

Attributing Authorship using a Contextual Network Graph

Aaron Coburn
National Institute for Technology and Liberal Education
acoburn@middlebury.edu

This approach combines word-use statistics and graph theory to identify the author of an unknown text. By comparing stylistic attributes of a document to a corpus of known texts, one can make a reasonable guess about its authorship. In the first step of this method, each text is reduced to word sequences. These are typically two-word sequences, but three- and four-word phrases work well for some very large texts. Next, the indexed collection is projected onto a graph in which the documents and term sequences are represented as an interconnected network.

For longer texts (typically greater than 3,000 words), the first stage is simple: word pairs are extracted and counted. If, for instance, the word pair "altogether too" appears more prominently in two particular texts, one can begin to associate these documents. By extracting commonly appearing word *sequences*, the resulting index tends to form a fingerprint of a document's style rather than its content.

Shorter texts are more difficult; there are typically too few word pairs that appear across the collection to provide any meaningful correlations. For these texts, I apply a part-of-speech tagger, and reduce nouns, adjectives and verbs each to a common token. Thus the phrase "I walk through the field at noon" becomes "I verb through the noun at noun". Then, the word pair frequencies are extracted as before.

With an index of word sequence frequencies for each document, these values are applied to the connected graph described above. Document and term nodes are connected by edges, and the value of each edge is determined by the frequency of a word sequence in a document. This *contextual network graph* produces a network in which similar documents are closely connected and dissimilar texts are less closely connected.

Once the graph is constructed, a measure of similarity can easily be determined between the document with unknown authorship and all other documents. Those documents with the highest level of similarity can then likely be identified as having the same author.

Still, this does not address the possibility that the text was written by none of the authors represented in the training corpus. To determine this, it is not sufficient to merely find the most closely connected documents. Instead, the overall *strength* of the similarity must be relatively high: there must be a sufficient number of word sequences shared by the sample document and the rest of the collection. If the sample document, however, is poorly connected to other documents, it can be guessed that the authorship remains unknown.

This method appears to work well with both English and some non-English texts. The part-of-speech tagging, however, can only be applied to languages for which a parser is available. For this series of texts, the tagger was only applied to certain English-language texts.