A Semi-Supervised Active Learning Algorithm for Information Extraction from Textual Data by Dr. William M. Pottenger & Tianhao Wu

Poster #17 Abstract: In this poster we present a semi-supervised active learning algorithm for pattern discovery in information extraction from textual data. The patterns are regular expressions composed of various characteristics of features useful in information extraction. Our major contribution is a semi-supervised learning algorithm that extracts information from a set of examples labeled as relevant or irrelevant to a given feature. The approach is semi-supervised because it does not require the labeling of the exact location of features in the training data. This significantly reduces the effort needed to develop a training set. An active learning algorithm is used to assist the semi-supervised learning algorithm in order to further reduce training set development effort. The active learning algorithm is seeded with a single positive example of a given feature. The context of the seed is used to automatically identify candidates for additional positive examples of the given feature. Candidate examples are manually pruned during the active learning phase, and our semi-supervised learning algorithm automatically discovers regular expressions for each feature. We have successfully applied this learning technique in the extraction of textual features from police incident reports, university crime reports, and patents. The performance of our algorithm compares favorably with competitive extraction systems being used in criminal justice information systems.

Biographical Information: **William M. Pottenger**, Ph.D., is an Assistant Professor in the Department of Computer Science and Engineering at Lehigh University in Bethlehem, Pennsylvania. Prior to coming to Lehigh, Bill completed his Ph.D. in Computer Science at the University of Illinois at Urbana-Champaign and worked as a Research Scientist at the National Center for Supercomputing Applications. Bill's research interests include the fields of machine learning as applied in text mining and parallel and distributed computing. **Tianhao Wu** is a Ph.D. candidate in Computer Science at Lehigh University. His major research area is textual data mining with a focus on information extraction. Tianhao received a B.S. in Computer Science from Beijing University. He received his M.S. in Computer Science from Lehigh University. Tianhao also has three years of work experience in China prior to coming to Lehigh University. Tianhao has participated in several text mining projects over the past three years. One of projects he has participated in was a collaboration with Lockheed Martin and the Pennsylvania State Police to create a prototype for automatic discovery of reduced regular expressions in textual data for use in entity extraction from police incident reports.