Adaptive Conversion of Web-Table Headers to Canonical Form Using XY trees

By

Ramana Chakradhar Jandhyala

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:

George Nagy, Thesis Adviser

Rensselaer Polytechnic Institute
Troy, New York

May, 2010
(For Graduation May 2010)

ABSTRACT

Tables can be laid out in several possible ways, but are usually divided structurally into a stub head, two header regions, and a body. A method is proposed for adaptively transforming the headers of web-tables to an algorithm-friendly canonical format, from which layout-independent information can be easily extracted for the purpose of interpretation. The topology of tables forms a hierarchy of rectangles obtained by alternating horizontal and vertical cuts, which can be represented using XY trees. Information about containment and adjacency of these rectangles can be coded into one-dimensional, parenthesized strings containing cell content and cell locations. These strings are used to determine whether corresponding header regions of two tables are similar to each other. If two such regions are deemed similar and one of them has already been transformed, then the other is automatically converted by adapting to the known transformation rules. Algorithms for extracting the linear parenthesized notation from tables and vice versa are implemented in Visual Basic for Excel (VBA), while the similarity and adaptive learning algorithms are implemented in Python.

Errors in the adaptive transformation are corrected by the user. The adaptive method reduces user intervention by automating repetitive transformations. The method is applied to column header regions of 200 web-tables sampled randomly from a larger corpus of 1005 tables drawn from nine different geo-political and scientific sources. The headers are copied into Excel worksheets and anomalies are interactively corrected. For the 28 non-trivial headers, the ratio of manual to automatic transformations is found to be 0.12.