セマンティクス自動抽出によるエンタープライズオントロジの構築

Enterprise Ontology Synthesis by Semantics Extraction from Web Documents

高梨　勝敏　Takanashi Katsutoshi
佐藤　俊也　Satou Shunya
塚原　朋哉　Tsukahara Tomoya

知識体系の可視化による情報共有システムの有効利用を目的として，エンタープライズオントロジの構築手法を提案した。本手法は，組織の知識には共有知識と個別知識が存在すると仮定し，両者の関係を規定することで共有－個別知識マップを作成するモデルに基づいている。このモデルに基づき，イントラネットでの知識体系共有システムのプロトタイプを作成し，オントロジに対しユーザの意思を反映させるための方式を検討した。プロトタイプを社内情報に適用したところ，オントロジ構築における従来の課題が改善され，その有効性が示唆された。

1. はじめに

近年，多くの企業では，情報共有システムによる業務効率・生産性向上に大きな関心が寄せられている。(株)日立東日本ソリューションズにおいても，2000 年度のナレッジマネジメント研究を契機として”知識交流システム inxs” [1]や”CoreExplorer” [2]の研究開発をおこない，業務に必要な情報の取得効率向上を支援してきた。

しかし，断片化された情報の共有だけでは業務効率の向上は難しく，情報間の関係を体系立てて共有することが課題である。このためには，システムが文書の単語情報だけでなく意味情報すなわちセマンティクスを扱えるようにすることが有効な手段と考えられる。

Web のコンテンツを知識として活用するための基盤として，セマンティック Web が注目されている [3]。セマンティック Web を構成する要素技術に，コンテンツの意味情報を記述するためのメタデータ記述言語およびメタデータ間の関係を記述するためのオントロジがある。特に企業等の組織では，各部署や業務プロセス毎のオントロジ(ドメインオントロジ)と，それらを統合し組織全体に拡張したエンタープライズオントロジが考えられる。エンタープライズオントロジを構築することにより，部署を越えた情報共有，業務理解を通して，組織の生産性向上が期待できる。

オントロジを構築する手法として，対話的構築ツールを用いた手動構築の研究 [4]や，Web コンテンツの収集・クラスタリングによる自動構築の研究 [5]がおこなわれている。前者では，人間がオントロジの構造を決定するので構築の精度が確保できるが，構築にコストがかかる，また日々出現する新しい情報を即時に反映させることが困難，等の課題がある。また，後者では精度の確保に課題がある。

一方では，これらと異なった立場で，人間が情報システムに対し行う操作から意味情報を抽出する研究がおこなわれている [6]。

本研究は，この立場から，ユーザの意思をオントロジに反映させるアプローチを採用した。このアプローチに従い，以下の項目を検討した。

(1) イントラネットや文書管理システム等で共有されている各種情報から，オントロジを自動的に導出するモデル
(2) オントロジに対してユーザの意思を反映させ，オントロジ構築の精度を向上させる手法

以下，モデルに基づいたプロトタイプの内容とユーザの意思を反映させる手法およびプロトタイプを組織のイントラネットに適用した時の評価を示す。

2. 共有－個別知識マップモデル

組織が持つ情報にはどのようなものがあるか，それらの情報からオントロジを自動的に導出するためにはどのようなモデルが考えられるかを検討した。

一般に，全ての知識を 1 個のオントロジで表現することは困難であり，個別の状況(コンテクスト)に従ったオ

ントロジの構築が重要とされている [7]。本研究では，この考え方に基づき，組織で共有している知識には共有知識 S(Shared Knowledge)と個別知識 I(Individual Knowledge)が存在すると仮定する。企業組織では，共有知識として使用技術の分類や用語体系等のエンジニアリング知識，業務のプロセス分類や各プロセスの入出力を定義したプロジェクトマネジメント知識等がある。
また，個別知識としては，各プロジェクトでの業務を通して個人が作成する設計文書や報告書等がある。これら知識間の関係から，共有－個別知識マップを作成するモデルを以下に示す。

組織が共有する全ての情報 K を，情報 k の集合として以下のように定義する。

$$K = \{k_1, k_2, ..., k_i\} \quad \text{.....(1)}$$

ここで，k は組織の共通用語，共有文書，共有文書が含む単語やメタデータ等である。あるドメインオントロジ m での共有知識 S_m と，個人 n の個別知識 I_n を次式で定義する。

$$S_m = \{s_m(k_1,k_2), s_m(k_1,k_3), ...s_m(k_j,k_k)\},$$
$$I_n = \{i_n(k_1,k_2), i_n(k_1,k_3), ...i_n(k_j,k_k)\} \quad \text{...(2)}$$

ここで，s_m および i_n は情報間の関連を表す。これは各知識での k 同士の関連の種類 r と関連の距離を表す関連度 d の組であり，以下のように表される。

$$s_m(k_p,k_q) = (r_{mpq}, d_{mpq}),$$
$$i_n(k_p,k_q) = (r_{npq}, d_{npq}) \quad \text{.....(3)}$$

関連の種類 r は既存の各研究でいくつかの定義がおこなわれているが，本研究では包含関係(is-a)，全体－部分関係(has-a)および類似関係(related-to)を用いる。また距離 d は，文献[9]に基づき親子関係，文書に単語が出現する確率および単語の包含関係により定義した。

知識を S と I に分けることにより，以下の式がオントロジ定義のために新たに付け加えられる。

共有知識間の関連　：

$$b(S_{m1}, S_{m2}, k_p, k_q) = (r_{m1m2pq}, d_{m1m2pq})$$

個別知識間の関連　：

$$b(I_{n1}, I_{n2}, k_p, k_q) = (r_{n1n2pq}, d_{n1n2pq})$$

共有・個別知識間の関連　：

$$b(S_m, I_n, k_p, k_q) = (r_{mnpq}, d_{mnpq})$$

.....(4)

これら 3 個の関係を用いて，それぞれドメインオントロジ間の対応関係，個人間の関係および個人がどのような業務に関わっているかを表すことができる。ここで b は階層構造を持っていないため，r は類似関係(related-to)のみで表現することができる。

共有知識 S_m と個別知識 I_n およびそれらの関連は，情報 k 間の関連の種類と関連度を表現した知識マップである。これは，ラベル付き有向グラフを用いて表現できる [8]。知識マップの概要を図 1に示す。関連を情報間の矢印として表す。各矢印には，関連の種類と関連度を表すラベル s，i が添えられている。関連度は矢印の長さでも表すことができる。これらが共有－個別知識マップの要素であり，異なる業務および個人，またこれらの間で情報がどのように関連しているかを知ることができる。

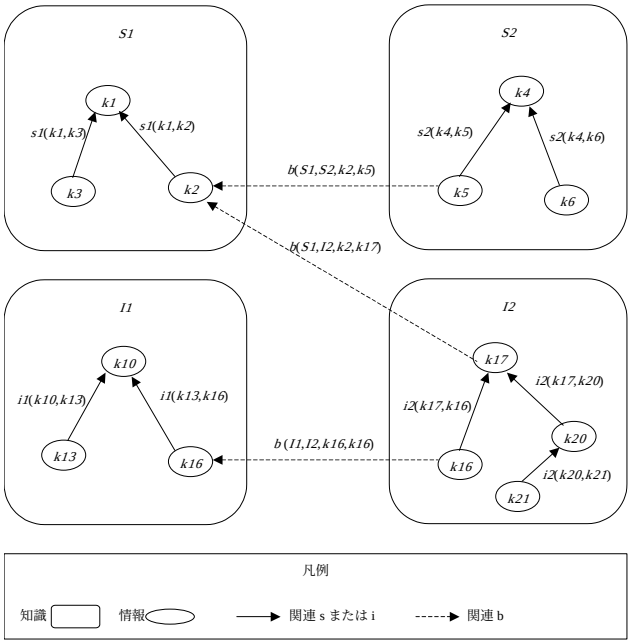

図 1　知識マップの概要

組織における知識マップの適用例として，エンジニアリングプロセスと組織階層の知識マップを図 2に示す。図 1では関連の種類と距離をラベルで表していたのに対し，図 2ではラベルの内容を図示するために，種類を矢印の形状，距離を矢印の長さで表している。例えば，開発プロセスは設計，プログラミングおよびテストのサブプロセスで構成されている(has-a 関係)。
また，設計は基本設計と詳細設計の上位分類と表現されている(is-a 関係)。

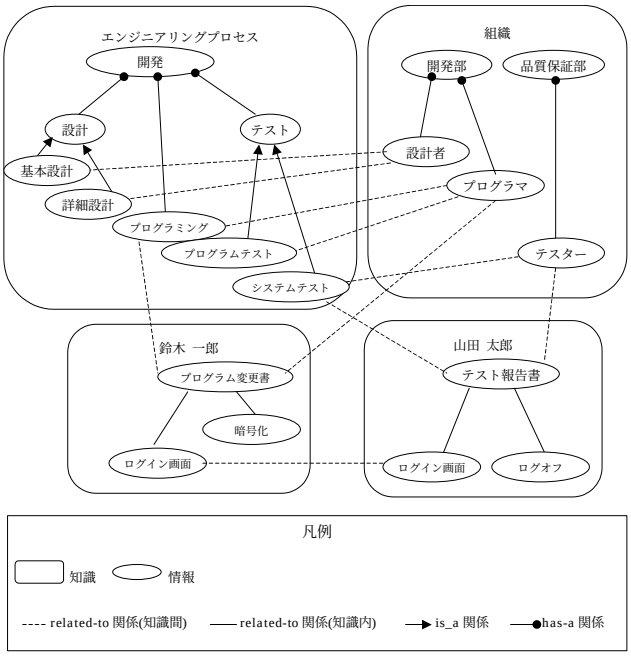

図 2　知識マップの例

3. プロトタイプの作成

以上のモデルに基づき，共有－個別知識マップを自動生成するためのシステムを検討し，プロトタイプを作成した。本プロトタイプでは，作成した知識マップに対しユーザの意思を反映させるための手法を検討した。オントロジ生成の流れは，以下のとおりである。

(1) 共有知識の収集
(2) 個別知識の収集
(3) (1)と(2)を用いた知識マップの生成
(4) 知識マップへのユーザの意思の反映

以下に各項目の内容を説明する。

(1) 共有知識の収集

イントラネットで公開されている表形式のデータを巡回し，共有知識を構成する各情報を自動収集する。この収集結果は，スキーマ形式 [3]のファイルに出力する。スキーマの例を図 3に示す。情報を<term>ノードで表現し，子ノードで情報の属性と，情報間の関係を表す。

(2) 個別知識の収集

組織で共有されている文書ファイルを巡回し，文書の表題，作成者，最終更新日等のプロパティと，本文のテキストを抽出する。この抽出結果と (1)のスキーマの文字列マッチングをおこなう。また，本文から tf-idf 法 [9]により特徴語を抽出する。これらを文書のメタデータとして，文書に付与する。付与結果は，RDF 形式 [3]のファイルに出力する。RDF の例を図 4に示す。<property>ノードにおいてメタデータの文字列をノードのテキスト部分に，文書との関連度 d を score プロパティに格納している。関連度は，文字列マッチングおよび tf-idf 法により求めたスコアである。

```
<?xml version="1.0" encoding="UTF-8"?>
<ladl xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#">
  <rdf:RDF>
  <term rdf:about="http://foo/bar.html#1">
    <name>仕様書</name>
    <accession>ladl:00000001</accession>
    <category>TOS</category>
  </term>
  <term rdf:about="http://foo/bar.html#2">
    <name>機能仕様書</name>
    <accession>ladl:00000002</accession>
    <is-a rdf:resource="ladl:00000001"/>
    <category>TOS</category>
  </term>
  <term rdf:about="http://foo/bar.html#3">
    <name>詳細仕様書</name>
    <accession>ladl:00000003</accession>
    <is-a rdf:resource="ladl:00000002"/>
    <category>TOS</category>
  </term>
  <term rdf:about="http://foo/bar.html#4">
    <name>テスト報告書</name>
    <accession>ladl:00000004</accession>
    <category>TOS</category>
  </term>
</rdf:RDF>
</ladl>
```

図 3 共有知識収集で作成したスキーマの例

```
<?xml version="1.0" encoding="SHIFT_JIS"?>
<ladf xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#">
  <RDF >
    <Description about="http://foo/ users/suzuki/product1/1_0/ds.doc">
      <property type="subject" score="1.0">○○システム ver. 1.0 詳細仕様書</property>
      <property type="group" score="0.8">ビジネス第 1 グループ</property>
      <property type="author" score="0.9">鈴木 一郎</property>
      <property type="cdate" score="1.0">2004 年 07 月 13 日</property>
      <property type="lmdate" score="1.0">2004 年 08 月 29 日</property>
      <property type="TOS" score="0.8">詳細仕様書</property>
      <property type="Product" score="0.75">○○システム 01-00</property>
      <property type="key" score="0.53">性能</property>
      <property type="key" score="0.21">ユーザビリティ</property>
      <body>
  ○○システム ver. 1.0 詳細仕様書　作成日:2004/07/13　更新日:2004/08/28　作成者:ビジネス第1グループ 鈴木　　1. システム前提 2　2. システム構成 3　3. データベース仕様 4　4. ファイル構成 5　4.1 ファイル仕様
*** 中略 ***
フォルダの移動・コピーを実行：フォルダの移動・コピーを実行し，実行完了メッセージを表示する。移動・コピー後のパス名が長すぎる場合，ルートフォルダを移動しようとした場合および，移動・コピーに失敗した場合は，エラーメッセージを表示する。
      </body>
  </RDF>
<ladf>
```

図 4　個別知識収集で作成した RDF の例

(3) 知識マップの生成

(1)の共有知識と(2)の個別知識を用いて, 知識マップを自動生成する。知識マップの概要を図 5に示す。知識マップの表示は，節2の共有－個別知識マップモデルに従

っている。モデルの式を用いて，以下に表示手順を示す。共有知識 S は，スキーマにおける親子関係から関連 s を求め，ツリー状に表示する。共有知識の各情報と文書の関連 b は文字列マッチングにより求め，文書を関連度の高い情報の近傍に配置する。文書とメタデータ間の関連 i は，文書間ではベクトル空間法 [10]により求め，メタデータと文書間では RDF に記述された関連度により求める。

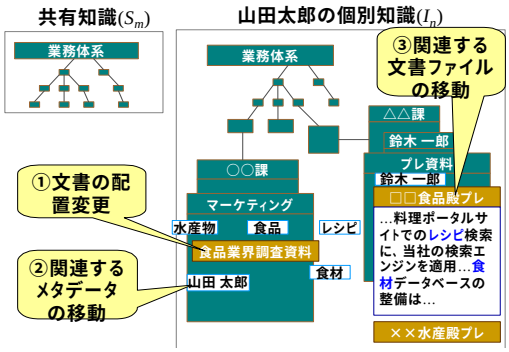

図 5 知識マップの生成と操作

(4) 知識マップへのユーザの意思の反映

知識マップでは，ユーザが文書の配置を変更することができる。変更操作により，式(4)により求まる文書と共有知識の関連 b に，ユーザの意思に基づいた分類の意思を反映できる(図 5の①)。

ここで，オントロジ構築の精度を向上するために，ユーザの操作に対するフィードバックの仕組みを検討した。ユーザが文書の配置を変更すると，文書と関連度の高いメタデータや他の文書の配置が，関連度 d が変化しないように移動する(図 5の②)。この結果，各文書やメタデータと共有知識間の関連 b が変化する。すなわち局所的な操作の影響が即座にオントロジ全体に反映されることで，自分が行なった操作の妥当性を検証することができる。

4. 社内情報への適用

プロトタイプを用いて当社の個別知識を共有知識で整理することを試みた。ここで共有知識には，4 つのドメインオントロジを用いた。すなわち，エンジニアリング規格，製品知識，技術知識，組織構造である。適用結果を以下に説明する。

(1) 対象情報の収集

まず，知識を抽出するための対象情報の選定をおこなった。各ドメインオントロジについては，イントラネットにおける各部署のサイトから選定した。また，個別知識として取得する文書群は，文書管理システムに格納されている設計文書，プレゼンテーション資料(以下プレ資料)および報告書等を採用した。

次に，これらのサイトと文書群に対し毎日 1 回のバッチ処理で情報を収集し，共有知識，個別知識およびこれらの関連を導出した。これら自動処理により，オントロジ構築のコスト低減と即時性を確保した。

(2) 知識マップの生成

(1)の結果得られた関連を表 1に示す。共有知識と個別知識の各組み合わせにおける関連度を，高，中，低の 3 段階に分類し，それぞれ”○”，”△”，”×”で示した。

表 1　共有知識と個別知識の関連度

個別知識＼共有知識		ドメインオントロジ			
		エンジニアリング規格	製品知識	技術知識	組織構造
文書	設計文書	○	×	△	○
	プレ資料	×	○	×	○
	報告書	×	△	○	○
	その他	×	○	△	○

表 1における関連度の高い各組み合わせについて，知識マップを自動生成した。例として，エンジニアリング規格と設計文書の知識マップを図 6 に示す。

共有知識の各情報(「提案資料」や「契約書」等)は楕円で表している。楕円の階層状の配置は，親子関係を表している。これは，共有知識における関連，すなわち節2における式(3)の s_m により求まる。関連の種類は，楕円を選択することにより画面左側に表示され，参照・変更が可能である。

また，個別知識の各情報は長方形で表している。その中で，文書は黄色，メタデータは水色である。長方形間の配置は，関連の距離を表している。関連の種類は，それぞれの長方形の近傍にある楕円を親とし，親の間の関係を用いて知ることができる。これは，個別知識における関連，すなわち式(3)の i_m により求まる。長方形間の距離や，長方形・楕円間の距離は，知識間の関連，すなわち式(4)の b により求まる。

以上の処理によって，組織で共有されているエンジニ

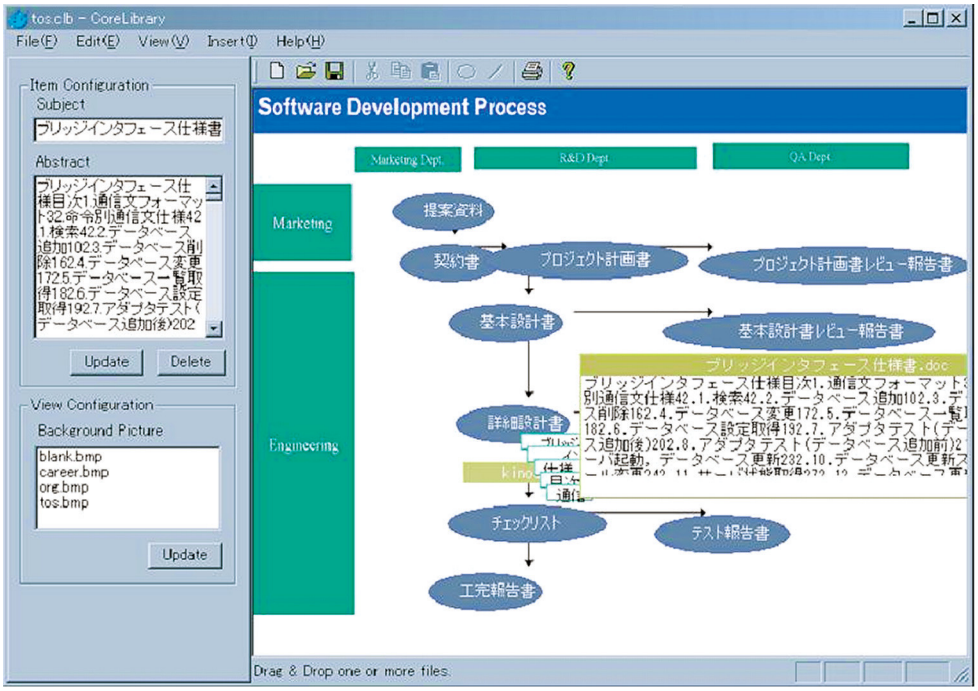

図 6 知識マップの適用例

アリング規格に従ったオントロジを自動的に作成することができる。

(3)ユーザの意思の反映

知識マップにユーザの意思を反映させ，オントロジの精度を向上させる手順は以下のとおりである。

図 6 は，ユーザが文書「ブリッジインタフェース仕様書」の位置を決める場合を表している。ユーザは当該文書が「詳細設計書」の楕円に分類されると判断し，楕円の近傍に文書を配置する。このとき，文書と関連度の高い他の文書やメタデータも関連度 d が変化しないように移動する。ユーザは，文書とメタデータの内容を参照して，詳細設計書に関連する情報であることを確認できる。この結果，3章の(4)で述べたフィードバックの仕組みによって，自分が行った配置の妥当性を検証することができる。

以上の手順により構築されたオントロジは，共有知識を基本としているが，各ユーザの意思がより反映されていると考えられる。本オントロジを用いて，個人が作成した文書を組織共通の分類体系で閲覧・検索することが可能となり，情報共有システムの有効利用に役立つことが期待される。

5. おわりに

組織の知識を共通知識と個別知識に分け，これらを関係付けることによるエンタープライズオントロジ構築のためのモデルを提案した。また，ユーザの意思をオントロジに反映させ，オントロジ構築の精度を向上させる方式を検討した。これに基づいたシステムのプロトタイプを作成し，有効性を確認した。

今回は，個人毎の知識マップ作成による知識共有を対象として研究を行った。今後は，以下の項目での発展が期待できる。

(1) 個人・グループの知識共有活動の組織への反映

個人の知識マップの操作が組織に及ぼす影響を把握することによるプロセス改善の仕組みを検討する。

(2) インターネットでの情報の体系化

対象となる情報をイントラネットからインターネットへ拡張し，複数組織にまたがる検索サービスを検討する。

参考文献

1) 高梨 他：ミクロ・コミュニティの知識交流システム，インタラクション 2002 ，IA-3，2002
2) 塚原：情報の視覚的検索方法，fit2003 一般講演論文集 第二分冊 pp.179-180，2003
3) 清野 他：セマンティック Web とオントロジ記述言語，情報処理学会誌， No. 43，no.7，pp.727-733，2002
4) 森田 他：セマンティック Web のツール，情報処理 No. 43，no.7，pp.734-741，2002
5) 鍾 他： Web マイニングおよび Web インテリジェンスに関する研究，電気通信普及財団研究調査報告書 第 17 号，pp. 576-591，2002
6) 小倉 他：セマンティック Web の応用システム，情報処理学会誌， No. 43，no.7，pp.742-750，2002
7) 中村 他：コンテクストアウェアコンピューティングとコンテクストの定式化，人工知能学会研究会資料 SIG-SWO-A402-03，pp03-01 – 03-08，2004-11
8) 猪口 他：多頻度グラフマイニング手法の一般化，人工知能学会論文誌，vol.19，No.5，pp368-378，2004
9) 廣田 他：文書情報からの分野オントロジ構築の支援，自然言語処理， vol.140，No.8，pp55-60，2000
10) 寺田 他：文脈情報を使用した略語の自動復元，研究報告-自然言語処理，vol.2001，no.69， pp39-45，2001

高梨 勝敏　　1995 年入社
生産技術推進センタ
ナレッジマネジメントシステムの企画・推進，研究開発

takana@hitachi-to.co.jp

佐藤 俊也　　1993 年入社
ナレッジソリューション G
CoreExplorer，テキストマイニングツールの拡販，コンサルティング

shu_sato@hitachi-to.co.jp

塚原 朋哉　　1997 年入社
ナレッジソリューション G
CoreExplorer，テキストマイニングツールの研究開発

tomo@hitachi-to.co.jp