社団法人 情報処理学会　研究報告
IPSJ SIG Technical Report

2003－FI－ 72　(7)
2003－NL－157　(7)
2003／9／29

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

大沼　宏行　　松平正樹　　渕上正睦　　森田幸伯

近年，Web 上にあるコンテンツの「意味」を取り扱う技術として，セマンティック Web が注目されているが，Web 上のコンテンツに「意味」をつけるメタデータの付与は課題になっている．

本稿では，HTML 文書にメタデータを付与する作業を支援するために，あらかじめ設定された RDF スキーマに基づいて，メタデータを半自動に付与する方法について述べる．すなわち，抽出したい属性を，クラスとその属性という関係で構成し，ある属性が他のクラスを指し示すという RDF Schema のクラス-属性の関係で表現する．そして，個々のコンテンツのどの部分が，それらの属性に対応しているのかを抽出する．Web コンテンツでは，個々の属性を表す見出し語がついていないことが多いため，それを考慮して属性抽出を行う．

講演会などのイベント情報について属性抽出を行い，その結果，適合率で 0.63，再現率で 0.56 が得られた．これらの指標をともに高めることが今後の課題である．

Ontology Construction and Extraction of Template Elements from Web Contents

Hiroyuki OHNUMA　Masaki MATSUDAIRA　Masachika FUCHIGAMI
and Yukihiro MORITA

Recently, semantic web is one of the most interesting topics as the technology for representing the "semantics " of web contents. In order to spread out semantic web, it is necessary to attach meta-data to web contents. However it is difficult to attach meta-data from existing web contents by hand.

We have analyzed web contents in order to construct ontology and developed the tool to support attachment of meta-data.

We can utilize the information extraction technology for attachment of meta-data. Namely, our tool assigns extracted named entities such as person name, place, date and so on, to attributes such as lecturers, meeting places and so on . The relations of attributes are represented as the RDF schema. As the evaluation results, we get a recall score of 0.63 and a precision score of 0.56.

1．はじめに

近年，Web 上のコンテンツの増大に伴い，利用者は必要な情報を得ることが難しくなってきている．そこで，コンテンツの「意味」を取り扱う技術としてセマンティック Web が注目されている[1]. セマンティック Web は, RDF M&S(Resource Description Framework Model and Syntax)等の所定の文法で，HTML 文書の意味を記述したメタデータと，メタデータの意味を記述した RDF Schema 等の各層によって構成され，計算機による Web 上の意味的な情報処理を可能にする．

沖電気工業株式会社研究開発本部
Corporate Research and Development Center,
Oki Electric Industry Co.,Ltd.

セマンティック Web においては，コンテンツに「意味」をつけるメタデータの付与が課題になっている．例えば，過去に蓄積された膨大なコンテンツにメタデータを付与する作業や，一般の情報発信者が RDF Schema を理解して記述する作業は困難である. したがって，HTML やテキスト文書に対してメタデータを半自動的に付与する支援ツールがあると便利である．

そこで本稿では, Web 上のコンテンツについて, RDF Schema に基づいて, メタデータを半自動的に付与する方法について述べる．すなわち，抽出したい属性をクラスとその属性という関係で構成し，ある属性が他のクラスを指し示すという RDF Schema のクラス-属性の関係で表現する．そして，個々のコンテンツのどの部分が，講演会等のイベント情報や製品情報の属性に対応しているのかを抽出する．

属性抽出については，テキストからの情報抽出技術が利用できる．従来の情報抽出技術では，文献[2][3]のように，抽出したい属性間の関係はフラットなテンプレートに表現されることが一般的であった．例えば，文献[2]では, 会告記事情報として記事種別, タイトル, 開催日，開催地，論文締切等の属性を抽出している．しかし，開催地は場所を示す情報なので，開催施設名の他にその郵便番号や住所等の属性が考えられるが，それらの属性のまとまりはフラットな構造では表現できない．

我々は，開催地は場所情報として表現され，場所情報には，施設名，郵便番号，住所等の属性があるといった属性のまとまりを，一つのクラスで表現することにより，属性抽出において，それらの情報が文書中に近接して記載されることを利用する．例えば，イベント情報として，イベント名を抽出している際に，開催地として施設名が見つかると，その近傍の住所や郵便番号は開催地の住所や郵便番号だと判断する．

また，属性抽出において，属性を判断する鍵になる情報として「見出し語」がある．属性のなかには，見出し語が記載されないことが多い属性と，見出し語が記載されることが多い属性がある．例えば，セミナー名や開催日には見出し語が記載されることが少ないが，その主催者には「主催」などの見出し語が記載されることが多い．したがって，属性抽出においては，各属性に付く見出し語を網羅することと，見出し語がなくても属性抽出ができるようにすることが望ましい．

本稿では，2 章で，Web コンテンツを分析し構築したイベント情報や企業情報を表すオントロジと，その属性抽出のための知識を示す．3 章では，見出し語の自動抽出方法について述べる．4 章では，オントロジに基づいた属性抽出方法を述べる. 5 章で見出し語の自動抽出と属性抽出の結果について述べ, 6 章で本稿をまとめる．

2．Web コンテンツの分析

2.1 オントロジ構築

イベント情報や企業情報にどのような属性があるか，数十の Web 上のコンテンツを人手で分析した. その結果，次に示す知見が得られた．

- ほとんどのイベント情報で，イベント名、日時，場所が記載される．
- イベント情報でも，シンポジウムと研究会では異なる属性が現れる．
- 開催場所については，開催施設名のほかに，その住所，郵便番号などの情報が記載される．また，場所情報は，企業の所在地などでも記載される．

以上の点を考慮してオントロジを構築する．図 1 にオントロジの一部を示す．丸で囲まれているオブジェクトがクラス，四角形で囲まれているオブジェクトが属性である．イベント情報には，「イベント名」，「主催」，「開催開始日」，「開催地」等の属性がある．開催地は，その値として場所情報が設定される．場所情報は，「郵便番号」，「住所」，「施設名」の各属性によって構成される．また，場所情報を企業の所在地でも再利用するように，企業情報の「本社所在地」から場所情報にリンクを設定する．

なお，以下の説明では，属性を指し示す際に，例えば「企業情報.本社所在地.住所」というように，ドットで区切った表記をする．

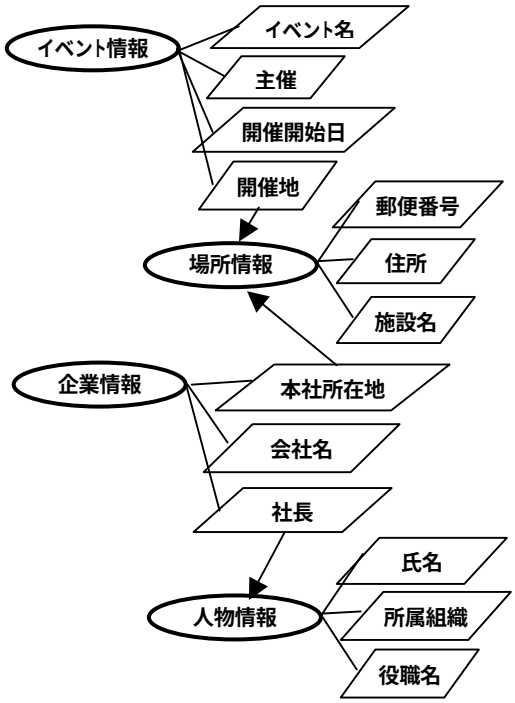

図 1　オントロジ（一部）

2.2 属性抽出のための知識

Web 上のコンテンツから個々の属性を抽出するための知識として，その属性値のタイプと，その属性の見出し語をあらかじめ設定する．図 1 の各属性に対する属性値のタイプと見出し語を表 1 に示す．属性値のタイプは表 2 に示す固有表現のいずれかのタイプに対応させる．例えば，表 1 の「イベント情報.イベント名」の属性値のタイプは，表 2 の固有表現「イベント」になる．また，見出し語がないとその属性に値を設定しないことを示す属性を用意する．例えば，「イベント情報.イベント名」や「イベント情報.開催開始日」などの属性は，見出し語がなくても，その属性に値を設定することがある．一方，「イベント情報.主催」等の属性は，見出し語がないと値を設定しない．

表 1　属性抽出のための知識

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	
場所情報			
郵便番号	郵便番号	〒	
住所	住所		
施設名	施設名, 組織名		
企業情報			
会社名	組織名		
本社所在地		所在地,本社	
社長		社長,代表取締役社長	◎
人物情報			
氏名	人名		
所属組織	組織名, サブ組織名		
役職名	役職		

(*1) その属性には見出し語が必須ならば◎.

表 2　固有表現の例

タイプ	タグ	例
人名	PERSON	ブッシュ，山田
役職	OCCP	大統領，社長，教授
電話番号	TEL	03-NNNN-NNNN
郵便番号	ZIP	NNN-NNNN
住所	ADDRESS	○○市××1-1
組織名	ORG	○○株式会社
サブ組織名	SUBORG	工学部，営業部
日付	DATE	2003年9月29日
時間	TIME	16:00
施設名	INST	○○ビル
イベント	EVT	自然言語セミナー

3．見出し語の自動獲得

見出し語は属性抽出において重要な役割を果たすので，インターネット上に公開されている文書において，個々の属性にどのような見出し語が使われているのかを調べる．HTML 文書で，<table>タグで囲まれる表部分について，図 2 に示す表のうち，いずれに該当するのかを決定し，黒塗部分を見出し語として抽出する．実際には，<table>タグは表を表すだけでなく，レイアウトを整える目的にも利用されるため，次のように見出し語かどうかを判断する．

[Step.1] HTML 文書について，<table></table>で囲まれた部分を検出する．

[Step.2] 検出した部分について，図 2 に示す表のいずれに該当するのかを次の条件で判断する．

(条件 1) <th>タグがある場合に，それが行方向または列方向に並んでいるならば，その方向に見出し語があると判断する．

(条件 2) 図 2 の黒塗部分に，見出し語を示す記号(○, ●,【】，：)等が存在するならば，その方向に見出し語があると判断する．

(条件 3) <th colspan=*n*>タグ(*n*≧2)によって，黒塗部分の各列や行で，項目数が異なっているならば，その方向に見出し語がないと判断する．

(条件 4) 図 2 の黒塗部分の個々の項目の文字数に基づいて，列方向と行方向でスコアリングし，文字数が全体に短い方に見出し語があると判断する．但し，スコアリング結果によって，見出し語がないと判断することもある．

[Step.3] 見出し語を出力する．

(形式 1) 最左列に縦方向に見出し語

イベント名	情報シンポジウム
開催日	2003年9月29日
会場	○○大学

(形式 2) 1 行目に横方向に見出し語

イベント名	開催日	会場
情報シンポジウム	2003年9月29日	○○大学
第1回例会	2003年9月29日	△△大学

(形式 3) 縦横方向ともに見出し語

	男性	女性
00年度	10人	8人
01年度	13人	10人

図 2　表構造の種類

4．属性抽出方法

システムは属性抽出を次の手順で実行する．

[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．この処理では福本らの固有表現抽出技術[4]を利用する．例えば，文書中に「山田太郎　（○○大学）」と記述があれば「<PERSON>山田太郎</PRESON>　（< ORG >○○大学</ ORG >）」のように人名タグや組織名タグを付与する．タグで囲まれた範囲が，対応する固有表現である．図 3 に固有表現抽出処理後の文書の例を示す．黒塗部分が，固有表現抽出処理によって，タグ付

けされた文字列である．

[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，場所情報等のうち，どの属性に対応するのかを決定する．各固有表現の前に見出し語があれば，それを利用して，対象となる属性を判別することは比較的容易である．しかし，図 3 のように，見出し語が必ずしも記載されているわけではない．見出し語が記載されていない場合にも属性を判断するために，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，表 1 の「見出し語」項目に一致し，かつ，その固有表現が「属性値のタイプ」項目に一致すれば，その属性に割り当てる．例えば，「主催：<ORG>◎◎学会</ORG>」と記載されていれば，「イベント情報.主催」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，表 1 の「見出し語」項目に一致し，かつ，図 1 において，その属性から他のクラスにリンクが設定されていれば，そのクラスにある属性に優先的に割り当てる．

例えば，「会場：<ORG>△△大学</ORG><INST>第 3 講義室</INST>」と記載されていれば，見出し語は「会場」なので，属性は「イベント情報.開催地」になる．さらに，「イベント情報.開催地」は場所情報である．「<ORG>△△大学</ORG>」は組織名であり，場所情報のうち「属性値のタイプ」項目として組織名があるのは，「イベント情報.開催地.施設名」であるので，この属性になる．

(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

例えば，図 3 の 11 行目の「<ZIP>123-4567</ZIP>」には見出し語がついていない．この場合には「属性値のタイプ」項目が一致する属性を見つける．「場所情報.郵便番号」が一致するが，場所情報は，表 1 において，「イベント情報.開催地」と「企業情報.本社所在地」からリンクされている．したがって，属性として，「イベント情報.開催地.郵便番号」（イベントの開催地の郵便番号）と「企業情報.本社所在地.郵便番号」（企業の本社の郵便番号）が考えられる．

そこで，近接した固有表現の属性は 1 つのクラスになる傾向を利用して，処理対象の固有表現の上の行に記載されている固有表現の属性をチェックする．「<ZIP>123-4567</ZIP>」の上の行の「<ORG>△△大学</ORG>」は「イベント情報.開催地.施設名」であるので，「イベント情報.開催地.郵便番号」を優先的に，この固有表現に対応する属性であると判断する．（「<ORG>△△大学</ORG>」の属性を「イベント情報.開催地．施設名」であると判断するためには，さらにその上の行の「<EVT>第 1 回○○セミナー</EVT>」の属性が，「イベント情報.イベント名」であることを利用する．）

結果として，図 3 の文書の属性抽出の結果は，図 4 のようになる．

```
1  <table border="0"><tr> <td>
2  <EVT>第 1 回○○セミナー</EVT>
3  </td><td></td></tr>
4  <tr><td>
5  <DATE>2003 年 7 月 23 日</DATE>
6  </td><td>
7  <TIME>14：00-18：30</TIME>
8  </td></tr><tr><td>
9  <ORG>△△大学</ORG><INST>第 3 講義室</INST>
10 <br>
11 〒 <ZIP>123-4567</ZIP><ADDRESS> ○○市××町
   1-1</ADDRESS>
12 </td><td></td></tr></table>
```

図 3 HTML 文書の例

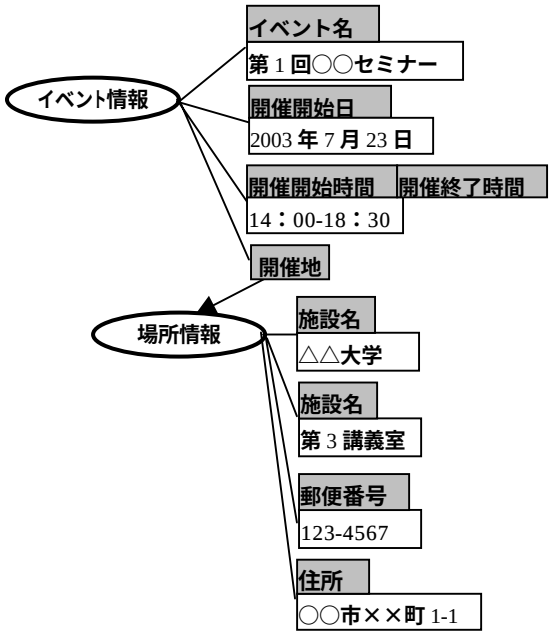

図 4 属性抽出結果

5.実験

5.1 見出し語の自動獲得

抽出したい属性を含みそうな文書を見つけるために，「会社概要」「シンポジウム」「製品紹介」をキーワードにしてインターネット検索エンジンを用いて検索した．検索結果文書について，3 章に示す方法を用いて，見出し語の自動獲得を実施した．これらの単語は，それぞれ「企業情報」「イベント情報」「製品情報」の

各属性の見出し語を得ることを期待して選択した．

表 3 に，各検索キーワードごとに獲得できた単語のうち，目視にて見出し語として適切だったものを示す．

結果として，「会社概要」では，少数の文書で比較的適切な見出し語を得ることができた．一方，「シンポジウム」や「製品紹介」では，見出し語と言えない文字列が数多く抽出され，見出し語を得るのが困難だった．特に「製品紹介」では，製品名や製品カテゴリが数多く抽出されてしまった．

しかし，例えば「事例紹介」という見出し語と類似した意味として，「導入事例」があることに気づくことができる等の利点があった．

見出し語の自動獲得では，単にキーワードで検索された文書の<table>タグから抽出するだけでは不十分であることが確認できた．<table>タグ内で，所望のクラスの属性を確かに含んでいるのかを判断する処理が必要になる．

また，個々の見出し語がどのような属性の見出し語に使われるかを判断するのは今後の課題である．

表 3　見出し語の自動獲得結果

検索キーワード	文書数	表を含む文書数	目視にて確認できた見出し語（一部）
会社概要	98	65	所在地，最新ニュース，本社所在地，採用情報，本社所在地　及び連絡先，商号，事業範囲，総店舗数，売上高，代表取締役社長
シンポジウム	988	573	会場，プログラム，参加費用，使用言語
製品紹介	300	206	事例紹介，導入事例，機種，価格(税別)

5.2 属性抽出

イベントに関する属性抽出の実験を，次の 2 種類の文書集合について行った．

(文書集合 1)　イベント情報のうち講演内容（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)

(文書集合 2) イベント情報を含みそうな HTML 文書を見つけるために，「イベント日程」をキーワードにしてインターネット検索エンジンを用いて検索した．検索結果文書について，セミナー等のイベント情報を含む文書だけを目視で選別した．100 文書中 9 文書がマッチした. その 9 文書に対して属性抽出を実施した．

(文書集合 1)，(文書集合 2)の属性抽出結果を表 4 に示す．各項目の計算式は次の通りである．

$$\text{適合率} = \frac{\text{正しく属性抽出できた固有表現数}}{\text{抽出した固有表現数}}$$

$$\text{再現率} = \frac{\text{正しく属性抽出できた固有表現数}}{\text{属性が割り当てられるべき固有表現数}}$$

$$\text{F 値} = \frac{2\times(\text{適合率})\times(\text{再現率})}{(\text{適合率}) + (\text{再現率})}$$

文書集合 1 について，講演日，講演者などの講演情報は適合率，再現率ともに高かった．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．

文書集合 2 については，固有表現抽出処理の段階で抽出できなかった場合が多かった．例えば，「イベント情報.イベント名」については，適合率は高かったが，再現率は低かった．これは，イベント名が，英語で記載されていたりして固有表現抽出処理で抽出できなかったためである．また，「イベント情報.主催」についても，主催者に対応する固有表現が，組織名として抽出できなかった．

文書集合 2 では，適合率が 0.63，再現率が 0.56 であり，2 つの指標をともに高めることが今後の課題である．

そのための方法として，次のことが考えられる．

1．固有表現抽出の精度向上

2．固有表現抽出ができなかった場合でも，見出し語を利用して属性抽出処理を行うようにする．例えば，「イベント情報.主催」は，主催者を表す見出し語が必ず記載されるので，その後に続く単語は，たとえ固有表現抽出に失敗しても主催者であると対応づけが可能だと考えられる．

また，今回の実験では，イベント情報を含みそうな文書を目視によって確認したが，なんらかの方法で自動分類することが必要である．

表 4　属性抽出の結果

オブジェクト名	属性名	文書集合1			文書集合2		
		適合率	再現率	F値	適合率	再現率	F値
イベント情報	イベント名	0.67 (24/36)	0.69 (24/35)	0.68	0.85 (22/26)	0.46 (22/48)	0.59
	開催開始日	0.72 (18/25)	0.62 (18/29)	0.67	0.63 (31/49)	0.70 (31/44)	0.67
	開始時間	0.95 (17/18)	1.00 (17/17)	0.97	0.43 (9/21)	0.9 (9/10)	0.58
	終了時間	0.85 (11/13)	0.92 (11/14)	0.88	- (0/0)	0.00 (0/10)	-
	主催	0.69 (9/13)	0.64 (9/14)	0.66	- (0/0)	0.00 (0/6)	-
イベント情報.開催地	施設名	0.29 (26/87)	0.81 (26/32)	0.44	0.66 (21/32)	0.48 (21/44)	0.55
	住所	0.85 (11/13)	0.92 (11/12)	0.88	1.00 (3/3)	1.00 (3/3)	1.00
講演情報	講演日	0.69 (24/35)	0.96 (24/25)	0.8	0.65 (11/17)	1.00 (11/11)	0.79
	タイトル	0.71 (86/121)	0.66 (86/131)	0.68	0.43 (3/7)	1.00 (3/3)	0.60
講演情報.講演者	人名	0.92 (171/185)	0.90 (171/188)	0.91	1.00 (1/1)	1.00 (1/1)	1.00
	役職	0.79 (76/96)	0.88 (76/86)	0.83	1.00 (1/1)	1.00 (1/1)	1.00
計		0.78	0.82	0.80	0.63	0.56	0.59

6．おわりに

本稿では，メタデータの作成を支援するための属性抽出方法について述べた．本属性抽出方法では，抽出したい属性をクラスとその属性という関係で構成し，ある属性が，他のクラスを指し示すという RDF Schema のクラス-属性の関係で表現した．属性抽出結果として，適合率が 0.63，再現率が 0.56 であり，2 つの指標をともに高めることが今後の課題である．

参　考　文　献

[1] 萩野達也,神原顕文,清水昇,豊内順一,細見格,津田宏,白石展久, 韋慶傑: セマンティック Web とは, 情報処理学会誌, Vol.43, No.7, pp.709-717 (2002).

[2] 佐藤円,佐藤理史,篠田陽一: 電子ニュースのダイジェスト自動生成,情報処理学会論文誌,Vol.36,No.10, pp.2371-2379 (1995).

[3] 岩爪道昭,白神謙吾,畑谷和右,武田英明,西田豊明：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,情報処理学会誌,Vol.38,No.3, pp.606-615 (1997).

[4] 福本淳一,下畑光夫,桝井文人: 固有名詞抽出における日本語と英語の比較, 情処研報,98-NL-126, pp.107-114 (1998).