Annual Report for Period: 08/2006 - 07/2007	**Submitted on:** 06/19/2007
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

Name: Jha, Piyushee

Worked for more than 160 Hours: Yes

Contribution to Project:

Graduate research assistant funded from this project

Name: Padmanabhan, Raghav

Worked for more than 160 Hours: No

Contribution to Project:

Graduate research assistant funded from this project

Undergraduate Student

Name: Murphy, Luke

Worked for more than 160 Hours: Yes

Contribution to Project:

Undergraduate research assistant funded from this project

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:

Research and Education Activities:

As reported last year, implementation of TANGO components at Rensselaer Polytechnic Institute did not begin until August 2006 because grant funding was confirmed only in September 2005, too late to engage a student. Furthermore, Nagy was on sabbatical leave in 2005-2006.

Graduate Research Assistant Piyushee Jha (female, US permanent resident) joined the project in August 2006. Her MS thesis (about 60% completed) is an interactive tool to convert web tables (HTML, PDF, or scanned bitmaps) to layout-independent Wang notation. The Wang notation represents a table in the form of a set of category trees and a list of leaf data cells with path information for each category. Jha's tool (called WNT) converts simple tables with very little operator intervention. The resulting table hierarchy is displayed to the operator for confirmation either in the form of indented text or as a reformatted table with clickable headers and content cells. Egregious (complex, oddly laid out, or ill-formed) tables often cause conversion errors, which can be conveniently corrected by the operator with a built-in graphical table editor. A fundamental design objective of WNT is to allow conversion to Wang notation of any table. Some tables will, of course, require considerable interaction. WNT has a logging component (described below) to quantify this aspect.

In April 2007 Jha began work on the proposed table processing ontology. A prototype ontology was created for Version 1 of the conversion tool. Further progress requires gaining more familiarity with the Object Oriented System Modeling tools developed by Prof. Embley's group. Jha will visit the BYU group in July.

During the coming months, Jha will spend about half her time on extending WNT to PDF and, via PDF, to scanned pixel-map tables. Based on previous research by others, we believe that tables in PDF documents can be analyzed most easily by first rendering them and then using image processing tools to extract and analyze the tables. However, instead of using OCR to recognize ASCII table entries, we expect to extract character locations, labels, typefaces, type sizes, and type styles directly from the PDF file. Scanned pixel tables would, of course, require incorporating OCR routines. However, fast and accurate conversion of PDF tables to Wang notation would be in itself a significant achievement that may attract considerable commercial interest.

RPI senior Luke Murphy (male, US citizen) was engaged as an undergraduate research assistant from December 2006 until he graduated in May 2007. He developed the logging component of WNT with Jha's help. The logger keeps track of every mouse click and every keystroke. The log is loaded into an Excel sheet, which has provisions for aggregation by (1) Type of interaction (2) Table and (3) Subject. It therefore allows identifying the bottlenecks in the interaction, which types of tables take longest, and how much variation there is between subjects with different skill levels. An off-the-shelf screen-capture routine was also integrated into the Wang Notation Tool. This allows complete replay of a table-conversion session, which may reveal opportunities for improvement that are not apparent from the quantitative log. The logger and screen capture functions workso far only with earlier versions of the conversion tool, and the spreadsheet based statistical analysis requires further development.

Graduate Research Assistant Raghav K Padmanabhan (male, Indian citizen) joined at the end of May 2007. His first task was the modification of some Java routines of the Table Interpretation of Sibling Pages (TISP) software developed by BYU researcher Cui Tao to allow interactive web table harvesting. TISP converts web pages to a Document Object Management Tree (DOMtree) and extracts tables. The Table Harvester (TH), based on TISP, extracts single or multiple tables from HTML source code, and converts them to an ASCII descriptor in the format expected by WNT. The conversion is based on the customary HTML constructs for tables. These are often abused: for instance, multi-column text and figures can be laid out with table tags. Therefore TH also requires an interactive component to ensure that only legitimate tables are extracted from each web page, and that the ASCII descriptor generated by TH is correct. This requires, in turn, a comparison of the original web table with the table represented by the ASCII descriptor, and an editor that can resolve any discrepancies. Here again the design objective is robust conversion, at the possible cost of considerable human intervention. Therefore TH also includes a logger. TH is currently about 25% complete.

In summary, during the past year we have made progress on (1) the conversion of a web table to a layout-preserving ASCII descriptor, and (2) the conversion of the ASCII descriptor to layout-independent Wang notation. Details of current status are contained in the student progress reports posted on the TANGO web site.

In June 2007 Professor Embley visited RPI for five days of intensive interaction with Nagy, Jha and Padmanabhan. We decided that interchange of tables described by Wang notation will take place in the form of XML documents that represent tables as OSM object sets. We therefore developed an OSM conceptual diagram for representing arbitrary tables. During the process, we eliminated some widely known limitations of the original Wang notation by including provisions to describe (1) the source of each table, (2) table title, (3) table caption (4) footnotes (in the title, table headers, or content cells), (5) annotations, and (6) aggregates (like total or partial sums, averages, minima or maxima). The issue is complicated by the fact that the scope of a footnote, annotation or aggregate must be preserved. The resulting OSM representation also preserves the important layout-independence property of the Wang notation. An existing tool allows the generation of an XML Schema descriptor of a Table Object, which in turn can be used to validate an XML document that purports to describe a specific table.

During Dr. Embley's visit we spent considerable time discussing evaluation of both individual components and of the end-to-end TANGO process. While further refinement of the evaluation is required, we decided on generalizing the current logger and statistical analysis routines built into WNT to other interactive project components. At this time the most promising approach appears to be the development of a tool similar to the profiling routines built into most compilers. This tool would allow specification of the lowest level of aggregation of timing information for mouse clicks and key presses. The necessary instructions would have to be automatically or interactively inserted into every interactive program to be logged, but the remainder of the data collection and statistical analysis could be quite general. The development of this tool would be of independent research interest to the document image analysis (DIA) community.

One major application of TANGO is responding to queries when the response must be generated from multiple tables, generally from diverse web sites. To this end, we examined the notion of formulating queries as empty tables, and using TANGO to fill in the content information. The necessary data could be found either in existing TANGO ontologies (in our restricted domain of geo-political information), or the TANGO ontologies could guide the extraction of relevant information from hitherto unseen web pages. Here again, we envision that the Query By Table (QBT) process will be interactive, therefore the evaluation will be based mainly on operator time compared to direct (i.e., unaided by TANGO) filling of the tables from multiple web pages.

Professor Embley's visit also provided an opportunity to firm up our definitions of table-processing terms such as table interpretation, table understanding, and table learning.

Findings:

Findings:

1. As expected, successive versions of the Wang Notation Tool, which require less and less operator interaction, greatly accelerate the conversion of tables into a form that can be used to generate mini-ontologies. Although in WNT Versions 1, 2, and 3, every table requires some interaction, in WNT Version 3 most operator time is due to the correction of conversion errors due to egregious tables. Therefore it is time to switch our attention from speeding up the table specification to speeding up the correction of errors. Our most important goal must remain 'learning' to ensure that the same type of errors do not occur repeatedly.

2. Most web tables contain important information that cannot be represented by the original Wang notation. It is therefore important to extend the table representation at least to (1) source, (2) title, (3) caption, (4) footnotes, (5) annotations, (6) aggregates. These aspects have received little attention in the vast existing table-processing literature.

3. Even though it is possible to devise and program algorithms for most of the wide variety of structural features found in web tables, there is a very long-tailed distribution of such features, and hundreds of thousands of tables would have to be tested to validate such programs. It appears preferable to provide a broad interactive editing capability, which in the limit allows entering manually the complete description of any table. It is expected that the distribution of user time will also be highly skewed, and only a very small fraction of web tables will require extensive interaction.

4. The new OSM representation of tables is a significant step towards both an ontological representation of table-processing, and of the entire TANGO process. It specifically allows documentation of every aspect of our project in a common (OSM) notation.

5. Query By Table (QBT) appears to be a valuable new method of information retrieval for semi-structured data from multiple tables at the same website or at multiple web sites.

6. In our context, table interpretation is defined as the construction of Wang notation for a table. Table understanding can be defined as the mapping of concepts from a new table to the existing ontology. Understanding is complete if all of the concepts from the new table can be mapped, and it is partial if only some of the concepts can be mapped. Correspondingly, table learning is the process of integrating concepts from new tables into the current ontology.

Training and Development:

TANGO has already introduced two graduate students and an undergraduate to the joys and challenges of research. Fundamental research skills that are better taught through frequent individual discussions than by graduate courses include:

1. Selective literature reviews based on our large and growing collection of table-processing literature;
2. Professional reporting of activities, with detailed written feedback on multiple iterations of student progress reports;
3. Problem formulation and prioritizing research tasks based on realistic predictions
of ultimate value, duration, chances of success, and fit with other tasks;
4. Feedback on demos and presentations to visiting experts in related and unrelated fields of research;
5. Efficient, courteous and productive team work based on technical interaction with local students and with researchers at BYU;
6. Honing of refereeing skills by comparing independent reviews of conference papers and articles with Nagy's reviews.

We believe that the student interactions and exchanges between BYU and Rensselaer Polytechnic Institute, two institutions of very different character and mission, will be valuable in expanding the students' horizons. We also interact extensively with faculty and students at Lehigh University, University of Nebraska Lincoln, and Boise State University, which contributes further to the integration of students into the wider research community.

Outreach Activities:

In conjunction with the School of Engineering and the Department of Electrical, Computer, and Systems Engineering, we hold open houses for prospective undergraduate and graduate students. We will also continue to interact with the New Visions high-school-on-campus. On the research front, we ensure that our many visitors meet and interact with the students on the project, who typically offer demonstrations and request a critique.

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

S. Veeramachaneni and G. Nagy, "Analytical results on style-constrained Bayesian classification of pattern fields", IEEE Transactions on Pattern Analysis and Machine Intelligence, p. 12180, vol. 29, (2007). 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

D. Goldin, R. Mardales, G. Nagy, "In search of meaning for time series subsequence clustering: Matching algorithms based on a new distance measure", (2006). Book, Published
Bibliography: Procs. 15th ACM Conference on Information and Knowledge Management (CIKM06), pp. 347-356, Arlington, VA,

D. Lopresti, G. Nagy, S. Seth, X. Zhang, "Multi-character field recognition for Arabic and Chinese Handwriting,? Conference Collection, accepted, Summit on Arabic and Chinese Handwriting", (2006). Book, Accepted
Editor(s): D. Doermann
Collection: Summit on Arabic and Chinese Handwriting
Bibliography: Proceedings of a Conference at College Park

G. Nagy, "Digitizing, coding, annotating, disseminating, and preserving document", (2006). Book, Accepted
Editor(s): P. Majumder
Bibliography: Int?l Workshop on Research Issues in Digital Libraries (IWRID), Kolkota, India,

Web/Internet Site

URL(s):

http://tango.byu.edu/

Description:

We post project-related publications, data, and progress reports here.

Other Specific Products

Product Type:

Software (or netware)

Product Description:

WNT v1, v2, v3: Programs to convert tables to Wang notation.
TH: Programs to convert HTML web tables to an ASCII representation.
LOGGER: Code to monitor and analyze interactive table, OSM, and ontology editing.

Sharing Information:

Software available on demand. Will be posted on TANGO website after thorough verification.

Contributions

Contributions within Discipline:

Novel methods of table analysis that preserve data in table headers, content cells, and various types of table annotations, as well as the relations between them.

Contributions to Other Disciplines:

TANGO will be evaluated on data collection and analysis in the geopolitical arena. However, there is nothing in its design specific to this (inter-)disciplinary area. If it is successful, we expect that it will be adopted in other areas where information in commonly presented in the form of tables. Researchers at BYU are already testing it in the biomedical domain, and at Rensselaer Polytechnic Institute we are planning to make use of it in two on-going projects: SkinScan, for analyzing photos of visible blemishes in terms of other information available both from the patient and from the sources of dermatological data on the web, and Cervix, a collaborative project with Lehigh University and NIH-NLM on segmenting cervigrams and analyzing tissue properties in the light of additional data from the web.

Contributions to Human Resource Development:

The direct contribution here is offering a valuable research experience to the students on the project.
A possible eventual indirect contribution is the expansion of access to web resources for a segment of the population which finds existing interfaces unsatisfactory.

Contributions to Resources for Research and Education:

The already collected test data is already publicly available on demand. We shall post it on the TANGO web site after additional annotation. We will also make most of our software available after thorough testing.the already collected test data publicly available.

Contributions Beyond Science and Engineering:

The collection and analysis of complimentary, overlapping or redundant semi-structured information scattered on the web extends to most scholarly activity in social science and humanities. Tables are ubiquitous.

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: