5B-1

辞書構築技術適用によるＤＢシステム再構築

○鹿島　理華†　　佐藤　彰洋†　　谷垣　宏一†　　山足　光義†

三菱電機株式会社　情報技術総合研究所†

1. はじめに

企業内の情報システムは、業務毎に独立したシステム、サブシステムを段階的に構築してきたため大規模で複雑なシステムとなり、データも各システムに分散し個別に管理されていることも多い。これは、同じ意味を示しているにもかかわらずデータ名称が異なる項目名のばらつき（例:CORPNAME と CORP_NM）やデータの二重持ちなどのデータの品質の低下、メンテナンス負荷増加、分析などのデータの二次利用が困難といった問題につながる。

これに対し一般的にデータ統合を行なうが、共通マスタ DB を構築するようなデータの統合はコストが大きいため、メタデータと呼ばれる“データに関するデータ”だけを統合する方式がある。ここで異なるシステム間で管理されている冗長データの把握が必要であるが、そこにわれわれの持つ辞書自動構築技術[1]とスキーママッチング技術[2]を適用する DB システム再構築の一方式を提案する。

2. 背景と課題

データ品質の課題を解決するために、システムの集合のデータ統合が解決策としてあげられる。

データ統合は２つのフェーズからなる。現状のシステム全体のデータベース構造を分析（As-Is 分析）する第一フェーズと、統合後のデータベースを設計（To-Be 設計）する第二フェーズである。As-Is 分析フェーズでは、異なるシステム間において同一内容を表しているが別々に管理されている冗長データ(例 kokyaku_id と custmer_id)を把握するなどの作業が必要になるが、仕様書がない、有識者がいない、データベース定義書を分析して冗長データを見つけようとしても単純な一致検索ではひっかからないなどの理由により、非常に大きな作業であるという課題があり、データ名称が統一されていないデータ間の関係性の抽出技術が必要である。

また、データ統合をあるべき姿にし、維持する To-Be 設計のフェーズでは、現状分析で抽出したメタデータをどのように活用すべきかが課題になる。例えば、現状分析をかけて抽出したメタデータはあくまでその瞬間の情報で、新規システムや新規データ連携の構築によって常にメタデータは変更が入り、何もしないとどんどんメタデータも陳腐化していき役に立たなくなるという課題がある。

3. スキーママッチング技術と辞書自動構築技術

3.1 技術概要

これらの課題に対し、われわれの保有する辞書自動構築技術とスキーママッチング技術の適用を提案する。これら技術の概要を説明する。

（１） 辞書自動構築技術

辞書自動構築技術は、自動的文書解析や人工知能のアプリケーション支援のために構築された汎用的な概念辞書である WordNet[3][4]と連携してシステムの概念辞書を自動構築する技術である。概念辞書は、単語の意味を示す「概念」を単位とし類義語、省略語を管理する辞書である。WordNet は、一般的な概念を網羅する辞書であるため多分野にわたる。例えば、SHIP は、「船」「役職」「出荷」と様々な意味を持つため、WordNet をそのまま辞書として使用すると曖昧性が増加してしまう。

辞書自動構築技術は、対象とするスキーマ情報における単語の使われ方に着目し、「似た使われ方（出現文脈）をする語群は、似た意味を持つ」ことを利用し推定を反復、曖昧性を解消しそのシステムの概念辞書を自動構築する。

（２） スキーママッチング技術

データベース間の項目対応を、項目名やデータ型、桁数等の定義情報の類似性に基づいて高精度推薦する。加えて定義情報の類似性判定では判断できない日本語ローマ字⇔英語等の照合を、概念辞書の情報を利用し実現する。

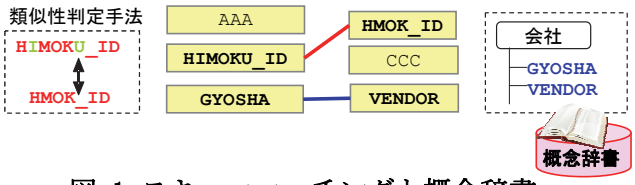

図 1 スキーママッチングと概念辞書

DB system rebuilding by the dictionary construction technology application

†Rika Kashima, Akihiro Sato, Koichi Tanigaki, Mitsuyoshi Yamatari,

Information Technology R&D Center, Mitsubishi Electric

3.2 データ統合層の提案

メタデータを統合し複数のサブシステムからなるシステムのデータ統合を行なうためのデータ統合層を提案する。データ統合層はメタデータと、メタデータを管理し活用するメタデータ管理機能群から成る。図２に構成案を示す。

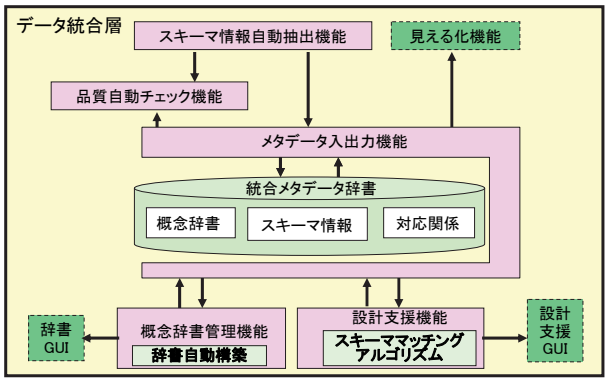

図２ データ統合層構成案

As-Is 分析フェーズと To-Be フェーズに分け技術適用について示す。

3.3 As-Is 分析フェーズへの適用

As-Is 分析フェーズでの現状システム分析での冗長データの把握は、２章で記した理由により大きな作業となる。この課題解決のために、辞書自動構築技術によりまずデータの使われ方に着目した対象システムの概念辞書を作成し、この概念辞書をスキーママッチングで用いることで、データ項目名にばらつきがあっても、データ項目名を「文字列」でなく意味でとらえることができ、項目名にばらつきがあっても効率よく冗長データを把握することができる。

3.4 To-Be フェーズへの適用

To-Be 設計のフェーズでは現状分析で抽出したメタデータをどのように活用すべきかが課題になる。本節では、ユースケースに分け技術適用について述べる。

（１） 初回のメタデータ投入

提案するデータ統合層のメタデータ辞書は図２に示すように①管理対象データベースのスキーマ情報、②概念辞書、③データ項目間の対応関係から成り DB に格納し管理する。概念辞書は[5]にて提案したドメイン辞書を想定している。ドメイン辞書は RDF モデル[6]を採用し RDB 形式で DB 上に保存する。しかし、データ統合層への適用では、概念と表現形式の関係が提供できれば RDF モデルにこだわらなくてもよいと考える。

項目間の対応はスキーママッチング技術により抽出する。スキーママッチングはデータ項目間の対応関係の候補を対応の確度の数値と共に示す。その数値をもとにした対応関係の確定を自動化するか人手でするかは適用する実システム側の要件次第といえる。

（２） 新規データや新規データ連携の追加

設計支援として、概念辞書を使った意味よせや、すでにそのデータが入っていないかの確認にスキーママッチング技術を適用する。

（３） メタデータの品質維持

新しいサブシステムをデータ統合の管理対象に追加する場合や、新規データや新規データ連携がメタデータ管理機能群の提供する設計支援機能を使わずに追加された場合、メタデータの品質が下がり、陳腐化し役に立たなくなる。そこで、管理対象の全データベースの最新のスキーマ情報を取り込み、スキーママッチング技術を適用した品質自動チェック機能で統合メタデータ辞書 DB の情報と整合性がとれているかのチェックを定期的に行う運用を行い、メタデータの品質を維持する。

4. おわりに

複数のサブシステムからなるシステムのデータを、メタデータだけを統合しデータの意味を明確にし、信頼性や再利用性を向上させるデータ統合方式に対する辞書自動構築技術とスキーママッチング技術の適用について報告した。現在この方式の実システムへの適用を検討しており、来年度に実装予定である。今後はこの実装開発を通し、技術適用効果の検証を行っていく予定である。

参考文献

[1] K. Tanigaki, et al. Density maximization in context-sense metric space for all-words WSD. In Proc. of ACL2013, pp. 884–893.

[2] 小出他 「学習データ量によるスキーママッチング精度向上効果評価報告」情報処理学会第74回大会 6B-4

[3] Princeton University, “About WordNet”, Princeton University, http://wordnet.princeton.edu, 2010

[4] 日本語 WordNet, http://nlpwww.nict.go.jp/wn-ja/

[5] 鹿島他 「ドメイン辞書のデータベース化へのRDF モデル適用の提案」, 2011, 第 74 回情報処理学会全国大会論文集,6B-3

[6] W3C, “RDF”, http://www.w3.org/RDF/