6B-3

ドメイン情報のデータベース化へのRDFモデル適用の提案

鹿島理華† 柴光輝† 谷垣宏一† 高山茂伸†
三菱電機株式会社 情報技術総合研究所†

はじめに

企業内情報システムは、業務毎に独立したシステム、サブシステムを段階的に構築してきたため、大規模で複雑なシステムとなっている。データも各システムに分散し、同じものを示しているにも関わらず別個の名称を持ったり、そのサブシステム内だけで通じる呼称を用いたりしている。そこで、企業内の情報を一元的に管理するための辞書としてドメイン辞書が必要になる。本稿では、ドメイン辞書を、ある属性を満たすことを条件に情報をグループ化し１つのドメインで使われる用語を集めたものと定義し、その構築の一方式を提案する。

背景と課題

分散したシステム間のデータ連携を連携アプリケーションが行っている場合のシステム更新コストの低減や、互いに連携するシステムのデータ品質確保のために、統合データベースの構築やデータ連携システムなどによるマスタデータ管理の製品やサービスが提供されている。このとき、データベース間のカラム対応関係が必要となるが、データ仕様書をもとに人手で連携関係を抽出する場合、品質が作業者のスキルに依存し、工数が多くなる。また、データ仕様書自体が存在しないことさえある。そこで、データベースのカタログ情報に対し、編集距離、部分一致、接尾、接頭、親子関係といった様々な手法による判別を行い、それを総合判断してデータベースのテーブル間のカラム対応関係の自動推薦を行う手法[1]を開発した。

しかし、カラム名に、業務システムごとに固有な呼称や省略形などを使っていた場合、データベースのカタログ情報からだけでは同一性の判断ができず、カラム対応関係の自動推薦精度が低くなるという課題があった。

ドメイン辞書

ドメイン辞書の構築の提案

この課題に対し、ドメイン辞書の構築を提案する。ここで単位とするのは「概念」であり、対象ドメインにおける「概念」に対する類義語や省略語の表記形式を登録し利用することで、表記のばらつきに対応することができ、カラム対応関係の自動推進精度を上げることができる。

ドメイン辞書の要件

一方、これら情報は企業内で使われているデータ（概念）の「情報」であり、今後、ドメイン辞書へ概念の意味や、概念間の関係性なども格納していくことで、表記のばらつきを解決するためだけでなく、企業内の情報を一元的に管理するための辞書として構築していくことができると考える。これらをふまえ、ドメイン辞書は次の要件を満たす必要がある。

- １つの概念は、複数の情報（表記形、説明、意味など）を持つことができること
- 概念は他の概念と意味的に結びつくことができること（is-a 関係や part-of 関係など。これら関係は絶対的なものではなく、ドメインごとに異なる）
- ドメイン辞書に格納する内容は、仕様を変更することなく拡張することができること

実装モデルの検討

スキーマが固定のリレーショナルデータベースや csv では、前述の要件を満たすことができない。このため、ドメイン情報を RDF (Resource Description Framework)[2]モデルを用いて表現することを検討した。RDF モデルは、リソースの関係を主語、述語、目的語（プロパティの値）という 3 つの要素で表現し、メタデータなどリソースの相互関係を、特定のアプリケーションを前提とせずに記述するための標準的な枠組みを提供する。また、RDF モデルでは、リソースに入れる値はあくまでも属性値であるので、将来ドメイン辞書への格納内容を拡張する場合も、リソース、プロパティ、リテラルの属性を追加することで対応できる。

RDF モデル適用例

ドメイン情報を、RDF モデルを使って表した例を図１に示す。概念間の関係を示すプロパティ(図上の矢印)の URI の名前空間には、メタデータ共通化のために国際標準となっている語彙

A proposal of the RDF model application to domain information
†Rika Kashima, Mitsuteru Shiba, Ko ichi Tanigaki, Shigenobu Takayama ・ Information Technology R&D Center Mitsubishi Electric Corp.

セットである DC(Dublin Core)と DC の拡張版の DC タームズを使用した。

図１では、概念「識別子」は、”ID”と”No”の２つの表記形を持ち、概念「設備識別子」をサブクラスとして持つ（is-a 関係）ことを示している。また、概念「設備」と、「設備識別子」および「設備年月」は part-of 関係にあることがわかる。また、各表記形には、説明やカラムの物理的な情報（図中、「カラム URL」と表記）を付加している。これにより、カラムという物理的な情報と、概念の情報を結びつけることができる。この例では、カラムの物理情報は、“http://localhost/スキーマ名/table/テーブル名/column/カラム名”といった URL で入れている。ただし、カラムを一意に特定できる形であればよく、その表記方法は、このドメイン辞書の情報を使用するシステムが、テーブル情報をどのような形で扱うかに拠るので、各システムに合わせた形式のカラムの識別情報を入れればよい。このように、RDF モデルでは、使用する側の仕様にあった形で情報を保持することができる。

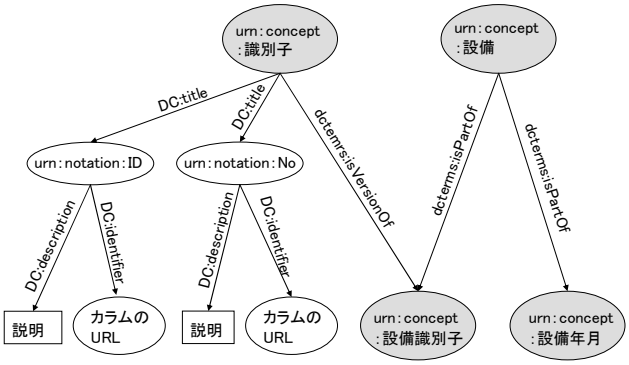

図 1　モデル化例

今後の課題と機能拡充

企業内システムは段階的に構築されているため、テーブルのカラム名は、そのサブシステム内では統一されていても、サブシステム間では英語表記であったりローマ字表記であったりする。これらを全てドメイン辞書で網羅するためには、ドメイン辞書へ入れる情報を自動的に生成するしくみが必要となる。

一方、自動的文書解析や人工知能のアプリケーション支援のために構築された概念辞書として、WordNet[3]がある。WordNet では概念にあたる synset と呼ばれる同義語のグループに分類され、簡単な定義や、他の同義語のグループとの関係が記述されており、Princeton WordNet に日本語を付与した日本語 WordNet[4]も公開されている。

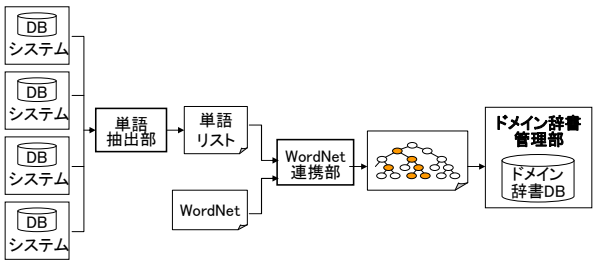

図 2　WordNet 連携

そこで図２で示す機能を現在検討中である。対象とするシステムのデータベースのカタログ情報から、例えば“HOSHU_DATE”, ”HOSHU_ID”であれば“HOSHU”,“DATE”,“ID”と単語分解し、抽出した単語を WordNet で定義された synset に関連付け、該当ドメインでの出現単語と WordNet を連携した辞書情報を構築する。synset には複数の同義語が定義されているため、出現する単語を単に synset に関連付けると、候補が増えかえって自動推薦精度が落ちるため、対象とするドメインでの単語の使われ方に着目した曖昧性解消手法を検討中である。このように構築したドメイン辞書情報は、一般的な概念情報とドメイン固有の情報を持っているため、カラム間の対応関係抽出の精度をさらに向上させることができると考える。

おわりに

企業内システムの連携時に必要となる、カラム間対応関係抽出の自動推進精度向上のための、RDF モデルを適用したドメイン辞書を提案した。また、このドメイン辞書に一般的な概念情報を加えるための WordNet 連携の構想について述べた。今後、この WordNet 連携による半自動的なドメイン辞書生成を実装し、商用システムへの適用を検討していく予定である。

参考文献

[1] 小出他 「学習データ量によるスキーママッチング精度向上効果評価報告」情報処理学会第74回大会 6B-4
[2] W3C, “RDF”, http://www.w3.org/RDF/
[3] Princeton University, “About WordNet”, WordNet, Princeton University, http://wordnet.princeton.edu, 2010.
[4] 日本語 WordNet, http://nlpwww.nict.go.jp/wn-ja/