Annual Report for Period:08/2005 - 08/2006 **Submitted on:** 09/06/2006
Principal Investigator: Nagy, George . **Award ID:** 0414854
Organization: Rensselaer Polytech Inst
Title:
Collaborative Research: TANGO: Table Analysis for Semiautomatic Generation of Ontologies

Project Participants

Senior Personnel

Name: Nagy, George
Worked for more than 160 Hours: Yes
Contribution to Project:

Post-doc

Graduate Student

Undergraduate Student

Technician, Programmer

Other Participant

Research Experience for Undergraduates

Organizational Partners

Brigham Young University

Other Collaborators or Contacts

Prof. Daniel Lopresti, CS Dept, Lehigh University, Bethlehem, PA
Dina Goldin, CS Dept, University of Connecticut, Storr, CT

Joint work on information retrieval and digital libraries: co-authored publications.

Activities and Findings

Research and Education Activities:

Project funding arrived at RPI in mid-September, too late to recruit a graduate student. I was on sabbatical leave in 2005-2006. No NSF funds were expended on the project until Sept 2006.

Piyushee Jhap (F, US cit.) was recruited in April and joined the project as a Graduate Reseach Assistant (20 hours per week) on August 2, 2006. Since then she has been working on programs to convert web tables to Wang notation, as discussed in the proposal.

For preliminary evaluation, I collected a set of tables from US and Canadian government statistics sites.

Visited Prof. Embley and Lonsdale at BYU for a week for project planning. Had discussions with their students and colleagues.

Visited Chinese Academy of Science (Beijing), University of Salerno, Queens University (Kingston, Ont), XEROX PARC, University of Nebraska-Lincoln, IRST (Trento, Italy), Ecole de Technologie Superieur and University de Montreal, University of Connecticut, University of Missisipi, and Lehigh University to discuss current methods of consolidating web information and to give seminars.

Participated and presented papers at Context05 (ACM, Paris), DIAL06 (Lyon), DAS (Nelson, NZ), and DR&R-SPIE (San Jose).

Co-authored with Embley, Hurst and Lopresti an invited reseach survey paper about table processing (with 99 references) for Int J Document Analysis and Processing (summary in Prof. Embley's NSF Annual Report).

Co-authored with Embley and Lopresti a paper (presented jointly with Lopresti at DAS (Int Workshop on Document Analysis Systems) about interactive tools for table processing.

Co-authored with Lopresti (and presented) a paper on document analysis for DIAL06 (Document Image Analysis for Libraries). We collaborated on an annotation interface for tables ('Tabbycat'), and on new ideas solicited be a US Government agency for non-Latin script recognition for Arabic and Chinese documents.

Wrote a paper for another workshop on document analysis for digital libraries in Kolkota, India (October 06). Although I cannot attend, I was invited to submit this position paper.

Co-authored a paper with Goldin on clustering time series for the ACM conference on Information and Knowledge Management (November 06).
Our experiments and analysis contradicts a widely accepted finding to the effect that the result of K-means clustering of time series subsequences (STS clustering) is independent of the time series that created it.

Collaborated and exchanged long visits with Veeramachaneni (Trento) to advance our undestanding of style-consistent classification, i.e., recognizing related objects together rather than in isolation. We submitted a journal paper and an invited chapter for a book on adaptive methods for document analysis.

Findings:

Most past research on table analysis addressed the recovery of the cell structure and content of tables. For information consolidation, it is essential to move to the discovery of the complex relationships between multi-dimensional headers and subheaders (i.e., 'categories') and leaf cells.

Tables on the web appear in a bewildering variety of formats: scanned bitmaps, text files, pdf files, and html files. XML annotation is increasing rapidly. Even when stuctured constructs are readily available, they are not always used for tables, and they are often used for non-tables (actually the line between tables and non-tables is not easy drawn). Furthermore, table interpretation requires broader context than machines can currently be endowed with (furthering the automation of context acquisition is indeed one of the major goals of TANGO). It is therefore unlikely that completely automatic table processing will be realized within a decade. The measure of success will be to what extent human interaction can be minimized. This, in turn, requires an informative dialog between the human and the automated algorithms.

The similarities between tables originating from the same, or related, sites have been barely exploited. Embley's work on sibling tables and our research on style constrained classification are steps in this direction. Parsimonious human inter-action throughout the interpretation process is much better than operator intervention only at the beginning and the end, e.g., framing the objects to be recognized or dealing with rejects.

Mobile wireless web access is increasingly common. Therefore interactive information extraction paradims must also address small display sizes, touch screens, and speech I/O.

My sabbatical year did not result in any software, but it increased my understanding - through both many personal contacts and much reading - of where table analysis is moving and what bottlenecks it is facing.

During the coming months we continue working - in close collaboration with the BYU group - on interactive extraction of Wang schema, detailed measurement of human effort, and on evaluation protocols for semi-automatically constructed ontologies based on tables.

Training and Development:

I was the only person at RPI who worked on the project during the last year. The external colleagues with whom I collaborated during my sabbatical year may have gained increased appreciation and understanding of structured information extraction.

Outreach Activities:

I conducted research and wrote several papers on digitizing, coding, annotating, disseminating and preserving documents in digital libraries. Perhaps this can be considered an outreach activity in science and technology.

Journal Publications

Yuri A. Tijerino, D.W. Embley, Deryle W. Lonsdale, and G. Nagy, vol. 6, #3, Springer-Verlag, September 2005., "Towards ontology generation from tables", World Wide Web Journal, p. 261, vol. 6, (2005). Published

D. Lopresti, D.W. Embley, M. Hurst, and G. Nagy, "
Table Processing Paradigms: A Research Survey", International Journal of Document Analysis and Recognition, p. 66, vol. 8, (2006). Published

Books or Other One-time Publications

D.W. Embley, D. Lopresti, and G. Nagy, "Notes on Contemporary Table Recognition", (2006). book chapter, Published
Editor(s): H. Bunke and A. L. Spitz
Collection: Document Analysis Systems VII, 7th International Workshop, Procs. DAS 2006
Bibliography: vol. 3872, LNCS, pp. 164-175, Springer,

G. Nagy and D. Lopresti, "Interactive Document Processing and Digital Libraries", (2006). Conference proceedings, Published
Editor(s): F. LeBourgeois
Collection: Procs. 2nd IEEE International Conference on Document Image Analysis for Libraries, Lyon, France, April 27-28,
Bibliography: Universite de Lyon

D. Lopresti, A. Joshi, and G. Nagy, "Match Graph Generation for Symbolic Indirect Correlation", (2006). Conference proceedings, Published
Collection: Procs. SPIE Symposium on Document Recognition and Retrieval, vol.SPIE 6067, San Jose, CA, SPIE/IST
Bibliography: SPIE

G. Nagy, S. Veeramachaneni, "Adaptive and Interactive Approaches to Document Recognition", (2007). Book, Accepted
Editor(s): S. Marinai, H. Fujisawa
Collection: Procs. SPIE Symposium on Document Recognition and Retrieval, vol.SPIE 6067, San Jose, CA, SPIE/IST, January
Bibliography: Springer Verlag LNCS

Web/Internet Site

Other Specific Products

Contributions

Contributions within Discipline:

We hope that our conference presentations, journal, book chapter and proceedings publications, and our seminars, have furthered the discipline.

Contributions to Other Disciplines:

Contributions to Human Resource Development:

Contributions to Resources for Research and Education:

We will make the already collected test data publicly available.

Contributions Beyond Science and Engineering:

Special Requirements

Special reporting requirements: None
Change in Objectives or Scope: None
Unobligated funds: $ 0.00

Animal, Human Subjects, Biohazards: None

Categories for which nothing is reported:

Any Web/Internet Site
Any Product
Contributions: To Any Other Disciplines
Contributions: To Any Human Resource Development
Contributions: To Any Beyond Science and Engineering