D-029

相関値差分比較方式によるマッピングモデル Split 判定

Schema Matching Method for String Split Using Correlation Value

細田 聖人† 楓 仁志† 高山 茂伸† 菅野 幹人†
Kiyoto Hosoda Satoshi Kaede Shigenobu Takayama Mikihito Kanno

1. はじめに

近年, ホストダウンサイジングによるコスト削減や, 新たな業務への対応のために, システム再構築が行われている. システム再構築を実施する際には, 連携元データベースと連携先データベースにおけるカラム名やデータの配置といった設計情報の差異を解消し, 異なるデータベースのテーブル間にてカラム対応関係を取ることで, データ連携を実現する必要がある. これをデータ統合と呼ぶ. このとき, システム間の類似したテーブルやカラムの対応関係を判別する技術は, スキーママッチング技術と呼ばれる.

スキーママッチング技術の基本的な手法としては, スキーマ情報(カラム名称, 型など)や, インスタンス情報(単語や値の出現パターンなど)を利用した分析方法がある[1]. 上記手法は, 基本的に 1 対 1 のカラム対応関係を判別するものであるが, さらに応用的な手法として, 複数カラム組間の対応関係を判別するものが挙げられる. ここで, 複数カラム組間の対応関係とは, あるカラムの組と別のカラムの組の対応関係である. 特に連携元カラム数が 1 で連携先カラム数が複数の場合, 本対応関係は, マッピングモデルSplitと呼ばれる[2]. Splitの具体例としては, 図 1に示すように, 連携元で氏名として 1 つのカラムで扱っていたものを, 連携先で姓と名に分割し 2 カラムで扱うといった例などが挙げられる.

本書では上記のマッピングモデル Split を対象とし, 判定方式を検討した. なお本書では, データ統合の連携元におけるカラム内容(データ)を特定の位置で分割した後, 連携先の複数カラムに配置される場合を対象とする.

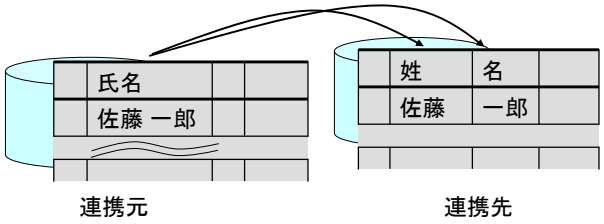

図 1　マッピングモデル Split の例

2. 従来手法と課題

本章では, マッピングモデル Split に関して, 従来手法による実現の方法と, 課題を示す. 従来手法としては, 人手により作成したサンプルデータを利用する手法, インスタンス分析を用いた自動化手法を挙げる.

2.1 サンプルデータ利用による Split 判定

マッピングモデルSplitの判定方式を実装した方式として, QuickMigプロジェクトなどによって研究されている方法が挙げられる[2]. これは, データ移行担当者が, 後段の分析に利用するサンプルデータを連携元に作成した上で, インスタンス分析・スキーマ分析により, Split判定している. 上記の方法は, 大規模データ統合においては, サンプルデータ作成時などの開発コストが増大するため, Split判定自動化の需要がある, 次節では, 自動化に着目した方式を説明する.

† 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

2.2 インスタンス分析を用いた Split 判定

本節では, インスタンス分析を用いた Split 判定の方法と, 自動化に係る課題を説明する.

2.2.1 インスタンス分析を用いた Split 判定の方法

Split 判定の自動化に着目した手法として, 頻度に基づくインスタンス分析手法を利用したものが挙げられる. 連携元インスタンスをある指定部分にて分割し, その前後部分を仮想的に 2 つのカラムとみなした上で, それぞれのカラムに対し, 登場するインスタンスの回数を集計し, 頻度の降順にソートする. また, 連携先の各カラムに対しても同様に, 出現するインスタンスを頻度の降順にソートする. 続いてソート済みインスタンスに対し, 連携元の上位 N 個と連携先の上位 N 個の中で一致数を算出することで, カラム単位にて連携元と連携先の一致関係を判定するものである.

2.2.2 自動化の課題

上記インスタンス分析を用いた Split 判定手法に関する課題点を述べる. Split 判定は, 連携元カラムにおけるデータを特定部分によって分割した前後 2 つの部分それぞれに対し, 連携先での対応関係を把握する. このとき, 連携先にて, 複数のカラムに同一インスタンスが重複して出現する場合, 対象カラムを特定できず, Split 判定が不可能となる.

自動 Split 判定達成のための技術的課題を整理する. 連携元の判定対象カラムを分割した 2 つのカラム間の関係を考慮し, 連携先に存在する複数の類似候補カラムから, 対応関係を判定する必要がある.

3. 提案手法

提案手法では, 上記課題の解決方法として, 以下(a)(b)(c)３つの手法を順次実施することで, Split 判定を自動化する.

(a)単独カラムインスタンス分析：カラム毎のデータ頻度に基づいて, 対応関係を比較し, 候補選別を行う. (a)は従来手法「インスタンス分析を用いた Split 判定」と同一である.

(b)複数カラム相関分析：連携元・連携先それぞれのテーブル内において, カラム間の相関値を計算し, 連携元と連携先を比較し結果を絞り込む.

(c)スキーマ情報利用分析：上記(a), (b)の情報により対象を絞り込んだ後, カラム設計時の情報として, テーブル内のカラムの並び順を利用し, Split 対象カラムを決定する.

以下では, 本 Split 判定方式の中核である, (b)複数カラム相関分析に関して, 説明する.

3.1 複数カラム相関分析

複数カラム相関分析は，相関値差分比較方式を用いることでカラムの組を連携元と連携先にて比較し，類似組を判定する．複数カラム相関分析の概要を図 2に示す．

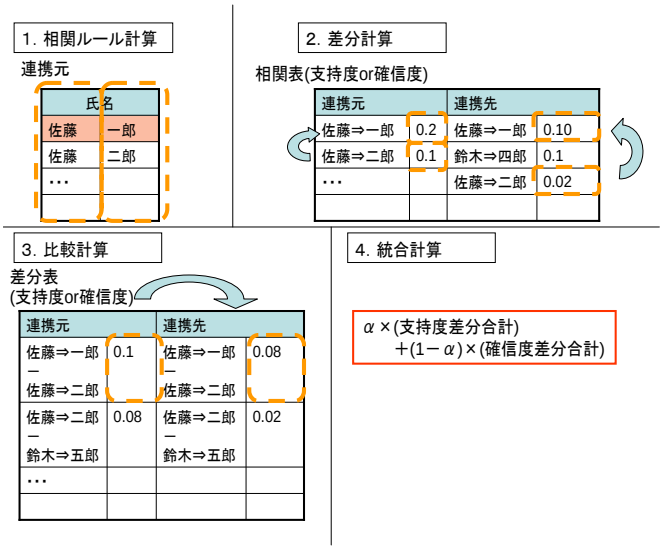

図 2　複数カラム相関分析

まず，1. 相関ルール計算では，連携元インスタンスを分割した前後の部分を，仮想的に 2 つのカラムとみなした上で，2 つのカラムにて相関表を作成する．ここで相関表とは，相関ルール(支持度・確信度)を算出し，降順にソートした表である．相関ルールとは，ある対象Ａと対象Ｂ(図では A が「佐藤」，B が「一郎」)の間の相関関係を，次の２つの指標にて示す．

確信度：Ａ選択者がＢを選ぶ確率

支持度：ＡとＢが同時に出現する割合

続いて，2. 差分計算では，この表に対して上位 K 位までの集合から ${}_K C_2$ の組合せを取り，上位の支持度(確信度)から下位の支持度(確信度)を減算し，組合せと差分値を示した差分表を作成する．図では，連携元の「佐藤⇒一郎」の値 0.2 から「佐藤⇒二郎」の 0.1 が減算される．なお，差分表は，支持度・確信度の両方が作成される．

続いて，連携先の相関表に対し同様の操作を実施するが，この際，連携元の ${}_K C_2$ 個の組合せに出現したものに対し計算実施する．ただし，連携先に登場しない相関ルールに対しては，支持度・確信度ともに 0 を割り当てる．図の例では，「佐藤⇒一郎」の値 0.10 から「佐藤⇒二郎」の 0.02 が減算されるが，連携元にて「鈴木⇒四郎」の組が存在しないため，連携先での計算には利用されない．

次に，3. 比較計算の手順を説明する．まず，作成した差分表に対し，(連携先の値)－(連携元の値)を実施する．図では，連携先の値 0.08 から 0.1 を減算している．同様の計算を差分表の全体に対して実施し，結果の総和を取った後，絶対値計算を実施する.最後に，組合せの一致数にて商計算を実施する．

上記内容は以下の式で示される．a_i は連携元相関表の i 番目の数値(例：佐藤＋一郎の支持度を降順に並べた際の i 番目の数値)である．b_i は，a_i に紐付けられたインスタンス組と b_i に紐付けられたインスタンス組(例：佐藤＋一郎の組)が同じである相関表の値である．ただし，連携先に a_i が存在しない場合は a_i=0 とし計算する．

$$\frac{1}{{}_k C_2}\left|\sum_{i=1}^{k-1}\sum_{j=1}^{k-i}\left\{\left(b_i - b_{i+j}\right) - \left(a_i - a_{i+j}\right)\right\}\right|$$

上記計算は支持度・確信度の双方に対して実施される．最後に，4. 統合計算による統合結果は以下のとおりとなる．

統合結果　＝　α×(支持度利用の結果)

＋　(1－α)×(確信度利用の結果)

4. 机上評価

マッピングモデル Split 判定を実現する提案方式に関して，実際の業務データから作成したデータセットを利用して机上評価した結果を示す．本業務データは，資産管理データであり，管理者と使用者の項目に格納される値に重複があるものを利用している．提案方式は頻度(ここでは支持度，確信度)の差分値を比較に利用しているため，Split 判定を実施する対象としては，各項目間の頻度に差分がないケースから，差分が最も大きいケースまでが想定できる．ここで，差分が最も大きいケースは初期状態のデータである．これをデータセット 1 とする．頻度差がないケースは，レコード等を削除し，比較に利用する項目間にてレコード間の頻度差分をゼロにしたものである．これをデータセット 2 とする．上記 2 つのセットを利用し評価実施することで，本提案方式による Split 判定の効果を示した．

提案アルゴリズム(a)(b)(c)を用いた判定結果を表 1に示す．表における数字は，方式がSplit対象であると判定したカラム組の数である．

表 1　判定結果

評価対象	従来手法	提案手法
データセット 1	4	1
データセット 2	4	1

実験の結果，データセット 1，データセット 2 ともに，従来手法にて，4 つの Split 対象候補を選出し，判定不可であったのに対し，提案手法では候補を 1 つに絞り込み，対象を決定することができた．

5. おわりに

本書では，マッピングモデル Split の判定方式を検討した．評価の結果，類似データカラムを複数含む業務データにて，従来法にて誤判定となる部分を改善し，自動 Split 判定方式の有効性を確認した．

今後の課題としては，今回は利用者が指定していた区切り部分に関して，自動特定・分割の方法を検討する必要がある．また別の課題として，大規模データに関しては，インスタンス分析手法や，性能向上について検討する必要がある．

参考文献

[1]E. Rahm, P. A. Bernstein, “A Survey of Approaches to Automatic Schema Matching”, the VLDB journal, Vol.10, No.4 (2001).

[2]C. Drumm, M. Schmitt, H.-H. Do, E. Rahm, “QuickMig : Automatic Schema Matching for Data Migration Projects”, In Proc. ACM CIKM (2007)