JGAAP: A Modular Software Framework for Evaluation, Testing, and Cross-Fertilization of Authorship Attribution Techniques

Patrick Juola, Duquesne University, juola@mathcs.duq.edu

Problem Statement

The question may be as old as scholarship ; given a manuscript of unknown source, what can one learn from studying its style? The idea of using statistical properties (such as word length) to quantify "style" dates back at least to the 19th century. What does not yet exist is a well-understood, and above all reliable method to do this quantification. What we have instead is an *ad hoc* collection of methods proposed by individual researchers but without systematic testing or general acceptance.

The JGAAP (Java Graphical Authorship Attribution Program) system is a framework to remedy this. Using standard Java 1.6 technology, a clear theoretical framework, and an easy, user-extensible interface, JGAAP provides for objective testing of proposed methods.

Structure of JGAAP

JGAAP uses a simple three-phase pipeline to perform authorship attribution. Each document is "canoncized" by eliminating uninformative sources of variance, such as normalizing whitespace. Documents are then broken into "event sets" such as words, characters, parts of speech, word N-grams, and so forth. Finally, inferential statistics such as principle component analysis, k-nearest neighbor (using a variety of distance measures), or support vector machines, to determine which author should be assigned to the unknown documents.

A key feature of JGAAP is the modular structure; like LEGO blocks, any set of canonicizers can be used with any event set, which in turn can be used with any statistical method. Similarly, the use of objects and inheritance makes field extension easily; any new class that purports to extend "Canonicizer" can be used as easily as a built-in class. We have even developed a way to auto-populate the GUI to add new (user-written) classes.

At this writing, we have incorporated a half-dozen canonicizers, approximately thirty different event sets, eight different major analysis techniques, and a dozen different distance/divergence measures for use with nearest neighbor analysis. Including all combinations and variations, the current system is capable of approximately 40,000 different types of analysis.

Testing JGAAP

One advantage of using Java is its advertised property of system independence. We have confirmed this using the NMI Build-and-Test suite and have encountered no major differences on any system in that cluster.

To test performance of the system across different settings, we used the Ad-hoc Authorship Attribution Competition corpus, containing thirteen problems, in a variety of lengths, styles, genres, and languages. This publically available problem-set has been directly into JGAAP, making it possible for anyone to reproduce this "standard" problem set in their own development work. We also use this corpus to test performance of the 40,000 different methods described above. Work on this is ongoing but we have some preliminary results to present here:

- OCR errors do not materially impact accuracy
- Asymmetry is a significant factor in distance-based attribution methods
- Algorithm performance dominates language or data size effects
- Linear separation on large numbers of words outperforms higher-overhead methods on fewer words
- Characters trump words for Chinese at current word segmentation technology
- Cosine and KL distances perform well (our current informal best practice candidates)

Impact of JGAAP JGAAP has become an important aspect of the field, holding the #3 spot (and highest ranked software project) in a Google search for "authorship attribution," and even displacing "Japanese Generally Accepted Accounting Principles" for the #1 spot in a search for "JGAAP."[1]. The current version of the program, licenced under GPL, as well as a wiki about the current program, can be accessed freely at `www.jgaap.com`; interested readers can also download the text of a book on authorship attribution. It has received international attention and will form the basis of a master class at CSSLLI 2010 this upcoming summer.

[1] August 4, 2009