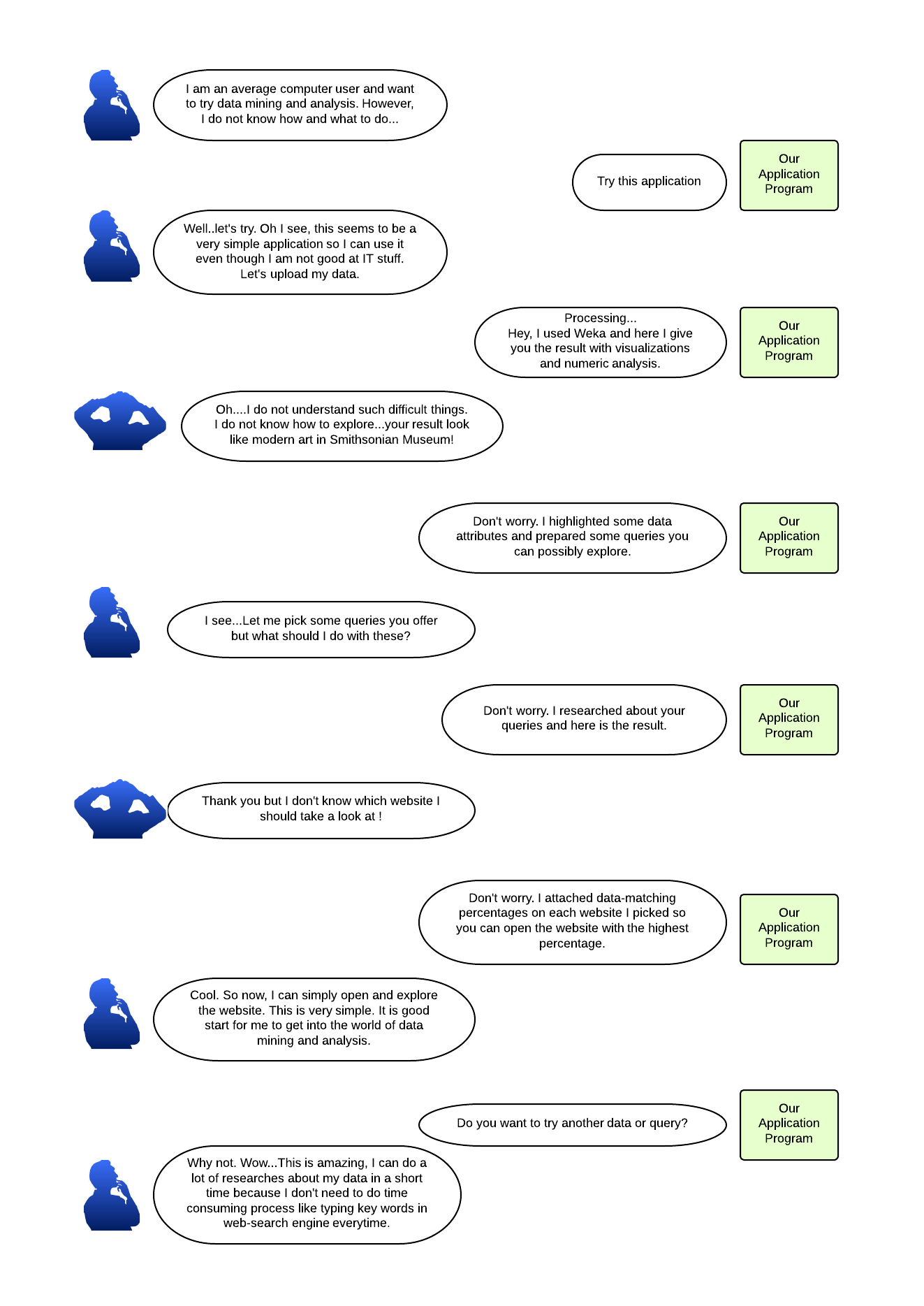
# Methodology

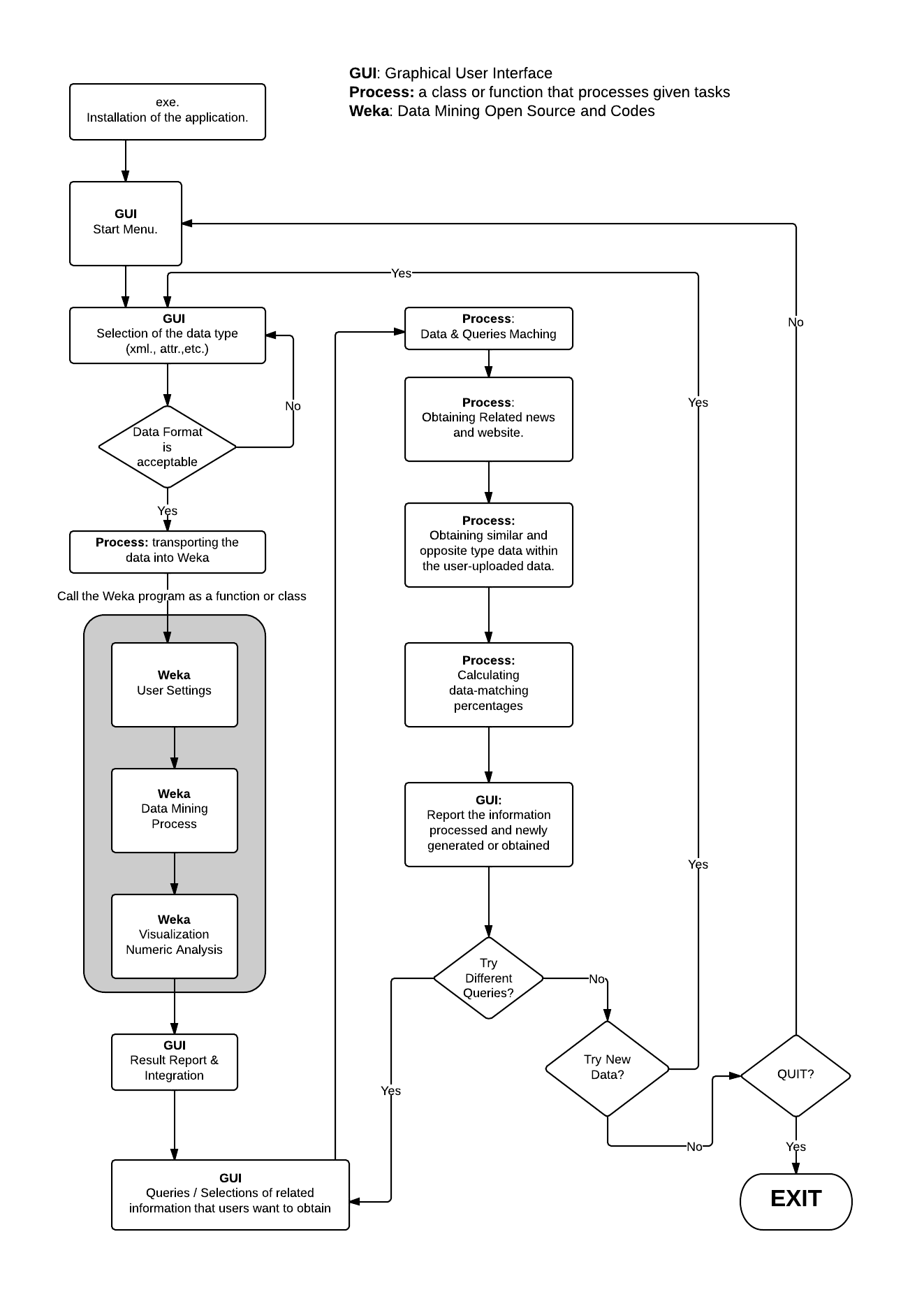
As for the methodology portion of the paper, there were two major methods carried out. For the background portion of the project, we researched texts, books, manuals, and any information related to the topic of data mining/analysis to gain an in depth knowledge of the subject. Due to the fact that data analytics is a large field pertaining to a variety of industries, the focus was shifted to business analysis in general. The second portion of the methodology consisted of actually running a variety of software that was similar to the objectives in the capstone project. Because the project focused on developing open source software, we researched and downloaded open source projects that carried out data analysis. The process of learning how to operate the individual software proved to be a challenge due to each of the software having its own methodology of carrying out the data analysis.

Most of the research regarding texts was carried out through the George Washington University’s library. Through research we were able to identify scholarly journals, business related articles, eBooks on the data mining software themselves, and texts on methodology of data analysis methodology. Through text material we were able to gain basic fundamental knowledge of how the data mining industry operated, along with basic concepts of how algorithms are constructed.

The second half of the research as mentioned before consisted of a hands on approach to learning the actual software themselves. This gave the much needed insight to how simple applications accepted data files and what steps were necessary to allow the software to attempt to analyze the data. Although there were variances from one software to another, the basic requirements for analyzing data stayed similar in terms of concepts.

# Conceptual Brief Workflow



Algorithm Map and ComponentsGUI (Graphical User Interface)   
  
 This is the part that is needs to be coded. The GUIs basically generate dialog boxes with menu buttons on the display and give users menu and choices. Based on the users’ decisions, the GUIs lead them to the next step and export user option data to other classes that process unique tasks required to complete the main class process. Most GUIs include processes but these are separated from the process part because the processes here are highly dependent on the users’ decisions.   
Types of GUI

### EXE.

**GUI**

This GUI appears when the user clicks the exe. file for this application and there should be several pages used for acquiring users’ agreement, brief explanation of how this application will be installed and uninstalled.

* Start Menu Start menu shows choices for users to do: start the application, read more information, quite, and option menu. Each menu directs the users to the different dialogue box for each task.
* Data Selection Menu  The users need to upload data here and the reason why we create an extra data selection menu even though Weka has the page which does the same task, is because we need to limit the number of available data formats to facilitate the process part we are going to code. In this part, the menu just asks users to upload data and if the data does not match to the format we specify such as Excel and attribute files, error and retry message will appear on the dialogue box.

### Report Menu 1

After exporting the uploaded data into Weka, calculated and visualized data will appear on the Weka application but these data are not user-friendly: these are quite difficult to understand and use. Therefore, our application makes the data generated by Weka user-friendly; all highlighted or remarkable numeric data will be sorted out and put into tables and other non-remarkable data will be sorted into another table. In other words, this GUI gives the users simplified data that are important in the database or sheet they uploaded.

* Query Menu In this menu, the users are asked to choose some key words that they want to obtain to compare with the generated data in the previous section. Perhaps, some users do not have any idea about what data they should bring up to compare with the generated data so offering choices top them in this part is considered to be helpful. Therefore, the users will choose the key words from the choices we prepare such as news, datasheet, database, and any other website.
* Report Menu 2 After the users decide what to compare with, the data-matching process will start behind the Query Menu. After all the calculations and data matching are done, the Report Menu 2 appears and the new results will be displayed. The result contains the list of the factors the users have chosen in the previous section. For example, if the user chose website and related data, Report Menu 2 will offer the list of websites and related data available online. Additionally, each item offered by the application will have the percentages about data-matching (the algorithm for this part is in a process). At this point, this menu is the end part of our application but we consider putting something additional here if we have enough time when the beta version for this application is completed.

# Weka

**Weka**

This part is composed by open source codes from Weka, a data mining application. This is the core part of our data mining application and we use two-dimensional data visualization and calculation. There is no additional coding into the Weka open-source code so we simply use the codes for the core complex data calculations and visualizations. We use Weka for the data mining core process and since Weka allows extracting Java files from its Jar file containing all the program files used in Weka, we are able to add our original GUI, class, and functions.

# Process

**Process**

This part is tasked to process the data before and after Weka process. Some of them will be coded from a scratch and others might have open-source code. The conceptual basic solution of all calculation part we are going to code is that we extract data-mined data from Weka and put them into the class and functions written by us. Eventually, the data will be compared with similar or related objects such as related news, database, and products. This information is obtained automatically by data-matching process part and consequently shows how closely and relatively all data match.

## Types of Process

### Transporting the data into Weka

This process is going to be used for uploading the data the user uploaded on the GUI we code and then transport the data into Weka. Since we want to limit data formats, we need to have this extra process to upload data.

### Data & Queries Matching

After the user chose automatically-generated or suggested queries in the Query Menu, this process refine the priorities among the queries and determine how and where to get related data. The key words that will be put on web search engines and used for obtaining similar and opposite type data within the user-uploaded data are determined in this process and the words are going to be passed to the next process.

### Obtaining related news and website

This process uses the string variables passed from the previous process and input them into web search engines such as Google, Yahoo, and bing. The numbers of websites that will be shown in our GUI is dependent on user’s choice but the default setting limits up to 5 websites from each search engine the users pick.

* Obtaining similar and opposite type data within the user-uploaded data This part is going to explore numeric attributes within the user-uploaded data. Then, the data that has numerical similarities and radical diffrences with the key attributes that are contained in the queries the users have chosen in the Query Menu GUI. All numerical data and related string variables explored in the process will be passed to the Report Menu 2 and be shown on the GUI.
* Calculating Data Matching PercentagesThe users might demand some tips to figure out what and which websites are more suitable and dependable for the particular key word they picked in the Query Menu. At this point, we are still researching about how to increase the accuracy in data-matching other than counting the number of the use of the key words on each website. After the certain calculations are done, numeric variables and string variables (URL and name) will be passed to the Report Menu 2 and sorted in tables.

# Decisions

**Decisions**

The users need to choose whether yes or not for the required action in each decision process. We haven’t decided whether we do make decision process in the above-described GUI or make a simple GUI for each decision.

**Terminator**

# EXIT (Terminator)

This is the end process of this application and the users can simply choose whether they quite using the application or not. When they choose to quite, the application ends processing and closes.

# Conclusion

We understand the difficulties in achieving success in our application development since we haven’t had enough software development skills and knowledge. Thus, we finally decided to development the application in different method from what we proclaimed at the beginning of the semester. Our current plan focuses on the better use of data mining application and enabling the users to explore further and related information of the data that are produced via Weka by highly sophisticated calculations and methodology. In other words, our tasks are to design user-friendly GUIs and some simple process that use the data-mined data from Weka and keep the data mining function. At the end the class file we develop will encapsulate the whole core programs of data mining. We have already started designing the GUI and this process should not be a difficult part. Some difficult is about calling functions and class files from Weka to our original GUI. Also, data transfer and variable conversions within every section requiring converting variables is also difficult part.   
 In conclusion, it could be said that our application development plan is improved since the virtual presentation. We refined our objectives, goals, application size, and algorithm; we made a project size much smaller than the first draft idea. We are quite excited in completing our project as we proposed in this rough draft.

# Resources

Araki, K.; Furukawa, Z.; Jingde Cheng, "A general framework for debugging," Software, IEEE ,

vol.8, no.3, pp.14,20.

Chou, Chen-Huei (2013). Design and Implementation of jAHP: A Java-based Analytic

Hierarchy Process Application. International journal of computer applications, 62 (15), p. 35 - 41

# Corbitt, T. (2003). Business intelligence and data mining. *Management Services, 47*(11), 18. Retrieved from http://search.proquest.com/docview/234332816?accountid=11243

# Frank, E., Holmes, G., Kirkby, R., and Hall, M. (2002). Racing committees for large datasets. In Proceedings of the International Conference on Discovery Science, pages 153–164. Springer-Verlag.

Froelich, J., Ananyan, S., & Olson, D. L. (2005, Winter). Business intelligence through text mining. *Business Intelligence Journal, 10*, 43-50. Retrieved from http://search.proquest.com/docview/222638608?accountid=11243

Hornik, K., Buchta, C., & Zeileis, A. (2009). Open-source machine learning: R meets weka. *Computational Statistics, 24*(2), 225-232. doi:http://dx.doi.org/10.1007/s00180-008-0119-7

Sauban, M. and Pfahringer, B. (2003). Text categorisation using document profiling. In *Proceedings of the 7th European Conference on Principles and Practice of Knowledge Discovery in Databases*, pages 411–122. Springer.

# 