SUMMARY

We often need to access and reorganize information available in multiple tables in diverse Web pages and traditional documents. To understand tables, we rely on acquired expertise, background information, and practice. Current computerized tools for extracting information from tables are inadequate because they seldom consider the structure and content in the context of other tables with related information. We propose to endow a computerized system with the knowledge and skills necessary to extract and organize information from a set of heterogeneous tables in a broad domain of interest. Understanding and exploiting the relationships in structured data requires thoughtful integration of several active strands of research in diverse disciplines.

In our NSF-funded TIDIE project (IIS-0083127) we extract information from various types of Web documents, and integrate and merge structured data from such sources as car ads, job announcements, university course listings, and similar topics. TIDIE work has not focused specifically on the processing of tabular data. Yet we have pursued non-TIDIE work in table-specific processing: identifying tables from text documents and OCR images, recognizing their lines and cells, and characterizing their layout and content.

We propose to address the table processing issue by developing an approach to table understanding called TANGO (Table ANalysis for Generating Ontologies). TANGO will further our previous work by allowing us to integrate the TIDIE data extraction approach with other sophisticated table-based recognition techniques. We will apply TIDIE's conceptual modeling extraction approach to: (i) understand a table's structure and conceptual content to the extent possible; (ii) discover the constraints that hold between concepts extracted from the table; (iii) match the recognized concepts with ones from a more general specification of related concepts; and (iv) merge the resulting structure with other similar knowledge representations (i.e. ontologies) for use in future situations. The result will be a formalized method of processing the format and content of tables while incrementally building a relevant reusable conceptual ontology.

As in TIDIE, TANGO assumes that the user will specify an initial ontology with a core set of concepts relevant to a particular topic of interest. For purposes of illustration and testing, we have chosen geographic information as the application domain—one that has a large and important field of knowledge where data often appears in tabular format. The concrete results will include: (i) a comprehensive ontology of geographic relationships produced from various tables; (ii) an ontology describing tabular formats, arrangements, styles, and content types; and (iii) a publicly available demonstration system deployed on the Web, which allows users to specify, search for, and extract information on a given topic (e.g. geography) that is contained in tables, and (iv) a publicly available collection (or corpus) of a wide range of sample tables in various formats that can serve for future testing and development. We will also develop and document useful metrics for testing and evaluating the success of our approach.

Our research team is housed in three departments (Computer Science, Linguistics, and Electrical, Computer, and Systems Engineering), and at two institutions (BYU and RPI); we have collaborated for several years. Our extensive combined experience in several areas is relevant to this work: ontology building, information retrieval, Web data extraction, natural language processing, computational linguistics, image processing, table understanding, document image analysis, OCR, geographic information systems, and computational geometry. The TANGO project focuses this experience and collaboration in new and groundbreaking ways. The interdisciplinary nature of this work will help provide exciting and novel learning opportunities for the students supported under this contract, who will be closely supervised by the principal investigators. TANGO will also appeal to a much wider group: anyone who is interested in extracting information from tabular formats and organizing the information in an ontology.