SUMMARY

As the Internet and other information outlets grow, people are becoming ever more swamped with large volumes of data. But people do not want volumes of data; they want critical information extracted expertly, organized automatically, and summarized smartly in a usable personalized view.

Making this valuable but "buried" information accessible is a huge challenge (possibly even a "grand challenge"). We propose a particular approach to this challenge, which we call TIDIE (Target-based, Independent-of-Document Information Extraction, pronounced "Tidy").

TIDIE embodies two key ideas: (1) a target description and (2) document-independent matching of source information with the target description. A TIDIE target description includes (1) a conceptual-model-based ontological framework over application-dependent objects and relationships and (2) a set of document-independent text-recognition specifications based on linguistic, geometric, and metatextual clues. The combination of constraints imposed by the ontology and clues provided by linguistic, geometric, and metatextual analysis allows us to locate, extract, organize, integrate, and summarize information in a specified user view.

Over the past few years we have experimented successfully with extracting data from multiple-record Web documents such as car ads, job ads, obituaries, course listings, and similar applications. We have also experimented successfully with converting scanned images of printed tables into electronic data. Currently, we are involved in several related projects: target-based information integration, business report data extraction, sentence boundary recognition, microfilm record recognition, data extraction from dynamically generated Web forms, ontology-based document filtering, and atomic information culling.

Based on our initial success and these current projects, we propose to work in four focus areas: (1) data extraction from data-rich Web documents, (2) revitalization of data in historical documents, (3) integration—target-to-source mapping generation, and (4) integration—data merging. Each of these focus areas includes a half-dozen resolvable issues, for example, high-precision filtering with respect to a target specification, matching filled-in tables and forms with target specifications, generating implied source object- and relationship-sets to support a target specification, and transforming values to a common target ontology. Given the high amount of overlap among the issues and the central theme provided by the two key TIDIE ideas, we believe that parallel, synergistic work in these four areas of focus will speed the resolution of all issues. As performance metrics, we propose the use of precision and recall over sets of human-countable size similar to what we have done in our initial experimentation.

Our research team leaders have complementary expertise in database theory, ontology building, information retrieval, pattern recognition, natural language analysis, and Web data extraction. Although we represent three different departments (Computer Science, Information Systems, and Linguistics), we have collaborated in various combinations for many years. Diverse departmental perspectives will expand educational opportunities for the students with whom we intend to work and will ensure a good mixture for sharing ideas and results both within our group and in the broader CS, IS, and Linguistic communities.