データウェアハウス構築におけるデータクリーニング処理方式

2W－1

林　剛　岸本　義一　長谷川　靖
ＮＴＴ情報通信研究所

1．はじめに

企業内には，業務分野ごとに構築された複数の基幹系データベース（ＤＢ）システムが存在している．これらの基幹系ＤＢの日々刻々と変化する時系列データを蓄積し，販売分析などの意思決定に活用するデータウェアハウス（ＤＷＨ）の構築が盛んになってきている．

複数の基幹系ＤＢから抽出されたデータをＤＷＨに格納する場合，次のような問題が存在する[1][2]．

①同一データ項目名で，データの意味の相違
②異なるデータ項目名で，データの意味が同じ
③データ属性（型）の相違
④データフォーマットや文字コードの相違
⑤データの重複や不整合の発生

①～④は，複数の基幹系ＤＢのスキーマ構造の相違に依存する問題であり，ＤＷＨとして規定した統一スキーマへのデータ変換処理により解決できる．我々は，このデータ変換処理を簡易に実現するDB-STREAM[3]を開発している．しかし，⑤は，基幹系ＤＢ間のデータ値の相違に依存する問題であり，これまではデータ不整合を検出・除去するＡＰを個別に作成する必要があった．

本稿では，⑤の問題を解決するデータクリーニング処理を簡易に実現する方式について述べる．

2．データクリーニング処理方式

2．1　データクリーニング処理モデル

複数の基幹系ＤＢから抽出されたデータ（レコード）の中から，不整合レコードを検出・除去する，データクリーニング処理として，次の三つの処理からなるモデルを規定する（図１）．

①レコード統一化：基幹系ＤＢから抽出されたレコードのデータ形式の統一化処理
②レコード突合：複数の基幹系ＤＢ間の複数レコードを突合し，統合レコードを作成する処理
③データ項目間制約条件チェック：統合レコード内のデータ項目間の整合性チェック処理

このように，データクリーニング処理を３処理に分割するメリットは，以下の通りである．

・データクリーニング処理内容に変更が発生した場合，処理内容の修正箇所が限定され，修正が容易である．
・各処理の再利用が容易である．
・不整合レコードの除去が段階的に可能であり，かつ，不整合原因を早期に特定できる．

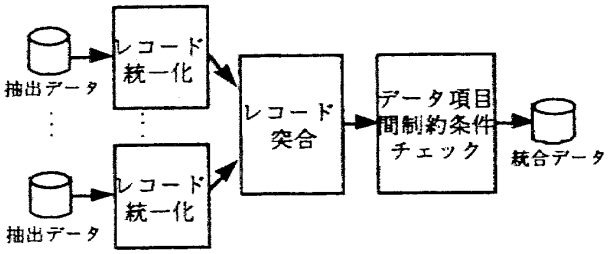

図１　データクリーニング処理モデル

2．2　所要機能

データクリーニング処理モデルの各処理として，具備すべき機能を以下に示す．

(1) レコード統一化
・データ項目の構造（属性・精度）の統一
・データ値のコード体系・表現形式の統一
・実体の表現方法が基幹系ＤＢとＤＷＨで異なる場合（実体とテーブルの関係が１：多，多：１の場合），ＤＷＨの構造に合わせたテーブル（レコード）の分解・結合
・ＤＷＨ構築に不要なレコード及びデータ項目の削除
・レコード統一化が失敗したレコードの分別出力
・レコード統一化処理結果に関する統計情報出力

(2) レコード突合
・レコードの主キー及び第２候補キーによるレコード突合と統合レコードの作成
・突合が失敗したレコードの分別出力
・突合処理結果に関する統計情報出力

(3) データ項目間制約条件チェック
・統合レコードのデータ項目間に矛盾がないことのチェック処理
・矛盾が検出されたレコードの分別出力
・データ項目間制約条件チェック結果に関する統計情報出力

2．3　実装方式

DB-STREAM は，データの変換・統合・分配に必要な機能を共通部品として提供し，ユーザが部品の組合せ方をシナリオで記述するだけで，データの変

A Study of Data Cleaning Method for Data Warehousing
Tsuyoshi Hayashi, Giichi Kishimoto, Yasushi Hasegawa
NTT Information and Communication Systems Laboratories
1-1 Hikari-no-oka, Yokosuka, Kanagawa, 239-0847, Japan

換などを実行するデータ流通プラットフォームである．

２．２に示した所要機能の多く（レコード統一化のデータ項目構造の統一，コード体系・表現形式の統一など）は，DB-STREAM の共通部品が活用できる．このことから，DB-STREAM の共通部品にデータクリーニング処理に具備すべき機能（レコード統一化が失敗したレコードの分別出力，各種の統計情報出力など）を追加し，ユーザがシナリオを記述するだけでデータクリーニング処理を実現できるようにした．

３．データクリーニング処理方式の適用評価

二つの基幹系ＤＢから顧客情報に関するＤＷＨを構築する際に，今回開発したデータクリーニング処理方式を適用評価した．

３．１　ユーザ要件

本適用事例は，ＤＷＨ格納用として，二つの基幹系ＤＢシステムから抽出されたレコードを顧客コード（電話番号）で統合し，顧客名が二つのＤＢで同じレコードと異なるレコードを分別出力する処理である．この処理に関するユーザ要件を以下に示す．

要件１：抽出データごとに異なる電話番号の表現形式を統一する．

（例）“0123-45-6789”，“01234-5-6789”などのハイフンを除去し，“0123456789”に統一する．

要件２：抽出データごとに表現形式が異なる顧客名を統一する．

（例）“（株）…”，“株）…”などの略称を，“株式会社…”に統一する．

要件３：電話番号をキーとして，二つのＤＢから抽出したレコードを統合する．

要件４：二つのＤＢの顧客名が一致するかチェックし，チェック結果を示すデータ項目を追加し，顧客名が一致する場合と異なる場合で別ファイルに出力する．

３．２　適用評価結果

(1) 実用性

３．１の要件１及び要件２はレコード統一化，要件３はレコード突合，要件４はデータ項目間制約条件チェックに関するシナリオを記述するだけで，所要のデータクリーニング処理が実現できた（図２）．

約８千件の実データについて，本処理を適用した結果，適用前の顧客名一致率が１２％であったのに対し，適用後は９７％まで向上し，本処理方式の有効性が確認できた．

(2) 残された課題

要件２（表現形式の統一）は，あらかじめ想定される表現形式を抽出し，変換シナリオとして記述する必要がある．今回の適用例では，表現形式のバターンが多く，すべての表現形式を事前に人手により抽出することは困難であった．表現形式の統一には，①文字列検索を行い，全表現形式パターンを発見する機能，及び，②発見した表現形式のパターンから変換シナリオを自動生成する機能が必要と考える．

４．おわりに

複数ＤＢからＤＷＨを構築する際に必要となるデータクリーニング処理に関する，処理モデル，所要機能，実装方式について示した．また，適用評価により実装方式の有効性を示した．今後は，表現形式のパターン発見の自動化及び変換シナリオの自動生成について検討を進める．

参考文献

[1] Surajit Chaudhur 他：An Overview of Data Warehousing and OLAP Technology, ACM SIGMOD Vol.26, No.1, 1997

[2] Alan Joch：データ・ウェアハウスを上手に構築する方法,日経バイト,April,1997

[3] 池田他：データ流通プラットフォームシステム：DB-STREAM,情処論文誌,Vol.38,No.12,1997

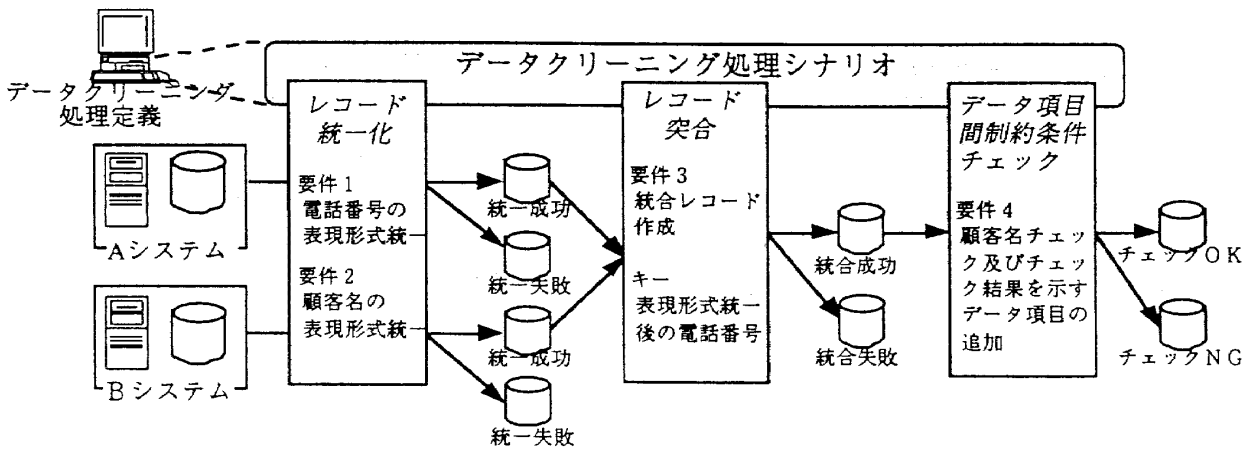

図２　データクリーニング処理の適用事例