CCSU

DEPARTMENT OF MATHEMATICAL SCIENCES

COLLOQUIUM

Thursday, April 26
3:00 – 4:25 PM
Maria Sanford, Room 101

APPLYING NATURAL LANGUAGE PROCESSING AND DOCUMENT CLASSIFICATION TO TEXT MINING RSS FEEDS IN ORDER TO CLASSIFY DOCUMENTS AS INTERESTING OR NOT TO AN ANALYST AT THE COMPANY, ALLIANT

MALCOLM HOUTZ

(Data Mining MS Thesis Presentation)

CENTRAL CONNECTICUT STATE UNIVERSITY

Abstract: Really Simple Syndication (RSS) is an internet-based technology allowing website publishers to easily and frequently broadcast relevant content updates to subscribers. Business analysts looking for content relevant to their needs have thousands of RSS feeds to chose from, with each feed potentially syndicating dozens of articles each day. Text mining techniques combined with classic data mining methodology can be used to help predict the relevance of RSS content, saving the analyst valuable time.

As the need for text mining applications grows with the overabundance of available data, text mining methodology has grown more and more complex, integrating the natural language rules of the spoken language. Natural Language Processing (NLP) programs are available in many languages, and some programs are specialized for particular industries, such as finance. This thesis proposes that tuning the Natural Language Rules in IBM's SPSS Modeler software for the interests of the direct marketing service provider Alliant, will improve predictions as to whether or not an RSS feed is relevant.

Tuning the Natural Language rules from the perspective of Alliant's interest was chosen because of the author's employment there as well as the availability of data. Articles syndicated by a variety of web sites related to the marketing and marketing analytics industries were collected for analysis. This corpus served as the basis for developing models predicting whether or not the article was interesting from the perspective of an analyst working at Alliant. After these predictions were made, the NLP rules were studies and modified, resulting in a measurable improvement in model accuracy.

For further information:

gotchevi@ccsu.edu 860-832-2839
http://www.math.ccsu.edu/gotchev/colloquium/