データ項目名等の意味解析による概念モデルの作成法

5F－3

中渡瀬　秀一　川下　満　中川　優
NTT情報通信網研究所

1　はじめに

近年、業務プロセスの再構築(BPR)が行われることが多くなってきている。それに伴いDBシステムも再構築されることが多い。その際に新たなDB概念設計が行われることになり、実体のマッピングという作業が発生する。従来この作業は経験的に行われており実体の抽出には方法と呼べるものは存在しなかったといえる。そのためモデルは属人化により不安定な品質となり、大量の処理も困難であった。

本稿では既存のDBシステムの設計資料（データ項目名等）を分析して実体を自動抽出、整理する手法を提案する。

2　概念モデル実体抽出手法

本手法は次の２つの工程から成る。

step１．データ項目名から実体とそれらの関連（＝概念マップ）を構成する。

step２．概念マップの中から主要な実体と関連の集合を抽出する。（この実体と関連の集まりが既存システムに対する概念モデルの部分を成す。）

以下に概念マップのイメージを例で示す。

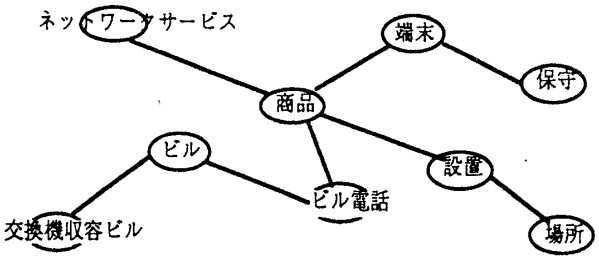


図１

2.1　概念マップ作成　(step1)

この工程では次のようにして実体、関連を構成しマップ化する。

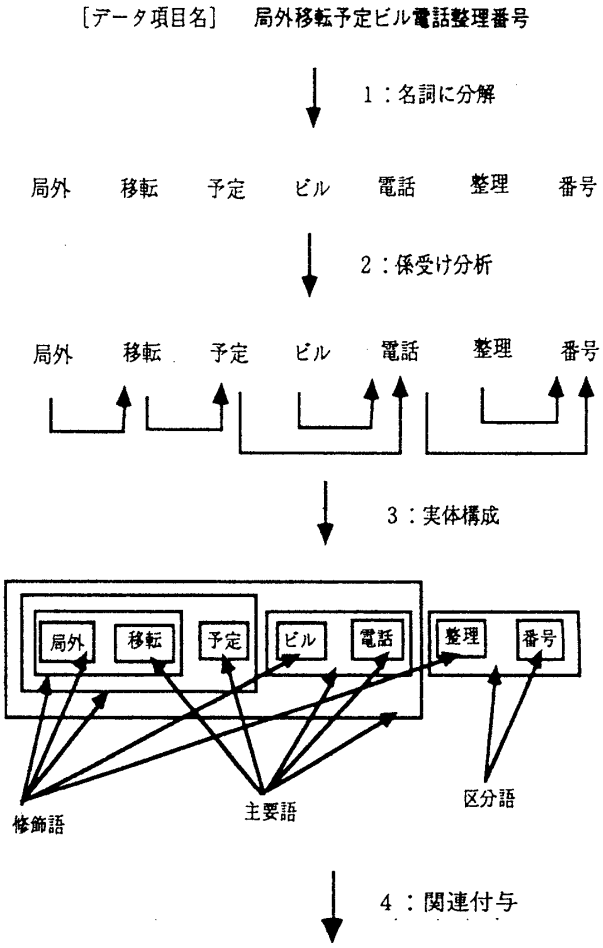
（１）実体の構成

概念は名詞によって表現されるがこれをデータ項目名の複合語から取り出すためにまずデータ項目名を修飾語、主要語、区分語に分け[1]それぞれを実体とする。ただし修飾語、主要語、区分語が内部構造を持つときはそれをさらに修飾語、主要語に分けそれぞれを実体とする。

（２）関連の付与

(１)で得られた名詞の示す概念同士の関係を得るためにそれらの係受け関係を調べる。これにより先に得た実体間に関連を付与する。

例：”局外移転予定ビル電話整理番号”から作成される実体と関連



A Design Method for Conceptual Data Model by Analyzing the Name of Data Items

Hidekazu Nakawatase, Mitsuru Kawashimo and Masaru Nakagawa

NTT Network Information Systems Laboratories

[1] 修飾語、主要語、区分語：Durell[1] の命名規則におけるデータ項目名を構成する単語の種類、修飾語（主要語の内容を分類する）、主要語（データ項目が表現する対象）、区分語（値の種類を表わす）の３つがある。

局外移転予定ビル電話 ——— 整理番号
局外移転予定 ——— ビル電話　整理 —— 番号
局外移転 —— 予定　ビル — 電話
局外 —— 移転

2.2　主要実体抽出　(step2)

「概念マップ作成」で抽出された実体はデータ項目と対応していてその量は膨大になる。よってシステム企画、設計時に人がこれに基づいて必要な実体を整理できない。そこで主要な実体（概念マップの縮図を表現できるような実体）を計算機により抽出する（実体ポートフォリオの作成）ための方法を考案した。

2.2.1　抽出アルゴリズム

抽出のターゲットとなるのは他の実体との関連の多い実体（データ項目として沢山の属性その他、関連実体を持つ実体は重要度が高い。）とする。これを数値的に評価するためモーメント形式を用いる。これを概念マップ上で定義すると次式のように表わせる。

実体ｉの概念マップＣ上での
モーメント $M_C(i)=\sum_{j\in G\ (i\neq j)} f\cdot d(i,j)$

G：実体の集合
ｆ：重み付け関数
ｄ：実体間の距離。

ここでｆ(x)を減衰関数，ｄ（ｉ，ｊ）を実体ｉ，ｊ間の最短距離とすると、各実体についてこの指標は近くに関連のある他の実体が多いほど高い値を示す。これを用いて主要実体抽出アルゴリズムを次のように定める。

１：概念マップ上のすべての実体についてモーメントを計算する。
２：モーメントの値が大きいものから順にＮ個（Ｎ：概念モデルのユーザが指定）取り出す。

3 実験

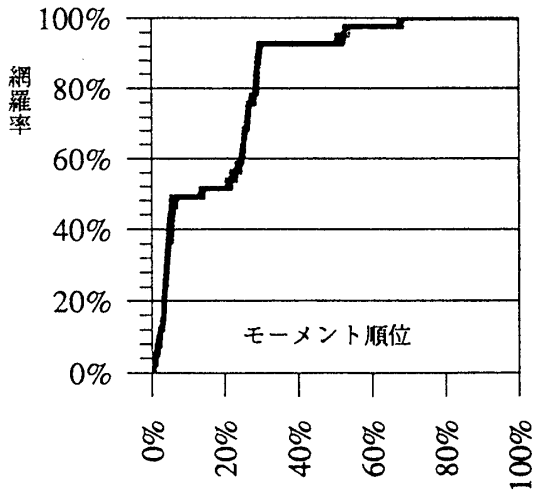
本手法の有効性を確認するために実験を行った。対象となる資料は社内で開発されたいくつかのシステムのデータ項目、及びその概念設計である。

概念マップ作成については「１：データ項目名を名詞に分解」「２：係受け分析」「３：実体構成」「４：関連付与」を手作業で行い実体を６００個余抽出した。主要実体の抽出については計算プログラムを作成しワークステーション上で実行させた。これによって実体に優先順位を付けそれが手作業で作られた実体にどれだけ高い重要度を与えられるかを調べた。なお減衰関数はｆ(x)=1/x　を使用した。

以下にその結果の１例を報告する。（他例は発表時）

3.1 結果

人が手作業で作成した概念モデルでは実体が４５個与えられている。これに対してデータ項目に含まれる名詞から実体を再構成する本手法では６００個余の実体を構成したがその中に上のうち４１個の実体を構成することができた。構成できなかった実体は直接係受け関係のない語によって構成されるものであった。さらにこれら６００個余の実体についてモーメントを計算しその値の高い順に整理しその順序が実態を反映しているかどうかを以下の図で示す。



モーメントの値の上位約５％中に実体の５０％、３０％中では９０％以上が含まれていることが分かる。

4 おわりに

本稿では概念モデルを効率良く作成するための手法を提案した。この手法はデータ項目の資料から機械的に実体を抽出するものであるが、従来人が経験的に抽出していた実体を十分に補足できるということがわかった。

今後の課題としては
１：データ項目名に含まれる語から表層的に構成できない実体を作る。（例えば「交換機」、「ケーブル」等の実体から実体「設備」を得る。）
２：ユーザの与えた観点を主要度に反映させる。（例えばサービス、生産等の観点から所内工事、開通試験を、サービス、営業等の観点から販売、受注を得るためのモーメント計算法）
ための本手法の拡張が考えられる。

参考文献

[1] Durell.R.W：データ資源管理、日経マグロウヒル、１９８７