SQL クエリ説明文を利用したテーブル検索

小菅　哲哉

データ分析は，収集したデータを用いて知見を得るために有効な手段であり，ビジネスなどで活用されている．今後もさらに加速するデジタル化により，データ分析はより幅広い分野で活用され，ますます重要性を増していくと考えられる．しかし，データ分析をする際に起こりうる課題として，目的となるデータがどこにあるかわからず，データ分析を始めるために時間がかかるということが考えられる．

そこで，本論文ではデータ分析に関連する技術の中でも，リレーショナルデータベースにおいてこの課題を解決するために，SQL クエリと，その SQL クエリを自然言語で説明する SQL クエリ説明文を利用して，リレーショナルデータベース上のテーブルを検索する手法を提案する．例えば，「8 月の占いのアクティブユーザー数」のような SQL クエリ説明文が与えられたとき，占いについてのユーザーの情報を管理するテーブルである fortune_user_log のようなテーブルを検索することを目的とする．これによって，リレーショナルデータベースを利用したデータ分析をする際に必要なテーブルを素早く探し出し，不要な時間を削減し作業効率の向上が期待できる．

本研究で取り組んだ手法は，テーブルの出現頻度によるアルゴリズム，マルチラベル分類，ランキング学習である．ここで，テーブルの出現頻度によるアルゴリズムとは，データセットの SQL クエリ中に登場する頻度が高いテーブルを返すアルゴリズムである．マルチラベル分類では，SQL クエリ説明文から bag-of-words によってベクトルを作成し，テーブル名をラベルとした．ランキング学習では，TFIDF や BM25 をはじめとした特徴量を利用し，テーブル名をランク付けした．また，本研究において SQL クエリ説明文に対して正解となるテーブルは，その SQL クエリ説明文に対応する SQL クエリ中に登場するテーブルとする．

実験には株式会社 Gunosy から提供された SQL ファイルをデータセットとして利用し，提案した手法をそれぞれ nDCG@k (k=1,3,5,10) を用いて比較した．実験の結果，3 つの手法のうちランキング学習が最も良い性能となった．また，ランキング学習に用いた特徴量がどれだけ性能へ影響しているかを明らかにするために，ランキング学習において特徴量を減らした実験を行った．実験の結果，ランキング学習における各特徴量の影響の大小を確認できた．特に，テーブル名と SQL クエリ説明文から得られる特徴量を減らすことで性能が非常に大きく低下することが明らかになった．

（指導教員　　加藤　誠）