2C-07

スキーマ構成文字列と主キー制約情報に基づくデータベース冗長カラムの推定

佐藤 彰洋† 鹿島 理華† 永嶋 規充†

三菱電機株式会社 情報技術総合研究所†

1. はじめに

データベースのデータ統合のためには、データベースのテーブル間参照関係を考慮しながら統合の候補となる冗長カラムを抽出する必要がある。しかし、一般的なデータベースでは主キー制約は定義されているが、性能等の理由から参照制約は付与されていないことが多く、テーブル間の参照関係を人手により定義する必要があり、大きな工数が必要となる。さらに、参照関係が明らかな場合であっても、性能向上のために参照先のテーブルに存在するカラムを参照元のテーブルにも持つ非正規化されたテーブルが存在するため、これを人手により抽出するための工数も必要となる。

2. 課題

対象データベースのテーブル定義情報を使用して、テーブル間で類似カラムを抽出するスキーママッチング技術がある[1]。しかし、スキーママッチング技術で抽出するのは複数のテーブル間で類似の項目であり、テーブル間の参照関係を考慮しておらず、冗長カラムの抽出にそのままでは利用できない。

上記の課題を解決するため、テーブル定義情報を用いたスキーママッチングに加え、主キー制約情報の組み合わせにより、参照関係を考慮した上で冗長カラムを抽出する。

3. スキーマベース冗長カラム推定方式

提案手法の構成を図1に示す。従来のスキーママッチング技術で抽出されるのは、図1中の P であるが、提案手法では参照関係の抽出と、冗長カラム抽出により、非正規化されたテーブルに存在する冗長カラム Q と、参照キーであり冗長カラムではないカラム R の抽出を実現した。提案手法は、スキーママッチング対象となるデータベースに関して、各テーブルのカラム間類似性をスキーママッチングにより算出するフェイズと、スキーママッチングの結果から参照関係を推定するフェイズ、冗長カラムを抽出するフェイズからなる。以降、各フェイズについて説明する。

3.1 スキーママッチング

対象となるテーブル定義情報が入力されると、任意のカラム名ペアを取得し、ペア間の類似度を算出する。類似度の算出には名称そのものの類似性と、シノニムを組み合わせたハイブリッド型のマッチャーである Name Matcher[2]を使用する。カラム名はトークンに分割し、トークンごとに 3-gram 単位での類似度および類義語辞書中の該当類義語同士の一致度に応じて類似度を算出し、総合してカラム名同士の類似度を 0 から 1 までの実数値で算出する。

Estimation of Duplications using Schemata and Primary Key Constraints

†Akihiro Sato、 Rika Kashima、 Norimitsu Nagashima、 Information Technology R&D Center、 Mitsubishi Electric Corporation

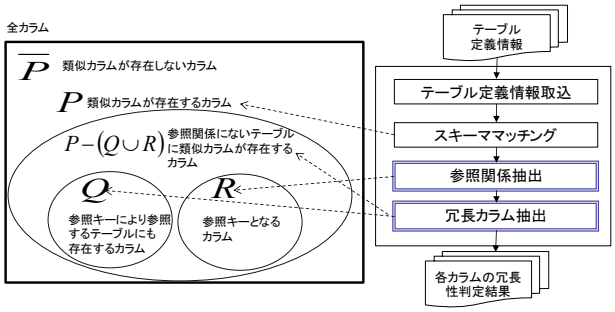

図1:スキーマベース冗長カラム推定方式

3.2 参照関係抽出

任意のテーブル A とテーブル B のカラム名ペアについて類似度を取得し、類似度 $Score(C_{An},C_{Bn})$ が閾値 θ 以上となるカラム名ペア(候補ペア)を抽出する。ここで C_{An} はテーブル A 中のカラム、C_{Bn} はテーブル B 中のカラムを表す。なお、閾値 θ が 1.0 の場合、テーブル定義情報が完全一致するカラムのみ候補ペアとすることに相当し、類似度 0.7 程度では名称の一部が異なるカラムも候補ペアに含めることに相当する。

そして、(C_{An},C_{Bn})に関して、以下の条件に合致する候補ペアの組み合わせ$\{(C_{A1},C_{B1}),(C_{A2},C_{B2}),\ldots(C_{Ai},C_{Bi})\}$が存在する場合、テーブル B から A に対する参照関係が存在すると判定し、候補ペアの組み合わせを参照キーカラムとして出力する。

- テーブル A の主キーとなるカラムをすべて含む。
- 同じテーブルに属するカラムが複数回登場しない。
- 候補ペアどうしのデータ型が同一である。
- 類似度の合計値が全組み合わせの中で最も大きい。

3.3 冗長カラム抽出

任意のテーブル A とテーブル B について類似度 $Score(C_{An},C_{Bn})$ が閾値 θ 以上となるカラム名ペア(候補ペア)を抽出し、以下の条件に基づき冗長カラム判定を行う。

- 条件1:テーブル B から A に参照関係が推定されており、かつ候補ペアが参照キーカラムに含まれない場合、冗長カラムである
  条件2:テーブル A と B 間に参照関係が存在しない場合、候補ペアは冗長カラムである

4. 評価実験

提案手法による冗長カラムの推定について有効性を確認するため、データベースのスキーマを用いて評価実験を行った。

4.1 実験条件

実験に用いたデータは、業務ワークフロー処理システムの実データベースから抽出したテーブル定義情報である。テーブル数およびカラム数、テーブルに設定されている制約の数を表 1 に示す。

表 1:実験データの条件

テーブル数	541
うち、主キーが存在するテーブル数	462
カラム数	8966
うち、主キーとなるカラムの数	1055
外部キー	0

ここで 462 テーブル中の任意の 2 テーブルに対してスキーママッチングを行い、得られたカラム名ペア間の類似度を用いて参照関係の推定を行い、スキーママッチングで得られた類似度および参照関係の推定結果を用いて冗長カラムの抽出を行った。なお、類似度閾値 θ は 0.7 と 1.0 を用いた。以上のようにして、全テーブル 541 の中の任意のテーブルの組み合わせに対して算出した 146,070 のスキーママッチング結果から、参照関係の推定および冗長カラムの推定を行った。ただし、冗長カラムの推定において、「最終更新日」「最終更新者」「削除フラグ」といった 6 カラムは、目視確認の結果ほとんどのテーブルに存在することから、テーブル定義のルール上必ず存在するカラムであり、冗長なカラムではないとみなし冗長判定の対象外としている。

推定結果の評価として，評価対象とした業務ワークフロー処理システムのテーブル定義情報のうち，システム設計者によりカラムの論理的な意味が与えられており、人手による参照関係の推定と冗長カラム判定が行われている 10 テーブルを精度評価対象テーブルとして用いた.

4.2 結果と考察

提案手法により冗長カラムの推定を行った結果、閾値 1.0 の場合には全 8,966 カラム中、4,270 カラムが冗長カラムと推定された。このうち、2.3 節の条件1に該当したカラム数は 508、条件2に該当したカラム数は 3,762 であった。条件1に該当したカラムは、参照先テーブルに存在するカラムにもかかわらず、参照元にも存在するという関係性であり、条件2に該当したカラムは、参照関係を持たないテーブルと共通項目を持っているという関係性である。類似度閾値と推定した冗長カラム数の関係を表 2 に示す。

表 2 類似度閾値と推定結果

類似度閾値	条件1に該当したカラム	条件2に該当したカラム	冗長カラム合計
1.0	508	3,762	4,270
0.7	554	4,177	4,731

推定の結果、領域 $\overline{P}$ と推定されたカラムが存在せず、すべてのカラムが領域 Q および領域 R と判定されたテーブルが 15 テーブル存在した。これらのテーブルは、すべてテーブル名が、「テーブル名_WK」や「テーブル名_BK」といった名称であり、ほとんど同じカラムを持つ別テーブルに対して参照関係を持つテーブルであり、一般的にテスト用やデータの一時退避用にあるテーブルをコピーして生成したテーブルであると考えられ、提案手法により機械的に抽出することが可能である。

次に、精度評価対象とした 10 テーブルに関して評価した結果について表 3 に示す。10 テーブルに含まれる全 134 カラムに対して、提案手法により推定された類似カラムは 11、そのうち参照キーとなるカラム（領域 R）は 10 であり、参照キーにより参照されているテーブルに存在するカラムが参照元にも存在すると判定されたカラム（領域 Q）は 1 であった。これは、カラムの論理的な意味から人手により推定したカラムの分類とすべて一致する結果となった。

表 3 精度評価対象テーブルに関する推定結果

類似度閾値	領域 $\overline{P}$	領域 P	領域 Q	領域 R
1.0	123	11	1	10
0.7	123	11	1	10

スキーママッチングの結果のみでは、類似カラム（領域 P）であるという判定しかできないが、提案手法では参照関係の推定を行ったうえで冗長カラムの判定を行うため、参照先のテーブルに存在するカラムを参照元のテーブルに持っているカラムを特定し、冗長なカラムであると判定することができる。

表 3 に示すように、本評価では、類似度閾値 θ を 1.0 から 0.7 へ低下させても推定結果への影響が表れなかった。これは、閾値を低下させても類似のカラムが少ない、すなわち対象テーブル定義の命名規則の統一性が高いと考えられる。ただし、類似度閾値 1.0 においても全カラム中の半分程度が冗長なカラムと推定されていることから、命名規則は統一されているが冗長なテーブル設計であると考えられる。

本評価で用いたような、命名規則は統一されているが冗長なテーブル設計のデータベースをデータ統合の対象とする場合、多数のカラムが冗長なカラムと判定されることが考えられる。このようなデータベースをデータ統合するためには、次に述べる課題の解決が必要となる。

データ統合においてはテーブルの再設計、正規化といった設計作業を行うが、多数のテーブルが存在する実システムでは設計作業の工数も多大なものとなる。そのため、提案手法により抽出した参照関係で繋がっているテーブル群のみ提示し、かつ条件1に該当したカラムは、参照キーカラムにより参照先のテーブルから取得可能なカラムであるため、これらを除外した形で提示するといった、設計作業の効率化を行う方式が必要と考えられる。

5. おわりに

本稿では、対象のテーブル定義を用いてスキーママッチングを行い、主キー制約情報と組み合わせることで、参照関係を推定し冗長カラムを抽出する方式を提案した。既存の業務ワークフロー管理システム管理システムのデータベースに提案手法を適用し、人手により抽出した参照関係および冗長カラム判定と同一となる推定結果を得ており、テーブル間の参照関係が明示でないデータベースに対しても本手法による推定が可能であることを示せたと言える。

今後は、テーブル再設計作業を支援する方式によりデータ統合作業の効率化を図るとともに、命名規則が統一されていないデータベースに対しても提案手法を適用し、冗長カラム推定の評価を行う予定である。

参考文献

[1] Rahm,E. and Bernstein,P.A.:A survey of approaches to automatic schema matching.VLDB J(10) pp.334-350,2001.

[2] Do, H.H. and Rahm,E.:COMA - A System for Flexible Combination of Schema MatchingApproaches,Proc. 28th Intl. Conference on Very Large Databases ,VLDB ,2002.