Interactive Data Visualization Program to Analyze Word Count Frequencies Over Time

Aryn Grause

March 8, 2011

1 Objective

The goal of this project is to build an interactive software tool which will produce a meaningful data visualization from the word frequencies for a given document over time.

2 Introduction and Background

Word frequencies or "word counts" are vital artifacts used to study human behaviors. Word frequency is the count of an individual word within a document. They are commonly used by the following fields: stenography, informational science, psychology, pedagogy, linguistics and other similar disciplines. These areas use frequencies to study behavior and structure. My tool is designed to be used by educational researchers. Specifically, Dr. Bandeen, Director of Technical Assistance, Education, and Support with the WorkSource Standards & Integration Division in Employment Security, Washington State, who will be exploring how language in educational policy has evolved over time. However, if successful it has the potential to have a broader impact, after all, "[word frequency are] one of the favorite and most traditional issues in the history of quantifying approaches to language" [5].

My project is composed of two distinct components: 1) Data gathering - receiving the input documents and storing the word frequency results within a database and 2) data visualization - portraying the data in a meaningful and effective manner. Both components will use a wide range of resources that will be highlighted in the next section. This project incorporates many

key fields in computer science, such as file manipulation, database management, data visualization, and human computer interaction.

The data gathering component of the tool will take multiple versions of the same document as its inputs. Although it could be used more generically, the first application will be to observe the modification of word use over time within a particular educational document. Each version of the document will be parsed into individual words or user specified phrases and the frequencies of these findings will be recorded. Naively calculating the word frequencies for one document is trivial, however a more sophisticated tool presents several challenges. One of these challenges will be to store the information in a meaningful manner for later analysis. Another will be to remove unimportant words, such as "the", "as", ect. Also, incorporating the ability to count user defined phrase will add complexity. Yet another challenge will be to distinguish the roots of words with multiple suffixes thus providing a more accurate frequency count.

A key tool that I will use to address these issues will be a database. This will allow me to incorporate archives of unimportant words, common tense structures, and provide an excellent platform to integrate with my visualization component. In order to ensure the efficiency of the database, it will be designed in Boyce-Codd Normal Form (BCNF) [8]. This will eliminate duplicates and anomalies within the dataset which allows for quick access through query statements.

The data visualization component will take the massive amount of information stored in the database and display it in an educational yet concise manner. Data visualization is a visual interpretation of data, simple examples include tables and graphs while a more complex example would be a dashboard. Designing a meaningful data visualization can be difficult. The designer has many choices about axis lengths, how much information to portray, layout, type, variables, etc. These choices can skew the users interpretation of the data. The research in the field of data visualization focuses on how to correctly display data and reduce the likelihood of incorrect interpretations of the data. Due to this concern, the design of the data visualizations will be careful and concise. By following the fundamentals and guidelines established by the field experts and incorporating a novel interactive user interface, I believe this tool will successfully be able to build an accurate visual representations of the evolution of word frequencies in a given document.

The fundamentals of data visualization were formally establish by two individuals: William Cleveland and Edward Tufte. Both have researched how to clearly communicate information in a static visual. Cleveland's research focuses on displaying scientific results objectively. His philosophy revolves around minimizing the effort a viewer must use in order to understand the visual while making differences large enough to perceive [2]. Tufte's philosophy is known for extravagant compositions. He is best known for popularizing the use of small multiples. Small multiples is a series of small graphs arranged together so the user can easily distinguish the differences [8]. Currently, small multiples is the most common approach to displaying variable

change over time. However, my research will focus on developing an interactive alternative to this small multiples approach.

The use of animation is a relatively new movement within data visualizations. Small multiples usually incorporates line graphs to display change in variables over time [4]. My idea is to use an animation to "flip" through these line graphs rather than displaying them in a small multiple. This will allow me to condense the information displayed in an visual while hopefully heightening the contrasts between document versions. The current design of the tool is hierarchical in form. The first level of visual will display the **n**th most frequent words (e.g. 50 most frequent words) in the first version of the document in a bar chart. A scroll bar will allow the user to animate through the other versions of the document to see how the bar chart changes over time. The second level displays a fine grain level of information, specifically it will allow the user to view information relating to an individual word. Later versions of the tool will try to incorporate and explore alternative animations and visuals.

3 Research Plan

As stated above, the goal of this project is to provide an efficient tool for observing word frequencies. I have already done a fair amount of research and have established an initial "blueprint". I plan to execute my project in the following phases:

Phase One: Build a prototype web interface in JavaScript and PHP that will be used to test the program (the real interface will be implemented with the data visualization). This interface will have the ability to upload files and display the returned visualizations.

Phase Two: Design an efficient database using MySQL. The current design calls for the database to contain five distinct elements: document information (edition, release date, etc), word frequencies, user defined phrases, unimportant words, and suffixes.

Phase Three: Develop the document parsing and word count program in Java. A key component will be to connect to the database and the user interface. I will rely on common software design practices to ensure flexibility and modularity.

Phase Four: Use Protovis [6], a common data visualization software tool, to explore different static visualizations of the information. In this phase, I will also develop the rudimentary animations to determine which visualization works best.

Phase Five: Implement the interactive and dynamic version of the visualizations. This will be done using JavaScript on the web interface. This will allow the user to quickly "flip" through the visualizations produced by Protovis.

Phase Six: Data visualization verification. For the scope of this project, I will be working closely with Dr. Bandeen to satisfy her needs. An extension of this project could include a more through user interaction testing.

4 Contributions

This project makes the following contributions:

- Other similar tools in existence, such as Humble Finance [6], Wordle [4], ect, to the best of my knowledge do not use interaction within their applications; nor do they allow the user to pass multiple documents as inputs. This greatly limits the user's ability to compare multiple versions of a document.
- Interactive data visualization is relatively novel approach to displaying information. This tool would add to the current state of the art in data visualization, while expanding the practicality of data visualization creators available on the wed.
- It will aid in Dr. Bandeen's research on word use in educational legislation.
- Finally, this tool will be beneficial to my future in computer science. This project will prepare me for graduate school in two ways: it will expose me to sophisticated research and disseminating my findings.

5 Conclusion

My proposed research into data visualizations will not only impact the viability and usability of the word frequency tool described above but has the potential to make a significant contribution to the field of data visualization. If successful, the tool I will develop will be the first available on-the-web application that incorporates an interactive animation for the compact display of complex information. It will establish a new standard and demonstrate that interactive data visualization tools are not only feasible by highly applicable.

References

[1] H. Akiba and K-L. Ma. An Interactive Interface for Visualizing Time-Varying Multivariate Volume Data, *Proc. EuroVis* 2007.

[2] W. Cleveland. The Elements of Graphing Data. Monterey: Wadsworth Advance Books and Software,1985.

[3] Feinberg, Jonathan, *Wordle*. Wordle tm, 2009. Web. 3 March 2011. < http://www.wordle.net/>

[4] S. Few. Data Visualization Past, Present, and Future, " [Whitepaper] Cognos Innovation Center, 2007.

[5] *HumbleFinance.* Humble Software Development, 2011. Web. 3 March 2011. < http://www.humblesoftware.com/>

[6] I. Popescu. Word Frequency Studies. Berlin: Mouton de Gruyter, 2009.

[7] *Protovis*. Stanford University, 2010. Web. 27 January 2011. < http://vis.stanford.edu/protovis/>

[8] E. Tufte. Envisioning Information. Cheshire: Graphics Press, 1998.

[9] J. Ullman and J. Widom. A First Course in Database System. Upper Saddle River, NJ, pages 88-92. Pearson Education, Inc. 2008.

[10] N. Zumel. Good Graphs: Graphical Perception and Data Visualization, www.win-vector.com, 2009.