2R-7

データ標準化を目的とした類似データの分類手法

川下　満　　関根　純　　鈴木　健司
NTT情報通信処理研究所

1．はじめに

企業の持つデータ資源を企業活動に有効活用するには、業務対応に個別に開発されてきたシステム間でのデータの流通を促進し、利用者が要求する情報を迅速に取り出せるようにすることが重要な課題となる。

個々に開発されてきたシステム間では、データの流通を阻害する原因として一般に次の問題を抱えている。

①同じデータ項目を互いに持っていても、異なる名称であるため発見が難しい。

②同じデータ項目であることがわかったとしても、異なる属性、異なるコード体系を持つため変換が必要である。

データ流通を促進するためには、これらに該当するデータ項目を事前に検出し、標準化により名称、属性、コード体系等を統一しておくことが必要である。しかし、調査の対象とすべきデータ項目の数は1つの企業で数千～数万に及ぶため、これを全て人手で行うには莫大な期間と工数が必要である。従って、全てのデータ項目を対象に類似のデータ項目（同一、もしくは、同じコード体系とすべきデータ項目）のグループに分類するツールが望まれる。

類似の名称を検索するツールは既にいくつか存在するが、従来のこの種のツールは指定の単語を含む名称を検索することを目的としており、上記の目的で用いるには効率が悪い。本稿では、データ項目をデータ項目の日本語名（以後、単に名称と記す）をもとに分類する方式を提案し、その効果を示す。

2．分類方式

2．1　考え方

名称はデータ項目の意味を簡潔に表したものであり、名称に同じ単語を含むものを類似のデータ項目とみなすことは妥当と思われる。但し、単純に同じ単語を含む名称を検索したのでは次の様な問題があり、人手の負荷の軽減には不十分である。

①同じ意味で用いられた単語（同義語）であっても別グループに振り分けられるため、同じグループへの統合を人手で行う必要がある。

②それほど類似していない名称も同じグループとして検索されるためそれらを人手で除く必要がある。

（例）「電話番号」と「郵便番号」は同じ「番号」でもコード体系は異なる。

筆者らは、①の問題については単語間の同義語の関係を定義した辞書（用語辞書）を持つことで、②の問題については同じ単語の組合せを持つものを類似名称とすること、及び、名称の類似の判定に寄与しない単語は除くことで解決を図った。

以下に、分類方式の詳細を説明する。

2．2　方式詳細

2．2．1　言葉の説明

（1）用語

名称は、複数の単語（用語）から構成される。

（例）普通預金口座番号＝普通＋預金＋口座＋番号

（2）用語の種類[1]

用語は次の用途に分けられる。

①区分語　データ項目が表す値の範疇、単位。データ項目の目的を示す。

（例）量、日付け、番号、時間、金額

②主要語　区分語を修飾する用語の内で特に重要なもの。データ項目が関与する対象の実体を表すものが候補となる。

（例）預金、支店、回線、局社、メール

③修飾語　区分語、主要語を修飾する。名称の類似の判定への寄与は少い。

（例）定期、最大、前、左、上、以上

（3）同義語

用語はシステム毎の都合で用いられているため、同じデータ項目でもシステムが異なれば異なる用語が用いられるが、用語を統制し標準用語を指定することは可能である。同じ標準用語に置き換えられる用語同志は互いに同義語の関係にある。

（例）顧客、利用者、ユーザ → 顧客

（4）類似名称

同じ主要語と区分語の組合せを持つものを類似名称とする。

（例）

修	主	区	
臨時	＋電話	＋番号	┐ 類似
新	＋電話	＋番号	┘
新	＋郵便	＋番号	― 非類似

2．2．2　検索方法の工夫

（1）構造体の扱い

設計の容易さやドキュメントの見やすさ等の理由により構造体の形で記述される名称が多い。構造体の場

Data classification method for data standardization
Mitsuru KAWASHIMO, Jun SEKINE, Kenji SUZUKI
NTT Communications and Information Processing Laboratories

合、個々の名称に主要語と区分語がそろっていなくとも、上位から下位までの名称を1つの名称ととらえると完全になるものがある。

（例）1　利用者　（主）
　　2　氏名　（区）
　　2　住所　（区）

そこで、この考え方に従って検索するが、構造体の各レベルの名称に主要語と区分語がそろっていると逆に不適当な組合せのグループに分類されることも多くなるので、次の制約を設けることでこれを抑止する。

①名称に主要語と区分語がともに存在するなら、構造体の上位の名称は参照しない。これは、最も密接な関係を持つ主要語と区分語の対で検索するためである。

②名称に主要語と区分語がともに存在しないなら、この名称は上位または下位の名称を修飾しているものとみなし検索対象としない。

（2）禁止語の排除

データの標準化を目的とした分類には、データベースやファイルの構造に関する用語（主要語）は無意味であるので、それらを禁止語として区分語との組合せより除く。

（例）データ、レコード、ブロック、エリア

3　試行例

3．1　試作ツール

試作したツールのデータ構造を図1に示す。

検索は、まず標準用語テーブルより区分語と主要語を順次取り出し、それらの用語を持つ名称を検索リストをもとに検索し、区分語と主要語の組合せの各グループに分類する。

このデータ構造には、次の特徴がある。

①あらかじめ標準名称を作成し、それをもとに検索リストを作成することで、同義語による検索を不要としている。

②名称が不完全で主要語または区分語がない場合、標準名称にこれらの用語を加えることで、名称の本来の意味を表現できる。

3．2　分類結果と考察

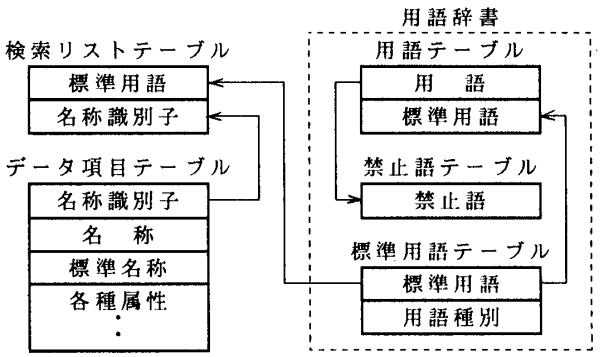
表1の諸元をもつデータベースを対象に、データ項目の分類を行った。結果を表2に示す。用語辞書は分類対象の全名称を単語に分解し、それを基に作成した。

（1）構造体の配慮とヒット率

$$\text{ヒット率} = \frac{\text{グループに分類された名称数（重複なし）}}{\text{グループ毎の名称数の総和}}$$

ヒット率が高いことは1つの名称が複数のグループに分類されることが少ないことを示すので、分類方式の良否を示す1つの目安となる。

提案の方式は構造体を配慮することで分類できない名称を40％から20％に削減したが、ヒット率が10％低下した。これは、主要語と区分語の組合せを複数持つ名称が増えたためであり、ヒット率を高めるには、無意味な組合せによる分類を抑止するため組合せに優先順位を付ける工夫が必要である。



用語　　：名称を構成する単語。
標準名称：名称を構成する各用語を標準用語で置き換えたもの。
検索リスト：標準用語とそれを含む標準名称を持つデータ項目との対応表。
→　　：参照関係を示す。矢印の付いている側は矢印の付いていない側に無い値を登録できない。

図1　試作ツールのデータ構造

（2）禁止語排除の効果

禁止語を除いたことにより、グループ数を4％、分類対象の名称を8％削減した。

（3）分類できない名称

主要語または区分語を持たない不完全な名称が20％あった。これらは標準名称に不足の用語を追加することで名称の意味付けを明確にする必要がある。

表1　諸　元

・ｼｽﾃﾑ数	3	・名称当りの用語数	2.4
・ﾚｺｰﾄﾞﾀｲﾌﾟ数（ﾃｰﾌﾞﾙ数）	216	・用語数	1003
・ﾃﾞｰﾀ項目数	5562	・禁止語数（主要語）	19
・構造体数	469	・標準用語数	814
・非構造体ﾃﾞｰﾀ項目数	752	（区分語	88）
		（主要語	556）
		（修飾語	170）

表2　分類結果

	提案の方式	従来方式（単語による分類）
ｸﾞﾙｰﾌﾟ数	1697	1003（＝用語数）
ヒット率	54 %	42 %

4　おわりに

今後は、ヒット率を更に高める工夫を行うと共に、分類結果をデータ標準化に適用し、問題点を探る。

[謝辞] ﾃﾞｰﾀﾍﾞｰｽの設計情報を提供して下さったNTTｵﾍﾟﾚｰｼｮﾝｼｽﾃﾑ開発ｾﾝﾀ・大沼主幹技師に感謝します。

文献：[1]William R.Durell（味村重臣・監修、IRM研究会・訳）「ﾃﾞｰﾀ資源管理」日経ﾏｸﾞﾛｳﾋﾙ社（1987）