Mining Massive Text Data and Developing Robust Statistical Tracking

Regina Liu[1]

[1] Department of Statistics, Rutgers University, Piscataway, NJ, USA

Abstract

We present a systematic data mining procedure for exploring large free-style text datasets to discover useful features and develop tracking statistics, often referred to as performance measures or risk indicators. The procedure includes text classification, construction of tracking statistics, inference under error measurements. An aviation safety report repository from the FAA is used to illustrate applications of our research to aviation risk management and general decision-support systems. Some specific text analysis methodologies and tracking statistics are discussed. A robust approach for incorporating misclassified data or error measurements into the inference for tracking statistics is proposed and evaluated.

Although most illustrations here are drawn from aviation safety data, the proposed data mining procedure with its robust inference framework applies to many other domains, including, for example, mining free-style medical reports for tracking possible disease outbreaks.

This is joint work with Daniel Jeske, Department of Statistics, UC Riverside.