情報システム大規模スキーマの分析方式

桐村 綾子† 高山 茂伸† 菅野 幹人†

†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

E-mail: †(Kirimura.Ayako@cw, Takayama.Shigenobu@db, Kanno.Mikihito@bc).MitsubishiElectric.co.jp

あらまし 本論文では，大規模システムにおける可視化を用いたスキーマ統合手法における，相関ルールをベースとしたスキーマデータの分析手法の提案と，本機能を実装した結果について報告する．大規模化した企業システムにおけるシステム開発時のデータ整合性確保や，企業合併などにおけるシステム統合の必要性が高まっている．我々は，スキーマ全体構造やスキーマデータの自動分析結果をグラフにより可視化し，データ統合を支援する手法について研究を行っている．冗長なテーブル，類似したテーブルを分析によって示し，さらに複数の分析手法の結果表示を切り替えながら，統合の候補を段階的に絞り込み，統合すべき対象の検討を行うことを目的としている．本論文では，スキーマデータの分析手法のひとつとして，テーブルに属するカラムの型タイプによるテーブルの類似性の自動分析において，相関ルールを応用した分析方式を提案し，実装した結果について述べる．

キーワード メタデータ，大規模スキーマ，システム統合

Analysis method of information system large-scale schema

Ayako KIRIMURA† Shigenobu TAKAYAMA† and Mikihito KANNO†

† Information Technology R&D Center, Mitsubishi Electric Corporation

5-1-1 Ohuna, Kamakura city, Kanagawa, 247-8501 Japan

E-mail: †(Kirimura.Ayako@cw, Takayama.Shigenobu@db, Kanno.Mikihito@bc).MitsubishiElectric.co.jp

Abstract We propose a schema matching algorithm for large scale enterprise information system based on the association rule. And we also present the implementation of this algorithm. Recently the necessity of system-integration in enterprise information system has increased. We research the process for supporting the system integration, and develop the supporting tool that visualizes the structure of the entire schema and some analysis algorithm results by graphs. On this supporting tool, we can try to redesign the data structure through finding the candidates of integration by combining results of analysis algorithms. One algorithm of them is our proposal which apply the association rule to automatically classify the table due to the type of the columns that belongs to the table, and find the candidates of integration in similar tables.

Keyword Metadata, Large scale schema, System integration

1. はじめに

企業における情報システム開発分野において，データの統合に対するニーズが高まっている．データ統合の目的としては，企業内におけるデータウェアハウス構築のためのデータ統合や販売・物流・生産管理システムなど複数システムの統合，企業間においては，企業の吸収・合併による異なる企業情報システム統合に伴うデータ統合や E-business 対応・連結経営対応のための関係会社や取引先とのデータ統合などが挙げられる．一方，企業における情報システムの利用が進み，既存システムへの追加開発を重ねるごとに個別最適・部分最適化されたシステムとなることで全体構成は複雑化しており，システム開発やメンテナンスコストが増大する傾向にある．また，情報漏えい対策や内部統制強化などの観点からも，情報管理の重要性が増している昨今では，各種システムのデータを一元的に管理し，全体最適なシステムを構築する必要性が増している[1][2].

全体最適なシステムを構築する上では全社システムを把握した上でデータを統合し一元的に管理する必要があるが，大規模な企業システムにおいては大規模ゆえにシステム全体像の把握が難しいという問題がある．

我々は，ユーザのニーズをインタラクティブに統合結果にフィードバックする仕組みを供えた統合支援ツールの開発を行ってきた[3]．これは，大規模なスキーマデータに対して，様々な手法を用いて分析・分類した結果をグラフ表示し，ユーザは複数の分析結果を切り替えながら統合の候補を段階的に絞り込み，さらにツール上で統合結果をシミュレーションすることを目的とするものである．本論文では，スキーマデータに

対する分析のうち，テーブルに属するカラムの型タイプによるテーブルの類似性の自動分析において，相関ルールを応用した分析方式を提案し，実装した結果について述べる．

本論文は以下のように構成される．２章では関連研究について言及する．３章では，スキーマデータ分類手法について説明する．４章では分類手法を実装した結果について報告する．5 章はまとめである．

2. 関連研究

データを一元管理する手法としては，スキーマ統合の各種アルゴリズムが提案されており，サーベイ論文にまとめられている[4] [5]．文献[4]では，統合手法の分類として，スキーマレベルの情報のみを用いるものとインスタンスデータ（データの内容）を用いるものに大別し，さらにスキーマレベルの情報を用いた分類は，スキーマの個々の要素（属性）に基づくものと，個々の要素を組み合わせたスキーマの構造に基づくものに分類している．文献[5]では，ルールベースのソリューションと学習ベースのソリューションに大別している．さらにそれらを組み合わせた手法にも言及している[6][7]．文献[6]では全てを自動的に統合するのではなく，統合の過程でユーザフィードバックが可能な手法を提言している．文献[7]ではシステム製品という観点で，様々なアルゴリズムを組み合わせてスキーマを統合する手法について述べている．文献[6]，[7]の手法は，基本的には２つのスキーマの類似性を様々なアルゴリズムで計算して Similarity matrix で表示するものであり，数学的に類似度を表示することが可能である．ただし，大規模な企業情報システムであれば数万項目にもおよぶスキーマの要素の全ての組合せに対して，特に全体を把握していない場合には，統合対象としてどの要素に絞れば良いかがわからず，ユーザがフィードバックをかけながら統合していくことは容易ではない．

3. スキーマデータ分類手法

3.1. 型タイプによるスキーマ分析

我々の開発する統合支援ツールでは，スキーマデータを，スキーマ全体のデータベース構造情報やテーブル参照関係，メタデータ解析アルゴリズム分析結果などの異なる視点から視覚的に提示することで，統合結果にユーザのニーズをインタラクティブに反映する．これまで，メタデータ解析機能として，テーブル名称類似度による分類機能を提供していた[3]．複数のアルゴリズムによるメタデータ解析結果を組み合わせることによってスキーマの類似性を表示するため，カラムの型タイプによる分類機能を新たに提供する．これは，テーブルに属するカラムの型の種別に対し，その組み合わせによってテーブルデータをグループ化し，グラフ表示することでテーブル同士の類似性を表示するというものであり，スキーマレベルの情報を用いた分類のひとつである．

テーブルをカラムの型タイプに基づいて分類する目的は，全社データベースの統合において冗長なテーブル，類似したテーブルを発見することである．全社データベースは一般には様々な部署が管理しており，テーブル生成の全社標準ルールがないことを想定している．どのような型タイプの組をグルーピングの候補とするかが課題となる．

3.2. 相関ルールによる型タイプ組み合わせ抽出

共起関係の頻出パターンを抽出する方法として相関ルール抽出[8]がある．相関ルール抽出はデータベースに蓄積された大量のトランザクションデータから，同時性や関係性が強い事象の組み合わせを導くといった用途に利用されている．テーブルをトランザクションと捉え，テーブルに属するカラムの型を事象と捉えることで，テーブルに属するカラムの型の頻出組み合わせを抽出する．

型タイプ組み合わせ抽出では，全てのテーブルをサーチして，テーブルを構成するカラムタイプを調べ，どのタイプのカラムがテーブル中に存在するかを数え上げ，支持度および信頼度によって多くのテーブルが属しているカラムの組を抽出する．支持度および信頼度は以下のように決定できる．

- 支持度によるカラムの抽出
  支持度とは，全テーブルの中であるカラムが使用されている割合を示す．最小支持度を定義し，支持度が最小支持度以上のカラムを抽出する．
- 信頼度によるカラムの組合せの抽出
  上記で抽出されたカラムに対し，組合せの信頼度を計算する．カラム B のカラム A に対する信頼度とは，カラム A が含まれるテーブルの中で，カラム B が含まれる割合である．
  最小信頼度を定義し，信頼度が最小信頼度以上のカラムの組合せを抽出する．

本手法では，出現頻度が多いものが自動的に抽出される．しかしながら，企業データベースの統合においては，重要な列と重要でない列が存在する．重要な列とは，テーブルのキーを構成するような列である．重要でない列としては，例えば備考欄のような列が考えられる．上記の方法では，例えば多くのテーブルに備考欄（CHAR 型 256 バイト）が存在すると，備考欄を持つテーブル同士が統合の候補として抽出される．しかし，これでは効率的に冗長テーブル，類似テーブルの発見はできない．また，グラフによる可視化を前提とすると，相関ルールによる分類を行う場合，テーブル，カラムの数が多い場合はカラムの組み合わせが爆発的に増大するため，ユーザにとって意味のある組み合わ

せのみを抽出・表示する必要がある.

3.3. カラムの順序を考慮した分類

上記の課題を解決するため，相関ルールによる分類を行う前にカラムの出現順によるグループ化によりテーブルを絞り込む 2 段階の分類を実施する.

通常，キーとなるカラムは先頭から定義されていることが多いため，先頭カラムのデータ型が共通であるテーブル同士の類似性は高いと考えられる．カラムの現れる順序に基づき先頭からN番目までのカラムが共通なテーブルをグループ化し，テーブルを絞り込んだ上で相関ルールを適用する．ここで，N を大きくしすぎると抽出グループ数が少なくなりすぎ，N を大きくしすぎるとパターンの絞込みができないので，N をいくつにするかが実装上の課題となる.

3.4. 型タイプのカラム数による分類

通常の相関ルールでは，データ型の組み合わせのみが考慮されるため，ひとつのテーブルに同じ型のカラムが複数含まれている場合にもそれが分類に影響することはない．例えば，Varchar(10)型のカラムがひとつしかないテーブル A と，10 個あるテーブル B，同じく 10 個あるテーブル C とでは，組み合わせによる分類では全て同じ分類となるが，一致度としては B と C は A に比べて類似度が高いといえる．このように，組み合わせだけではテーブル同士の類似度が充分には考慮できないという課題があった.

我々は，相関ルールによって抽出される頻出組み合わせに対し，数の概念を加える．型タイプ別の組み合わせと，同じ型タイプのカラムの数により組み合わせを作成し，テーブルを分類する．ただし，組み合わせのみでは分類グループ数が爆発的に増えるため，不要な組み合わせを削除して類似度が高い分類グループを残す必要がある．不要な組み合わせとは，以下のようなものである.

- 属するテーブルがひとつしかない組み合わせ
- 複数のテーブルがまったく同じ組み合わせで属している組み合わせ
  - 要素数の少ない組み合わせは要素数の多い組み合わせの部分集合であるため
- 組み合わせ要素数が少ない組み合わせ
  - 一致度の判定基準となる閾値が必要であるが，カラム数はテーブルによって大きく異なり，カラム数の少ないテーブルの属する組み合わせは相対的に一致度が低いと判定されるため，実装上の課題となる.

4. 評価

本分析方式を評価するために，プロトタイプシステムを実装した.

4.1. 手順

分析の手順は以下の通り.

1. 全カラムの支持度を計算し，支持度が最小支持度以上のカラムを抽出する.
2. テーブルに対し，先頭のカラムタイプが同一のものをグループ化する．グループ化の結果テーブル数があらかじめ設定した閾値以上の場合，次に出現するカラムについて同様にグループ化する．その際抽出されたカラムは支持度が最小支持度以上のカラムの組み合わせとする.
3. カラム出現順によってグループ化されたテーブルに対し，相関ルールによりカラム型タイプの組み合わせを作成する．ここでも，支持度によって抽出されたカラムを組み合わせ，信頼度によって組合せをさらに抽出する.
4. カラム型タイプの組み合わせに属する全てのテーブルから，それぞれのカラム型タイプのカラム数を抽出し，最大カラム数のバリエーションで組み合わせを作成する.
5. テーブルがひとつしかない組み合わせを削除する.
6. 複数のテーブルがまったく同じ組み合わせで属している場合，要素数が多い方を残して削除する.一致度の判定基準（ここでは例として 3 要素を閾値としている）より要素のカラム数の和が多い組み合わせのみ残して残りを削除する.

4.2. 分析結果

図 1 のように,分析の結果テーブルに含まれるカラムの数が分類結果のグループノードに表示される．図 1 上部では varchar[80]，varchar[10]，varchar[30]，varchar[1]の組み合わせによるグループが①～③がある．それぞれは型のタイプは同じであるが，数が異なるため別のグループとして作成されている．通常の相関ルールによる分類では，この 3 グループはひとつのグループとしてしか判別できなかった．しかし，カラム数を分類に加えることにより，③のグループは最も共通するカラムが多いことになり，一致度が高いことが分かるようになった.

図 2 は，グループに含まれるテーブルを展開したものである．②に属する 2 テーブルと③に属する 2 テーブルが①に含まれており，この 3 グループのみを見ると，①が②，③を含むような包含関係になっていることが分かる．しかし，これらのテーブルはこの 3 グループ以外のグループにも属しているため，包含関係になっているからとはいえ，一概に階層化することはできない.

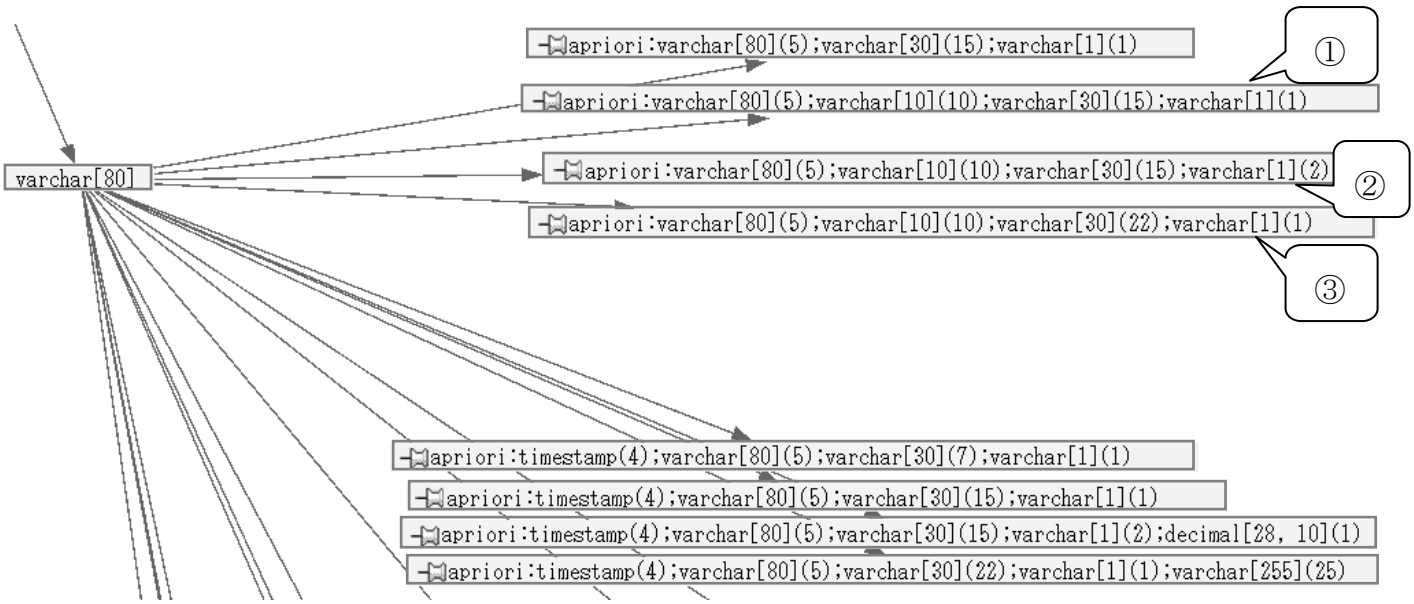

図 1　型分類結果(1)

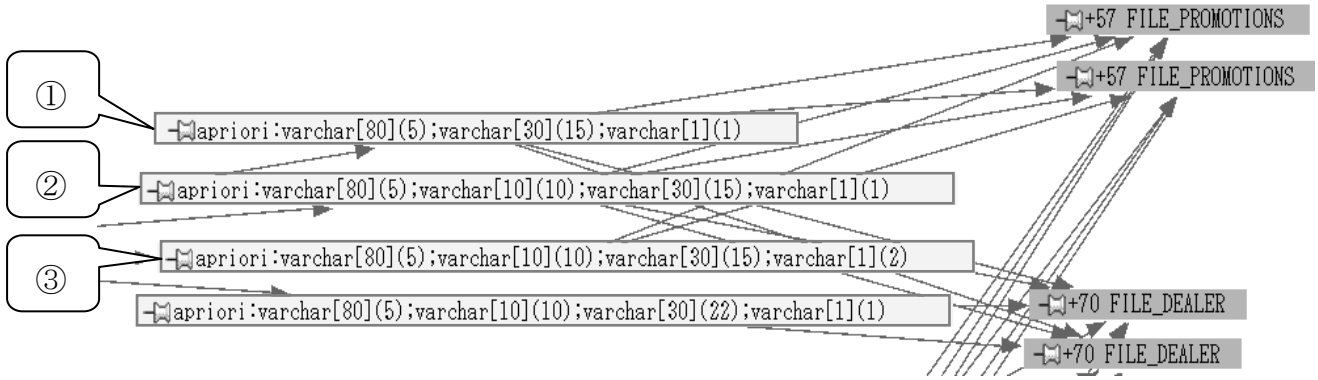

図 2　型分類結果(2) グループの展開

5. まとめと今後の課題

本論文では，テーブルに属するカラムの型タイプによるテーブルの類似性の自動分析において，相関ルールを応用した分析方式を提案し，実装した結果について述べた．

今後の課題として以下を挙げる．

- 分類アルゴリズム改良
  カラム数組み合わせ作成の高速化．
  カラム数組み合わせ作成時の固定的な閾値による切捨て方法の改善．
- 包含関係の可視化アルゴリズム改良
  図 2 のような包含関係を判別しやすくするようなレイアウトの開発．
- 類似度の表示
  分類結果の可視化の際に類似度によって表示方法を変える(強調色など)ことにより，ユーザの気付きを促す．

参　考　文　献

[1] 経産省：情報技術と経営戦略会議報告書,(2003)

[2] 経産省：新経済成長戦略骨子,(2006)

[3] 桐村,高山,菅野：情報システム大規模スキーマの分析・統合方式の実装, DEWS2008

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

[5] Doan, A. and Halevy,A.Y.: Semantic Integration Research in the Database Community: A Brief Survey, AI Magazine 26(1) pp.83-94(2005)

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

[7] Bernstein,P., Melnik,S., Petropoulos,M. and Quix,C.:Industrial-strength schema matching , SIGMOD Record ,Vol.33, No.4(2004)

[8] R. Agrawal, T. Imielinski, A. Swami "Mining Associations between Sets of Items in Massive Databases", Proc. of the ACM-SIGMOD Int'l Conference on Management of Data, (1993)