Assisting Reading and Analysis of Text Documents by Visualization

Ross J. Maloney
BE, MEngSc, BAppSc(Maths), GradDipComp, GradDipMaths

This dissertation is presented in fulfilment of requirements for the degree of Doctor of Philosophy of Murdoch University

August 2005

School of Information Technology
Murdoch University

I declare that this dissertation is my own account of my research and contains as its main content work which has not previously been submitted for a degree at any other tertiary institution.

_____

Ross James Maloney

Abstract

The research reported here examined the use of computer generated graphics as a means to assist humans to analyse text documents which have not been subject to markup. The approach taken was to survey available visualization techniques in a broad selection of disciplines including applications to text documents, group those techniques using a taxonomy proposed in this research, then develop a selection of techniques that assist the text analysis objective. Development of the selected techniques from their fundamental basis, through their visualization, to their demonstration in application, comprises most of the body of this research. A scientific orientation employing measurements, combined with visual depiction and explanation of the technique with limited mathematics, is used as opposed to fully utilising any one of those resulting techniques for performing complete text document analysis.

Visualization techniques which apply directly to the text and those which exploit measurements produced by associated techniques are considered. Both approaches employ visualization to assist the human viewer to discover patterns which are then used in the analysis of the document. In the measurement case, this requires consideration of data with dimensions greater than three, which imposes a visualization difficulty. Several techniques for overcoming this problem are proposed. Word frequencies, Zipf considerations, parallel coordinates, colour maps, Cusum plots, and fractal dimensions are some of the techniques considered.

One direct application of visualization to text documents is to assist reading of that document by de-emphasising selected words by fading them on the display from which they are read. Three word selection techniques are proposed for the automatic selection of which words to use.

An experiment is reported which used such word fading techniques. It indicated that some readers do have improved reading speed under such conditions, but others do not. The experimental design enabled the separation of that group which did

decrease reading times from the remaining readers who did not. Measurement of comprehension errors made under different types of word fading were shown not to increase beyond that obtained under normal reading conditions.

A visualization based on categorising the words in a text document is proposed which contrasts to visualization of measurements based on counts. The result is a visual impression of the word composition, and the evolution of that composition within that document.

The text documents used to demonstrates these techniques include English novels and short stories, emails, and a series of eighteenth century newspaper articles known as the *Federalist Papers*. This range of documents was needed because all analysis techniques are not applicable to all types of documents. This research proposes that an interactive use of the techniques on hand in a non-prescribed order can yield useful results in a document analysis. An example of this is in author attribution, i.e. assigning authorship of documents via patterns characteristic of an individual's writing style. Different visual techniques can be used to explore the patterns of writing in given text documents.

A software toolkit as a platform for implementing the proposed interactive analysis of text documents is described. How the techniques could be integrated into such a toolkit is outlined. A prototype of software to implement such a toolkit is included in this research. Issues relating to implementation of each technique used are also outlined.

Acknowledgments

To:

My parents, Pat and Jim Maloney, who made undertaking this research possible;

Erica Daymond of Edith Cowan University, for her assistance in formulating the contents of the reading experiment so as to achieve the required objectives;

Prof. Tamàs Gedeon who as a supervisor provided initial guidance, particularly into the necessary background for this research that I had not considered;

Susan Alexander who through her direct and indirect support was instrumental in ensuring this research was brought to completion;

Dr Graham Mann whose interest, encouragement, and assistance in this research both preceded, and then continued through, him assuming the role as a supervisor;

Dr Andrew Turk who saw the research through from inception to completion and during that time provided guidance, encouragement, and constant timely reviews, which in combination demonstrated supervision skills second to none.

Thank you all.

Contents

List of Tables

List of Figures