Scientists devise means to test for phony technical papers

April 24 2006

Authors of bogus technical articles beware. A team of researchers at the Indiana University School of Informatics has designed a tool that distinguishes between real and fake papers. It's called the Inauthentic Paper Detector -- one of the first of its kind anywhere -- and it uses compression to determine whether technical texts are generated by man or machine.

"This is a potential problem since no existing systems, the Web for example, can or do discriminate between content that is meaningful or bogus," says assistant professor Mehmet Dalkilic, a data mining expert. "We believe that there are subtle, short- and long-range word or even word string repetitions that exist in human texts, but not in many classes of computer-generated texts that can be used to discriminate based on meaning."

Joining Dalkilic on the IPD project are Assistant Professor Predrag Radivojac, informatics doctoral student James Costello, and Wyatt T. Clark, who will graduate in May with a bachelor's degree in informatics.

The IPD system is based on a combination of compression algorithms that reduce the amount of data to save space and speed transmission time.

To begin their study, the team identified two kinds of texts they would analyze. "Authentic text" (or document) is a collection of several hundreds or thousands of syntactically correct sentences that are wholly

meaningful. "Inauthentic text" (or document) is a collection of several hundreds of thousands of syntactically correct sentences that, taken all together, have no meaning.

The researchers' work is documented in the very authentic paper, "Using Compression to Identify Classes of Inauthentic Texts," which they presented at the Society for Industrial and Applied Mathematics Conference on Data Mining in Bethesda, Md., this weekend.

The informatics study largely was inspired by a prank pulled by three Massachusetts Institute of Technology students, who in 2004 developed a computer program that churned out randomly generated fake computer science language, essentially a four-page compilation of gibberish. They submitted it as a research paper to an international conference on computer science and informatics – and it was accepted without review.

Radivojac, whose research expertise is machine learning, says the IPD easily detected numerous inauthentic technical papers tested, including the MIT students' spurious submission.

"We hypothesized we could build a reliable and fast model that recognizes fake papers automatically," says Radivojac. "We combined these with machine-learning methods to build a predictor of these kinds of papers."

In general, identifying meaning in a technical document is difficult, Dalkilic says. "We don't claim we have found a way to distinguish between meaning and nonsense, but we do emphasize that there are many nontrivial classes of inauthentic documents that can be easily distinguished based on compression algorithms."

Source: Indiana University School of Informatics