

---

# **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 demonstrate 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

<b>1</b>	<b>Making pictures from words</b>	<b>1</b>
1.1	Overview . . . . .	2
1.2	The problem and its importance . . . . .	7
1.3	Methodology of this research . . . . .	11
1.3.1	Assumed interactions . . . . .	12
1.3.2	Research hypotheses . . . . .	13
1.3.3	Assumptions . . . . .	15
1.3.4	Approach . . . . .	17
1.3.5	Relationships between the concepts in this research . . . . .	19
1.4	Implementation testbed – ‘Serine’ . . . . .	21
1.4.1	Graphical user interface for control . . . . .	24
1.4.2	Implementation . . . . .	25
1.5	Summary . . . . .	27
<b>2</b>	<b>Some literature which guides visualization</b>	<b>30</b>
2.1	A taxonomy for text document visualization . . . . .	31
2.2	Scientific visualization . . . . .	41
2.2.1	Specific examples . . . . .	45
2.3	Information visualization . . . . .	50
2.3.1	Specific examples . . . . .	52
2.4	Visualization of high dimensions . . . . .	56
2.4.1	Specific examples . . . . .	57
2.5	Text visualization . . . . .	59
2.5.1	Specific examples . . . . .	60
2.6	Consolidating the visualization techniques . . . . .	63

2.7	Non-visual text analysis techniques . . . . .	65
2.7.1	Comparing between texts . . . . .	65
2.7.2	Examination of a stand-alone text . . . . .	69
2.8	Use to be made of existing techniques . . . . .	71
2.9	Summary . . . . .	72
<b>3</b>	<b>Visualization via document type</b>	<b>74</b>
3.1	Markup . . . . .	76
3.1.1	Markup occurs in two forms . . . . .	77
3.1.2	Interpretation of markup . . . . .	79
3.2	Plain text . . . . .	80
3.2.1	Defining plain text . . . . .	81
3.2.2	Significance of such a definition . . . . .	82
3.3	Approaches to document classification . . . . .	82
3.3.1	Classification by subject content . . . . .	84
3.3.2	An alternative – classification by functional objects . . . . .	87
3.3.3	Linking appropriate analysis to classification . . . . .	92
3.4	Summary . . . . .	93
<b>4</b>	<b>Analysis of documents by word lists</b>	<b>96</b>
4.1	Text transformation . . . . .	98
4.1.1	Isolating sentences and paragraphs . . . . .	98
4.1.2	Algorithm for isolating sentences from paragraphs . . . . .	100
4.1.3	Algorithm for isolating words . . . . .	102
4.1.4	Software implementation . . . . .	107
4.2	Zipf’s laws . . . . .	110
4.2.1	Word frequency verses rank . . . . .	111
4.2.2	Frequency of occurrence of word numbers . . . . .	115

---

4.2.3	A resulting message principle . . . . .	117
4.3	Word frequency lists . . . . .	118
4.3.1	Variability in word frequency lists . . . . .	119
4.3.2	Derivation of a reference word frequency list . . . . .	123
4.3.3	Influence of the type of document on word frequency . . . . .	128
4.4	Analysis applications using word frequencies . . . . .	133
4.4.1	Methods using word frequency for classification . . . . .	134
4.4.2	Influence on classification of text sampling method used . . . . .	139
4.5	Top-Tail Truncation Technique ( $T^4$ ) . . . . .	143
4.5.1	Design . . . . .	143
4.5.2	An example of positioning frequency regions . . . . .	146
4.5.3	Implementation of $T^4$ as interactive software . . . . .	147
4.5.4	Fading frequent words . . . . .	151
4.6	Summary . . . . .	152
<b>5</b>	<b>Effectiveness of word fading</b>	<b>155</b>
5.1	Positioning of this experiment . . . . .	156
5.1.1	Text reduction experiments . . . . .	156
5.1.2	Speed of reading . . . . .	163
5.1.3	Delivery medium . . . . .	163
5.2	Design of experiment . . . . .	166
5.2.1	Hypothesis . . . . .	166
5.2.2	Method . . . . .	167
5.2.3	Implementation . . . . .	170
5.3	Preliminary result processing . . . . .	174
5.3.1	Verification of experiment's length . . . . .	174
5.3.2	Resolution of dataset anomalies . . . . .	175
5.4	Reading speeds using all participants . . . . .	176



---

5.4.1	Summary of the data . . . . .	177
5.4.2	Analysis of results . . . . .	181
5.4.3	Discussion . . . . .	182
5.5	Comprehension errors using all participants . . . . .	184
5.5.1	Summary of the data . . . . .	184
5.5.2	Analysis of results . . . . .	187
5.5.3	Discussion . . . . .	188
5.6	Those whose reading time did decrease . . . . .	189
5.6.1	Summary of the data . . . . .	190
5.6.2	Analysis of results . . . . .	192
5.6.3	Discussion . . . . .	194
5.7	Summary of experimental findings . . . . .	196
5.8	Summary . . . . .	197
<b>6</b>	<b>Visualizing numeric data</b>	<b>200</b>
6.1	A classic author attribution dataset . . . . .	201
6.1.1	Background . . . . .	202
6.1.2	Patterns to observe . . . . .	203
6.2	Parallel Coordinate plots . . . . .	208
6.2.1	Effect of noise and scaling on the Parallel Coordinate plots . . .	210
6.2.2	The number of Parallel Coordinate plots required . . . . .	215
6.3	Visualization of the <i>Federalist Papers'</i> variables . . . . .	220
6.3.1	Parallel axes plots . . . . .	222
6.3.2	Star plots . . . . .	226
6.3.3	Colour Maps . . . . .	229
6.3.4	Reducing the number of variables by Sammon plots . . . . .	233
6.3.5	Summary of multi-dimensional visualization section . . . . .	237
6.4	Fractal dimension – a toolkit number . . . . .	237

6.4.1	Fractal dimension theory . . . . .	238
6.4.2	Algorithm used . . . . .	242
6.4.3	Some practical issues in fractal dimension measurement . . . . .	249
6.4.4	Summary of this fractal dimension section . . . . .	253
6.5	Variation trending . . . . .	254
6.5.1	An overview of QSUM . . . . .	255
6.5.2	Environment used to examine QSUM . . . . .	256
6.5.3	Handling of Shewhart and Cusum charts . . . . .	260
6.5.4	Author attribution using QSUM and fractal dimension . . . . .	265
6.5.5	Concluding remarks on QSUM . . . . .	273
6.6	Summary . . . . .	273
<b>7</b>	<b>Visualizing document composition</b>	<b>276</b>
7.1	Two styles of sentence diagramming . . . . .	278
7.2	‘Sentence stick’: A sentence visualization . . . . .	281
7.2.1	Word class selection and refinement . . . . .	281
7.2.2	Special word classes . . . . .	287
7.2.3	Representation of the word classes . . . . .	288
7.2.4	Implementation . . . . .	289
7.3	Calibration of this visualization . . . . .	292
7.3.1	Sentences used in this tuning . . . . .	293
7.3.2	Occurrences of the word classes . . . . .	295
7.3.3	Direction assignment . . . . .	298
7.4	Extension for paragraph visualization . . . . .	305
7.4.1	Two approaches to visualizing sequential paragraphs . . . . .	305
7.4.2	Implementation in the ‘Serine’ software . . . . .	307
7.5	Two examples of potential applications . . . . .	311
7.5.1	Difference in writing style . . . . .	312

7.5.2	The <i>Federalist Papers</i> . . . . .	315
7.6	Summary . . . . .	320
<b>8</b>	<b>Conclusions and future directions</b>	<b>323</b>
8.1	Critique of this research . . . . .	324
8.1.1	Placing the visualizations considered into perspective . . . . .	325
8.1.2	Evaluation against the research hypotheses . . . . .	326
8.1.3	Beyond the limits of this research . . . . .	335
8.2	New in this thesis . . . . .	337
8.3	Future directions . . . . .	338
8.4	Overall conclusion . . . . .	341
<b>A</b>	<b>Snap shots of the visualization literature</b>	<b>343</b>
A.1	Examples of scientific visualization . . . . .	343
A.2	Examples of information visualization . . . . .	354
A.3	Examples of high dimension visualization . . . . .	365
A.4	Examples of text visualization . . . . .	371
<b>B</b>	<b>Directions given to experiment participants</b>	<b>377</b>
<b>C</b>	<b>Reading experiment questions</b>	<b>379</b>
C.1	Text - Clarke . . . . .	379
C.2	Text - Tritten . . . . .	382
C.3	Text - Haldeman . . . . .	386
C.4	Text - Banks . . . . .	390
<b>D</b>	<b>Examples of experiment's reading screens</b>	<b>392</b>
<b>E</b>	<b>Auxiliary Parallel Coordinate plot Information</b>	<b>395</b>
E.1	Representation of a point in Parallel Coordinates . . . . .	395

E.2 Parallel Coordinate plots showing linear dependency . . . . . 396

E.3 All 6 Parallel Coordinate plots for the *Federalist Papers'* variables . . . . 399

**References** **406**

# List of Tables

1.1	The hypotheses base of each Chapter . . . . .	15
2.1	Revised Shneiderman(1996) taxonomy for information visualization . .	34
2.2	Proposed taxonomy for Information Visualization . . . . .	37
2.3	A sample of scientific visualization techniques . . . . .	47
2.4	Distribution in proposed taxonomy of sample scientific visualizations .	49
2.5	A sample of information visualization techniques . . . . .	53
2.6	Distribution in proposed taxonomy of sample information visualizations	55
2.7	A sample of techniques for visualizing high dimensions . . . . .	58
2.8	Distribution in proposed taxonomy of sample visualizations of high di- mensions . . . . .	59
2.9	A sample of text visualization techniques . . . . .	61
2.10	Distribution in proposed taxonomy of sample text visualizations . . . .	62
2.11	Distribution of all the examples across the taxonomy with text visual- izations highlighted . . . . .	64
3.1	Functional objects present in various types of documents . . . . .	90
3.2	Techniques of document analysis considered and the functional objects used . . . . .	92
4.1	Parameters for examples using frequency verses rank considerations . .	112
4.2	Parameters for examples using frequency verses number considerations	115
4.3	Lists of frequently occurring words in children’s books . . . . .	122
4.4	Summary of the British National Corpus unlemmatised list information	125
4.5	Summary of the British National Corpus data used . . . . .	126
4.6	50 most frequently occurring words in British English . . . . .	127
4.7	Word variability in the British National Corpus . . . . .	129

4.8	Word occurrences in various forms of British English . . . . .	131
4.9	Occurrence percentages of the words of interest in the sample documents	135
4.10	Correlation coefficients ( $r$ ) obtained when classifying sample documents	136
4.11	Sample document percentage word occurrence explained by doc. type .	137
4.12	Changing classifications of documents through their length . . . . .	141
4.13	Changing classification of text through two sample documents using a smaller sample size . . . . .	142
5.1	Summary of word reduction experiments cited from the literature . . . .	162
5.2	Relationship between factors and data filling for the experiment . . . . .	168
5.3	Sources of texts used in these experiments . . . . .	171
5.4	Words remaining unfaded in texts after the cut off conditions used . . .	172
5.5	Summary statistics for ANOVA reading time cells using all participants	179
5.6	Summary statistics for ANOVA comprehension error cells using all par- ticipants . . . . .	185
5.7	Summary statistics for Group A and B reading times and comprehen- sion data . . . . .	191
5.8	Summary statistics for ANOVA reading time and comprehension error cells for Group A . . . . .	191
6.1	The literature's allocation of authors to the <i>Federalist Papers</i> . . . . .	203
6.2	Variables used for analysis of each Federalist paper . . . . .	206
6.3	Correlation coefficient matrix of the <i>Federalist Papers'</i> data set . . . . .	207
6.4	Sequences of vertices for use in Parallel Coordinate plots . . . . .	218
6.5	A 9 vertex edge matrix in use constructing a Hamiltonian circuit . . . . .	220
6.6	Successful QSUM <i>habits</i> included in Farrington (1996) . . . . .	257
6.7	Numbers of the <i>Federalist Papers</i> that employ reference citations . . . . .	260
6.8	QSUM results employing fractal dimensions for selected Federalist pa- pers . . . . .	271

7.1 Inclusion of the 50 most frequently occurring words in British English  
by the proposed word classes . . . . . 285

7.2 Colour assignments used for the word classes . . . . . 292

7.3 Test sentences used to determine the word class direction allocation . . . 294

7.4 Word classes that appear in the test sentences . . . . . 296

7.5 Word class directions tested . . . . . 298

  

8.1 Relationship between the visualizations of this research and the litera-  
ture . . . . . 325

# List of Figures

1.1	Examples of text document visualizations considered later in this thesis	4
1.2	Meta-level interactions assumed in this research . . . . .	6
1.3	Association map outline of the components of this research . . . . .	18
1.4	Analysis process with concepts and visualizations used in this research .	20
1.5	Logical interconnection between the elements of the ‘Serine’ toolkit . . .	23
1.6	Making selections in ‘Serine’ while displaying the text to be analysed . .	25
1.7	Progress map from Chapters 1 to 2 . . . . .	28
2.1	Visualization of astrophysics results after Keller & Keller(1993) page 66 .	39
2.2	Visualization of pollutant measurements after Keller & Keller(1993) page 53 . . . . .	40
2.3	Visualization of dance steps after Tufte(1990) . . . . .	51
2.4	Progress map from Chapters 2 to 3 . . . . .	73
3.1	Progress map from Chapters 3 to 4 . . . . .	94
4.1	State-machine for grouping characters into generalised words . . . . .	103
4.2	Interconnection of implementation software modules in ‘PerlBreaker’ and intermediate files it generates . . . . .	109
4.3	Word frequency verses rank for email examples . . . . .	113
4.4	Word frequency verses rank for article and novel chapter examples . . .	114
4.5	Frequency verses word numbers for email and novel chapter examples .	116
4.6	The words present in Federalist paper 36 . . . . .	148
4.7	Screen shot of ‘Serine’ top/tail truncation technique module applied to Federalist paper 36 . . . . .	150
4.8	Screen shot of ‘Serine’ frequent word fade module applied to Federalist paper 36 . . . . .	152
4.9	Progress map from Chapter 4 to 5 . . . . .	154



---

5.1	Strip-chart plot of measured reading times against word fading type (including outliers which were removed before further analysis) . . . . .	177
5.2	Strip plot of reading time data used in analysis against text being read .	178
5.3	Histogram of all reading time data . . . . .	179
5.4	Strip plot of reading time data against comprehension error . . . . .	180
5.5	Strip-chart of percentage comprehension errors against word fading . .	184
5.6	Histogram of percentage comprehension errors . . . . .	186
5.7	Reading time and comprehension error histograms for improved and non-improved groups . . . . .	190
5.8	Scatter plots of reading times and comprehension error against word fading in the group with decreasing reading times (Group A) . . . . .	192
5.9	Progress map from Chapter 5 to 6 . . . . .	198
6.1	Examples of data plotted as Parallel Coordinate and Star plots . . . . .	209
6.2	Effect of change in scaling of axes on Parallel Coordinate plots . . . . .	212
6.3	Parallel Coordinate plots of simulated data with noise added . . . . .	214
6.4	Parallel Coordinate plots of Federalist papers – arrangement 1 . . . . .	223
6.5	Star plots of Federalist paper’s statistics – self scaled . . . . .	228
6.6	Colour Maps of the variables measured in the Federalist papers . . . . .	232
6.7	Sammon plot of the <i>Federalist Papers</i> data . . . . .	236
6.8	Changing length of a line due to measurements taken . . . . .	239
6.9	Positioning of boxes over data to determine their fractal dimension . . .	243
6.10	Construction for determining the box cover of a line . . . . .	244
6.11	A method of determining whether a line enters a box . . . . .	245
6.12	Log/log plot of box counts for the straight line example . . . . .	250
6.13	Plot of data that produced the wiggly line example . . . . .	251
6.14	Log/log plot of box counts of the wiggly line example . . . . .	252
6.15	Relationship of software and files used in this part of this research . . . .	258
6.16	Plot of Shewhart and Cusum charts for the same input data . . . . .	261

---

6.17	Box covering data obtained for the Cusum chart . . . . .	263
6.18	Box covering data used to find Cusum chart fractal dimension . . . . .	264
6.19	QSUM plots of sentence length, 23lw, and 23lw + ivw Federalist paper 17 . . . . .	266
6.20	QSUM plots of sentence length, 23lw, and 23lw + ivw Federalist paper 37 . . . . .	267
6.21	Plot of data used to calculate the fractal dimension of Federalist paper 17 . . . . .	269
6.22	QSUM plots of appended texts from selected <i>Federalist Papers</i> . . . . .	270
6.23	QSUM plots of appended texts from selected Federalist papers for au- thor attribution . . . . .	272
6.24	Progress map from Chapter 6 to 7 . . . . .	275
7.1	Similarity of ‘sentence stick’ and molecular visualizations . . . . .	277
7.2	Sentence diagramming using syntactic parse-tree and Reed-Kellogg di- agram . . . . .	278
7.3	A Read-Kellogg diagram with sentence recovery difficulties . . . . .	280
7.4	The Octagon of directions and their corresponding word classes . . . . .	288
7.5	Change in stick model visualizations resulting from direction alloca- tions to reduce model spread . . . . .	300
7.6	Change in stick model visualizations resulting from word class direc- tion allocations to minimize spread and diagram complexity . . . . .	304
7.7	Visualization without rotation of example text containing multiple sen- tences and paragraphs . . . . .	306
7.8	Diagrammatic representation of stacking multiple paragraphs for visu- alization . . . . .	309
7.9	Visualization with rotation of example text containing multiple sen- tences and paragraphs . . . . .	310
7.10	Visualization without rotation of emails of sample emails . . . . .	313
7.11	Visualization with rotation of emails of sample emails . . . . .	314
7.12	Extracts from ‘sentence sticks’ for the <i>Federalist Papers</i> known to be writ- ten by Hamilton . . . . .	317

---

7.13	Extracts from 'sentence sticks' for the <i>Federalist Papers</i> known to be written by Madison . . . . .	318
7.14	Extracts from stick diagrams for the <i>Federalist Papers</i> for which authorship is disputed . . . . .	319
8.1	Meta-level interactions in 'human in charge' text document analysis . .	324
8.2	Use of the 'Serine' software toolkit . . . . .	334
D.1	Clarke screen with no word fading . . . . .	393
D.2	Clarke screen with Type 1 word fading . . . . .	393
D.3	Clarke screen with Type 2 word fading . . . . .	394
D.4	Clarke screen with Type 3 word fading . . . . .	394
E.1	Parallel coordinate representation of linear dependency . . . . .	397
E.2	Parallel Coordinate plots of the <i>Federalist Papers</i> – arrangement 1 . . . .	400
E.3	Parallel Coordinate plots of the <i>Federalist Papers</i> – arrangement 2 . . . .	401
E.4	Parallel Coordinate plots of the <i>Federalist Papers</i> – arrangement 3 . . . .	402
E.5	Parallel Coordinate plots of the <i>Federalist Papers</i> – arrangement 4 . . . .	403
E.6	Parallel Coordinate plots of the <i>Federalist Papers</i> – arrangement 5 . . . .	404
E.7	Parallel Coordinate plots of the <i>Federalist Papers</i> – arrangement 6 . . . .	405