Wang Notation Tool: A Layout Independent Representation of Tables

by

Piyushee Jha

An Abstract of A Thesis Submitted to the Graduate

Faculty of Rensselaer Polytechnic Institute

in Partial Fulfillment of the

Requirements for the degree of

MASTER OF SCIENCE

Major Subject: ELECTRICAL ENGINEERING

The original of the complete thesis is on file
In the Rensselaer Polytechnic Institute Library

Approved by:
Dr. George Nagy, Thesis Adviser

Rensselaer Polytechnic Institute
Troy, New York

May, 2008

ABSTRACT

The Wang Notation Tool (WNT) is a semi-automatic, interactive tool that converts tables from HTML pages to Wang notation and corresponding XML representation. Both are layout independent representations of tables where all relationships between cells are recorded in an abstract form that does not rely on the physical structure of tables. WNT requires minimal interaction to delineate the categories in a table, from which an intermediate category tree describing the relationships within each category is determined. The category trees are shown to the user for correction and/or approval. User correction at this step makes WNT robust because the user can modify the automatically generated category tree in almost any way. The approved category trees are used to generate a description of the relationship between each delta (content) cell and the categories as well as an XML representation of tables based on an ontology describing general trees. With current training methods, layout independent representations were generated for 98% of all tables, and were generated correctly for 71% of all tables. Evaluation indicates that with further training, most users will be able to rapidly and correctly generate a layout independent representation of tables using WNT.