Finding Meaning in an Ocean of Data

Eitan Pechenick
Dr. Peter Dodds
Department of Mathematics and Statistics
University of Vermont

April 26, 2011

Abstract

Classic literature, contemporary blogs, and a strand of DNA all share something in common: They each contain strings of information, which grow to incredible lengths. In turn, each can be considered in terms of "phrases"—or sequences of "words"—determined in accordance with a given alphabet. It is computationally straight-forward to examine a large data source and determine the frequencies of every word found and every phrase up to a given length. Moreover, this sort of analysis can shed a great deal of light on the overall structure of a given language—and even changes in that structure over time. To delve even deeper into this analysis requires a sense of which phrases are meaningful. A simple example in the English language is: "I am." In fact, not only does this particular combination of words create an intuitive sense of meaning, but the word "am" does not generally appear without the word "I." This means the words are highly correlated. The goal of this presentation is to explore new and existing methods for determining phrases of highly correlated words. Compiling lists of significant phrases has applications in data mining, in tracking social trends through the examination of blogs, and could be used as a tool in the analysis of genetic information.