Improving Similarly Joined Algorithms Using Vertical Clustering Techniques

Lisa Tan
Department of Computer Science
Wayne State University

Tuesday, September 15, 2009
3:00pm Rm. 110 Purdy-Kresge Library

Abstract:

String is a primary data format in majority of applications. With the rapid growth of diverse data driven applications in the current information era, retrieving string data from heterogeneous structured sources becomes more and more significant and challenging. The main concern is duplicate records are created when data is integrated from heterogeneous sources. Those duplicate records represent the same real-world entity because of inconsistent values and naming conventions, incorrect or missing data values, or incomplete information. Existing approaches make the assumption that group of related attributes will participate in the similarity join operation. However, in this paper we propose a pre-processing technique to improve existing similarity join techniques. Assuming relational data sources, our approach is to identify groups of related attributes that when similarity join is applied, we reduce false positives and false negatives, and increase precisions and F-measure.

Biography:

Ms. Lisa Tan is currently a part-time Ph.D. candidate in the Department of Computer Science at Wayne State University. She received her Masters Degree in Computer Science at WSU, after which, she has been working for Computing & Information Technology as Lead Systems Integrator. She has been working on several applications at C&IT such as Pipeline, Enterprise Directory and STARS application. Her research interests include Information Retrieval, Data Mining and eLearning in Pedagogy. She has published a refereed international conference paper. Ms. Tan will be an editorial review board member on Third International Conference of the Applications of Digital Information and Web Technologies (ICADIWT 2010).