マルチラベル分類と知識ベースの埋め込みを用いたテーブル中の列の概念決定

竹岡邦紘*, 小山田昌史*, 中台慎二*, 岡留剛**
*日本電気株式会社 データサイエンス研究所, **関西学院大学（問合せ先：k-takeoka@az.jp.nec.com）

\Orchestrating a brighter world
NEC

Introduction

Q. 大量のテーブルデータの中から
欲しいテーブルを見つけることはできるか？

- 列名の曖昧性によって単純な単語のマッチングでは見つけられない
- 列名の欠損などの理由でマッチングすら難しいこともある

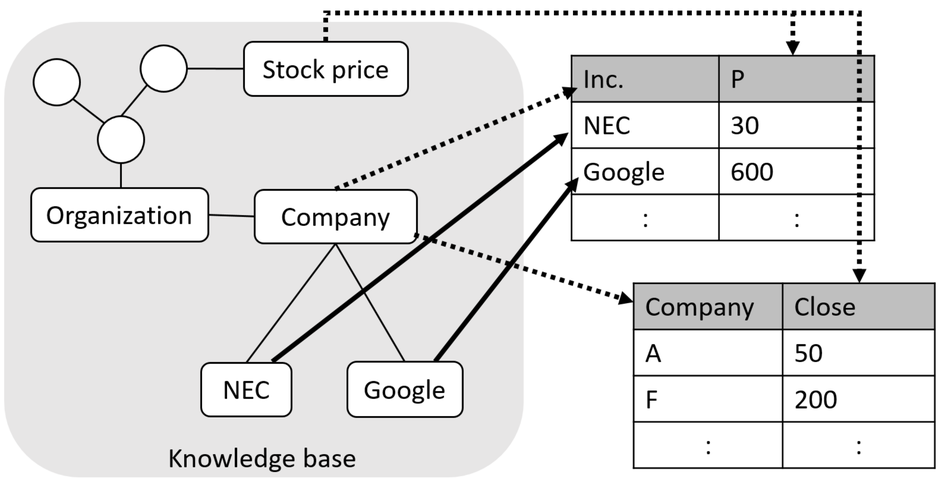
Venetis et al. PVLDB 2011.

テーブルデータの中身から
各列が何を表しているか推定する

各列と知識ベースの“**エンティティ**”を対応付け（アノテーション）

列の種類
- “NE列”: 列中の各要素が知識ベースのエンティティと対応付けられる列
- “リテラル列”: 列中の各要素が対応付けられない列（数値など）

（例）企業名が入っている列はNE列で、株価が入っている列はリテラル列



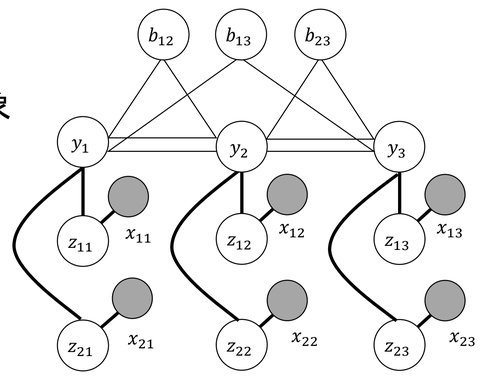
数値を要素とする列を含めた
テーブルの列に対応するエンティティ推定手法の提案
- **ポテンシャルとしてマルチラベル分類器の出力を利用**
- **列間の関係をモデル化し推定精度を向上**

Related work

Limaye et al. PVLDB 2010.

複数の“NE列”に対する同時アノテーションを行う手法

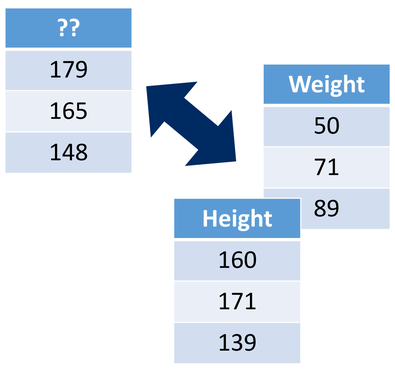
特徴
1. 単語のみで構成される列が対象
2. 列間の相互関係を利用
3. 2列間の関係も同時に推定



課題
- 計算コストが非常に高い
- 数値や記号など各要素が
知識ベースに対応付けられない列に対応できない

Pham et al. ISWC 2016.

数値列を含む列に対するアノテーションを行う手法

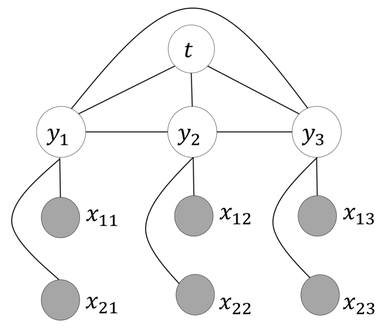


特徴
1. 複数の類似度を同時に利用
2. それらの重要度も学習
3. NE列もリテラル列も区別なく扱う

課題
- 単位変換に弱い
- 列間の関係が利用できない

Proposed method

- ◆ マルコフ確率場によるテーブルのモデル化
- ◆ ポテンシャル関数にマルチラベル分類器を利用
- ◆ 特徴量に知識ベースの埋め込みを利用



同時分布

$$p(\boldsymbol{X}, \boldsymbol{y}, t) = \frac{1}{Z} \prod_c \phi_{yx}(y_c, \boldsymbol{x}_c) \phi_{yy}(y_c, \boldsymbol{y}_{-c}) \phi_{ty}(t, y_c).$$

ポテンシャル関数

1. Column-cell potential: $\phi_{yx}(y_c, \boldsymbol{x}_c) = f_{y_c}(\boldsymbol{x}_c)$
マルチラベル分類器がy_cを出力する「確率」
セルの値と列に対応するエンティティのペアを学習

特徴抽出
- NE列：候補集合の各要素に対して知識ベースの埋め込みを用いて得たベクトル集合の平均と標準偏差
- リテラル列：列内の要素の複数の統計量と文字列的な特徴量

2. Column-column potential: $\phi_{yy}(y_c, \boldsymbol{y}_{-c}) = g_{y_c}(\boldsymbol{y}_{-c})$
マルチラベル分類器がy_cを出力する「確率」
どのような組み合わせでエンティティが共起するのかを学習

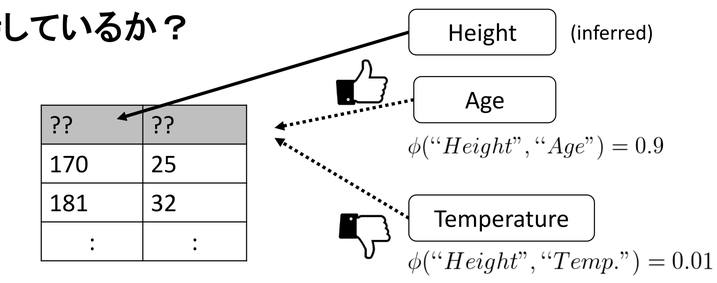
特徴抽出
c 列目以外の列の概念に対して知識ベースの埋め込みを用いて得たベクトル集合の平均と標準偏差

3. Title-column potential: $\phi_{ty}(t, y_c) = \exp\{-d(t, y_c)\}$
埋め込み空間上での
タイトルのエンティティと列のエンティティの距離を利用

目的関数 $\max_{\boldsymbol{y}, t} \prod_c \phi_{yx}(y_c, \boldsymbol{x}_c) \phi_{yy}(y_c, \boldsymbol{y}_{-c}) \phi_{ty}(t, y_c).$

与えられたテーブル$\boldsymbol{X}$を固定した下で最適な潜在変数値の推定
直接最適化することが困難なため、ギブスサンプリングで近似

何を期待しているか？



Experiments

- ◆ 人手でアノテーション済みのUCI Machine Learning Repositoryの183テーブル
- ◆ 5 foldのcross validationで評価
- ◆ 知識ベースとしてWordNetを利用

Method	MAP@5	nDCG@5	Sim@5
Limaye + Pham	0.225	0.291	0.480
Proposed (w/o interdependency)	0.351	0.413	0.537
Proposed (w/ interdependency)	**0.464**	**0.741**	**0.635**

列ごとに独立に推定した場合も、NE列やリテラル列に限定して推定した場合も、
既存手法に比べて推定性能が向上

Conclusion

本研究の貢献：既存研究の課題を解決する手法の提案
1. 多様なセルの値に対応できる
2. 推定性能が高い
3. 新たな種類のデータへ拡張が容易

References

- ◆ Limaye et al. Annotating and searching web tables using entities, types and relationships. PVLDB, 3(1-2): 1338-1347, 2010.
- ◆ Pham et al. Semantic labeling: A domain-independent approach. In ISWC, pages 446-462, 2016.
- ◆ Venetis et al. Recovering semantics of tables on the web. PVLDB, 4(9): 528-538, 2011.