A Rectangle Mining Method for Understanding the Semantics of Financial Tables

Xilun Chen*, Laura Chiticariu†, Marina Danilevsky‡, Alexandre Efthimiopoulos‡ and Prithviraj Sen‡
‡: IBM Research - Almaden
* Cornell University

Structures and Semantics of Tabular Data

ReMine is a two-stage iterative algorithm. Larger rectangles are built by iteratively combining smaller rectangles until convergence. Feature partial order ≼ is utilized to make decisions.

Definition 2. A rectangle 𝑟 is minimal within a list of rectangles if under partial order ≼, if 𝑟′ ∈ after 𝑟 in it, it satisfies 𝑟 ≼ 𝑟′.

*We only work on minimal rectangles in each iteration because they are "safe" to be combined or attached with non-potential children remaining in 𝑅.

-The Combination stage combines consecutive rectangles with equal features, and all the header rows of the combined rectangles become headers of the new rectangle.

- The Attachment stage attaches a rectangle 𝑟 to the preceding rectangle 𝑟1 as a child if 𝑟1 ≻ 𝑟, and the header row of 𝑟 remains as the header of the new rectangle.

ReMine Example

ReMine starts with each row as a rectangle.

iteration 1 Combination:
- Minimal rectangles: row 3-17
- Row 3,4 will be combined into a single rectangle Rect(3,4); similarly, we have Rect(6,10), Rect(12,16)

iteration 1 Attachment:
- Minimal rectangles: Rect(3,4), Rect(6,10), Rect(12,16), Rect(5), Rect(11), Rect(17)
- Rect(3,4) is attached to Rect(2) becoming Rect(2,4)

iteration 2 Combination: no adjacent rectangles with equal features

iteration 2 Attachment: Rect(2,4) continues to consume Rect(5) becoming Rect(2,5)

iterations 3 & 4 are similar to iteration 2: Rect(2,5) further consumes Rect(6,10) and Rect(11)

- At the end of iteration 4, Rect(211) has feature isEmptySection and stops taking children

iteration 5: Rect(2,11) is attached to Rect(1)

iteration 6: Rect(11,1) consumes Rect(12,16)

iteration 7: Rect(17) is attached to Rect(1,16)

Experiments and Results

Table II: Leave-one-company-out cross validation results. Section III defines Direct and Transitive evaluation schemas.

Datasets

- Financial Table Dataset
- 2015 Q3 financial statements of 6 companies
- 72 tables; manually labeled semantics

ICDAR 2013 Table Competition Dataset
- Only 7 tables exhibit non-trivial row hierarchy structures

Experimental Results

- Results shown in Table II
- Baseline methods
  - A SVM pair-wise classifier trained to predict parent-child relations
  - RIPPER, which learns rules of logical expressions of features, and is more transparent (explainable) than SVM

Evaluation Metrics

- Direct: Used in literature and evaluates only direct parent-child relations
- Feature Abstraction
  - Shown in Table III
  - When removing both rectangle features, transitive F1 decreases to 82% (down 5%)