

Visual Analytics for Effective Interdisciplinary Collaboration

Computational & Visual Analytics Tools for Educational and Applied Multidisciplinary Computer Science

RICHARD A. ALÓ PhD., ERIN HODGESS PhD., HOOMAN HEMMATI, DUBER GOMEZ-FONSECA,
SARAH JENNISCA, LILIAN ANTUNES, TIA PILAROSCIA

Center for Computational Sciences
University of Houston Downtown
One Main Street Suite 722-S
Houston, Texas 77002
UNITED STATES OF AMERICA
ralo@uh.edu

Abstract: In the data deluge of the modern era, interdisciplinary collaboration is an absolute necessity for identification, extraction and visualization of pertinent data. By creating context sensitive tools, it is possible to significantly expedite data analysis with tools optimized for an individual's knowledge base and expertise. R is a free software environment for statistical computing and graphics that provides remarkable flexibility, which unfortunately requires familiarity with the scripting language used to use effectively. Tools developed through this project are designed with end-users disciplines in mind and thus provide a customized GUI interface to assist in extracting relevant data from a pool of information through specialized algorithms and visualization tools.

Key-Words: Visual Analytics, R, Interdisciplinary Computational Science

1 Introduction

The modern era has been sometimes referred to as the information age due, more than all else, to the scientific advances that have increased data generation in almost all areas, ranging from physical attributes of tangible objects to characteristics of human relations and speech. The next frontier and the invisible wall that all scientists and engineers face today is no longer the limitation of equipment sensitivity, storage or computational power, but instead the extraction of meaning from aggregate data collected meticulously from various sources and represented in various mediums. In essence humanity is shifting into the age of knowledge, the prerequisite of which is meaningful collaboration across disciplines and digestion the massive quantity of available information and observations into meaningful concepts.

In theoretical science it has become pertinent to collaborate with experts across various specializations to detect meaningful patterns from various datasets and mediums and extrapolate the proverbial whole picture. A simplistic analogy

is that of a simple stone tossed in a pond. Instead of just measuring amplitudes and wavelengths of ripples to approximate the location where the it fell, it is important to analyze the sound produced, the distorted reflections on the water surface and the disturbances in sediment and the bottom of the pond to full understand what dropped in the pond and why. Simple single disciple analysis is limited by the data it can process and explain, but through simultaneous analysis and collaboration of various experts on datasets in different mediums, it is possible to surpass these limitations and understand events. One of the many obstacles to collaborative work is the quick meaningful analysis of data in every discipline and effectively communicating results across disciplinary and language barriers. A first step in resolving this obstacle is effective integration of visual analytics across all disciplines.

Data analysis and visualization, particularly those pertaining to statistical data characteristics, play a key role in deducing meaning from the modern data deluge. These tools allow effective representation of results that can be communicated independent of language. Visual models

and graphs, especially interactive mediums such as animated and 3D, allow interpretation of results with greater understanding.

This project tackles this problem through the development of discipline specific tools that will allow users to perform 3 critical tasks: 1) Filter and aggregate information into usable data, 2) Identify the underlying patterns and phenomena of the data and 3) Use simple but powerful multifaceted visualization tools to quickly and effectively analyze the data and its underlying characteristics. The R is a well established and versatile programming language and statistical tool with package based enhancement system that allows for specialized and extended functionalities. It is the best choice to utilized R packages for our purposes, since it will allow for a diverse set of base functionalities and is easily expendable allowing for the creation of customized tool-sets. This project takes advantage of the Affinity Research Group (ARG) structuring to best approach the problem.

2 Affinity Research Groups

Affinity research groups (ARGs) are interdisciplinary research teams composed of undergraduate researchers and advised by faculty members active in the computational research fields. For this project, the research team is composed of undergraduates majoring in the fields of statistics, mathematics and computer science. Some members also have additional background in biology and economics, which helps with identifying certain unique requirements for real world applications. Even though the disciplines represented by the group may not be sufficient to create comprehensive tools that are applicable to all fields, they do ensure that the resulting solutions will feature a fundamental level of flexibility required for a core structure that can later be customized and specialized through further interdepartmental consultation and research.

3 Background

The main tool used in this research is the R programming language. R is a name shared by both the programming language and software environment for statistical computing and graphics. The R language is a standard tool used by statisticians for the development of statistical software and data analysis. Although the windows binary and

GUI interface are used for this project, R is available on various operating systems in both source code and binary format. This project takes advantage of existing R features and plug-ins to build a better overall suite of statistical visualization and analysis tools.

3.1 Plug-ins: RCmdr and RGL

R does not incorporate a graphical user interface (GUI) for statistical functionalities. Even though R features commands that are generally console based, it still allows for the development of GUI interfaces. Rcmdr is one such interface developed to provide a multitude of statistical GUI features in addition to the console functionality. It is menu driven and capable of importing, modifying and visualizing data. Visualizations with Rcmdr are generally limited to 2D plots and diagrams. In order to take advantage of 3D models, the RGL plug-in will also be used. RGL is a C++/OpenGL based real time rendering tool that will allow for interactive 3D models to be integrated into the project.

3.2 Future Upkeep and Maintenance

An important aspect of the project is to ensure reliability, accessibility and forward compatibility. On all accounts, R proves to be the optimal candidate as the foundational technology for the project. Utilizing existing packages plug-ins as pre requisites promotes stability and reliability while taking advantage of future improvement in prerequisite components. The dissemination mechanism for R packages allows users to download only the packages (and prerequisites) necessary for the task at hand, thereby enhancing accessibility and ease of use, and sidestepping the difficulties associated with using distributing and updating monolithic solutions. A modular approach may also prove with future implementations associated with technological breakthroughs such as readily available cloud computing.

4 Method

The project includes two phases and multiple segments. Phase I is an introductory phase designed to introduce participants to aspects of computational statistics and R. Phase II is the algorithm and software development phase. Seg-

ments of this project include research, algorithm development, algorithm implementation, GUI design and testing. The project segmentation applies to both phases but is more pronounced in phase 2.

4.1 Tools

The primary tools used for the project are the R environment, Rcmdr and RGL. Other software tools, such as Microsoft Visual Studio may be used for testing and component development (C/C++ only), however this will be kept to a minimum to ensure forward compatibility and to avoid using binaries that will not benefit from future upgrades to R.

4.2 Phase I Approach

As mentioned previously the approach taken was in two phases. The first phase revolved around allowing researchers to acclimate to the R environment while simultaneously establishing the limitations of available functionalities in R. With consideration to the research objectives, sample statistical problems were solved computationally with R and available plug-ins. By creating and testing standalone functions for data processing, aggregation, analysis and visualization many conclusions regarding design and implementation limitations were reached resulting in a dynamic set of development guidelines. These guidelines are not strictly enforced and evolve with the project; however they encourage researchers to create code that can be easily used and understood by others in the group. During this phase, research was also conducted regarding possible needs and requirements for real world statistical problems. Of particular interest were problems and solutions relating to data set sizes, the computational capacity at users disposal and the possibility of noisy data. Some of these points were simulated in sample problems during this phase.

4.3 Phase II Approach

The approach in this phase is highlighted by overlapping subgroups. The research team created problem solving and research groups or committees. Though every member contributes, this organization allows for efficient use of time and expertise. The subgroups created consisted of the following:

- **Software Development:** Focused almost entirely on implementation and encoding of developed algorithms as well as integrating coded solutions implemented by other members.
- **Statistical Analysis:** Primary team researching suitable techniques for identifying statistical characteristics of particular data sets.
- **Algorithm Development:** This subgroup functions to develop general purpose and/or specialized mathematical algorithms as they are required.
- **Specialization Research:** Team member within this sub-group specialize in identifying program aspects that change with specialization such as key words, data set styles and more.
- **GUI Tools and Design:** Members of this subgroup work with the specialization research team to design a suitable user interface.

Subgroups work independently but results are presented to other members on a regular basis. These presentations allow for feedback and ideas that can be incorporated into the projects. During this phase, research is more segment specific and is carried out by the subgroups.

5 Progress and Results

The first phase of the project has been complete and the second phase is well under way. Current results include an outline of technical requirements for the final product, partial results in algorithm development and some completed visualization tools and GUI components. Limited testing of tools is currently underway, though upon completion tools will be distributed to collaborators in different institutions and departments for thorough applied testing.

5.1 Technical Requirements

Results from research show that the developed tools will almost always be utilized with large data sets. A 64bit variation of R would allow for memory usage of over 4 GB, however even though Rcmdr and RGL may be compatible, not every plug-in is 64bit ready. It would then be impractical to assume that all users will have 64bit

systems, sufficient RAM, or that they will only use R with tools developed by the project. As a result data aggregation is an absolute necessity to ensure reliability and utility. Other research results indicate that a significant portion of data in most disciplines can be classified as some form of time-series data and also that statistical information about the data is used in almost all cases. It was also observed that data analysis and visualization can become extremely discipline specific, thereby require a great deal of consideration when discipline specific tools are designed. However it should be noted that general visualizations such as histograms and plotted graphs, however customized, must be retained but some discretion should be exercised to avoid unnecessary options and tools that reduce usability of the final product. Finally it was noted that denser representations of data, such as interactive 3D models and animation, are desirable in almost all cases. Such visualizations are ideal for quick analysis and summarization of data and can communicate results more effectively. Figure 1 is an example one such visualization of data.

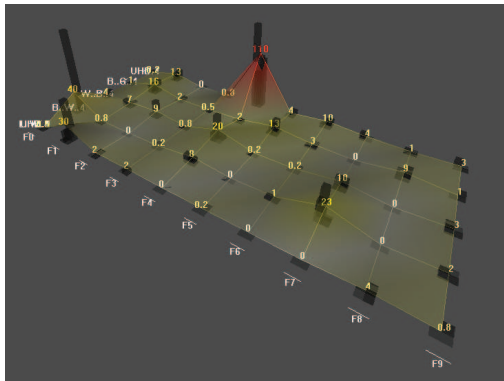


Figure 1: Clustered bar representation along with "mean" mesh

5.2 Algorithm Development

Most algorithms developed and/or research related to data processing. Aggregation was the primary aspect of the research, especially methods that could be applied to a stream of data. This approach would allow the program to effectively aggregate data as it is being read from the input source, thereby reducing processing time and memory costs. Rectangular window functions were the first and most generally favorable aggregation mechanism tested. A variant was created, by using small overlapping ranges for the functions used and applying successive layers to

aggregate results in an accurate fashion. The process resulted in points that traced the line more accurately, provided that the data points were ordered. Further testing is required to identify benefits of this method versus various others in un-ordered (or random) data sets.

Other algorithm being investigated are optimized sorting methods. Though processes such as quick-sort have favorable speed, they require complete datasets to be effective and as such do not conform to the original technical requirements imposed by memory limitations. Multi-pass algorithms are being developed and optimized to allow sorting data without pre-loading large data-sets into memory. Sorting algorithm are important due to the potential savings in computing time, however it should be noted that in many cases, such as time-series, data should not be reordered.

5.3 GUI Design

The project supplements the GUI provided by Rcmdr through the addition of discipline specific menus. Each menu item is a drop-down revealing more options with discipline specific terminology. Select options provide pop-up windows allowing for a more refined control over the analytical process. The customized vocabulary creates a user friendly environment and also allows for a degree of internationalization, since a majority of professional engineers and scientist are familiar with the english nomenclature of concepts particular to their field of expertise.

A final requirement that is outlined by the outcome of GUI research is the problem of visual strain. This can be remedied through better hardware, such as larger screens, however it is necessary to integrate features for cases where that is not an option. The R color scheme is generally acceptable, but menus for adjustment of colors (such as a high contrast scheme) and font size will be added. Also of note is that the correct management of multiple interfaces will also heavily impact this issue.

5.4 Visualization

Researchers are also researching discipline-customized visualizations that are clear while communicating a more complete and informative overview of data they represent. Current works are predominantly focused on interactive 3D models such as the one depicted in figure 1.

In this example one set of data is represented as clustered bars while another set pertinent to the context of research is presented as mesh overlay allowing users to compare and contrast information very quickly. In general the tool-set is designed so that diagrams such as this can be generated with only a few clicks rather than the copious amounts of scripting normally necessary for customized graphics. Further efforts are being made to create more enhanced visualization such as displaying geo-coded data on maps and introducing animations.

6 Future Work

Figure 2 illustrates the final objective of these tool sets as it would pertain to disciplines for which customized packages are created. To accomplish this various tasks must be undertaken. These tasks can be categorized as those contributing to short term goals and those contributing to the overall progress of the project. The short term goal is a working prototype for a generic tool. During phase II the general GUI structure and design will be completed, including GUI and general visualization tools. Additionally data aggregation and sorting mechanics should also mature and be tested. Testing for the prototype will

be done using pandemic, financial and environmental data. Long term goals for the project include finalized packages for various disciplines with specialized tool-sets and algorithms. Design and testing for these components will be done as a modification and extension of the prototype. These enhancements will require extensive inter-departmental work with students and experts to ensure suitability of the tool for its intended discipline. Final testing will be done externally by allowing other researchers to utilize products in actual research environments and provide feedback. Finally it is the intention of this project to provide extensive documentation so that packages can continue to be maintained and updated beyond the duration of this particular undergraduate research project.

7 Acknowledgements

Material generated by this project is based upon work supported by the U.S. Department of Homeland Security's International Center for Command Control and Interoperability - Visual Analytics for Command, Control and Interoperability Environments (VACCINE) Center under Award Number 2009-ST-061-CI0001.

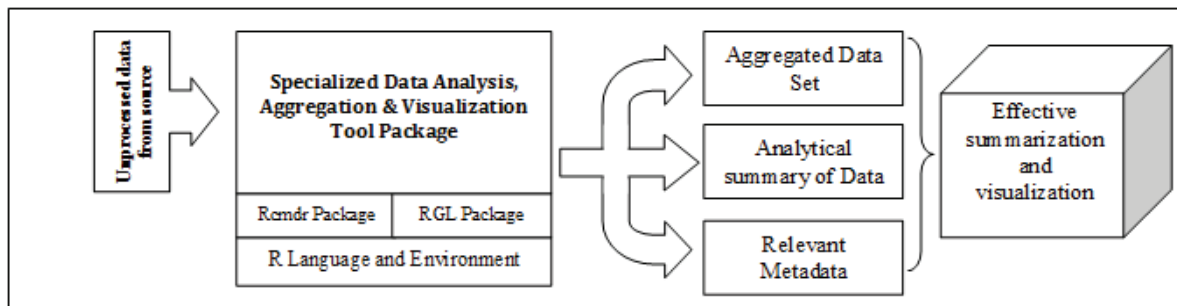


Figure 2: Life cycle of data processed using project tools

Visual Analytics for Effective Interdisciplinary Collaboration

Computational & Visual Analytics Tools for Educational and Applied Engineering on an International Scale

Richard A. Aló, PhD.¹, Erin Hodgess², PhD., Hooman Hemmati³, Duber Gomez-Fonseca⁴, Sarah Jennisca⁵, Lilian Antunes⁶, Tia Pilaroscia⁷

¹Center for Computational Sciences at the University of Houston Downtown, Houston, USA, ralo@uh.edu

²Center for Computational Sciences at the University of Houston Downtown, Houston, USA, hodgesse@uh.edu

³Center for Computational Sciences at the University of Houston Downtown, Houston, USA, hemmatih@uhd.edu

⁴Center for Computational Sciences at the University of Houston Downtown, Houston, USA, fonsecad@uhd.edu

⁵Center for Computational Sciences at the University of Houston Downtown, Houston, USA, sjennisca@yahoo.com

⁶Center for Computational Sciences at the University of Houston Downtown, Houston, USA, antunesl@uh.edu

⁷Center for Computational Sciences at the University of Houston Downtown, Houston, USA, tpilaroscia@gmail.edu

INTRODUCTION

In the data deluge of the modern era, an efficient method for identification, extraction and visualization of pertinent data is an absolute necessity. To this end interdisciplinary and international collaborations are a definite necessity; however there are many obstacles that prevent effective communication between engineers and researchers in such an environment. Communicative devices such mathematical diagrams and computer models serve to mitigate the problems arising from disciplinary and language barriers.

The goal of this project is to create discipline-specialized tools that take advantage of an individual's knowledge base and expertise in an interdisciplinary collaborative environment and allowing them to present the result of their analytical work using expressive models. These tools are intended to significantly improve the speed and efficiency of data analysis.

This project takes advantage of an interdisciplinary undergraduate research group led by two faculty advisors, also known as an Affinity Research Group (ARG), to investigate and implement a toolset for visual analysis of statistical data in various disciplines and areas. As depicted in figure 1, the end result of the project will be tools that allow for effective extraction of information from raw data.

METHOD

The research is conducted in two phases. The first consisted of researchers familiarizing themselves with the intended development environment while investigating its potential applications and limitations. The second phase is focused on the development, testing and eventual

distribution of the toolset. This phase was executed by subdividing the research team into task specific subgroups, with each individual contributing to three separate subgroups. This allowed researcher to remain task oriented while remaining aware of the interdependencies of separate components. Results from each subgroup were periodically distributed to other members to allow for deeper collaboration and constructive feedback.

TOOLS

The primary tools used for the project are the R statistical environment and its Rcmdr and RGL packages. This project takes advantage of existing R features and plugins to build a better overall suite of statistical visualization and analysis tools. The project takes advantage of the Graphical User Interface (GUI) provided by the Rcmdr package and the RGL interactive 3 dimensional (3D) rendering tools in the RGL package.

Utilizing existing R packages as prerequisites takes advantage of future improvement in these components. Also the dissemination mechanism for R packages allows users to download only updates for the packages/prerequisites necessary for the task at hand, thereby sidestepping the difficulties associated with updating monolithic solutions. Future implementations may be updated to take advantage of technologies such as cloud computing.

PROGRESS AND RESULTS

The first phase of the project has been complete and the second phase is well under way. Results are categorized as technical requirements, algorithms and software

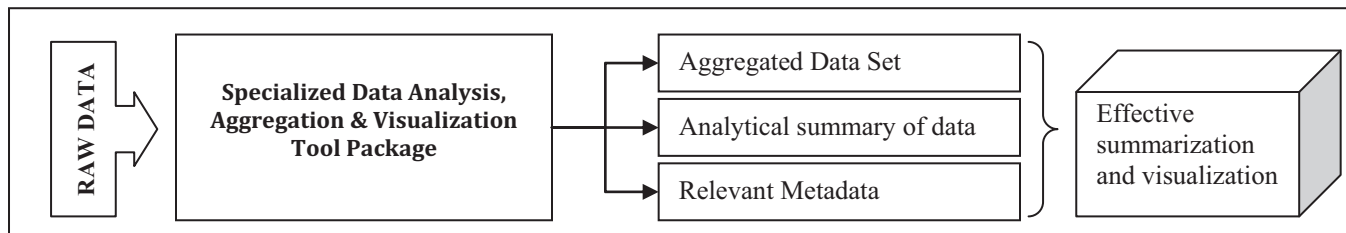


Figure 1 - Overview of toolset application

