ANNUAL REPORT FOR AWARD # 0414644

David W Embley ; Brigham Young University
Collaborative Research: TANGO : Table Analysis for Semiautomatic Generation of Ontologies

Participant Individuals:
Co Principal Investigator(s) : Yuri A Tijerino ; Deryle Lonsdale
Graduate student(s) : Cui Tao ; Zonghui Lian
Undergraduate student(s) : Chris Hathaway
Graduate student(s) : Reema Al-Kamha

Participants' Detail

Partner Organizations:
Rensselaer Polytechnic Institute : Collaborative Research ; Personnel Exchanges

Our grant is a joint, collaborative grant between BYU and RPI.

Other collaborators:

George Nagy -- Department of Electrical, Computer and Systems Engineering, Rensselaer Polytechnic Institute, Troy, New York. Co-PI on this joint, collaborative grant between BYU and RPI.

Yuri Tijerino -- Having left BYU, Professor Tijerino should be considered as a collaborator, rather than a Co-PI.

Daniel Lopresti -- Department of Computer Science and Engineering, Lehigh University, Bethlehem, Pennsylvania.

Stephen W. Liddle -- Department of Information Systems, Brigham Young University, Provo, Utah.

Activities and findings:

Research and Education Activities:

1. Development of a table-interpretation system

   a. Research survey on table processing paradigms

   b. Initial notes on table recognition

   c. Construction of the initial manual system for table interpretation

   d. Sibling-page technique for table interpretation

2. Development of a system to convert canonicalized tables to mini-ontologies

3. Development of a system to integrate mini-ontologies with a growing ontology

4. Use of XML as an exchange medium between the subsystems in our project

    a. Generalization/Specialization in XML

    b. Augmenting traditional conceptual models to accommodate XML

Findings:

1a. Research survey on table processing paradigms

Tables are a ubiquitous form of communication. While everyone seems to know what a table is, a precise, analytical definition of ``tabularity'' remains elusive because some bureaucratic forms, multicolumn text layouts, and schematic drawings share many characteristics of tables. There are significant differences between typeset tables, electronic files designed for display of tables, and tables in symbolic form intended for information retrieval. Most past research has addressed the extraction of low-level geometric information from raster images of tables scanned from printed documents, although there is growing interest in the processing of tables in electronic form as well.

Recent research on table composition and table analysis has improved our understanding of the distinction between the logical and physical structures of tables, and has led to improved formalisms for modeling tables. This review, which is structured in terms of generalized paradigms for table processing, indicates that progress on half-a-dozen specific research issues would open the door to using existing paper and electronic tables for database update, tabular browsing, structured information retrieval through graphical and audio interfaces, multimedia table editing, and platform-independent display.

1b. Initial notes on table recognition &
1c. Construction of the initial manual system for table interpretation

The shift of interest to web tables in HTML and PDF files, coupled with the incorporation of table analysis and conversion routines in commercial desktop document processing software, are likely to turn table recognition into more of a systems than an algorithmic issue. We illustrate the transition by some actual examples of web table conversion. We then suggest that the appropriate target format for table analysis, whether performed by conventional customized programs or by off-the-shelf software, is a representation based on the abstract table introduced by X. Wang in 1996. We show that the Wang model is adequate for some useful tasks that prove elusive for less explicit representations, and outline our plans to develop a semi-automated table processing system to demonstrate this approach. Screen-snapshots of a prototype tool to allow table mark-up in the

style of Wang are also presented.

1d. Sibling-page technique for table interpretation

The longstanding problem of automatic table interpretation still illudes us. Its solution would not only be an aid to table processing applications such as large volume table conversion and information extraction, but would also be an aid in solving related problems such as ontology learning and semi-structured data management. In this paper, we offer a solution for the common special case in which so-called sibling pages are available. Sibling pages, which are the pages commonly generated by underlying web databases, are compared to identify and connect nonvarying components (category labels) and varying components (data values). We tested our solution using more than 2,000 tables in source pages from three different domains---car advertisements, molecular biology, and geopolitical information. Experimental results show that the system can successfully identify sibling tables, generate structure patterns, interpret different tables using the generated patterns, and automatically adjust the structure patterns as needed.

4a. Generalization/Specialization in XML

XML is an effective universal data-interchange format, and XML Schema has become the preeminent mechanism for describing valid XML document structures. Generalization/specialization and its constraints are fundamental concepts in system modeling and design, but are difficult to express and enforce with XML Schema. This mismatch leads to unnecessary complexity and uncertainty in XML-based models. In this paper we describe how to translate various aspects of generalization/specialization from a conceptual model into XML Schema. We also explore what needs to be added to XML Schema to handle the other aspects of this fundamental modeling construct. If XML Schema were to include our proposed constructs, it would be fully capable of faithfully representing generalization/specialization, thus reducing the complexity of the XML models that rely on generalization/specialization.

4b. Augmenting traditional conceptual models to accommodate XML

Although it is possible to present XML Schema graphically such as in .NET or XMLSpy, these representations do not raise the level of abstraction of XML Schema in the same way traditional conceptual models raise the level of abstraction for data schemata. Traditional conceptual models, on the other hand, do not accommodate several of the XML Schema content structures. Thus, there is a need to enrich traditional conceptual models with features present in XML Schema but missing in traditional models. After establishing criteria for XML conceptual modeling, we propose an enrichment to represent the XML features missing in traditional models. We argue that our solution can be adapted generally for traditional conceptual models and show how it can be adopted for two popular conceptual models.

Training and Development:

Project participants have developed research and teaching skills and

experience in the following ways.

1. Weekly research group meetings:
   * discussion of research papers
   * presentations of student work
   * coordination of prototype development

2. Face-to-face visits with collaborators at RPI:
   * student presentations of their work to RPI visitors
   * RPI visitor presentations in our research group

3. International conference/workshop presentations:
   * at the Workshop on Document Analysis Systems, 'Notes on Contemporary Table Recognition' by George Nagy, February 2006.

4. Student presentations at the annual BYU College of Physical and Mathematical Sciences Spring Research Conference:
   * 'Table Structure Understanding by Sibling Page Comparison' by Cui Tao, March 2006.
   * 'Semi-Automatic Generation of Mini-Ontologies from Canonicalized Relational Tables' by Chris Hathaway, March 2006.
   * 'A Tool to Support Onotlogy Creation Based on Incremental Mini-Ontology Merging' by Zonghui Lian, March 2006.

5. Colloquium talks:
   * at BYU, 'Concepts, Ontologies and Project TANGO' by Deryle Lonsdale, October 2005.
   * at the Technical University of Vienna, 'Semantic Understanding: An Appproach Based on Information Extraction Ontologies' by David W. Embley, October 2005.

6. Other presentations:
   * TANGO-related term paper, 'Table Extraction Using MaxEnt' by Zonghui Lian, December 2006.
   * PhD dissertation proposal, 'Toward Making Online Biological Data machine Understandable' by Cui Tao, June 2005.

7. Doing research and data analysis for and writing papers: see publications section of this report for a list of publications and authors.

Journal Publications:

Yuri A. Tijerino, David W. Embley, Deryle W. Lonsdale, and George Nagy, "Towards Ontology Generation from Tables", World Wide Web: Internet and Web Information Systems, vol. 8, (2005), p. 261. Published

David W. Embley, Matthew Hurst, Daniel Lopresti, and George Nagy, "Table Processing Paradigms: A Research Survey", International Journal on Document Analysis and Recognition, vol. 8, (2006), p. 66. Published

Book(s) of other one-time publications(s):

Reema Al-Kamha, David W. Embley, and Stephen W. Liddle, "Representing Generalization/Specialization in XML Schema", bibl. Proceedings of the Workshop on Enterprise Modeling and Information Systems Architectures (EMISA '05), Klagenfurt, Austria, 24-25 October 2005, 250-263., (2005). Proceedings Published

David W. Embley, Daniel Lopresti, and George Nagy, "Notes on Contemporary Table Recognition",

bibl. Proceedings of the Seventh International Association for Pattern Recognition Workshop on Document Analysis Systems, (DAS 2006, LNCS 3872), Nelson, New Zealand, February 2006, 164-17, (2006). Book Published

Cui Tao and David W. Embley, "Table Interpretation by Sibling Page Comparison", bibl. Technical Report, Department of Computer Science, Brigham Young University, July 2006, (2006). Conference Proceedings Submitted

Other Specific Products:

Internet Dissemination:

tango.byu.edu

This is the homepage of the NSF project.

Contributions:

Contributions within Discipline:

1. We have written a survey or the current state-of-the art of table processing paradigms.

2. We have shown that table interpretation by sibling page comparison is possible. Further, we have tested our solution using more than 2,000 source tables from three different domains---car ads, molecular biology, and geopolitical information. Experimental results show the system (with near 100% accuracy) can identify sibling tables, generate structure patterns, interpret different tables using the generated patterns, and automatically adjust the patterns as needed while processing tables from a web site.

Contributions to Education and Human Resources:

* Since the award was granted, one student working on the project has graduated.

* Since the award was granted, students have participated as coauthors of several published papers. In addition, students are coauthors of several papers currently in press or under review.

  - Students are coauthors of 1 published paper.
  - Students are coauthors of 1 submitted paper.

* Since the award was granted, in forums with the public invited, students have given 3 presentations at college-sponsored workshops.

* Two students in our TANGO research group are female. Both are seeking a PhD. None of the other students working on the project is in an under-represented or minority group.

Contributions to Resources for Research and Education:

BYU has provided $2,192 of the $15,000 it promised as cost sharing for the project. We have used this money to purchase new computers and associated equipment.

Special Requirements for Annual Project Report:

Categories for which nothing is reported:
Outreach Activities
Products: Other Specific Product
Contributions to Other Disciplines
Contributions Beyond Science and Engineering
Special Reporting Requirements
Animal, Human Subjects, Biohazards

Submit Return

We welcome comments on this system