人工知能学会研究会資料
SIG-SW&ONT-A302-08

Web コンテンツの分析に基づくオントロジー構築および情報整理の試み

松平　正樹　　上田　俊夫　　大沼　宏行　　森田　幸伯

我々は，インターネットやイントラネット上の雑多な情報（非構造化情報）と，データベースや Web サービスのような構造化された情報（構造化情報），そして，イントラネット上の仕様書や製品カタログといったフォーマットはある程度固定されているがそれぞれの項目が明示的に構造化されていない情報（半構造化情報）をオントロジーによって統合し，利用者に必要な情報を収集・抽出して提供するシステムを開発している．本稿では， IT 技術に関するイントラネットおよびインターネット上の Web コンテンツの分析に基づくオントロジー構築について説明し，開発しているシステムについて概説する．また，システム内で重要なモジュールのひとつである属性抽出処理について説明し，イベント情報の属性抽出実験について報告する．

キーワード：オントロジー，情報抽出，Web サービス，構造化情報，半構造化情報

Ontology Construction Based on Web Contents Analysis And Information Arrangement Using It

Masaki MATSUDAIRA　　Toshio UEDA　　Hiroyuki OHNUMA　　Yukihiro MORITA

We develop a system, which collect, extract and combine non-structured, structured and semi-structured information based on ontology. This paper explains how to construct ontology about IT domain based on web contents analysis, and describes an outline of the system developed. An information extraction process, which is one of important processes in the system, is detailed, and an experiment report on information extraction from web pages is also introduced.

Keywords: ontology, information extraction, web service, structured information, semi-structured information

1.　はじめに

近年，インターネットの発達により大量の情報が流通する中，利用者が必要な情報を容易に得ることは困難になってきている．例えば，ある技術の概要を知りたい，あるいは，その技術を利用した代表的な製品について調査したい，技術を研究している大学や公的機関を調べたいといった様々な要求に，Yahoo!やGoogleに代表される検索エンジンは答えているだろうか？　実際にキーワードを入力して検索すると，必要な情報が多くの「ゴミ」に埋もれてしまうため，検索結果をひとつひとつ調査しなければならず，多大な手間がかかってしまう．また，イントラネットに必要としている情報が存在する場合でも，ナレッジマネジメントシステムやコンテンツマネジメントシステムによってすべての情報が組織的に管理されていることは一般的には少なく，インターネット検索エンジンと同様の問題が起こる．

この問題に対して，必要な情報を選択あるいは統合するために，セマンティック Web やトピックマップ等の技術が開発され，利用され始めている．

例えば，我々は，LSI 製品に関するオントロジーを構築し，イントラネット上の情報を統合・整理して提供する実験システムを試作した [松平 03]．そのシステムにおいて，オントロジーの構築は，イントラネット上にどのような情報が存在するか調査した結果をもとにボトムアップ的におこない，各概念に対応する情報を人手で関連づけした．初期コストの課題が残るものの，製品情報を統合して出力できる点は利用者から評価されている．また，Pepper らは，トピックマップで表現された情報を整理して表示するシステムを開発し，XML Conference や XML Europe 等の論文，キーワード，著者，組織等の情報をオントロジーとして構築している．

一方，インターネット上の情報を整理する研究もおこなわれている [岩爪 97, 武田 01, 前田 97, 仲川 01]．岩爪，武田らの提案したシステム IICA は，

沖電気工業（株）　研究開発本部
Oki Electric Industry, Co., Ltd. Research and Development Division
E-mail: matsudaira564@oki.com

オントロジーおよび対象領域特有の言語表現パターンに基づくヒューリスティックスを利用して，必要な情報を自動収集・分類・統合化するシステムである．弱構造化オントロジーという概念間の連想的な関係のみを記述したオントロジーを採用している．前田らが試作した CM-2 は，雑多な情報をゆるやかに関連づける連想構造というデータ構造を用いて，情報を収集し，整理する過程を支援するシステムである．また，仲川らは，検索のたびに，ユーザの検索目的に適した小規模なカテゴリ構造を提供することで検索作業を支援する手法を提案している．いずれの研究も，膨大な情報の中から必要な情報を取得するために情報を分類して整理するという方向性は，興味深い．

しかしながら，それらの研究は，Velardi ら [Velardi 01] が定義した Vertical relations としての Broader（上位下位関係;いわゆる isa 関係）や PartOf（部分全体関係），InstanceOf（クラス・インスタンス関係）および Horizontal relations としての Predication（属性項目）や Relatedness（(汎用的な）関連概念）等の概念間の意味関係がオントロジーとして明確に定義されておらず，構造化された情報と統合しにくいという問題がある．すなわち，前田らが例として挙げている「奈良」から「奈良公園」「大仏」「鹿」が想起されるという連想構造は，一般的に「奈良」が持つ自治体情報，地図，交通情報，宿泊情報といった地名や都市名に共通する情報が表現しにくく，それらの情報を格納したデータベースとの関連づけが困難になる．

このような問題に対して，我々は技術情報をターゲットとして，インターネットやイントラネット上の雑多な情報（非構造化情報）と，データベースや Web サービスのような構造化された情報（構造化情報），そして，イントラネット上の仕様書や製品カタログといったフォーマットはある程度固定されているがそれぞれの項目が明示的に構造化されていない情報（半構造化情報）をオントロジーによって統合し，利用者に必要な情報を収集・抽出して提供することが重要であると考え，それを実現するシステムを開発している．図 1 に全体像を示す．構造化情報としては，論文データベース，特許データベース，書籍情報等，インターネット上にも数多くあり，Web サービスとして提供されているものもある．例えば，Amazon は，書籍情報をタイトル，著者等で検索する機能を提供しており，XML 形式で情報を取得することができる．非構造化情報および半構造化情報は，情報抽出技術を用いてオントロジーに対応した情報を抽出する．

本稿では，まず 2 章でインターネットおよびイントラネット情報の分析について述べ，3 章で構築したオントロジーについて説明する．4 章では，システム全体について概説し，5 章では，オントロジーを用いた属性抽出方式とその実験結果について報告する．最後に，結果の分析ならびに今後の課題について 6 章で議論する．

2. Web 情報の分析

2.1. 分析対象

我々は，IT に関する技術情報について，インターネット，イントラネット上にどのような情報が存在するかを調査し，内容を分析した．対象としたのは，表 1 に示すように，インターネットでは「オントロジー」「XML」「VoIP」の各技術キーワードについて Yahoo!検索結果の上位ページ合計 300 件，イントラネットでは，「オントロジー」「XML」「VoIP」に加え，「MPEG」「CTstage」「ML2870」のイントラ検索結果の上位ページ合計 376 件である．CTstage および ML2870 は，それぞれ弊社の CTI ソリューション，音声 LSI の名称である．イントラネットで技術キーワードを追加した理由は，「オントロジー」に対する情報があまりにも少なかったことと，自社製品に関する社内情報を必要とする場合が多いと想定されるためである．

2.2. 分析結果

まず，内容によって主観的な分類クラスを作成し，

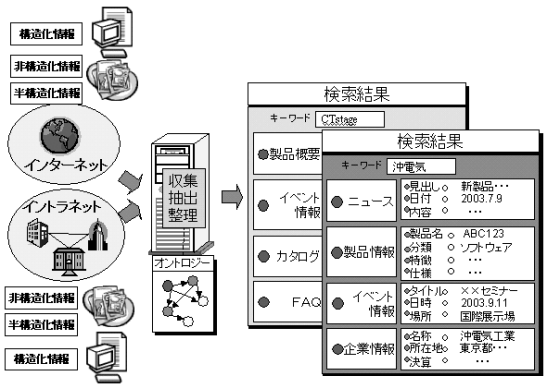

図 1　全体像

表 1　調査対象

リソース	技術キーワード	件数
インターネット	VoIP	100
	XML	100
	オントロジー	100
小計		300
イントラネット	VoIP	100
	XML	50
	オントロジー	4
	MPEG	100
	CTstage	100
	ML2870	22
小計		376
合計		676

表 2　インターネットの分類結果

分類クラス	オントロジー	XML	VoIP
技術解説	**13**	**35**	**21**
論文	**26**	0	0
イベント	7	0	0
組織･団体	3	7	2
書籍	3	**11**	1
ニュース	6	3	**16**
製品	4	8	**43**
ポータル	0	**12**	0
教育	5	3	1
エラー	10	1	9
その他	23	20	7

インターネット情報の Web ページを分類した．結果を表 2 に示す．表 2 において，エラーはリンク先ページの存在しないものを表し，その他は FAQ や個人ページ，市場調査報告書等である．また，表には記述していないが，上述した十数種類の分類クラスに対して，さらに細分類を付与している．例えば，論文の細分類として，論文の PDF ファイルや，学会のアブストラクトのページ，発表スライド，研究会プログラムの論文タイトル一覧等があり，組織・団体としては企業のページや学会，コンソーシアム等の団体のページがある．

結果から，各技術キーワードともに技術解説に関するページが多く見受けられるが，「オントロジー」については論文，「XML」についてはポータルや書籍，「VoIP」については製品，ニュースに関する情報が比較的多く存在し，技術キーワードによってバラツキがあることがわかる．これは，技術がある程度確立されているものは製品，今トレンドの技術はニュースや製品，今後発展するであろう技術は論文というように，技術的な成熟度に対応して中心となる情報が異なるためと推測できる．しがたって，技術キーワードに対する情報の一般的な分類を考えた場合，この分類クラスをすべて用意しておくことが，妥当であろう．技術キーワードとこの分類クラスの関係は，技術解説や組織・団体については，Velardi らが定義した Predication（属性項目）に相当し，論文やニュース，教育は，逆方向の Predication（属性項目）の関係と見ることができる．また，細分類については，論文とそのアブストラクトの関係は，PartOf（部分全体関係）あるいは Predication（属性項目），組織・団体と企業の関係は Broader（上位下位関係）を表している．

次に，我々は URL とタイトルから，この分類クラスに Web ページを分類するのためのヒューリスティックルールを作成した．例えば，書籍情報に分類するためのルールは

- URL に「book」あるいは「amazon」を含む

あるいは

- タイトルに「書籍」を含む

となる．このルールに対して，調査対象でのクローズドな簡易評価実験をおこなった．例えば，上記書籍情報に分類するための条件でクエリを作成して検索結果を解析すると，

- 再現率 12/15=0.80
- 適合率 12/12=1.00

という結果が得られる．いくつかの分類クラスについてのヒューリスティックルールと再現率，適合率を表 3a，表 3b に示す．

この結果から，いくつかの分類クラスについては，URL とタイトルからある程度の精度で分類が可能であるが，十分な精度が得られない分類クラスも多いことがわかる．

一方，調査対象の Web ページから，入力した技術キーワードに関連する技術キーワードを抽出することができる．例えば，「オントロジー」に対して「セマンティック Web」や「エージェント」，「VoIP」に対しては「SIP」や「コールセンター」，「QoS」等である．これらの技術キーワード間の関係は，Velardi らの定義した Relatedness（関連概念）に相当する．この関連概念を用いて，前田らの連想概念と同等の機能を持つことができる．

イントラネット情報についても同様に調査，分析をおこなった．その結果，インターネットと大きく

表 3a　ヒューリスティックルール

分類	ヒューリスティックルール
論文	URL={paper} TITLE={論文}
イベント	URL={event,seminar,symposium} TITLE={イベント,セミナー,シンポジウム,研究会}
書籍	URL={book,amazon} TITLE={書籍}
ニュース	URL={news, press*release} TITLE={ニュース,プレスリリース}
製品	URL={product,solution,service} TITLE={製品,商品,サービス}
教育	URL={lecture,syllabus} TITLE={教育,研修,講義}

表 3b　簡易評価結果

分類クラス	再現率	適合率
論文	4/26=0.15	**4/5=0.80**
イベント	3/7=0.47	3/8=0.38
書籍	**12/15=0.80**	**12/12=1.00**
ニュース	15/25=0.6	**15/17=0.88**
製品	26/55=0.47	**26/27=0.96**
教育	6/9=0.67	**6/8=0.75**

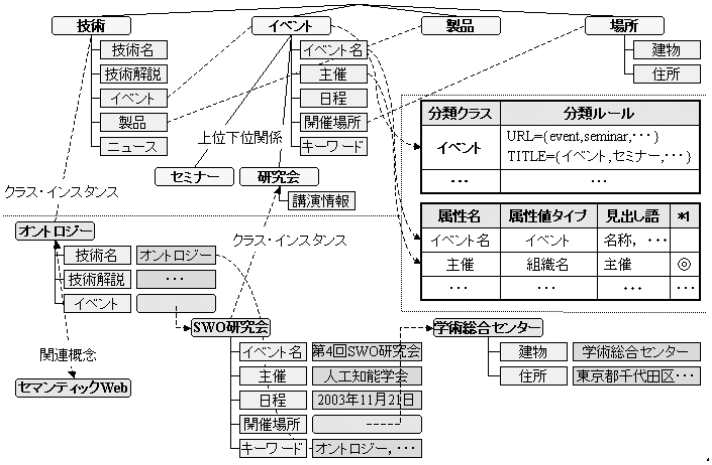

図 2　オントロジー辞書

異なる点は，細分類に関して製品情報の属性項目が多いことが挙げられる．これは，インターネットでは製品の概要のページが検索されることが多いのに対して，イントラネットでは，製品の仕様や顧客への説明用資料，開発管理情報等，より詳細な情報が検索結果として出力されるためと考えられる．

3.　オントロジー辞書

2 章で分析した結果をもとに，収集・抽出した情報を利用者に提示する情報モデルを構築するための知識として，オントロジー辞書の構築を試みた．オントロジー辞書の概念図を図 2 に示す．

本研究でのオントロジー辞書は，クラス階層とインスタンス，および情報分類ルール，属性抽出ルールから構成される．クラス階層は，2 章で説明した分類クラスを最上位とし，各分類クラスは，属性項目および部分全体関係を広義の属性項目として持つ．例えば，＜技術＞クラスの属性項目として＜技術名＞＜技術解説＞＜製品＞＜イベント＞等があり，＜イベント＞クラスは＜イベント名＞＜日程＞＜開催場所＞等の属性項目を持っている．これらの属性項目にはタイプがあり，どのような値が入るかを規定する．例えば，＜イベント＞クラスの＜開催場所＞属性項目は，＜場所＞クラス型と定義している．また，クラスには上位下位関係があり，＜イベント＞クラスの下位クラスとして＜研究会＞や＜セミナー＞がある．下位クラスは上位クラスの属性項目等を継承する．

インスタンスは，クラス階層の概念のうち実体化されたものである．例えば，＜研究会＞クラスにはインスタンスとして「SWO 研究会」があり，＜場所＞クラスには「学術総合センター」がある．

オントロジーの各クラスには，2 章で説明した Web ページを分類するための情報分類ルール，および属性値を抽出するための属性抽出ルールを割り当てている．属性抽出ルールは，その属性値のタイプと見出し語を設定している [大沼 03]．＜イベント＞分類クラスについての属性抽出ルールを表 4 に示す．属性値のタイプは，文字レベルの出現パターンによって判定する人名、組織名、日付等の固有表現のタイプに対応している．また，見出し語を必須とする属性（＊1）を用意している．属性抽出の詳細については，5 章で説明する．

最後に出力結果の情報モデル構築の概略を示す．先に属性項目の概念タイプを定義したが，インスタンスを連結することで情報モデルを構築することができる．例えば，図 2 下で「オントロジー」の属性項目「イベント」には「SWO 研究会」が入り，「SWO 研究会」の属性項目「開催場所」には「学術総合センター」が入ることでインスタンスが関係づけられ，オントロジーに関する情報モデルが構築できる。ただし，現在のオントロジー辞書には，インスタンス間の関係は明示的に記述していない．イントラネット，インターネット情報から抽出することによって，関係づけをおこなっている．

今回，インスタンスとしての技術名，製品名，組織名を用語集やリストから数百抽出し，情報分類ルールは, 300 程度のヒューリスティックルールを作成した．オントロジー辞書は RDF で記述し，リポジトリとしてオランダ Aduna 社（旧 Aidministrator 社）が開発している Sesame [Sesame] を利用して構築した．

4.　システム概要

我々は，構築したオントロジー辞書をもとに，非構造化情報、半構造化情報としてイントラネットお

表 4　属性抽出ルール

属性名	属性値のタイプ	見出し語	*1
イベント情報			
イベント名	イベント	名称	
主催	組織名	主催	◎
開催開始日	日付	日時,日程,開催日時,開催日程	
開催地		場所,会場,開催会場	

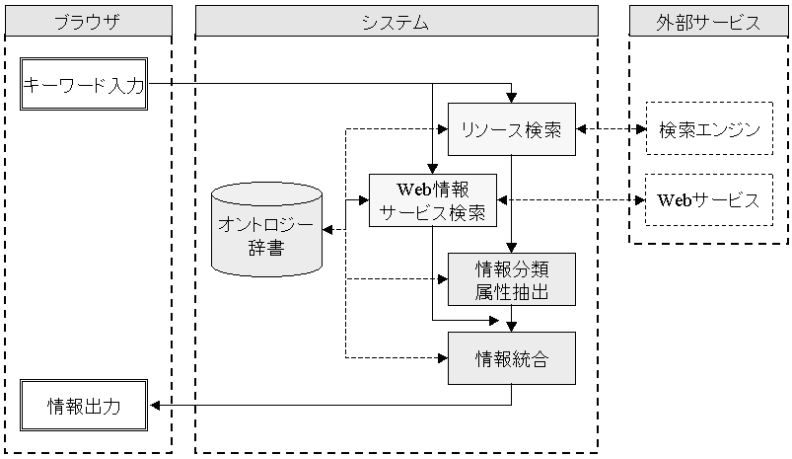

図 3　システム構成図

よびインターネットから情報を収集し，また構造化情報として Web サービスから情報を取得し，利用者に必要な情報を属性項目ごとに情報を抽出し整理して出力するシステムを開発している．システムの全体構成を図 3 に示す．

利用者は Web ブラウザによってシステムにアクセスする．利用者が技術キーワードを入力すると，システムは，辞書，情報抽出・属性抽出，情報統合の各内部モジュール，および検索エンジン，Web サービス等の外部サービスを利用して情報を収集・抽出し，ブラウザに出力する．

システムの各モジュールの処理を以下で説明する．

[Step1] リソース検索処理

イントラネット検索エンジンおよびインターネット検索エンジンを利用して，入力したキーワードに関する情報（URL，タイトル，ページ概要等）を取得する．その際，入力された技術キーワードの意味分類クラスに応じて属性項目をオントロジー辞書から取得し，その属性項目に関する情報を取得するための検索キーワードの拡張をおこなっている．インターネット上の検索は、Google 検索エンジン [Google] を利用している．

[Step2] Web サービス検索

リソース検索と並行して，インターネット上の Web サービスから特定の属性項目に関する情報を取得する．現在のシステムでは，書籍情報として Amazon の Web サービス [Amazon] を利用している．

[Step3] 情報分類・属性抽出処理

リソース検索で取得した URL から HTML ソースを取得し，オントロジー辞書から情報分類および属性抽出のためのルールを取得し，ルールを用いて情報を分類するとともに属性項目に関する情報を抽出する．

[Step4] 情報統合処理

Web サービスで取得した情報と，情報抽出・属性抽出処理で取得した情報を統合し，属性項目ごとに整理して出力する．

各モジュールは，JSP および Java Servlet として実装しており，利用者とのインターフェースモジュールによって呼び出される．画面出力例を図 4 に示す．現在のシステムでは，属性項目が分類クラスである場合は，その分類クラスの属性項目が抽出できるときはそれを出力するようにしている．例えば，技術キーワード「VoIP」に関する属性項目としての書籍情報は，分類クラスとしての書籍情報の属性項目としてのタイトル、著者、価格を出力している．

5.　属性抽出

この章では，システム内で重要なモジュールのひとつである属性抽出処理について詳細に説明し，イベントに関する属性抽出実験について報告する．

5.1.　属性抽出処理

属性抽出処理は，オントロジー辞書から取得した情報分類ルールによって分類クラスが判定された情報から属性項目を抽出する処理である．例えば，書籍の情報が記載されている Web ページからタイトル、著者、価格等の属性項目，あるいはイベント情報が記載されているページからイベント名、日時、場所、主催者等の属性項目を抽出する．処理は，次の手順でおこなう．

[Step.1] 固有表現抽出処理

文書中の固有表現にタグを付与する処理である．文書中の語句に人名タグや組織名タグ等、属性値のタイプを付与する．

[Step.2] 属性決定処理

固有表現抽出処理で抽出した各固有表現が，イベント情報，場所情報等の分類クラス

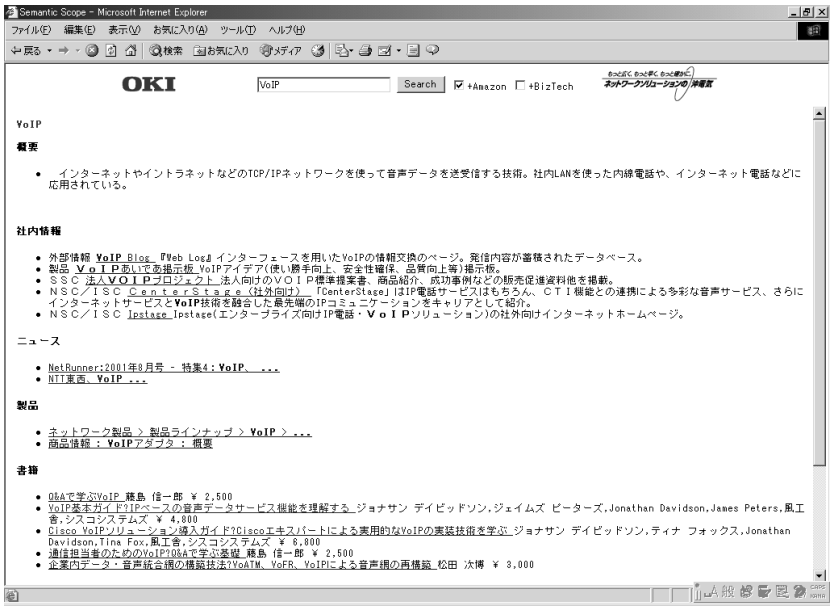

図 4　画面出力例

のどの属性項目に対応するのかを決定する．見出し語が記載されていない場合にも属性項目を判断するために，各固有表現について，次の条件を設定した．

(条件 1-1)　固有表現の前に見出し語が記載されている場合：
　その見出し語が，属性抽出ルールの見出し語に一致し，かつ，その固有表現が属性値のタイプに一致すれば，その属性項目に割り当てる．

(条件 1-2)　固有表現の前に見出し語が記載されている場合：
　その見出し語が，属性抽出ルールの見出し語に一致し，かつ，その属性項目が分類クラスであれば，その分類クラスにある属性項目に優先的に割り当てる．

(条件 2)　固有表現の前に見出し語が記載されていない場合：
　この場合には属性値のタイプが一致する属性項目を見つける．近接した固有表現の属性項目は１つの分類クラスになる傾向を利用して，処理対象の固有表現より前の行に記載されている固有表現の属性項目をチェックし，優先的に同じ分類クラスの属性項目であると判断する．

5.2.　属性抽出実験

　イベントに関する属性抽出の実験を，次の文書集合について行った．
　(文書集合) イベント情報のうち講演・発表（講演情報）を含んでいる文書を人手で収集し属性抽出を実施した．(12 文書)
　属性抽出結果を表 5 に示す．
　講演日，講演者などの講演情報は再現率，適合率ともに高かった．これは，講演情報では＜タイトル＞＜講演者＞＜講演時間＞などが文書中に近接して出現するために，抽出しやすいからと考えられる．また，＜イベント情報.開催地.施設名＞の適合率が低かったのは，この属性が，ある団体への加入組織一覧の部分に誤って対応づけられてしまったからである．見出し語を使わずに属性抽出をする場合には，想定外の記載がされている場合に，誤った対応づけが起こりやすい．また，日時を表す情報は固有表現抽出処理の精度は高かったが，＜イベント情報.開催終了日＞を誤って＜イベント情報.開催開始日＞に対応づけてしまうなどの誤りが見られた．

表 5　属性抽出実験結果

分類クラス	属性項目	再現率	適合率
イベント	イベント名	24/35=0.69	24/36=0.67
	開催開始日	18/29=0.62	18/25=0.72
	開始時間	17/17=1.00	17/18=0.95
	終了時間	11/12=0.92	11/13=0.85
	主催	9/14=0.64	9/13=0.69
イベント.開催地	施設名	26/32=0.81	26/87=0.29
	住所	11/12=0.92	11/13=0.85
講演情報	講演日	24/25=0.96	24/35=0.69
	タイトル	86/131=0.66	86/121=0.71
講演情報.講演者	人名	171/188=0.90	171/185=0.92
	役職	76/86=0.88	76/96=0.79

6. 考察

構築したオントロジー辞書は，主観的な分類クラスおよび属性項目をもとに構築しているが，境界が曖昧なものもある．研修や授業等，継続的な学習に関する情報，例えば「VoIP 基礎」というタイトルの社内研修コースの案内は教育とし，「第 133 回自然言語処理研究会」のような展示会や研究会等はイベントに分類したが，「VoIP 最新技術セミナー」等の教育セミナーはどちらに分類すべきか（あるいは双方に分類するか）といった曖昧性が残っている．この分類クラスについては，統計的にカテゴリ構造を構築する研究［仲川 01］等がおこなわれており，今後は統計的手法を統合することを検討していきたい．

属性項目については，イベントの属性項目のひとつとして開催場所を定義し，場所の属性項目として建物名や住所を定義しているが，住所をさらに郵便番号，都道府県，市区町村と細分化することもでき，どの程度の粒度が実用的かといった問題がある．あまり粒度が粗いと情報間の関係がわかりにくくなり，粒度が細かすぎると人手でメタ情報を付与するコストが高くなるという問題や，情報分類や属性抽出の精度が落ちるという問題がある．一方，設計対象となる人工物（問題解決のドメイン）の機能的側面から設計知識をオントロジーとして体系化した興味深い研究がある［來村 02］．我々が開発しているシステムを業務アプリケーションのコンポーネントとして利用する場合，個々の業務タスクを機能モジュールとして捉え，このような機能的側面を考慮したアプローチが必要になるだろう．また，構文パターンを利用した上位下位関係の抽出［Amann 02］や Web ページから HTML タグを利用した階層関係の抽出［間瀬 03］等，自動抽出の研究も多く見られる．しかし，我々のシステムでは，属性項目の異なる分類クラスに分類できれば充分であり，必要以上に深い階層化を求めるものではない．SUO［Niles 01］のような汎用的なオントロジー，RosettaNet Technical Dictionary［RNTD］のような分野固有のオントロジーを部分的に流用することも考えられるが，以前の調査からアプリケーションによっては適合しない場合があることがわかっている［松平 03］．

情報分類については，2 章で説明したように URL およびタイトルから情報を分類する精度は，分類クラスによってかなり異なる．この問題に対して，システムでは書籍やニュース等の分類精度の高いルールはそのまま使用し，分類精度の低い分類クラスについては，例えば，イベントの場合は「主催」「開催日時」「会場」等，その分類クラスの属性項目に対応する属性抽出ルールとして記述した見出し語が本文中に出現する，といった条件を追加している．しかし，いずれもヒューリスティックなルールに頼っており，スケーラビリティの点で限界がある．定量的な評価をおこない，オントロジー辞書の構築とあわせて統計的手法の統合を検討していく．

情報分類や属性抽出は，非構造化情報あるいは半構造化情報から必要な情報を取得するための手法であるが，一方で Amazon の Web サービスのような構造化情報を提供するサービスが今後普及すると予測される．このような Web サービスの統合に関する研究としては，例えば，音楽情報，チケット情報，天気，地図等の各社が提供する構造化情報を関連づけて横断的な検索を可能にする TAP Semantic Search の研究が挙げられる［TAP, Guha 03］．ただし，Web サービスが提供する構造化情報は，信頼性が高く，使い勝手が良いという反面，情報が一部に限定されてしまうという懸念がある．登録した企業や製品の情報しか提供されなかったり，必要な属性項目が不足していたりする場合である．また，イントラネット上には構造化されていない重要な情報が埋もれている場合も多く，その情報を活用したいという需要は大きい．したがって，構造化情報としての Web サービスと，非構造化情報，半構造化情報から情報分類および属性抽出によって取得した情報との融合は，引き続き重要な課題である．

最後に，情報を知識として蓄える知識ポータルへの応用を議論したい．現在のシステムでは，インスタンス間の関係はオントロジー辞書に記述せず，情報分類および属性抽出によって関連づけている．しかし，網羅性（再現率）や信頼性（適合率）の点で問題があり，利用者が定義あるいは修正，確認した関係をオントロジー辞書に反映するための機能が必要である．この点については，Meadche らがモデル化したオントロジーの学習プロセス，すなわち，Import/Reuse→Extract→Prune→Refine と進み，再び Import/Reuse に戻るプロセスが参考になる［Maedche 01］．我々の開発しているオントロジーを用いた情報収集・抽出システムに，このようなオントロジーを成長させるプロセスを組み込むことによって，次世代の知識ポータルとしての主要な役割を担うことができるであろう．

7. まとめ

IT に関する技術情報について，イントラネット，インターネット上の情報を分析し，分類クラスおよび属性項目、上位下位関係を抽出した．それをオントロジー辞書として構築し，インスタンスとして技術名，製品名，組織名を用意した．

構築したオントロジー辞書をもとに，非構造化情報、半構造化情報としてイントラネットおよびインターネットから情報を収集し，また構造化情報として Web サービスから情報を取得し，利用者に必要な情報を属性項目および関連概念ごとに情報を抽出し整理して出力するシステムを開発している．システムの概要，およびリソース検索，Web サービス検索，情報分類・属性抽出，情報統合の各モジュールについて説明した．

システム内で重要なモジュールのひとつである属性抽出処理について，イベントに関する属性抽出実験をおこない，講演日，講演者などの講演情

報は再現率，適合率ともに高い性能が得られた．

今後は，システム全体の評価，および統計的手法を利用した性能向上，知識ポータルあるいは業務アプリケーションへの利用の検討等をおこなっていく予定である．

参考文献

[Amann 02] Amann, B. et al, Ontology-Based Integration of XML Web Resources, International Semantic Web Conference 2002 (ISWC2002)

[Amazon] Amazon.co.jp, Amazon Web サービス, http://www.amazon.co.jp/exec/obidos/subst/associates/join/webservices.html/

[Davies 02] Davies, J. et al., OntoShare: Using Ontologies for Knowledge Sharing, The Eleventh International WWW Conference (WWW2002) Semantic Web Workshop, 2002

[Google] Google, Google Web APIs, http://www.google.com/apis/

[Guha 03] Guha, R. et al., Semantic Search, The Twelfth International World Wide Web Conference (WWW2003), 2003

[Kietz 00] Kietz, J.U. et al., A Method for Semi-Automatic Ontology Acquisition from a Corporate Intranet, 12th International Conference on Knowledge Engineering and Knowledge Management (EKAW'2000) Workshop "Ontologies and Texts", 2000

[Maedche 01] Maedche, A. and Staab, S., Ontology Learning for the Semantic Web, IEEE Intelligent Systems Vol.16 No.2, 2001

[Niles 01] Niles, I. and Pease, A., Origins of The IEEE Standard Upper Ontology, Seventeenth International Joint Conference on Artificial Intelligence (IJCAI-01), 2001

[Pepper 02] Pepper, S. Garshol, L.M., The XML Papers: Lessons on Applying Topic Maps, XML Conference & Exposition 2002 (XML2002), 2002

[RNTD] RosettaNet Technical Dictionary, http://www.rosettanet.org/RosettaNet/Rooms/DisplayPages/LayoutInitial

[Sesame] Sesame, http://sesame.aidministrator.nl/

[TAP] TAP Semantic Search, http://tap.stanford.edu/

[Velardi 01] Velardi, P. et al., Using Text Processing Techniques to Automatically enrich a Domain Ontology, International Conference on Formal Ontology in Information Systems (FOIS-2001), 2001

[阿辺川 03] 阿辺川他, 機械学習による科学技術論文からの書誌情報の自動抽出, 情報処理学会 情報学基礎／自然言語処理合同研究報告 FI-72/NL-157, 2003

[岩爪 97] 岩爪他, オントロジーに基づく広域ネットワークからの情報収集・分類・統合化, 情報処理学会論文誌 Vol.38 No.3, 1997

[大沼 03] 大沼他, Web コンテンツの分析に基づくオントロジ構築および属性抽出の試み, 情報処理学会 情報学基礎／自然言語処理合同研究報告 FI-72/NL-157, 2003

[來村 02] 來村他, オントロジー工学に基づく機能的知識体系化の枠組み, 人工知能学会論文誌 17 巻 1 号, 2002

[武田 01] 武田, 人工知能におけるオントロジーとその応用, 2001

[仲川 01] 仲川他, 可変なカテゴリ構造を用いた文書検索支援手法, 情報処理学会論文誌 Vol.42 No.10, 2001

[廣田 99] 廣田他, オントロジー主導による情報抽出の検討, 情報処理学会 自然言語処理研究報告 NL-133, 1999

[前田 97] 前田他, 連想構造を用いた情報整理システム, 情報処理学会論文誌 Vol.38 No.3, 1997

[間瀬 03] 間瀬, 山田, Web ページ集合からの階層的知識の構築, 人工知能学会 第 17 回全国大会, 2003

[松平 03] 松平他, Topic Maps を用いた LSI 製品情報システムの開発, 第 2 回情報科学技術フォーラム（FIT2003）, 2003