# Schema Extraction for Tabular Data on the Web

Marco D. Adelfio Hanan Samet

Department of Computer Science
Center for Automation Research
Institute for Advanced Computer Studies
University of Maryland
College Park, MD 20742, USA

VLDB 2013

# Data Tables

- Many structured datasets never stored in database
- Instead, data stored in manually created spreadsheets or data tables within larger documents (HTML, PDF, DOC)
  - character strings positioned in a two-dimensional grid format
  - more *data dense* than prose
  - often communicate tabular structure using implicit, visual cues
  - structure not explicit in document storage formats
- Other datasets exist in private databases, published as spreadsheets or data tables
- Collection of tables on web useful as large, distributed database of relational data

# Simple and Complex Tables

Simple

| Last | First | Gender | Age |
|---|---|---|---|
| Doe | John | Male | 29 |
| Doe | Jane | Female | 30 |
| Smith | Kate | Female | 42 |

header row
followed by
multiple data rows

# Simple and Complex Tables

Simple

| Last | First | Gender | Age |
|---|---|---|---|
| Doe | John | Male | 29 |
| Doe | Jane | Female | 30 |
| Smith | Kate | Female | 42 |

header row
followed by
multiple data rows

---

Complex

| | | |
|---|---|---|
| **Patent Applications by Residents** | | |
| Data Source: worldbank.org | | |
| (showing top countries in each continent) | | |
| **Country** | **Residents** | **Applications** |
| ***North America*** | | |
| United States | 307,007,000 | 224,912 |
| Canada | 33,739,900 | 5,067 |
| Mexico | 112,033,369 | 822 |
| | *N.A. Total* | *230,801* |
| ***Asia*** | | |
| Japan | 127,557,958 | 295,315 |
| China | 1,331,380,000 | 229,096 |
| South Korea | 48,747,000 | 127,316 |
| | *Asia Total* | *651,727* |
| | | |
| Note: data from 2009 | | |

title
notes
header
group headers
data rows
aggregates
blank

## Related Work

- Many existing methods use data from web tables
  - WebTables – exposing HTML tables to search queries [Cafarella et al. WebDB'08, VLDB'08]
    - Detects relational HTML tables, determines which tables have headers
    - Rule-based classifier with real valued features
    - Emphasizes recall over precision
  - Entity resolution of table columns [Limaye et al. VLDB'10]
  - Knowledge base expansion [Yin et al. WWW'11]
  - Incorporate table data into "knowledge taxonomy" [Wang et al. VLDB'11]
- Existing methods for web data not designed to handle more complex table structures found in many spreadsheets and substantial subset of HTML tables.
- Want a method for determining function of individual table rows
  - WebTables does this, but only to detect if the first row is a header row.
    - Q: Can this be extended to other rows and other row classes?
  - Method using conditional random fields (CRFs) exists for classifying rows of ASCII characters that serve as table rows in plain-text documents [Pinto et al. SIGIR'03].
    - Q: Are CRFs effective for table classification in other formats?

# Row Class Definitions

- **Header (H)**
  - cell values describe those contained in subsequent data rows within the same column
- **Data (D)**
  - data records (corresponding to relational *tuples*)
- **Title (T)**
  - describes the entire data collection found in the data table
- **Group Header (G)**
  - provides category for subsequent rows. For example a table containing demographic data about cities may be grouped by country.
- **Aggregate (A)**
  - summaries (typically numeric) of preceding rows, such as totals/subtotals
- **Non-relational (N)**
  - notes, clarifications, or any text that does not contribute data or structure to the data table
- **Blank (B)**
  - contains only empty cells

# Row Classification Process

**Problem:** Given new table, classify rows so that data and structure information can be cleanly extracted.

**Approach:**

1. Extract Cell Attributes
   - Cell attributes include cell formatting information, fonts, alignments, etc.
   - Available cell attributes vary across table formats
     - e.g., HTML tables can use the <TH> HTML tag to indicate a header cell, but spreadsheets have no header cell indicator
     - Use general attributes – visual properties common to all table formats.
2. Compute Row Features
   - Transform cell attributes into row features suitable for passing to a machine learning algorithm
   - Use feature representation that captures human table-processing observations
3. Classify Rows
   - Conditional random field model used as sequence classifier
   - Incorporate row class transition statistics, rather than classifying rows independently

## Cell Attributes

- **Style**
  - IsBold?, IsItalic?, IsUnderlined?, IsColored?, Font, Format

- **Value**
  - IsEmpty?, IsNumeric?, IsDate?, IsShortText?, IsLongText?, IsTotal?

- **Layout**
  - IsMerged?, Alignment

- **Neighboring**
  - MatchesNeighborAboveX?, MatchesNeighborBelowX? (where X is one of the Format, Value, or Layout attributes)

# Row Features

- Explicit table attributes describe *cells*, but we are classifying *rows*
- WebTables, ASCII CRF classifier encode cell attributes into continuous row features
  - If $x$ of $y$ cells in row $r$ have attribute $a$:

$$f_a(r) = \frac{x}{y}$$

- Hypothesis: We can achieve better generalization of our training data by discretizing features along both row width and attribute frequency
  - Could do this by setting

$$f_{a:x \text{ of } y}(r) = \begin{cases} 1 & \text{if } x \text{ of } y \text{ cells in row } r \text{ has attribute } i \\ 0 & \textit{otherwise} \end{cases}$$

  - Result: lots of features, with limited generalization for large (i.e., wide) tables

# Logarithmic Binning of Features

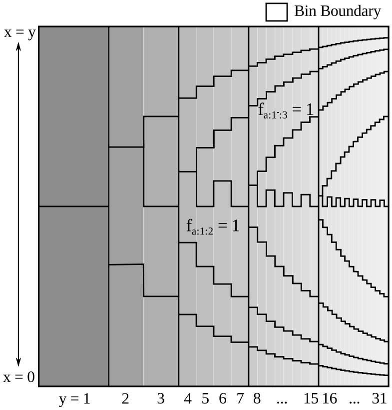
- Developed logarithmic binning encoding for row features
- For row $r$, in which $x$ of $y$ cells exhibit attribute $a$ and $x \leq y/2$

$$f_{a:u:w}(r) = \begin{cases} 1 & \text{if } u = \lfloor log_2(x) \rfloor \\ & \text{ and } w = \lfloor log_2(y) \rfloor \\ 0 & \textit{otherwise} \end{cases}$$

- And for $x > y/2$:

$$f_{a:\bar{u}:w}(r) = \begin{cases} 1 & \text{if } u = \lfloor log_2(y - x) \rfloor \\ & \text{and } w = \lfloor log_2(y) \rfloor \\ 0 & \textit{otherwise} \end{cases}$$

- Assign special value to represent $log_2(0)$.



# Row Classification

- What sequence of row classes best describes a table?

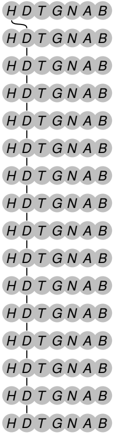
| | | | |
|---|---|---|---|
| **Patent Applications by Residents** | | | H D T G N A B |
| Data Source: worldbank.org | | | H D T G N A B |
| (showing top countries in each continent) | | | H D T G N A B |
| **Country** | **Residents** | **Applications** | H D T G N A B |
| ***North America*** | | | H D T G N A B |
| United States | 307,007,000 | 224,912 | H D T G N A B |
| Canada | 33,739,900 | 5,067 | H D T G N A B |
| Mexico | 112,033,369 | 822 | H D T G N A B |
| | *N.A. Total* | *230,801* | H D T G N A B |
| ***Asia*** | | | H D T G N A B |
| Japan | 127,557,958 | 295,315 | H D T G N A B |
| China | 1,331,380,000 | 229,096 | H D T G N A B |
| South Korea | 48,747,000 | 127,316 | H D T G N A B |
| | *Asia Total* | *651,727* | H D T G N A B |
| | | | H D T G N A B |
| Note: data from 2009 | | | H D T G N A B |

- Many row class sequences are possible.

# Row Classification

- What sequence of row classes best describes a table?

| | | |
|---|---|---|
| **Patent Applications by Residents** | | |
| Data Source: worldbank.org | | |
| (showing top countries in each continent) | | |
| **Country** | **Residents** | **Applications** |
| ***North America*** | | |
| United States | 307,007,000 | 224,912 |
| Canada | 33,739,900 | 5,067 |
| Mexico | 112,033,369 | 822 |
| | *N.A. Total* | *230,801* |
| ***Asia*** | | |
| Japan | 127,557,958 | 295,315 |
| China | 1,331,380,000 | 229,096 |
| South Korea | 48,747,000 | 127,316 |
| | *Asia Total* | *651,727* |
| | | |
| Note: data from 2009 | | |

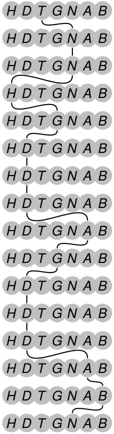



- Many row class sequences are possible.
- Possible row class sequence, if this were a simple table

# Row Classification

- What sequence of row classes best describes a table?

| Patent Applications by Residents | | |
|---|---|---|
| Data Source: worldbank.org | | |
| (showing top countries in each continent) | | |
| **Country** | **Residents** | **Applications** |
| ***North America*** | | |
| United States | 307,007,000 | 224,912 |
| Canada | 33,739,900 | 5,067 |
| Mexico | 112,033,369 | 822 |
| | *N.A. Total* | *230,801* |
| ***Asia*** | | |
| Japan | 127,557,958 | 295,315 |
| China | 1,331,380,000 | 229,096 |
| South Korea | 48,747,000 | 127,316 |
| | *Asia Total* | *651,727* |
| | | |
| Note: data from 2009 | | |




- Many row class sequences are possible.
- True row class sequence for this table
- Example of common row class transitions:
  - **T** in first row
  - **G** transitions to **D**
  - **D** transitions to **D**
  - **D** transitions to **A**
- Classification method needs to capture influence between successive rows.

# Table Grammar

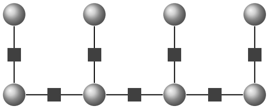
- Most common row class patterns:

| | Spreadsheets | HTML Tables |
|---|---|---|
| | *Spreadsheets* | *HTML Tables* |
| | THD | HD |
| | HD | THD |
| | TBHD | HDA |
| | THDN | THDA |
| | HDN | H(GD)* |
| | TBHD(BN)* | H(BD)* |

- Consecutive instances of same row class are omitted to make patterns more obvious
- Repeated subsequences denoted with (. . . )*
- HD, THD very common for both table formats
- B and N more common in spreadsheets—other common patterns not shared
- Can we utilize common patterns when classifying new table rows?
  - Basis of one of the methods we evaluate (B+A)

# Conditional Random Fields - Overview

- Undirected graphical models, commonly employed in NLP and IR settings



- Linear Chain CRFs useful as sequence classifiers
- For tables, CRF model used to identify most likely sequence of row classes
- Estimated probability of a row class sequence (***Y***), given a sequence of observed features (***X***) has two primary components:

$$exp\left(\sum_j \lambda_j f_j(\boldsymbol{Y}_{i-1}, \boldsymbol{Y}_i, \boldsymbol{X}, i) + \sum_k \mu_k g_k(\boldsymbol{Y}_i, \boldsymbol{X}, i)\right)$$

Reward likely row class transitions and penalize unlikely ones.

Reward likely feature/row class pairs and penalize unlikely ones.

# Dataset

| | Spreadsheets | | HTML | |
|---|---|---|---|---|
| Annotated documents | 1117 | | 1204 | |
| Annotated tables | 2259 | | 13789 | |
| Relational tables | 1048 | (46%) | 928 | (7%) |
| Non-relational tables | 1211 | (54%) | 12861 | (93%) |
| Annotated rows | 435160 | | 20537 | |
| Header rows | 1479 | (<1%) | 978 | (5%) |
| Data rows | 425195 | (98%) | 18906 | (92%) |
| Other row classes | 8486 | (2%) | 653 | (3%) |
| Relational tables: | | | | |
| "Simple" schema | 257/1048 | (25%) | 632/928 | (68%) |
| Multiple header rows | 157/1048 | (15%) | 63/928 | (7%) |
| Other row classes | 784/1048 | (75%) | 263/928 | (28%) |

- Tables sampled from the Web using targeted search engine queries
- 16,048 hand-annotated tables
- Human judge labeled each table as *relational* or *non-relational* and each row with the appropriate *row class*
- Spreadsheets much more likely to be relational
- Relational HTML tables much more likely to have simple schemas

# Methods

- **WT**
  - Uses row features and rule-based classifier developed for “Header Detection” task in original WebTables paper.

# Methods

- **WT**
  - Uses row features and rule-based classifier developed for “Header Detection” task in original WebTables paper.
- **B+A**
  - “Bayes + Automaton” method incorporates global table structure using automaton to enforce common row class patterns. Chosen assignment of row classes has highest overall likelihood (using Bayesian estimation) that satisfies common row pattern.

# Methods

- **WT**
  - Uses row features and rule-based classifier developed for “Header Detection” task in original WebTables paper.
- **B+A**
  - “Bayes + Automaton” method incorporates global table structure using automaton to enforce common row class patterns. Chosen assignment of row classes has highest overall likelihood (using Bayesian estimation) that satisfies common row pattern.
- **CRF-C**
  - CRF classifier using our cell attributes and a continuous feature encoding

## Methods

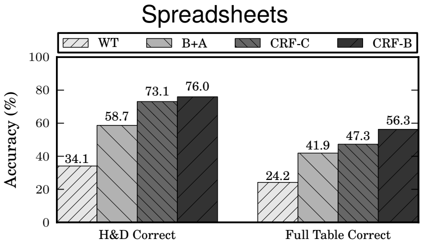
- **WT**
  - Uses row features and rule-based classifier developed for “Header Detection” task in original WebTables paper.
- **B+A**
  - “Bayes + Automaton” method incorporates global table structure using automaton to enforce common row class patterns. Chosen assignment of row classes has highest overall likelihood (using Bayesian estimation) that satisfies common row pattern.
- **CRF-C**
  - CRF classifier using our cell attributes and a continuous feature encoding
- **CRF-B**
  - CRF classifier using our cell attributes and features encoded using logarithmic binning

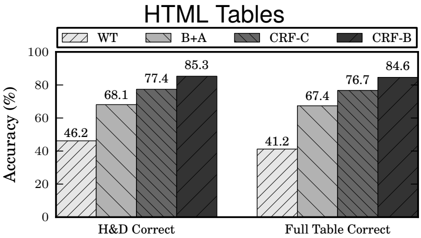
# Row-Level Accuracy

| | WT | B+A | CRF-C | CRF-B |
|---|---|---|---|---|
| Spreadsheets | 97.6% | 96.7% | 99.3% | 99.3% |
| HTML Tables | 92.3% | 92.7% | 98.2% | 98.1% |

- Measured % of row classes from classifier that match true row classes
- Experiments conducted using 10-fold cross validation on relational tables
- All methods achieve $>$ 90% accuracy on the row classification task for both spreadsheets and HTML tables
- Spreadsheet accuracy rates slightly higher
- Data rows account for very high percentage of all rows, so it's expected that all methods do fairly well on a per-row basis

# Full Table Accuracy




- Two measures of full table accuracy
  1. "H & D Correct" measures percent of tables in which all H and D rows are correctly classified
     - Many applications predominantly concerned with header and data rows
  2. "Full Table Correct" measures percent of tables in which rows of all row classes are correctly classified
- CRF-B achieves highest full table accuracy for both H&D and full tables.
- Accuracy higher on HTML tables for all methods, despite lower row-level accuracy. Likely due to the higher proportion of "simple tables" in the HTML dataset.

# Experimental Analysis

- CRF-C vs CRF-B
  - Row-level accuracy approximately equal, but CRF-B better on full tables
    - Improved accuracy with CRF-B for H, T, A.
    - Decreased accuracy for G, N
- Row class ambiguity
  - In spreadsheets
    - True D classified as N (0.16% of spreadsheet rows), G as N (0.14%)
  - In HTML tables
    - True D classified as H (0.34% of HTML table rows), G as N (0.32%), H as D (0.24%), A as D (0.24%)
- Application to Existing Dataset
  - Tested CRF-B method on publicly available dataset of nearly 6,000 HTML tables [Limaye et al. VLDB'10]
  - Simple tables accounted for between 78% and 98% of the tables in three collections of tables that we examined.
  - CRF-B method achieved full table accuracy rates between 89% and 99% on these collections, higher than percentage of simple tables in each case.

## Conclusions

- Using conditional random fields to classify table rows allows segmentation of tables by row function.
- Logarithmic binning improves row classification accuracy by generalizing row features.
- Using CRF-based method to pre-process data tables
    1. increases the pool of available tables, since complex tables and spreadsheets need not be discarded, and
    2. improves the accuracy of table processing methods by isolating segments of tables that are not relevant to the application.
- Automated schema extraction makes data and structure of data tables available to search engines.
- Future Work
    - Take advantage of common patterns in column attributes to extract column-level schema information
    - Spreadsheet search engine with column and row-level predicates

## Acknowledgements

- Thanks to:
  - Google Research
  - National Science Foundation

# CRF-C vs CRF-B

| Row | | CRF-C | | CRF-B | | Change in |
|---|---|---|---|---|---|---|
| Class | Count | Precision | Recall | Precision | Recall | F-Measure |
| **Spreadsheets** | | | | | | |
| D | 425376 | .999 | .999 | .998 | .998 | −.001 |
| H | 1486 | .937 | .915 | .945 | .915 | +.007 |
| B | 3792 | .874 | .862 | .908 | .974 | +.071 |
| T | 702 | .739 | .756 | .766 | .822 | +.046 |
| G | 1312 | .669 | .480 | .758 | .385 | −.048 |
| N | 1877 | .576 | .709 | .446 | .639 | −.111 |
| A | 615 | .965 | .703 | .991 | .890 | +.123 |
| **HTML Tables** | | | | | | |
| D | 18920 | .988 | .995 | .991 | .995 | +.001 |
| H | 979 | .921 | .908 | .911 | .939 | +.011 |
| B | 214 | .852 | .719 | .984 | .953 | +.188 |
| T | 154 | .702 | .717 | .875 | .913 | +.184 |
| G | 112 | .667 | .353 | .545 | .176 | −.195 |
| N | 69 | .667 | .095 | .120 | .143 | −.037 |
| A | 89 | .059 | .074 | .706 | .444 | +.479 |

# Confusion Matrix for Spreadsheets

| | Row label (assigned) D | H | B | T | G | N | A | Row Sum |
|---|---|---|---|---|---|---|---|---|
| Row label (true) D | 97.54% | 0.00% | 0.04% | 0.00% | 0.01% | 0.16% | 0.00% | 97.75% |
| H | 0.02% | 0.31% | 0.00% | 0.00% | 0.01% | 0.00% | 0.00% | 0.34% |
| B | 0.01% | 0.00% | 0.84% | 0.00% | 0.00% | 0.01% | 0.00% | 0.87% |
| T | 0.00% | 0.00% | 0.00% | 0.13% | 0.00% | 0.03% | 0.00% | 0.16% |
| G | 0.04% | 0.00% | 0.00% | 0.00% | 0.12% | 0.14% | 0.00% | 0.30% |
| N | 0.05% | 0.01% | 0.04% | 0.04% | 0.02% | 0.28% | 0.00% | 0.43% |
| A | 0.01% | 0.00% | 0.00% | 0.00% | 0.00% | 0.00% | 0.13% | 0.14% |
| Col Sum | 97.66% | 0.33% | 0.93% | 0.18% | 0.15% | 0.62% | 0.13% | |

# Confusion Matrix for HTML Tables

| Row label (true) \ Row label (assigned) | D | H | B | T | G | N | A | Row Sum |
|---|---|---|---|---|---|---|---|---|
| D | 91.64% | 0.34% | 0.02% | 0.00% | 0.02% | 0.03% | 0.08% | 92.12% |
| H | 0.24% | 4.47% | 0.00% | 0.03% | 0.02% | 0.00% | 0.00% | 4.76% |
| B | 0.05% | 0.00% | 0.99% | 0.00% | 0.00% | 0.00% | 0.00% | 1.04% |
| T | 0.00% | 0.05% | 0.00% | 0.68% | 0.02% | 0.00% | 0.00% | 0.75% |
| G | 0.08% | 0.02% | 0.00% | 0.03% | 0.10% | 0.32% | 0.00% | 0.55% |
| N | 0.19% | 0.03% | 0.00% | 0.03% | 0.03% | 0.05% | 0.00% | 0.34% |
| A | 0.24% | 0.00% | 0.00% | 0.00% | 0.00% | 0.00% | 0.19% | 0.44% |
| Col Sum | 92.45% | 4.91% | 1.00% | 0.78% | 0.18% | 0.41% | 0.28% | |