7H−5

データ標準化ツール（DBprompt／NAME）を用いたデータ分析手法

町原　宏毅　　黒川　清　　関根　純
（ＮＴＴ情報通信網研究所）

1．はじめに

最近のデータベースシステムの構築においては、新規にデータベースを作成し、データ投入から始めるというよりも、既存システムの更改であったり、既存システムからのデータの流通を図るケースが増えてきている[1]。

この様なデータ流通を考える場合のネックは、データ項目名やその属性、桁数などが各システムで統一されていないことである。

ここでは、既存のシステムからデータを抽出し、データ流通を図るために、それらのデータ項目名の統一と標準化のための分析手法と支援ツールについて述べる。

2．データ流通の問題点

複数システム間のデータの流通においては、データ項目に関して以下のような事象が生じやすいため、利用元のシステムを熟知していないとうまく流用できなかったり、思わぬデータ属性の変換などが必要になったりして多くの稼働がかかり、データ流通の妨げになっている。

①同じデータ項目名でも異なる定義（属性、桁数、説明、など）をされている。
②本来同じものを表現しているデータ項目が、異なるデータ項目名で定義されている。
③意味が曖昧なデータ項目名がある。
④実マシン上の定義名と帳票上で管理されている日本語名が対応していない。

3．データ標準化の原則

ここでは以下の原則に基づいて、データ項目名の標準化を行うことにする。

①同じデータには、同じ名前を付与する。
②異なるデータには、異なる名前を付与する。
③命名規則[2]に従った名前を付与する。
④統一された用語を使用する。

命名規則とは、データ項目の持つ、内容を正しく、誰にでも分かりやすいように表現するための規則である。データ項目名を構成する用語に関しては社内の統一用語を設け、同じ実体を表現するものについては、同じ用語を用いることにする。

4．データ分析手法

２節の問題を解決するために、データ項目の分析手法と、その結果に基づく標準化手法について以下に示す。全体のフローを図１に示す。

4．1　データの抽出

計算機上で実際に使われる定義名については、業務プログラムのコーディングの容易化をねらって、英数字で簡略化されたものが多く、データ項目の内容を表現した日本語名との対応は帳票で管理していることが多い。このため日本語名の情報は帳票から抽出（ＤＢprompt／ＲＴＲＶというツールで実現）し、また定義名については、マシン上に実際に定義されているスキーマ情報から抽出し、ディクショナリに格納する。

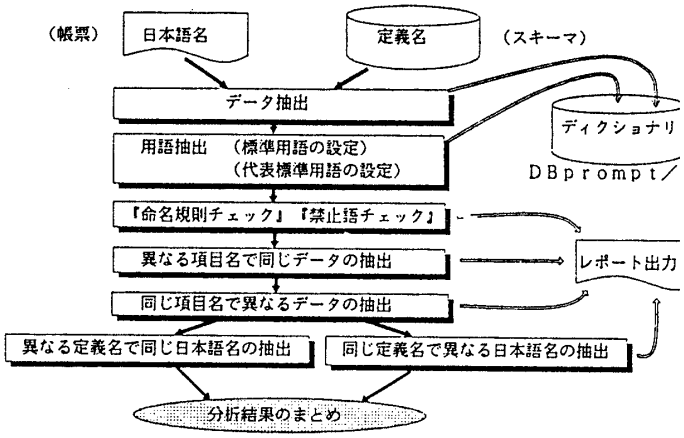


図１　データ分析の手順

4．2　用語の抽出と管理法

ディクショナリに格納されたデータ項目のチェックや検索が柔軟にできるように、用語単位に分解する。ディクショナリの中には、社内標準の用語が定義されており、その用語を使用していないデータ項目については、レポートに出力する。

用語には、社内で標準化された共用の用語辞書と、プロジェクト独自に定義するローカルの用語辞書が存在し、さらにそれぞれの中で、「用語」「標準用語」「代表標準用語」の３つにグルーピングし、用語間の関係を付ける。この用語分類により、４．４で示すような類似のデータ項目の検索が可能になる。

4．3　命名規則のチェックと禁止語チェック

ここで行うデータ項目名の命名規則に関するチェックというのは、データ項目を構成する用語の品詞に応じてその用語の並びを規定したもので、その規則に従っていないものを抽出する。

命名規則
（修飾語）＋（主要語）＋（区分語）
・修飾語：主要語の意味を補う用語（任意）
・主要語：主語となる用語（必須）
・区分語：値を具体的に表す用語（必須）

A Data Analysis Method Using Data Standardization Tool (DBprompt/NAME)
Hiroki MACHIHARA, Kiyoshi KUROKAWA, Jun SEKINE
Network Information Systems Labs., NTT

また、データ項目を構成する用語の中で、その意味が曖昧で具体性のない用語に関しては、禁止語として用語辞書に定義し、その禁止語が含まれているデータ項目を抽出する。

例えば、「情報」や数字などの用語を使用している場合は、曖昧なデータ項目としてエラーとする。

4．4　類似データの分類

ここでは、異なるデータ名で同じデータが存在しないかをチェックする。

同じ用語から構成されていないデータ項目でも、同じものを表現している場合がある。そこで用語間のグルーピングを行うことにより、関連した用語を使用しているデータ項目を抽出する〔3〕。

例えば「顧客名」というデータ項目に類似しているデータ項目を検索することを考える。図２に示すように用語「顧客」は、代表標準用語として用語「加入者」を持つ。また類似の用語としては、「ユーザ」や「ＵＳＲ」などが存在するようにディクショナリに定義されているものとする。このような場合、「顧客名」というデータ項目の類似データを分類すると、図３に示すように同一グループにある用語を使用しているデータ項目を表示させる。

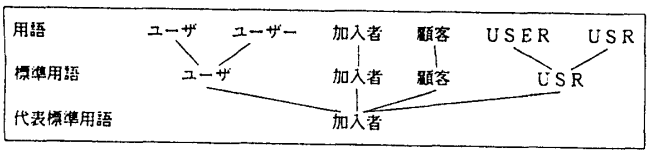


図２　用語『加入者』の構成例

データ項目名	テーブル名	システム名
ユーザ名	tab01	A
顧客名	tab11	B
加入者名称	tab33	A
顧客名	tab21	C
:	:	:

図３　『顧客名』の類似データの分類結果

4．5　属性、桁数の標準化

ここでは、同じ項目名で異なるデータ（属性、桁数、説明、等が異なる）のチェックを行う。

これによりどのシステムで同じようなデータ項目をどのように定義しているかを知ることができるため属性、桁数を含めたデータの標準化を行うことができる。

システム名	テーブル名	データ項目名	属性	桁数
B	tab11	顧客名	N	30
C	tab21	顧客名	NC	20
:	:	:	:	:

図４　同じ項目名で異なるデータ

4．6　定義名と日本語名のチェック

同じデータを表現する日本語名であっても、コーディングの問題でかなりいい加減に定義名が付与されている場合が多い。そのため、定義名を見ても、それが実際に何を表しているかを理解できない場合が存在する。

そこで、まず同じ定義名で異なる日本語名を使用していないかどうかを調べる必要がある。

次に同じ日本語名であるのにもかかわらず、異なる定義名を付与している項目を抽出する。

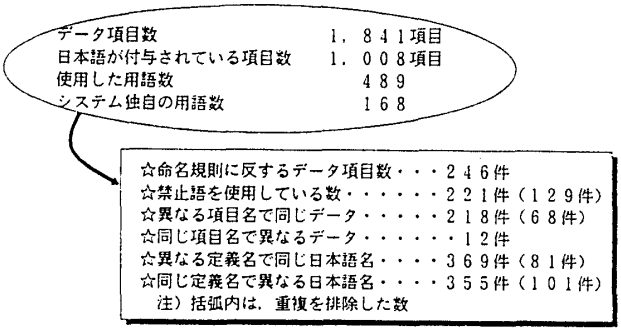
4．7　分析結果の使い方

ここで示した分析は、システム内で扱うデータ項目の不具合を示すものである。実際は、この分析結果に基づいてデータ項目の名称を修正し、標準化していく必要があるが、運用中の場合は、その更改時に変更するようにする。

具体的には、「命名規則チェック」で出されたものについては、その命名規則にあった形に修正する。また「禁止語チェック」で抽出された項目については、禁止語以外の用語に置き換える必要がある。「異なる項目名で同じデータ」で抽出された項目同士の場合、それが本当に同じデータであるかどうかについて担当者に確認し、同じであれば同じ項目名に統一する。「同じ項目名で異なるデータ」については、実際に同じデータを表している場合は、属性、桁数など異なる箇所を合わせるし、それが実際には異なるデータである場合は、異なる項目名になるように修正する。定義名と日本語名については、それらの整合性をたもつように修正する。

5．社内システムへの適応結果

ある既存システムのデータ項目について、４章の機能を実現したデータ標準化ツール（ＤＢprompt／ＮＡＭＥ）を用いて分析した結果を図６に示す。分析対象システムのデータ項目数は１、８４１項目であるが実際に対応する日本語名が帳票上に記述されていた、１，００８項目に対して行った分析を示す。



社内システムの分析結果

6．終わりに

本稿では、データ流通に伴う、データ項目の標準化のためのデータ分析の手法とその分析ツールについて紹介した。本手法によりデータを分析し、その結果に対して適宜修正を加え、標準化を行うことで、情報の流通を容易に行うことができるようになる。ＤＢprompt／ＮＡＭＥでは、製品固有のデータ項目の文字数など定量的な制限に関するチェックは実現していない。また、データのオカレンスに関しての一貫性チェックも考慮していない。今後はこのようなチェックも何らかの形で必要になるであろうと考える。

参考文献

〔1〕味村、山田、堀内：ﾃﾞｰﾀﾍﾞｰｽｼｽﾃﾑの設計と開発、ｵｰﾑ社、1983
〔2〕William R.Durell:ﾃﾞｰﾀ資源管理，日経ﾏｸﾞﾛｳﾋﾙ社，1988
〔3〕川下、関根：「ﾃﾞｰﾀ標準化を目的とした類似ﾃﾞｰﾀの分類法」、情処全国大会、第37回