6B-06

情報システム開発におけるRDB特異データ抽出方式を活用した現行データ調査

木村　誠†　　橋本　康範‡　　山口　潔†

株式会社 日立製作所 情報・通信システム社 生産技術本部†

株式会社 日立製作所 研究・開発グループ 横浜研究所 エンタープライズシステム研究部‡

1. はじめに

近年の情報システム開発は，殆どが何らかの現行システムと関連しており，全くの独立した新規システムを開発することは稀である．現行システムは多くの場合，更改・統廃合によりデータ仕様が複雑化したり，手順外運用の影響で異常なデータを含んだりしている．また仕様書の変更漏れがあるケースも多い．このため，正しく設計するには現行データ調査による仕様の明確化や異常の把握が欠かせない．

調査には多大なコストを要する．しかし近年システム開発は短期化しており，上流工程で十分に調査することは難しい．一方，調査不足により設計に抜け漏れ・誤りが生じると，実際に現行データを使用してテストする段階まで問題の発覚が遅れ，修正コストが大きくなる[1].

このような背景から日立製作所では，現行データ調査を素早く低コストに行うため，RDB特異データ抽出方式(以下，本方式)の研究を進めてきた[2][3][4]．一方，実案件で本方式による現行データ調査を行った場合に見込める効果は評価できていなかった．本論文では，本方式による現行データ調査の効果について評価する．

2. 本方式の概要

本方式はRDB上の現行データを分析し，仕様調査や異常の発見を助ける特徴情報を分析するものである．分析には以下の2種類がある．

(1)単カラム分析[2]…各カラムのデータパターンを抽出する．

(2)カラム間分析[4]…2つのカラムの値同士の依存関係を分析し，依存関係パターンを抽出する．

どちらもTable 1に示すように，パターンと，パターン別の頻度や例を抽出する．頻度が高いパターンは仕様上のデータ，低いものは異常データである可能性がある．最終的には現行仕様書との照らし合わせや業務有識者の判断でこれらを切り分け，設計への反映や，データクレンジングの実施に繋げる．

Table 1 分析結果のイメージ(単カラム分析)

カラム名	カラム区分	パターン	頻度	例
社員名	姓名	姓＋名	60%	日立太郎
		姓＋名＋敬称	39%	日立花子様
		<<特異値>>	1%	日立次郎_2
社員番号	コード	英1数3-英1数4	98%	A154-Y2349 A248-Z0192
		英1数3_数1	2%	N113_6
社員区分	少データ種	1	68%	1
		2	31%	2
		<<空文字>>	1%	<<空文字>>

3. 現行データ調査の効果評価

<3-1>評価目的

実案件で本方式による現行データ調査を実施するにあたり，異常データの存在を把握してから計画や設計を行うことで未然に防止できる可能性のある問題の割合を過去事例から推定する．

<3-2>評価方法

過去の案件で作成された問題管理票の内容を人手により精査し，仮に本方式による現行データ調査の結果を設計前に確認していれば，原因となったデータを未然に発見できていた可能性があるかどうかを判断した．

精査対象は，2011年～2014年に開発を行った2つの案件で発生した問題管理票のうち，データ不良及び設計不良が原因で発生した計456件である．各問題管理票には，発生事象，原因及び対策が記載されている．各問題管理票を，内容に応じて以下の3グループに分類した．

[A.単カラム分析] 及び[B.カラム間分析]…2章(1)及び(2)で述べた各分析結果から、原因データを未然に発見できた可能性があると判断した問題．

[C.スコープ外]…本方式による分析結果からは予見できないと考えられる問題．

Data condition survey utilizing the RDB-specific abnormal data extraction method.

†Production Engineering Department, Information & Telecommunication Systems Company, Hitachi, Ltd.

‡Enterprise System Research Department, Research & Development Group Yokohama Research Laboratory, Hitachi, Ltd.

分類は，人手による内容の目視と判断で行った．例えば「マスタテーブルの社員区分カラム("1"または"2"を取る)に，稀に不正値(空文字)が格納されている」という原因で発生した問題は，単カラム分析結果(Table 1)から頻度の低いパターンとして抽出可能であるので，[A.単カラム分析]に分類する．

一方例えば，「マスタテーブルの一部データが作業の伝達漏れのため格納されていなかった」という原因により発生した問題は，現行データ調査により防止できないため，[C.スコープ外]に分類する．

<3-3>評価結果

評価結果を Fig. 1 に示す．A.単カラム分析とB.カラム間分析の比較では，前者の効果が大きい．全体に占める割合は両者を合わせて 121 件(27%)となった．1 章で述べたとおり，これらの問題は現行データを使用してテストする段階まで発覚せず，修正コストが大きくなりがちであるため，効果は大きいと考える．

今後，より効果を得るには C.スコープ外の問題(335 件(73%))にも取り組む必要がある．C.に分類した問題は，次の 2 種類に大別できた．

①リレーショナルモデル上の制約に違反したために発生した問題：

例えば「受注テーブルの商品コード列の値は商品マスタテーブルにも存在しなければならない」という制約である．これについては，本方式のカラム間分析で行われるデータ依存関係のパターン抽出[4]を，テーブルやレコード間の依存関係抽出に応用することで今後対処が可能と考えられる．

②ビジネスルールから生じるデータ制約に違反したために発生した問題：

例えば「ある同一顧客の契約(レコード)について，契約期間に重なりがあってはならない」という制約である．本方式はナレッジや経験則を活用したものであり，このような個別性の強い制約を機械的に抽出することは困難である．しかし形式手法[5]等を活用して，制約を機械処理可能な形で記述し，現行データに対しての当てはまりを確認しながら制約の修正を繰り返すという手順を経れば，これらの問題も防止できると考える．但し形式手法を利用するには形式言語の専門知識が必要となり，人材確保が難しい問題がある．この問題については，制約ルールを形式言語で直接記述するのではなく，データ制約の記述に特化した簡易的な表現で記述し，それを形式言語に置き換える等の対策が考えられる．類似の例として[6]が挙げられる．

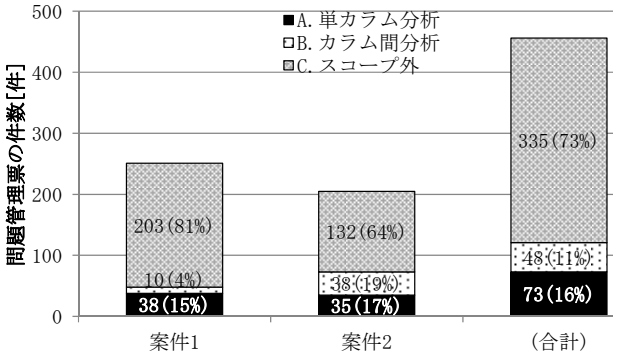

Fig. 1 問題防止効果の事前評価結果

4. まとめと今後の課題

RDB 特異データ抽出方式による現行データ分析を実際の案件で実施した場合の問題防止効果を，過去事例を用いた机上評価により推定した．過去の 2 案件の問題管理票を分析した結果，27%の問題に対する防止可能性があると判断した．

今後の課題として次の 2 点が挙げられる．

<4-1>実績による効果の評価

本方式による現行データ調査は既に実際の案件で実施している．今後，各案件での問題の改善実績に基づいて効果を評価する．

<4-2>本方式のスコープ外の問題への対応

3 章で本方式のスコープ外に分類した問題を，本方式の拡張や形式手法等の活用で防止していく．

＜参考文献＞

[1]松村知子, 門田暁人, 森崎修司, 松本健一, "マルチベンダ情報システム開発における障害修正工数の要因分析", 情報処理学会論文誌, Vol.48, No.5, pp.1926-1935, May 2007.

[2] 橋本 康範, 大島 敬志, ほか, "RDB の仕様詳細理解に向けたカラム単位特徴分析技術の提案", 平成 25 年 電気学会 電子・情報・システム部門大会, 2013.

[3] 橋本 康範, 大島 敬志, ほか, "情報システムのデータ移行効率化に向けた RDB 特異データ抽出方式の提案", 電気学会 第 59 回 情報システム研究会, 2014.

[4]橋本 康範, "データ仕様復元に向けたパターン検出技術の適用検討", 国立情報学研究所トップエスイー2014 年度修了制作, 2014.

[5]Dependable Software Forum, "形式手法活用ガイド【改訂版】", 独立行政法人情報処理推進機構, 2012, (URL : http://sec.ipa.go.jp/reports/20120928.html).

[6]伊藤 信治, 佐藤 直人, ほか, "SAT ソルバを活用した決定表作成・検証方式", 電子情報通信学会技術研究報告, 信学技報 113(489), 7-11, 2014.