Synthesizing Union Tables from the Web

Xiao Ling
University of Washington

Alon Halevy, Fei Wu, Cong Yu
Google Research

Overview

http://www.publicschoolreview.com/county_schools/stateid/AR/county/5007

Home > Arkansas > Benton County Public Schools

Benton County Public Schools

There are 57 public schools in Benton County, Arkansas, serving 37,224 students. You can narrow this list by selecting school levels above the table below, or specifying additional search criteria.

Benton County High Schools - Arkansas

Show: All Schools, **High Schools**, Middle Schools, Elementary Schools | Private Schools

Town	School	# Students	Grades
Bentonville	Bentonville High School	3333	9-12
Decatur	Decatur High School	120	9-12
Gentry	Gentry High School	417	9-12
Gravette	Gravette High School	526	9-12

http://www.publicschoolreview.com/county_schools/stateid/MA/county/25003

Home > Massachusetts > Berkshire County Public Schools

Berkshire County Public Schools

There are 46 public schools in Berkshire County, Massachusetts, serving 17,581 students. You can narrow this list by selecting school levels above the table below, or specifying additional search criteria.

Berkshire County High Schools - Massachusetts

Show: All Schools, **High Schools**, Middle Schools, Elementary Schools | Private Schools

Town	School	# Students	Grades
Adams	Berkshire Arts And Technology Charter Public School *(Charter school)*	216	6-12
Cheshire	Hoosac Valley High School	692	7-12
Dalton	Wahconah Regional High School	628	9-12

http://publicschoolreview.com/state_special_education_schools/stateid/MN

Home > Minnesota > Minnesota Special Education Schools

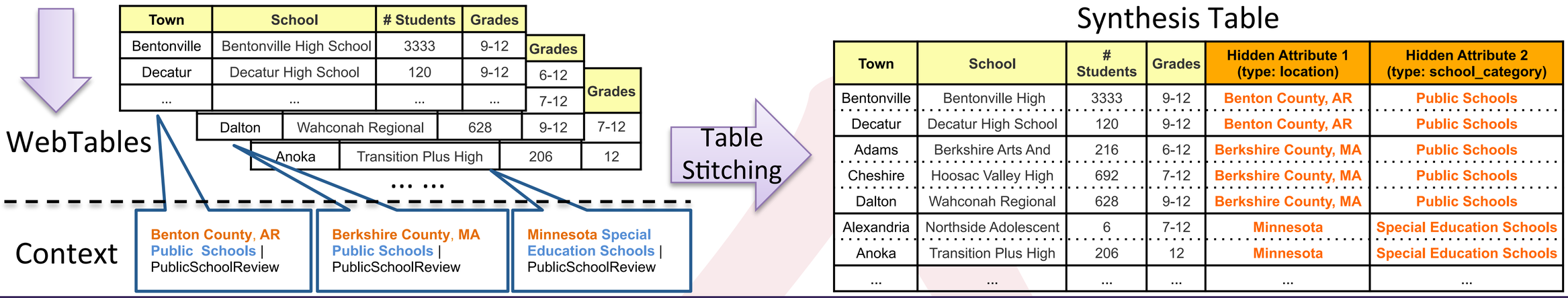
Minnesota Special Education Schools

There are 276 special education schools in Minnesota, serving 15,100 students. You can narrow this list by selecting school levels above the table below, or specifying additional search criteria.

Minnesota Special Education High Schools:

Show: All Schools, **High Schools**, Middle Schools, Elementary Schools

Town	School	# Students	Grades
Alexandria	Northside Adolescent School	6	7-12
Andover	Bridges High School	91	12
Anoka	Transition Plus High School	206	12
Apple Valley	917 Intra-dakota Educational Alternative	85	KG-12



Summary

- Goal: **structurally organizing individual tables with necessary context**
- Method: a **segment-based multiple sequence alignment algorithm** for extracting **hidden table attributes** from the **table context** in the form of word sequences. Given candidate segments from different heuristics as input, the algorithm seeks an optimal alignment of multiple sequences and determines the proper segmentations.

Key Ideas

- No direct supervision
- *Jointly* predicts segmentation and alignment
- The same candidate segment from *multiple* sources more likely to be useful

Candidate Segments

- (SEP) Punctuation/Tag Separators
- (LCS) Longest Common Subsequences
- (WK) Wikification Entities

Table Context

- Web page titles
- Surrounding text of the tables

Segment-based Multiple Sequence Alignment

- Let $score(s_1, s_2) \in \{\lambda_{h_1}, \ldots, \lambda_{h_n}, \lambda_{gap}, 0\}$ where h_i is the ith heuristic.
- Pair-wise Alignment:

Input: Two sequences of tokens T_1 and T_2 of size n_1 and n_2 and two sets of candidate segments S_1 and S_2 respectively.
Output: The best alignment of segments in T_1 and T_2.
Initialization: A chart C of size $(n_1 + 1) \cdot (n_2 + 1)$ where $\forall i, C(i, 0) = i \cdot \lambda_{gap}$, $\forall j, C(0, j) = j \cdot \lambda_{gap}$
for $i \leftarrow 1$ **to** n_1, $j \leftarrow 1$ **to** n_2 **do**
  for $s_1 \in S_1^i$, $s_2 \in S_2^i$ *where* $S^i_{l\in\{1,2\}} = \{$*candidate segments ending at* $T_l^i\}$ **do**
    Update the chart at
    $C(i, j) \leftarrow \max(C(i, j), score(s_1, s_2) + C(i - |s_1|, j - |s_2|))$;
  end
end

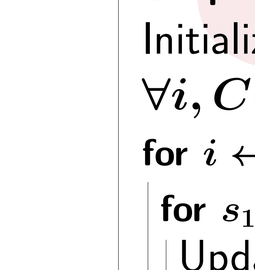
- For multiple sequences, we keep a **profile** of existing results and iteratively compute the best alignment between the profile and the rest.

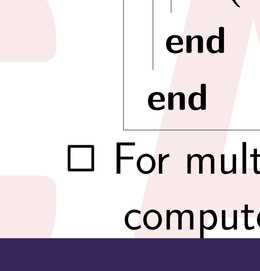
Experiments

Data Set: In a corpus of 130M WebTables grouped by their headers, we sampled 20 groups across 10 different websites (10 tables/group).

Hidden Attribute Extraction: We carry out leave-one-out experiments and evaluate on both cell and column levels.

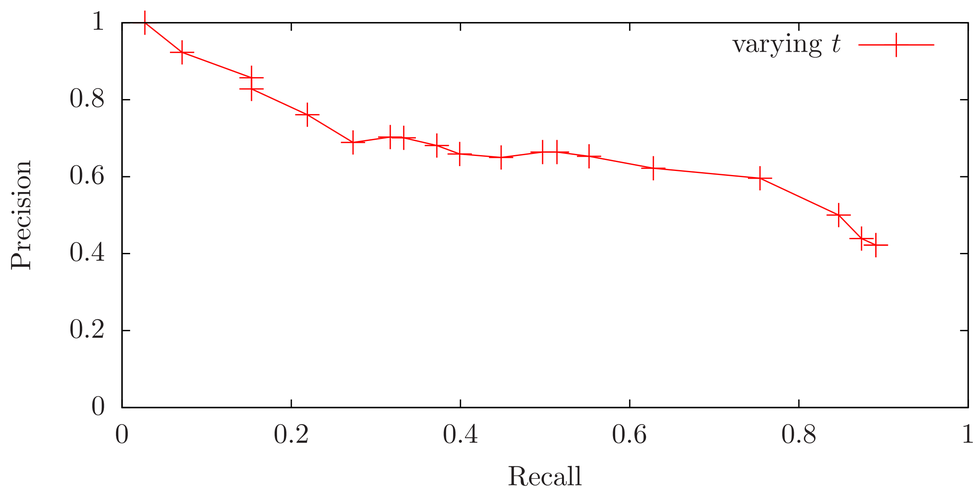
Cell-level performance			
Candidate Segments	Precision	Recall	F1
SEP	0.458	0.260	0.332
LCS	0.630	0.478	0.543
SEP+LCS	0.551	0.484	0.516
LCS+WK	0.650	0.516	0.575
SEP+LCS+WK	0.627	0.703	0.663
Column-level performance			
SEP+LCS+WK	0.387	0.436	0.410





- A combination of syntactic (LCS,SEP) and semantic (WK) candidates yields the best results.

Hidden Attribute Types: We match the values in the extracted cells to an existing database of **isA relations**. If a significant number ($t\%$) of values in a column get mapped to a common class in the isA database, we use the class name as the attribute name. The value of t is varied to get the following curve.



†Work done during the first author's internship at Google. xiaoling@cs.washington.edu, {halevy, wufei, congyu}@google.com