Authorship Attribution, Similarity, and Noncommutative Divergence Measures

Patrick Juola // Duquesne University, Pittsburgh PA // juola@mathcs.duq.edu

Mike Ryan // Duquesne University, Pittsburgh PA // michaelryan@acm.org

A common method of performing authorship attribution (or text classification in general) involves embedding the documents in a high-dimensional feature space and calculating similarity judgments in the form of numeric "distances" between them. Using (for example) a *k*-nearest neighbor algorithm, an unknown document can be assigned to the "closest" (in similarity or distance) group of reference documents. However, the word "distance" is ill-defined and can be implemented conceptually in many different ways. We examine the implications of one broad category of "distances" and find that symmetry or commutativity is very important for accurate authorship attribution.

This notion of distance can be generalized to dissimilarity judgements without previous embedding in a space. An example of this is the Juola-Wyner implementation of cross-entropy (Wyner, 1996; Juola,1997) which calculates an information-theoretic dissimilarity measure between two event streams where the events are not necessarily independent and thus cannot be directly tabulated as simple histograms. This kind of "distance" can easily be incorporated into a text classification system.

To a topologist, a "distance" is a numeric function D(x,y) between two points or objects, such that

- D(x,y) is always nonnegative, and always positive if x != y
- D(x,y) = D(y,x)
- D(x,y) + D(y,z) >= D(x,z)

However, there are many useful distance-like measures (technically known as "divergences") that do not have all these properties. In particular, divergences such as cross-entropy, its close cousin Kullback-Leibler divergence and vocabulary overlap are not commutative in that D(x,y) != D(y,x). [N.b. that 100% of {a} is contained in {a,b}, but only 50% of {a,b} is contained in {a}.]

Does this asymmetry have a significant impact on the accuracy of authorship attribution judgments? A standard trick for enforcing commutivity is to calculate D*(x,y) = [D(x,y)+D(y,x)]/2, but this will approximately double the time and cost of calculating distances and hence halve system performance? To test this, we are in the process of applying several noncommutative distance functions to a standardized corpus [the AAAC corpus (Juola, 2004)] of authorship attribution problems using the JGAAP framework (Juola, this conference; Noecker and Juola, this conference).

Preliminary results using cross-entropy indicate that enforced commutativity is *never* harmful and will on some problems increase accuracy by up to four-fold. We plan to continue this work using other noncommutative divergences; if this finding continues to hold, we consider this to be an important step to eliminating some of the "ad-hoc-ness" of the current state of authorship attribution, as we will be able to provide some steps to analyzing not merely what methods perform best, but what properties (such as commutivity) the best-performing methods have.