機械学習を利用したデータベースの再構成

～構造が悪くなったデータベース表をよみがえらせる～

応用情報科学研究科　応用情報科学研究科専攻

◎D3・濱地康平（はまじこうへい）、教授・中本幸一（なかもとゆきかず）

キーワード　**関係データベース、リファクタリング（再構成）、機械学習アルゴリズム**

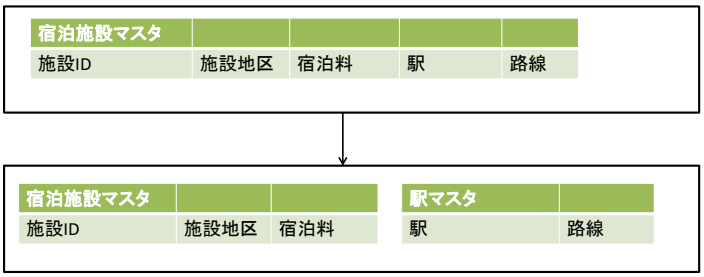
研究概要

企業システムなどで利用されたデータベースの表は，最初は「きれいに」設計した表が，時間がたってくるにつれて，利用環境が変化して新たにカラムが追加されて，だんだんと構造が悪くなり「汚くなってくる」ものです．本研究では、こうしたデータベースの表を，機械学習技術である k 近傍法と呼ばれる類似の情報の塊を見つける方法を使って，きれいな表に再構成（リファクタリング）することを支援するソフトウェアを研究しています（右図）．

システムの実行イメージ

アピールポイント

構造が悪くなったデータベースの表の特徴として，
C1）NULL データ（何もはいない）セルが至るところにある
C2）表のカラム名に異質なものが混ざっている
という 2 つに注目しています．
通常，このような汚くなったデータベース表の該当部分を，専門家は直観により見出しています．本ソフトウェアでは k 近傍法と呼ばれる，機械学習の手法を用います．これを使って，
C1）については何も入っていないセルの分布具合，
C2）についてはカラム名の語の関係度合（右図）
を調べ，どの部分を再構成すればいいかを利用者に提示し，データベース表の再構成を支援します．
企業システム等，一般のデータベース，csv データに適用可能です。



カラムリファクタリングの例