

東京大学大学院 工学系研究科 システム創成学専攻
早矢仕 晃章

データ市場における知識共有による異分野データ連携と利活用を促進する技術の開発

モデル1: データの概要情報の類似度の高いデータ同士は類似しており、似た変数ラベルを有している。 データの概要 (OD) 変数ラベル (VL)

類似度 ($D(I_1, D_1), D(I_2, D_2)$)

変数ラベル未知のデータの概要

静岡県庁が提供する街路灯に関する設置データ ($D(I_1, D_1)$)

東京都における年齢別の人口割合 ($D(I_2, D_2)$)

類似度 ($D(I_1, D_1)$)

東京都が設置した街路灯の位置情報 ($D(I_1, D_1)$)

緯度
経度
光束

地域名
人数
性別
年齢

モデル2: 変数ラベルには同じデータに同時に登場する頻度が高いものが存在する (共起性)

性別 住所 緯度 人口 市町村名 天気 年齢 体系 身長 緯度 日付 光束 気温 事故の件数 事故の種類 街灯種類

事故の種類

事故の件数

性別 体系 身長 年齢 人口

類似度を考慮 (モデル1) 共起性を考慮 (モデル2)

第1ノード (VL) 第2ノード (VL) 第3ノード (VL) 第4ノード (VL)

単語-VL行列 $E (=MR^T)$ VL共起性行列 $C (=RR^T)$ 単語-VL行列 $EC (=MR^T R R^T)$

単語-VL行列 $E (=MR^T)$

VL共起性行列 $C (=RR^T)$

単語-VL行列 $EC (=MR^T R R^T)$

● String Matching ● Matrix E ● Matrix EC

799件OD (総変数ラベル数: 4160, 総単語数: 30767, 変数ラベルの種類: 3216, 単語の種類: 1935) を訓練データとして、静岡県オープンデータポータルから取得した50件のテストデータの変数ラベル推定を実施。

Hayashi, T., "Inferring Variable Labels Considering Co-occurrence of Variables in Data Jackets," IEEE-ICDM Workshops 2016, Data Market for Co-evolution of Sciences and Business (MoDAT), pp.783-787, 2016.

Hayashi, T., Ohswa, Y., "Matrix-based Method for Inferring Variable Labels Using Outlines of Data in Data Jackets," The Pacific-Asia Conference on Knowledge Discovery and Data Mining 2017 (PAKDD2017), Advances in