異データベース間におけるデータマッピング手法の提案（2）

6R−6 ― データ項目間の類似性に着目したマッピング手法*1 ―

大沼　守一*2　石黒　正典*3　坂田　哲夫*4
NTT情報通信網研究所*5

1　はじめに

近年，1企業内で個別に開発された，複数のデータベース上のデータを統一的に取り扱いたいと言う要求が高まっている．しかし，この要求の実現のためには，複数データベース間において本来同一と見做せるもの（以降「同義」と呼ぶ）の特定が必要となる〔1〕．この同義であるものの特定は，スキーマやドキュメント等の設計情報を人手で分析することにより行われているが，これまで，同義か否かを判定するための尺度はなく，個人のノウハウに依存していた．そこで我々は，同義である可能性を判定するための尺度として「類似性」を提案し，上記分析を支援する手法について検討した．

本稿では，データベースで意味の有る最小単位であるデータ項目（ここでは，ネットワークモデルにおける用語ではなく，データベースにおいて意味を持つ最小単位として使用している）に着目し，データ項目間の類似性を明確にする手法を提案する．

2　データ項目間の類似性の基本要因

異なるデータベース間において同義であると見做せるデータ項目間には，次のような関係が成り立つと考えられる．

（1）　名称が類似している
（2）　値域が類似している

現実世界の事物（オブジェクト）をデータベースでモデル化する際に，設計者の個人差や対象業務の特性により事物の捉え方が変わる．例えば，図1において，オブジェクトのクラスである「ビル」は各々のデータベースA／Bにおいて，データ項目A／Bとして表現され，名称「局舎」／「営業所名」が付与される．また，「ビル」のインスタンスとなる「横須賀支店」は，各々データ項目の値として「横須賀支店ビル」／「横須賀」と表現される．この時，個々のデータベースに着目すると，オブジェクトをモデル化した際に付与された名称がそのデータベース内においてはオブジェクトの特徴を最も端的に表現していると考えられる．このため，名称を手掛かりとして，複数データベース間での類似性を判定することは，同義なオブジェクトの発見の為の有効な手段であると判断できる．

我々は，名称に基づいてデータ項目の類似性を判定することを「名称の類似性判定」と呼ぶ．

また，値に対して上記判定方法を適用し，値域間がどのような関係にあるかを導くことによりデータ項目間の類似性を導出する．我々はこれを「値域の一致性判定」と呼ぶ．

3　名称の類似性判定手法

3-1　名称の正規化

個々のデータベースで付与されている名称間で直接類似性判定を行っても，名称間の異種性〔1〕により，正しい判定ができるとは限らない．そこで，データ項目名称を異種性が生じる以前の，現実世界のオブジェクトを一意に識別できる名称に変換し（正規化），正規化された名称間で類似性判定を行う．正規化は，以降の節で述べる「継承化」「抽出化」という手順で行う．

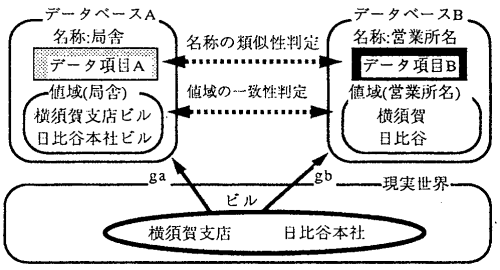


図1　名称の類似性判定と値域の一致性判定の関係

3-1-1　継承化

一般に，データ項目名称は暗示的にテーブル名称を修飾語と仮定して命名されている．その為，1つのデータベース内に同一の名称が複数存在してしまう．そこで，名称に一意性を与えるために，テーブルの名称に使用している用語をデータ項目に引き継がせる手法（継承化）を適用した．

表1の複数テーブル「フレーム」「パッケージ」の「製造会社」という同一名称のデータ項目に対し，この手法を適用すると，「フレームを製造する会社」と「パッケージを製造する会社」となり，相互の概念的な差である「製造する対象」が明確に区別できるようになる．

表1　継承化の例

データ項目名称	テーブル名称	継承化データ項目名称
製造会社	フレーム	フレーム製造会社
製造会社	パッケージ	パッケージ製造会社

3-1-2　抽出化

継承化により，データベース内で名称に一意性を持たせることができる．さらに，名称が示すオブジェクトをより明確化するために，ドキュメントに記述されている「データ項目説明」から抽出した用語を名称に対して追加・置換し，自然言語解析により最適な並べ替えを行う手法（抽出化）を適用した．

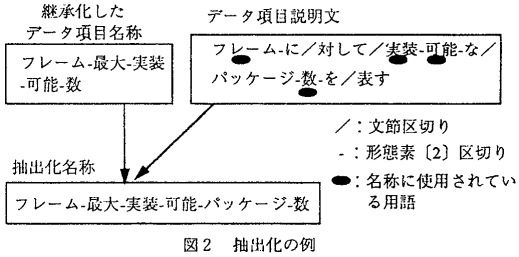


図2　抽出化の例

例えば，図2の継承化したデータ項目名称「フレーム最大実装可能数」に対しこの手法を適用すると，データ項目説明文の形態素の用語に名称内の用語を対応させ，説明文の中にしか出現しない用語｛-に，対して，-な，パッケージ-，-を，表す｝が得られる．この用語の中で，データ項目名称を構成できる用語は｛パッケージ-｝だけであり，この用語の係り先〔2〕である｛-数｝の前に追加することで「フレーム最大実装可能パッケージ数」という名称に変換できる．

---

*1 A Data Mapping Method among Heterogeneous Databases based on Similarities between Data Items .
*2 Syuuichi OHNUMA . *3 Masanori ISHIGURO .
*4 Tetsuo SAKATA .
*5 Network Information Systems Labs., NTT

3-2　類似性判定のための用語分析

一般にデータ項目名称は，複数の用語の組み合わせからなる複合語という形式を取る．複合語間の類似性を判定するには，複合語を構成する用語間で比較対象となるものどうしを決める必要がある．そこで我々は，Durellの命名規則〔3〕と格文法〔4〕を用いて，名称を統一的に解釈するための規則を作成した．

3-2-1　Durellの命名規則

Durellの命名規則では，管理する対象を示す用語（主要語），値の種類を表す用語（区分語）がデータ項目名称内に必須であるとしている．また，この２つでは名称が一意にならないとき意味を区別する用語（修飾語）を複数補って識別性を高める．

データ項目名称＝｛修飾語・・修飾語，主要語，区分語｝

我々は，データ項目名称を構成する各用語に対し品詞と意味カテゴリ〔2〕を解析することで，主要語／区分語／修飾語の区分を明確にした．

3-2-2　データ項目名称の格要素

上記区分により主要語間，区分語間，修飾語間を比較対象とすることができるが，修飾語は主要語に対して多様な役割を持ちうる．修飾語と主要語による構造は，文法的には連体修飾構造であり，格（Case）による分析が可能である．

格とは，文構造において用言と名詞の関係を表すものであり，格を表す構成素を「格要素」という．また，格要素は文中では，「体言句」と「格標識」で構成される．例えば，

”会社がパッケージを製造する”という文では，格要素は

｛会社-が｝　｛パッケージ-を｝であり，｛会社｝　｛パッケージ｝が体言句，｛が｝　｛を｝が格標識であり，用言｛製造する｝に係る．

データ項目名称を命名する際，命名者は現実世界のオブジェクトに関する知識を用いる．この知識は，上記の例文のように｛会社が・パッケージを→製造する｝という文の形式で表現できる．しかし，データ項目名称は複合語で構成されるので，｛会社｝を後置し｛パッケージを→製造する→会社｝とし，さらに格標識｛を，する｝を体言句の概念に隠蔽することにより｛パッケージ→製造→会社｝として，「パッケージ製造会社」というデータ項目名称の連体修飾構造を得る．

そこで我々は，上記の過程の逆方向に分析し，データ項目名称の連体修飾構造の隠蔽された格標識を再現することで，格要素を明確にする．

3-2-3　格要素と関係詞化

3-2-2節の手法では，複数の格要素の候補となりうる．格要素を特定するために以下に示す「関係詞化」〔5〕を適用する．

文中の格要素を用言に後置して，それを用言の係り先体言とする連体修飾構造を作った時，連体用言と係り先体言の間に元と同じ格関係が自然に読み取れるならば，その格要素は関係詞化が可能であるという．この関係詞化により，データ項目名称における連体修飾構造の格要素の候補に優先順位を与えることができる．

3-3　名称間の類似性判定

以上により，データ項目名称間において比較すべき用語どうしの対応が明確になる．類似性判定における評価規準としての用語の重要度は次の順であるとし，その順に従って比較する．修飾語における比較は前述したように，同一の格要素である用語間のみを比較する．

主要語間　＞　区分語間　＞　修飾語間

用語間の類似性は図３のように分類される．

名称全体の類似性は，上記の重要度と用語間の類似性により求められ，重要度の高い用語間の類似性が高いほど名称間の類似性も高いと判断できる．

同一　≧　同義　≧　略語同義　≧　カテゴリ一致

同一　　：用語の意味が一致し，表現も等しい
同義　　：用語の読みは同一だが，平仮名・片仮名等の表現が異なる
略語同義　：一方が他方の省略形として表現されている
カテゴリ一致：意味カテゴリが一致している

図３　用語間の類似性の尺度

4　値域の一致性判定

4-1　値への名称の類似性判定の適用

値間に名称の類似性判定を適用するには，値を構成する用語の意味カテゴリを明確にする必要がある．

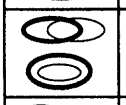
例えば，「局舎」というデータ項目の値には「日比谷本社ビル」や「横須賀支店ビル」が存在する．この値の用語構成には構文的な制約として「＜地名＞＜組織＞ビル」が存在する．＜地名＞＜組織＞は意味カテゴリである．

意味カテゴリの構成を分析し，同一の意味カテゴリの用語間で名称の類似性判定を行う．

4-2　値域の対応関係

値間で名称の類似性判定を行うことにより，値域の一致性を判定する事ができる．この時，類似性判定のソース側とターゲット側の値域間には表２に示すような関係がある．

表２　ターゲット側／ソース側値域間の対応関係

図	関係	説明
	同一	ターゲット側がソース側に完全に一致している
	包含	ターゲット側がソース側に満たされるが，冗長な要素が存在する
	交差	ターゲット側の一部だけがソース側に満たされる
	乖離	ターゲット側を満たすソース側の要素が存在しない

：ターゲット側の値域　：ソース側の値域

値域間の対応関係が「同一」「包含」であるならばそのデータ項目は同義であると見做すことができる．逆に「乖離」の関係ならば，同義と見做すことはできない．「交差」の場合，ソース側と重ならないターゲット側の値域については，その領域にある値の出現頻度や重要度により判断する必要がある．

5　おわりに

異なるデータベース間のデータ項目間の類似性を名称と値域をもとに導出する手法について提案した．今後，本手法を実際に適用し，その有効性を検証する．

参考文献

〔1〕石黒他,"異データベース間におけるデータマッピング手法の提案(1)",情報処理学会第４５回全国大会論文，1992
〔2〕宮崎,"係り受け解析を用いた複合語の自動分割法",情報処理学会論文誌,No 6，vol 25，1984
〔3〕Durell．W．R，Data Adminstration，MaGraw - Hill，1985
〔4〕野村,"自然言語処理の基礎技術"，社団法人電気通信学会，1988
〔5〕山端,"用言格要素の関係詞化の可能性"，自然言語処理，No 40，1991