1A2-1

スキーマ構成文字列と主キー制約情報に基づく外部参照関係の推定

Estimation of Relationships using Schemata and Primary Key Constraints

佐藤 彰洋[*1] Akihiro SATO　鹿島 理華[*1] Rika KASHIMA　谷垣 宏一[*1] Koichi TANIGAKI　山足 光義[*1] Mitsuyoshi YAMATARI

[*1] 三菱電機株式会社 情報技術総合研究所
Information Technology R&D Center, Mitsubishi Electric Corporation

For the understanding of the schema structure, it is necessary to understand the relationship between the unity that is meaningful, not only the structure (table if database schema), but if the relationship is not explicit, for large schema by extracting the connectivity of the Te manually is difficult. In this paper, we describe a method based on the similarity of the characters that make up the schema, the primary key information in the schema, to estimate the reference relationship in the schema.

1. はじめに

データベースのデータ統合やデータ連携では，テーブルに含まれるカラムの情報をどのテーブルのカラムから抽出すればよいかというデータ連携関係を設計するために，テーブル間の外部参照関係が重要である．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ものの，性能等の理由から外部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は明示的でないため，これを人手により定義する必要があり，多大なコストを要する．

これに対し，対象データベースのテーブル内の実データを読み込み，データ間の関係性およびテーブル間の結合関係を推測して，テーブル間の参照関係を推定する方式が提案されている[Ling 2001]．このように実データ自体を処理するような方式の場合，もっとも正しいデータ参照関係を出力するためには実データ全件が必要となり，大量のデータ処理が必要となるため，対象範囲を絞り込んだ上での適用にならざるを得ない．

一方，対象データベースのテーブル定義情報を使用して，データベーステーブル間で類似のカラムを抽出する既存技術として，スキーママッチング技術がある［Rahm 2001］．テーブル定義情報を利用するため，実データと比較するとデータ量が圧倒的に少なく，探索範囲を広く設定可能であるが，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およびテーブル間の連携関係を考慮しておらず，データ連携の設計にそのままでは利用できない．

本稿では上記の課題を解決するため，一般的なデータベースから取得可能な情報であるスキーマ定義情報および主キー成約情報を用いたスキーママッチングにより，外部参照関係を推定する方法を提案する．さらにデータベースシステムのスキーマデータを用いた評価実験の結果と考察について述べる．

2. スキーマベース参照関係推定方式

本稿で述べる手法は，スキーママッチング対象となるスキーマに関して，各テーブルの主キーとなるカラムのみを抽出し，スキーママッチングを行うフェイズと，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からなる．以降，各フェイズについて説明する．

連絡先：佐藤 彰洋,三菱電機株式会社 情報技術総合研究所,
〒247-8501 神奈川県鎌倉市大船 5-1-1,
Sato.Akihiro@ea.MitsubishiElectric.co.jp

2.1 スキーママッチング

まずテーブルに属するカラム間の類似度を算出するため，スキーママッチングを行う．対象となるスキーマデータおよび主キー制約情報が入力されると，スキーマ中に存在する各テーブルに関して，主キーのみからなる中間テーブルを生成する．さらに，中間テーブルと，中間テーブルの元となるテーブル（元テーブル）の任意のカラム名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，類義語辞書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[Do 2002]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に文字 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算出する．

2.2 参照関係推定

次にテーブル間の外部参照関係を推定するため，任意の中間テーブル A と元テーブル 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} は中間テーブル中のカラム，C_{Bn} は元テーブル中のカラムを表す．

そして，ある中間テーブル A と元テーブル B の候補ペア (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，参照関係の推定結果として出力する．なお，中間テーブル A の元となったテーブルと，元テーブル B が同一の場合は抽出対象としない．

- 中間テーブル A の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. 評価実験

提案手法による外部参照関係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．

3.1 実験条件

実験に用いたデータは，受発注管理システムの実データベースから抽出したスキーマ定義情報である．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	339
うち，主キーが存在するテーブル数	322
カラム数	22378
うち，主キーとなるカラムの数	1826
外部キー	0

ここで主キーが存在する 322 テーブルから主キーのみを抽出して中間テーブルを生成し，主キー情報を含む元テーブル 339 テーブルとの間で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に対して類似度閾値を 0.5 から 1.0 まで 0.1 刻みで設定し，参照関係の推定結果を得た．以上のようにして，中間テーブルと元テーブルの組み合わせである 109,158 個のスキーママッチング結果から，外部参照関係の推定を行った．

推定結果の評価として，評価対象とした受発注管理システムのスキーマ定義情報のうち，システム設計者により外部参照関係が示されている 8 テーブルを精度評価対象テーブルとし，これらテーブル間の 7 つの参照関係を正解データとして用いた．なお，本参照関係はデータベースの外部キーとしてではなく，設計書より抽出している．推定された外部参照関係と正解データである7つの参照関係との一致数を算出し，適合率と再現率の調和平均である F 値を参照関係推定精度として評価した．

3.2 結果と考察

提案手法により参照関係の推定を行った結果，閾値 1.0 の場合には 386 の参照関係が推定された．これはマッチャーによりカラム間の類似度が 1.0 と判定されたペアのみから成る参照関係であり，例えばあるテーブルの主キーである「ORDER_ID」および「JUCHU_ID」に対して別のテーブルの「ORDER_ID」および「JUCHU_ID」が外部参照関係にある，といった関係性である．類似度閾値と推定した参照関係の数の関係を表 2 に示す．

表 2 類似度閾値と参照関係推定結果

類似度閾値	推定した参照関係の数	うち精度評価対象テーブルに関する参照関係の数
0.5	1,953	25
0.6	1,427	20
0.7	800	9
0.8	752	9
0.9	386	9
1.0	386	9

8 つの精度評価対象テーブルに関して算出した類似度閾値と参照関係の推定精度の関係を図1に示す．類似度閾値を下げることで，推定される参照関係の数は増加するが，一方で正しくない参照関係が推定結果に含まれることにより推定精度が低下している．例えば類似度閾値 1.0 の場合の参照関係推定結果は，正解データに対して評価した F 値は 0.75 であり，類似度閾値を 0.5 とした場合の F 値は 0.44 であった．

正解データに含まれない参照関係を推定したケースでは，カラム名を構成するトークンについて，単語および登場順が同じになるケースが頻出しており，カラム名を構成するトークンの重要度を一律とし，類義語の登場回数で類似度を評価する NameMatcher では誤検出を起こしたと考えられる．例えば，「受注履歴番号」と「注文履歴番号」という属性のカラムがそれぞれ”JUCHU_RIREKI_NO”と”ORDER_RIREKI_NO”と命名されている場合，これらと” SHUKA_RIREKI_NO”との間の類似度

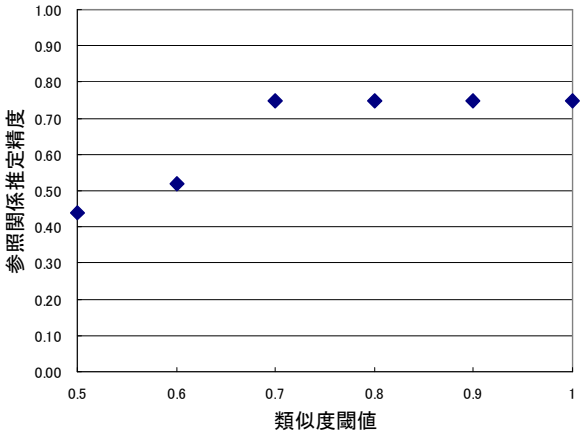

図1 類似度閾値と参照関係推定精度

は同程度となり，類似度閾値を低くするといずれも参照関係の候補として抽出される．

また，正解データではテーブル A,B,C 間の参照関係が「テーブル A から B」および「テーブル B から C」であるのに対して，参照関係の推定結果は「テーブル A から B」「テーブル B から C」「テーブル A から C」の3つが推定された．テーブル A のあるカラム群がテーブル B の主キー群と合致している場合，テーブル B から参照されるテーブル C の主キー群とも合致しているという関係であり論理的には必ずしも誤りとは言えないが，テーブル間の関係性を人間が判断する際には過剰な情報であり，本来除外すべき関係と考えられる．

4. おわりに

本稿では，対象スキーマの主キー情報を用いてスキーママッチングを行うことで，スキーマ内テーブル間の参照関係を推定する手法を提案した．既存の受発注システム管理システムのスキーマに提案手法を適用し，類似度閾値 1.0 にて F 値 0.75 を得ており，テーブル間の参照関係が明示でないスキーマに対して本手法による参照関係の推定効果が評価できたと言える．

ただし，スキーマ中の8テーブルの範囲で類似度が同程度の候補ペアが複数存在し誤検出が発生したことを考えると，大規模なスキーマへの適用時には誤検出が多くなると予想される．そのため，提案手法ではマッチャーとして NameMatcher を採用したが，スキーママッチングを行う際に，トークンの重要度を一律とせず，頻出するトークンの登場順番や、頻出単語については重み付けを小さくし、トークン間の重み付けに偏りを持たせることでスキーママッチングの精度を向上させ，より類似のカラムが多いケースで提案手法による評価を行いたい．

参考文献

[Ling 2001] Ling Yan and Renée J. Miller and Laura M. Haas and Ronald Fagin , Data-Driven Understanding and Refinement of Schema Mappings, SIGMOD '01 Proceedings of the 2001 ACM SIGMOD international conference on Management of data , 2001.

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

[Do 2002] 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.