

An Integrating Suite of Tools to Assist Investigation of Non-Originality

Fintan Culwin
London South Bank University
fintan @ lsbu.ac.uk

Thomas Lancaster
University of Central England
Thomas.Lancaster@uce.ac.uk

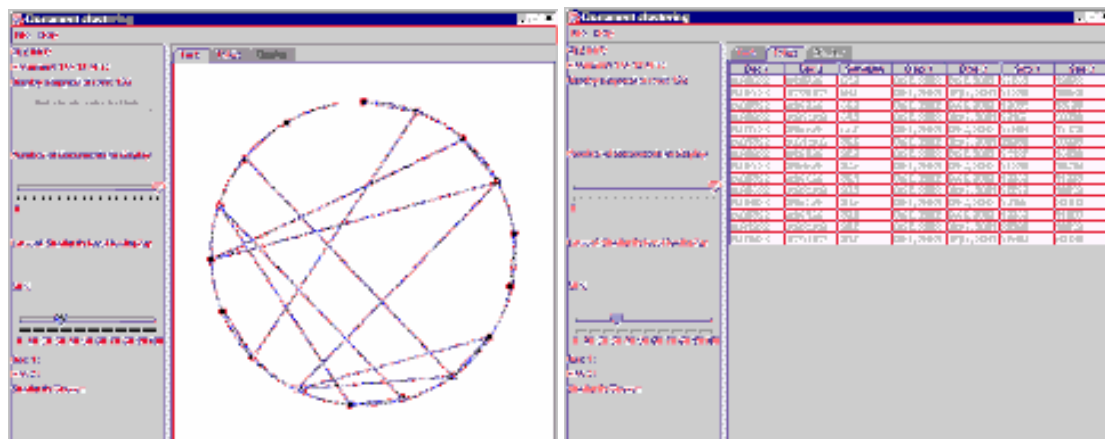
Introduction

This paper will introduce the design and implementation of a suite of tools intended to assist in the detailed investigation of non-originality. Culwin & Lancaster's four stage model identifies the third stage, confirmation, as being the one where the synergy between human and computer is most crucial. The preceding stages, collection and analysis, can be largely automated and the succeeding stage, investigation, is non-tool dependant. Hence the major design intention in these tools has been to provide effective and efficient interactive representations of the similarity intersections of two, or more, documents. However, the tools also support the analysis phase. They are described as integrating rather than integrated as they were initially developed as a series of separate tools and are currently being packaged together. All tools are written 100% in Java and so are readily available on the Web for all environments. They are free of charge and descriptions of the precise algorithms used can be obtained from the author. The tools have been developed at the Centre for Interactive Systems Engineering (CISE) and can be located at <http://cise.lsbu.ac.uk/tools.html>.

Plotted Ring of Analysed Information for Similarity Exploration (PRAISE).

In its current implementation PRAISE is used for intra-corporal investigation. That is, given a set of text documents, typically a set of student submissions all written to the same assignment specification, it will analyse them to compute a gross similarity index between every possible pair of documents. This information can then be represented graphically on an interactive torc representation, or as a table view of all documents, or as an alternative table view of a cluster of similar documents.

Figure 1. PRAISE showing the torc view (left) and full table view (right)



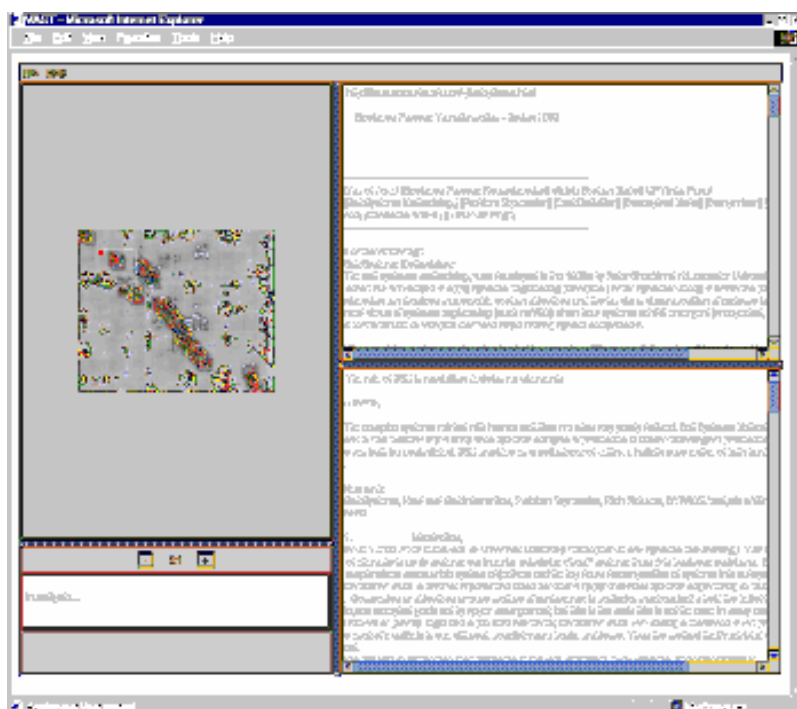
PRAISE is illustrated in Figure 1. The torc view has controls that allow the number of documents shown and the degree of similarity shown to be varied. These settings then constrain the amount of information shown on the tables. When a single document on the torc is selected the cluster of documents connected to it is shown, and that cluster can be investigated in the cluster table view. When a pair of documents are selected the extent of similarity within the pair can be investigated using the VAST tool as described below.

PRAISE is currently being enhanced to allow all the URLs contained in the set of documents to be automatically extracted and resources that they point to downloaded and included in the analysis. The belief here being that although an individual student who has illicitly included material from a Web source in their submission is unlikely to cite that URL, it is possible that it has been cited in another submission. This enhancement should be complete by the time of the conference. A further enhancement is to allow an OrCheck style visualisation, as described below, to be spawned from the torc cluster view in the same way that a VAST visualisation is currently spawned from a torc pair-view.

Visualisation and Analysis of Similarity Tool (VAST).

VAST supports the interactive exploration of a pair of documents. Two documents may be believed to be similar in some way, but the extent and location of the similarity may not be immediately obvious from a visual inspection. VAST makes a fuzzy match between all parts of one document and all parts of a second document, highlighting those parts that have a greater degree of similarity. This representation can then be interactively explored with the corresponding parts of the two documents highlighted.

Figure 2 VAST in operation



VAST is illustrated in Figure 2. The VAST visualisation is shown at minimal magnification on the left of the tool and the two documents that are being investigated in the upper and lower text panes on the right. The contents of the upper document are plotted left to right in the visualisation, and the contents of the lower document top to bottom. The extent of similarity at every possible pair of locations is shown, with darker shades indicating greater similarity. Hence the upper left of the image represents that start of both documents whilst the lower left, for example, is the start of the upper document and the end of the lower document. The 45 degree diagonal represents identical physical locations in both documents.

The interpretation of the visualisation shown is that upper document is longer than the lower one, and that the entire contents of the lower document are contained within the upper document, with some original material towards the start and the end. The documents can be investigated by zooming the visualisation in and out whilst marking areas of interest with the rectangular marquee. The parts of each document indicated by the marquee are shown highlighted in red in the panes on the right.

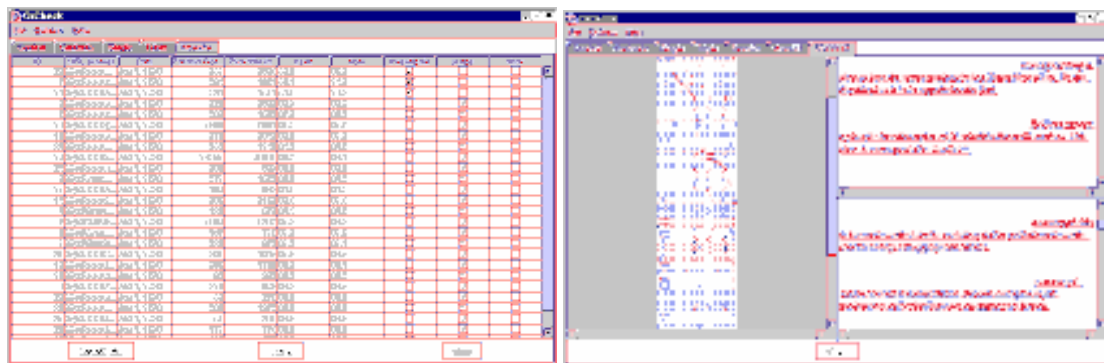
Currently VAST has been integrated into the PRAISE tool although it can also be used standalone to investigate any pair of documents. The only improvement planned is to enhance the identification of similarity in the document panes. VAST uses fuzzy matching so that it can see through attempted disguise but this also means that in some circumstances it is difficult to locate the similarity that it is indicating.

Originality Checker (OrCheck).

OrCheck is used to investigate a single document where it is suspected that it, or parts of it, may have been obtained from Web sources. It makes programmatic use of the Google search engine to automatically conduct a series of searches, using search terms supplied by the user. The documents located by the search engine are then automatically downloaded and analysed. The results of the analysis are presented to the user in a table from where documents which have been shown to have no similarity can be discarded, a collection of similar documents can be further analysed and the results presented in an interactive visualisation, or a single document can be shown alongside the target document with areas of similarity highlighted.

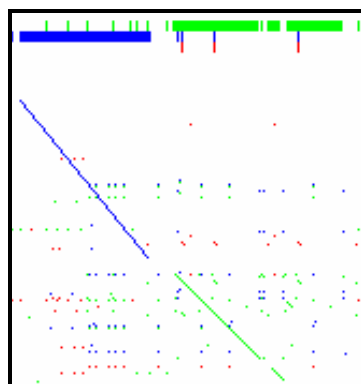
OrCheck is illustrated in Figure 3. The table on the left indicates the amount of material from each downloaded document that is essentially identical to what proportion of material from the document under investigation. OrCheck also uses fuzzy matching to see through attempted disguise, although the matching is far less fuzzy than that used by VAST. This table is suggesting, via the checkboxes on the right of the table, that only the top three documents may have significant similarity and the remaining downloaded documents could be ignored.

Figure 3 OrCheck table view (left) and interactive visualisation (right)



The right hand part of Figure 3 shows OrCheck's main interactive capability. This view of the tool is reached from the table view by agreeing to discard all but the top three matches and requesting that they be analysed in detail to obtain the visualisation shown on the left of the window. An OrCheck visualisation, like a VAST visualisation, plots the contents of the document under investigation left to right across the representation. All the documents which have been shown to be similar are then plotted from top to bottom, with the similar areas of each document shown in a different colour. An enlarged section of the top left of an OrCheck visualisation is shown in Figure 4.

Figure 4. Enlarged detail of an OrCheck visualisation



The band of lighter and darker colours at the top of the image show that two sources have been identified. The extent of the each line is showing where the source has been used in the document being investigated. The very first part of the target document is identical to the very first part of the darker source, as indicated by the diagonal line; and the next part is almost identical, a few words have been changed in the middle of the extract, to part of the lighter source.

In the context of the interactive tool a diagonal line can be selected with the mouse pointer and the corresponding parts of the two documents are shown in the upper text pane on the right of the tool with the similar sections highlighted using the appropriate colour. The lower

text pane provides a list of pairs of fragments from the document showing all located matches.

The OrCheck tool currently also contains a concordance which will be removed at some stage and presented in an enhanced manner wither as a concordance tool in its own right or included in the FreeStyler tool, as described below. It also contains a capability for a simple pair of documents to be selected from the table view, shown in Figure 3, to be examined side by side with the similar sections highlighted. The interactive capability of the OrCheck visualisation, allowing the similarity of a number of documents to be investigated, will be made available from the PRAISE torc cluster view.

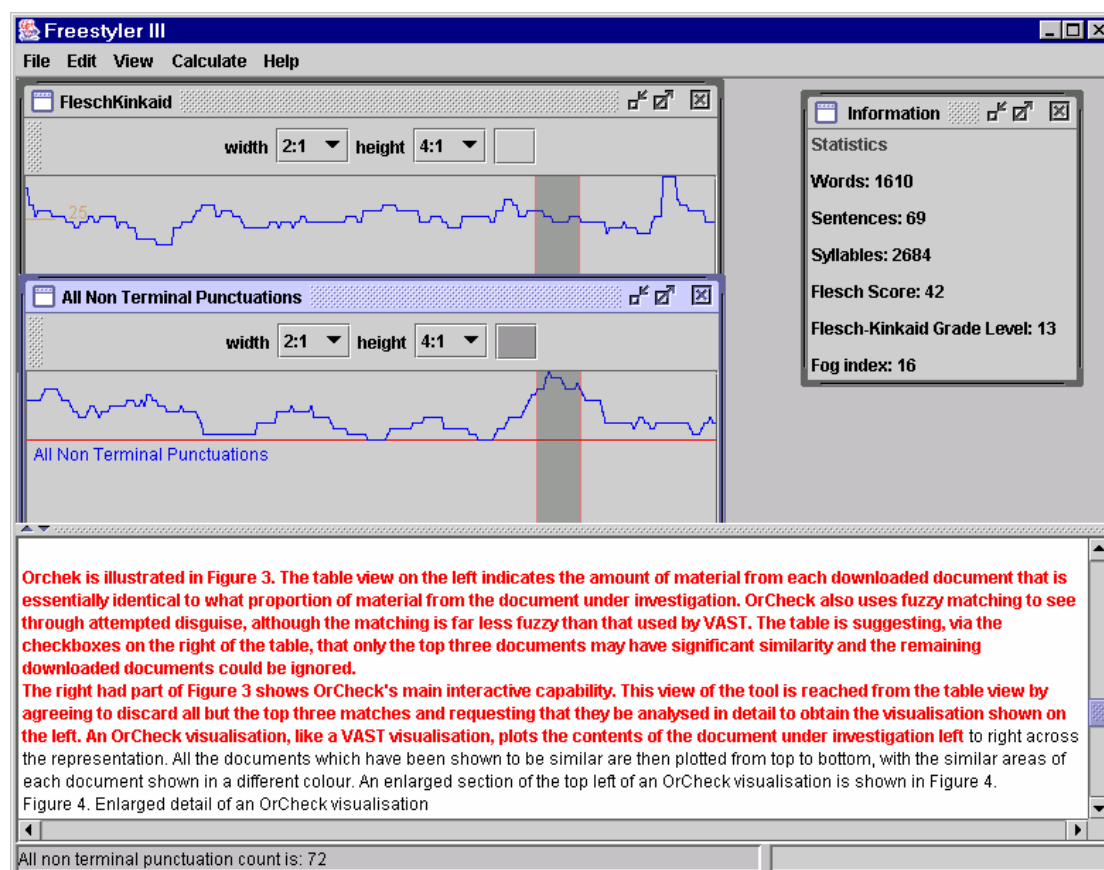
FreeStyler

The FreeStyler tool is intended to be used where it is suspected that a document which is supposed to be sole authored shows signs of multiple authorship. It analyses a document using a number of metrics and plots rolling average graphs of how that metric changes through the document. Available metrics include: word length, sentence length, various reading age metrics, use of non-terminal punctuation, use of active voice and arbitrary letter sequences (e.g. 'ise' or 'ize'). The intention being that an individual's writing style is defined by a combination of these metrics so if it can be shown that they are used differently in different parts of the document there is some evidential basis to suggest multiple authorship.

Figure 5 illustrates FreeStyler in operation with it showing the FleschKinkaid reading age and all non-terminal punctuation graphs of the document contained in the lower text pane. The information window to the right of the graphs gives some metric information for the document as a whole. An area of interest can be selected on any graph, which then caused the corresponding areas on all other graphs to be highlighted, as well as the section of the document that they represent.

The tool is intended to be enhanced by additional, more complex, metrics including: common american or UK spellings, proportionate use of common words, depth of sentence parse, etc. Although the tool was intended to detect distinctly different styles within a document it is also being used to ensure consistency of style throughtout a document whilst it is being produced. For example Figure 5 is showing graphs of an early draft of this document and indicates that the reading age is rather high at the start, where for the benefit of the readers it should be at its lowest, but is fairly consistent thereafter apart from a blip towards the end that might benefit from some simplification. The tool is currently highlighting a part of the document where, ther might be, undue, use of commas; or other non-terminal punctuation.

Figure 5 FreeStyler



Other tools on the drawing board

The concordance part of the OrCheck tool will be removed as it has not proved useful in assisting in the composition of Google search terms. It might be extracted, enhanced to show key words in context (KWIC), and presented either as a tool in its own right or included within FreeStyler, replacing an existing simple concordance capability.

The interpretation of JISC non-originality reports is fairly straightforward when the submissions are of a moderate size, say up to 5,000 words. However, when they go significantly beyond that it can become difficult to interpret and integrate the extent and significance of the non-originality shown. A utility, rather than a tool, called JiscView has been developed which gives a high-level, non-interactive coloured 'map' of the report. The map is produced by plotting a pixel for every word in the document using either white for words which are deemed original or the appropriate colour for those that have been shown to be non-original. The non-original parts of the map then stand out as bands of colours and patterns indicate exactly how the different putative sources have been used. It is intended to make the map interactive so that clicking on a band will scroll the report, shown in another pane, to the corresponding part.

Two versions of a document under development exhibit dissimilarity where material has been added, removed or relocated. Although the current generation of word processors provide support for tracking revisions the tools have to be used from the outset, are complex and hence cumbersome with a steep learning curve and are not always available. Faced with a troublesome student who was repeatedly submitting very slightly changed versions of chapters from his thesis and insisting on comments, the OrCheck tool was jury rigged to highlight the differences and then only those parts were considered and commented. It is intended to make this capability available in a dedicated tool.

The analysis capability of the PRAISE tool operates at the gross, document, level. Accordingly it is possible for a relatively short but highly similar part of a document to be hidden from the tool. Work is just beginning in investigating how the similarity of sentences and other parts of documents can be effectively and efficiently analysed.

Acknowledgements

Many of the tools have been largely developed by a succession of students, who are named in the tools' About dialogs. Particular mentions are due to Thomas Lancaster for most of the development of the VAST tool and the first version of FreeStyler, whilst he was working as a research assistant at CISE, and to Jonas Altin for the first version of PRAISE.