3V－8

データ項目名等の意味解析による既存のDB中の実体型抽出アルゴリズム*

中渡瀬秀一　山室雅司　鈴木源吾

NTT情報通信研究所

1.はじめに

近年，企業におけるデータの意味を一元的に管理するために概念レベルのデータモデルが用いられている．このデータモデルを作るには，その構成要素である実体型，実体型間の関連と属性を調べなければならない．それを行うために例えば業務機能やデータベースのデータ定義などの分析作業が行われる．その際，実体型を求めるための材料になるのは人間の知識，既存のドキュメント類である．ところがモデリングの対象となる世界が巨大である場合，モデルに含まれる実体型の量も膨大であり，従来のようにこれらを人手のみで識別するのは困難である．このため実体型分析のための機械的支援方法が望まれていた．そこで我々は既にDBシステムの設計資料（データ項目名）を分析して実体型を抽出，整理する手法を提案し，簡単な実験を行った[1][2].

その後，この手法についてより詳細な実験を行い，本手法の長所を確認した．本稿ではその結果を報告し，さらに本手法の改良法に関する考察も述べる．

2.実体型抽出手法

我々が既に提案した手法の概略は以下のようなものである．

Step1. データ項目名に含まれる語を使って概念の名称（概念名）を生成する．
Step2. それらの概念間の関係を表現する．
Step3. 概念間の関係より実体型の候補となる概念を決める．

これらを実現する方法として既に我々は以下のような方法を提案した．

Step1. データ項目名を名詞に分解し，名詞間の係り受けを調べ，その結果から一定の規則にしたがってその名詞を組み合わせたものを概念名とする．

Step2. Step 1で得られた概念名，係り受け関係をそれぞれノード，リンクとする意味ネットワークを作成する．

Step3. 意味ネットワークの中で他の概念とのリンクの多さによって実体型の候補としての優先度を計算する．

我々は[1]でこの手法を評価するための実験を行った．実験では，ある分野における実体型（総数45：当該分野に多少知識のあるモデル設計者が識別した）に対するデータ項目（総数約300）からこの手法で実体型を抽出し，それと設計者が識別した実体型とを比較した．その結果，本手法によって生成し優先度付けした実体型の上位5%を取ると，人間が識別した実体型の約50%が捕捉できることを確認した．

今回はさらに多量のデータ項目を用い，同時に本手法の優先度付け方法の妥当性を確かめるために，モデル設計者に識別した実体に優先度を付けてもらい，本手法による優先度との比較を行った．また捕捉できない実体型の発生する原因の考察を行った．

3.実験

社内のあるシステムのデータ項目およびモデル設計者が識別した実体型を用いて本手法の効果を調べた．

（実験に用いた資料）
項目数　　：約１８００項目
モデル設計者が識別した実体型の総数
　　：１８５
設計者が与えた優先度別の実体型数内訳

（設計者が与えた）優先度	個数
大	２９
中	１３６
小	２０

ここでモデル設計者による優先度は以下のような方法で定めた．

1：設計者がモデルに必要であると判断する実体型をすべて識別する．（総数185個）
2：モデルで使用できる実体型の数を制限（30～40個くらい）したときに設計者がなおも必要であると判断した実体型にチェックする．この実体型の優先度を大とする．
3：その中で実体型として疑わしいものをチェックする．この実体型の優先度を小とする．
4：最初に識別された実体型（185個）のうち優

---

* An Algorithm for Identifying Entity Types of Conceptual Data Model by Analyzing the Name of Data Elements.
Hidekazu NAKAWATASE, Masashi YAMAMURO and Gengo SUZUKI.
NTT Information and Communication Systems Laboratories

先度が大でも小でもないものを優先度中とする．

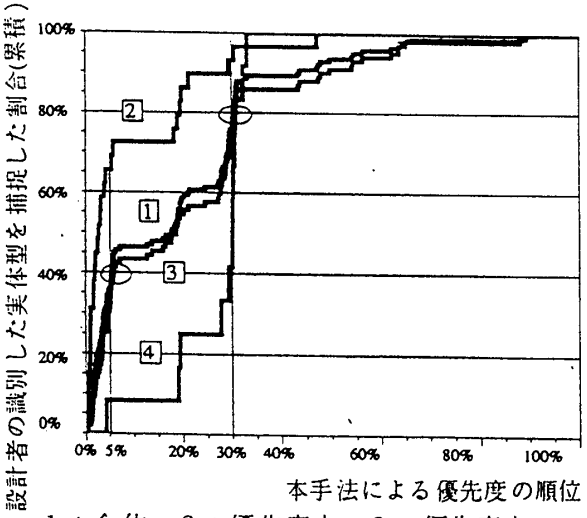
3.1 結果

まずStep 1. によって与えられたデータ項目から２７００概念名を生成した．この中に設計者が識別した実体型名１８５種のうち１４０種が含まれている．それらの設計者が与えた優先度別の内訳を以下に示す．

（Step 1で生成できた実体型名の数）

優先度	個数（設計者）	個数（本手法）
大	２９	２９
中	１３６	９９
小	２０	１２
合計	１８５	１４０

次にStep 2によって意味ネットワークを作成してからStep 3による実体型候補の優先度を計算した．その結果を以下に図示する．



１：全体，２：優先度大，３：優先度中，
４：優先度小

図１：本手法による優先度の上位Ｎ％に含まれる概念名が捕捉する実体型の割合

4. 考察

(１)Step 1で生成できない実体型名について

本手法のStep 1により生成できなかった実体型名は45個（総数185個中）であった．その原因はモデル設計者が付けた実体型名のなかに，データ項目名として使われていない語が用いられているためであった．これは以下のように分類できる．

（以下
本手法による名称＜－－＞設計者による名称）

原因１：類義語によるもの
これにはたとえば
工事担当社員　＜－－＞　工事担当者
のように語彙そのものが違うものや
ダイヤルイン　＜－－＞　Ｄイン
のように略語の関係にあるものがある．

原因２：語の修飾の程度の違いによるもの
これにはたとえば
修理進捗　＜－－＞　故障修理進捗
などがある．

原因３：総称した名称によるもの
これにはたとえば
オーダ送信　＜－－＞　オーダ送受信
オーダ受信

出張費用料金　＜－－＞　保守料金
管理費用料金
技術費用料金
などがある．

原因１に関しては，意味的には本手法によって同等な実体型が類義語として生成されているので問題ないといえる．原因２でデータ項目名の修飾の程度が原因である場合には，データ項目名にさらに多くの修飾語が必要である．この修飾語を補うには，多くの資料を用いる，知識ベースを用いて欠けている修飾語を推論するなどの方法が考えられる．原因３の場合はデータ項目名の字句からの変換（本手法）では妥当な実体型名は得られない．したがって概念辞書による援用方法が必要である．

(２)Step 3による優先度の妥当性について

実験の結果，生成された実体型名候補全体から優先度の高い順に上位5％をとると，その中にモデル設計者の識別した実体型の約40％，上位30％中に約80％が含まれていることがわかった(図１)．さらに実体型名の設計者による優先度別では優先度が大きい実体型ほど，本手法によっても高い優先度が計算されていることがわかった．つまりこれは計算した優先度が人間の認識に近いことを示唆している．つまり本手法の優先度は実体型抽出において非常に有効だといえる．

5. おわりに

本稿では，既に提案した概念データモデリングにおける実体型識別を支援する方法について，より詳細な実験を行いその結果を報告した．考察では，実験において本手法で抽出できなかった実体についてその原因を調べた．これによって本手法の課題を明確にした．今後はそれらの点を考慮して本手法を改良して行く予定である．

参考文献

[1]中渡瀬秀一，川下満，中川優：データ項目名等の意味解析による概念モデルの作成法．情報処理学会第４８回全国大会，Vol.4,pp.247-248,1994

[2] Nakawatase, H., Yamamuro, M., Kawashimo, M. and Nakagawa, M.:"A bottom-up Design Method for an Enterprise-wide Conceptual Data Model", The 13th International Conference on THE ENTITY-RELATIONSHIP APPROACH (ER'94) (Dec., 1994) (to appear)