The University of Chicago
Department of Statistics

Seminar Series

REGINA LIU

Department of Statistics
Rutgers, The State University of New Jersey

Mining and Tracking Massive Text Data

MONDAY, April 9, 2007 at 4:00 PM
133 Eckhart Hall, 5734 S. University Avenue
Refreshments following the seminar in Eckhart 110.

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 and risk management. The main difficulty in deriving this inference scheme is the accounting for misclassification errors, for which we propose two types of approaches: "plug-in" and "projection" methods. We also consider the bootstrap calibration for fine tuning. Finally, as an illustrative example, the proposed data mining procedure is applied to analyzing an aviation safety report repository to show its utility in aviation risk management.

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

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

Please send email to Mathias Drton (drton@galton.uchicago.edu) for further information. Information about building access for persons with disabilities may be obtained in advance by calling Karen Gonzalez (Department Administrator and Assistant to Chair) at 773.702.8335 or by email (karen@galton.uchicago.edu).