地方自治体ウェブページからの地域情報獲得

藤田 茂 †, 今野 将 ††

† 千葉工業大学情報科学部情報工学科 †† 千葉工業大学工学部電気電子情報工学科

地方自治体のウェブページは，地域住民のための情報が数多く掲載されており，その効果的な利用が望まれている．しかしながら，地方自治体のウェブページのデザインには自治体固有の表記に従っており，標準的なウェブページ構成が存在しない．そのため，その自治体について多くのことを知りたい転入者のように，もっとも情報を求めている利用者にとって使い難いものとなっている．また，ウェブページ自体が人が閲覧することを前提としているために，他のウェブサービスと組み合わせて，情報を利用することが困難である．我々は，地方自治体から機械処理可能な情報を抽出することを目的として，その手法を検討しいくつかの評価実験を行ったのでそれを報告する．

Getting Inter-Organizational Information from Multiple Municipality 's Web Site

Shigeru Fujita † Susumu Konno ††

† Department of Computer Science,
Faculty of Information and Computer Science, Chiba Institute of Technology

†† Department of Electrical,Electronics and Computer Engineering,
Faculty of Engineering, Chiba Institute of Technology

There is knowledge on municipality 's web site to serve information for inhabitants. But, these web sites are much different design to use web pages on each other. Therefore new comer has a hard time to get right knowledge from the municipality 's web site. Of course, some method to retrieve knowledge from the web site are proposed and used as useful. But, some web sites are not applied exiting method, because web site informs knowledge by multiple web pages. We discussed knowledge and design of municipality 's web site and classified it. We propose a new method for get inter-organizational knowledge from multiple web pages in this paper. The proposed method is evaluated for twenty regions in Kantou.

1 はじめに

これまで多くのウェブコンテンツ，特にhtmlで記述された情報を対象として，情報獲得手法が提案されてきている[1, 2, 3, 4, 5, 6]．我々は地方自治体のウェブページを対象として，これまでの手法では困難であった，ページ遷移を前提として情報提供を行うウェブサイトからの情報獲得手法を提案し，動作を確認する実験を行ったので，これを報告する．

2 地域情報ウェブページと既存の情報取得手法

2.1 地域情報

地方自治体がウェブを介して公開している情報は，多岐に渡り有用なものが多いが，機械処理を前提としていないために，二次的な利用が困難である．

たとえば，著者らが属する研究グループで開発を進めている" 子供や高齢者の見守り支援システム[7] " など，人々の生活を対象としたサービスを提供する場合，イベント開催，道路工事，犯罪発生率，そして交通事故などの情報が必要である．

これらの情報は，自治体毎にウェブページのデザインが異なるため，それぞれの書式や記載されている内容などが異なり，既存のウェブを対象とした手法を全てのウェブページに適用することは困難である．

2.2 地域情報ウェブページの構成

ウェブページのデザイン上の構成は，その機械処理による利便性よりも，むしろ見た目分かりやすさやインパクトを重視するために，多種多様な形態を取っている．本稿では，地方自治体により地域情報が提供されるウェブページを対象として，ウェブページのテンプレート形式を，以下の3形

式に分類した．

シングル・インスタンス型： ウェブページ内にイベントや工事について，一件だけの地域情報が掲載されている形態

マルチプル・インスタンス型： ウェブページ内に同一のカテゴリについて，複数件の地域情報が掲載されている形態

ウェブページ移動型： マルチプル・インスタンス型において，それぞれの地域情報に対応するハイパーリンクが設置され，ハイパーリンク先にシングル・インスタンス型で詳細な地域情報が掲載される形態

これらのうち，先に述べたシングル・インスタンス型とマルチプル・インスタンス型については，野口らの研究[1]で述べられているテンプレート形式であり，イベントや工事の名称や開始日，終了日など，一つの事柄を表す情報の単位を“インスタンス”とし，ウェブページ中に記載されているインスタンスの数で分類すると，この2種類に分けられる．しかし，地域情報を取り扱うウェブページを観察すると，シングル・インスタンス型により地域情報を記載しているウェブページは，必ずその上位ウェブページにマルチプル・インスタンス型の形式で，それぞれのシングル・インスタンス型のウェブページへのハイパーリンクが設置されており，ウェブページ移動型と考えられる．そのため，地域情報を扱うことを考慮した場合，新たに定義したウェブページ移動型を，一つの種類として扱うことが必要となる．

2.3 ウェブラッパーによる情報取得

構造化文書として利用できないHTMLから，HTMLのタグに注目し，目的とする情報を含むタグを指定することで，取得したい情報を見つけ出す，WebWrapperの研究がある[2]．しかしながら，現状のウェブサイトでは，作成者によって利用されるHTMLタグの種類や用途が異なる．そのため，各ウェブページに合わせたラッパーを生成する必要があり，情報を取得するウェブサイトごとや，ウェブサイトの構成が変わるたびに，新たにラッパーを生成するのは現実的でない．また，ラッパーを自動生成する研究もあるが，ハイパーリンク先へ移動することで，詳細な情報が得られるような構造を持つウェブサイトが存在するため，そのようなウェブサイトでは，ウェブラッパーを用いることができない．

2.4 文字列の類似度に基づく情報取得

ウェブページの情報を取得する研究として，梅原らの研究[3]がある．この手法では，あらかじめ事例として文書を与え，その文書のテキストブロックに意味を付加し，その文書のテキストブロックと取得対象の文書中のテキストブロックとの類似度を算出することで，語の意味的な把握により，情報取得を行う手法を提案している．

しかし，1つの文章に複数の情報が記述された場合と，単語の類似度と情報の意味に関連が無い場合には，十分な情報の取得が困難になる．

3 地域情報取得手法の提案

3.1 地域情報取得の概要

本稿では，既存の情報取得手法では適用困難なウェブページからの地域情報の取得を行うために，既存手法に，属性語利用とウェブページ移動処理を加えた地域情報取得手法を提案する．

属性語は，情報となる語に対して付加されて意味を包含する特徴ある語である．地方自治体のウェブページは，様々な人が閲覧することが考慮されているため，属性語の数が限定されて用いられている．そのため，あらかじめ属性語とその意味を事例として与えておくことで，意味の特定を行うことが可能となる．また，属性語が付加されていない場合は，梅原らの手法[8]を利用することで対応することが可能となる．

ウェブページ移動型によって情報提供がなされているウェブサイトに対しては，新たにウェブページの移動処理を行うことで，単一ウェブページに限定せずに，情報の記述された範囲を特定し機械的な情報獲得を行うことが可能になる．

地域情報の取得を行うために，提案した手法に加えて前処理となる部分を含めた地域情報取得の流れを図1に示す．

図1で示した，HTMLは処理対象としてのHTML文書であり，Iterative Structure Analysisは繰り返し構造解析を示し，Part of Documentsは，HTML文書から切り出された部分を示す．これにCase base(事例)を用いて，出力として日付，タイトル，場所の情報を獲得する．

マルチプル・インスタンス型の文書では，単に複数の情報が記述されているだけでなく，記述される情報は，同一の形式を保った状態で提供されて

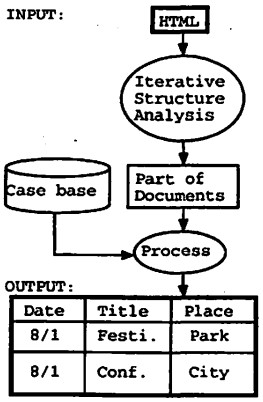

図 1　地域情報取得手法の流れ

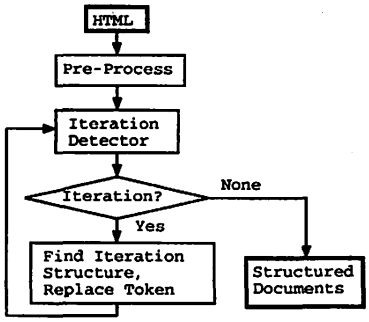

図 2　繰り返し構造解析のフローチャート

いると推定できる．同一の形式であるために，同じ HTML タグを用いて文書を構成するので，HTML タグの組み合わせが繰り返し構造となっている箇所を，各インスタンスの範囲を決定することができる．この結果，一組の情報が記述された範囲を絞り込むことができる．

この処理を，地域情報取得の前処理として行い，シングル・インスタンス型と見なせる範囲に分割する．そして，得られたシングル・インスタンス型の文章，それぞれに対して，地域情報の取得を行うことで，目的の情報取得を行う．

このとき，地域情報取得の際に用いる事例情報として，3.3 にて述べる属性語とその対象となる属性値を，あらかじめ与える．

3.2　繰り返し構造の解析

HTML 文書の繰り返し構造の解析には，HTML 文書をボトムアップに構造化を行う南野らの手法 [4] を用いる．南野らの手法は，最もプリミティブな繰り返し構造の検出を繰り返すことで，HTML 文書の構造化を行う．処理の過程として，まず，現時点で最もプリミティブな繰り返し構造を検出し，検出された繰り返し構造が存在する部分を 1 つのトークンとして置き換える．1 つのトークンに置き換えることで，新たなプリミティブな繰り返し構造は，より大域的な構造となる．

この際，繰り返し回数のみが異なるだけで，繰り返しを構成する基本単位が同じ繰り返し構造は，同一のトークンで置き換える．これにより，繰り返し回数は異なるが，繰り返しを構成する基本単位が同じであるため，構造の解析処理では同一視される．この HTML 文書の構造化手法のフローチャートを図 2 に示す．

まず，入力された HTML 文書に対して前処理 (Pre-Process) を行い，繰り返し構造化を行うために不要な箇所を除去することや，開始タグと終了タグの対応関係の修正を行う．その後，前処理を終えた文書に対して，最もプリミティブな繰り返し構造の検出処理 (Iteration Detector) を行う．検出された繰り返し構造について，トークンの置き換え (Find Iteration Structure, Replace Token) を行い，さらに大域的な繰り返し構造の検出を行う．検出された繰り返し構造をトークンに置き換えることで，繰り返し構造が検出されなくなった (Ieration?− >None) とき，HTML 文書の構造化 (Structured Documents) が完了となる．

図 2 の手法では，文書の構造化のみを目的としているため，地域情報を推定することはできない．中野らの手法 [5] では，HTML タグ入れ子の回数である深さを基準として重要箇所を判定している．しかし，コンテンツ部より HTML タグが深くなっている場合には，適切に重要な箇所を抽出できない場合がある．これに対して，HTML タグの繰り返し構造が，文書中において繰り返し範囲が最長となる繰り返し構造を探し，その範囲から情報を取得する手法 [6] がある．本論文ではこの二つの手法を組み合わせ，重要箇所を判定する．

3.3　属性語と属性値の取得

3.3.1　属性語と属性値の定義

情報の取得を行う際に用いる属性語は，一覧や箇条書きをする箇所では，“○○○文化ホール” や “○○会館” などの語と合わせて記述し，一般化して情報の意味を表す “会場” や “開催地” などの語を用いる．これを属性語と呼ぶ．また，属性語と主に連続して記述され，属性語と共に記述されることで情報単位となり，“○○○文化ホール” や “○○会館” などの情報自体を表す語を属性値と呼ぶ．

表 1　属性語の取得に用いた HTML タグと修飾文字

HTML タグ:	TD,TH,LI,DT,DD,B,STRONG,FONT,TINY,EM,TT
接頭修飾:	*, ＊, ●, ○, ■, □, ・, ◆, ◇, ★, ☆, ○
括弧類:	(-), [-], <->, (-), [-], [-], <-> , (-)
接尾修飾:	:, ：, ／, /, =

この際，属性語として満たす条件は，吉永らの手法[9]を用いて属性語を発見する．

以下に，属性語の定義を述べる．まず，文書において，表1に示したHTMLタグ内において，示されているいずれかの文字修飾がされている部分文字列，もしくは，文字修飾されていない場合は，そのタグ内の文字列全体を属性語の候補とする．

この条件を満たす属性語候補のうち，さらに以下の条件を全て満たす文字列を属性語とする．

- 空白文字は入らない連続した文字列である
- 接頭修飾は，行頭や文字列の先頭に存在，もしくは直前が空白文字である
- TD，THタグ内で修飾文字を持たない属性語候補の場合，一行目と一列目のセルに対応する箇所である
- 文字修飾がされてない場合，事例情報として与えられた属性語の事例と一致する

接頭修飾の制約において，「行頭や文字列の先頭に存在，もしくは直前が空白文字」を満たさない場合，“詳細・申し込み”のような文字列が考えられる．この場合，接頭修飾文字（ここでは“・”）が，前方の“詳細”の文字列と連続しており，属性語の接頭修飾としては不適切である．しかし，接尾修飾では，“会場：○○○文化ホール”のように，接尾修飾文字より後方にも文字列は連続する場合があるため，接頭修飾のような制約は設けない．また，TD，THタグの制約に関しては，表中で属性語が記述されるのは，主に一行目と一列目のセルであるという観察[10]から，それらのセルからのみ属性語とする制約を設けた．

このようにして定義した属性語のうち文字修飾された属性語の取得手順を図3に示す．

まず属性語の取得には，3.2にて述べた繰り返し構造を取得する際に利用した“Text”トークンのよ

```
Function: GetAttributeWord
Input: TextBlock
BeginProcedure:
    (Prefix,Brackets,Suffix) :=
        DivisionWithWhiteSpace(TextBlock);
    (start,stop) :=
        AttributeWordPosition(
            Prefix,Brackets,Suffix);
    Word :=
        getWord(start,stop);
EndProcedure:
```

図 3　属性語の取得処理手順

うに，HTMLタグを区切りとしたテキストブロックのみに対して属性語の取得処理(DivisionWithWhiteSpace)を行う．次に，表1に示した文字修飾がされているテキストの推定(AttributeWordPosition)を行い，さらに，属性語の制約を満たす文字列を推定(getWord)する．

3.4　語と属性語の類似度による情報取得

事例として与えた属性語と属性値が，発見された属性語と属性値のセットに類似している場合，そのセットは，事例と同じ意味を持っていると考える．ウェブページ内を3.2にて限定した範囲において，HTMLタグを区切りとして分割したテキストブロックについて，梅原らの手法を用いて各事例情報との類似度 $Sim(V_i, V_j)$ を算出する．その後，それぞれのテキストブロックについて，属性語の一致比較を行うために，属性語の発見を行う．本手法で取り入れた属性語の発見には，吉永ら[9]の手法を用いて，属性語獲得に用いた文字修飾されている語を属性語としている．属性語を探す範囲として，主に，同テキストブロック内，もしくは直前のテキストブロックにあること仮定し，その範囲での属性語の発見を行う．また，表については，一列目と一行目に属性語が記述される場合も多い[9]ことから，それらからも属性語の取得を試行する．また，テキストブロック中に，属性語の文字修飾がされている箇所がない場合，そのテキストブロック全体を属性語候補として扱い，事例として与えられた属性語と比較して一致した場合，属性語としている．発見された属性語を，その直後に現れる語を属性値とし，属性値の類似度を算出することで，情報の意味的な把握を行う．

求めた類似度に総合的な意味の類似度を求めるため，事例との属性語の一致についての比較を行

う．属性語を考慮した総合的類似度 $S(T_i, T_j)$ を以下に示す．

$$S(T_i, T_j) = \begin{cases} Sim(V_i, V_j) + \alpha & if\ \text{属性語が一致} \\ Sim(V_i, V_j) & \text{その他} \end{cases}$$

V_i，V_j は T_i，T_j の項ベクトル，αは，属性語が一致した場合にのみ与えられる定数である．この $S(T_i, T_j)$ が大きい値を示すほど，2つの情報は意味的に類似していると考える．また，本稿では $\alpha = 2.0$ とした．

3.5 情報の補完

それぞれの情報について，類似度により取得したことで，イベントの名称や開催地が取得できていることが期待される．しかし，開催日時に関して，それぞれに絞り込んだ対象範囲には，年や月が省略されている場合がある．そこで，共通の情報として書かれていると判断し，ウェブページ全体から年と月の情報を得て，年と月の情報を補完している．

4 実験

4.1 実験の概要

本稿で提案した地域情報取得手法において，実際の地方自治体ウェブページを対象としたときの有効性を確認するために，提案手法を実装したシステムを作成し，実験を行った．ただし，今回の実験は，HTMLタグを使ってウェブページ上に情報発信をしている20の市区を対象に行われている．これは，実験対象としなかった市区のウェブページでは，例えば〈br〉改行タグのみのように，ごく限定されたHTMLタグのみを使ってページが構成されていたり，PDFでの情報発信が中心であって，本手法が適応できないためである．このため，実験に用いられたデータには偏りがある．

実験のためにあらかじめ与えた24種類の事例情報の例を表2に示す．

評価実験の対象とするウェブページは，イベント開催情報についてのウェブページを16件，道路工事情報についてのウェブページを4件選び，合計20件のウェブページについて行った．

提案手法の評価基準として，本稿では“正確度(%)”[3]を

$$\text{正確度} = \frac{a}{b} \times 100 \qquad (1)$$

を用いた．対象とするウェブページには，複数のインスタンスが存在しているため，aは，入力した地域情報ウェブページに対して，取得した全ての情報（名称や開始日などそれぞれを一つの情報として扱う）において，正しいテキストデータを持つ情報についての総数である．一方で bは，入力した地域情報が記述されたウェブページ中における，全ての情報についての総数である．

表2 実験に用いた事例情報の一部

情報種別	属性値	属性語
イベント名称	小学生音遊び教室	行事名
イベント場所	青少年会館	場所
工事場所	船橋市宮本町 2 丁目	工事場所
終了日	2007 年 12 月 2 日	～
月補完	11 月のイベント情報	(属性語なし)
年補完	平成 19 年 12 月のイベント情報	(属性語なし)

本稿においては，名称や場所を一組としたインスタンス一つの中に，取得すべき情報の種類数が1つの地域情報につき4情報であることが大部分であるため，マルチプル・インスタンス型のウェブページに含まれるインスタンス数と，ウェブページ中における全ての情報総数の関係は，

$$\text{全情報総数} = \text{インスタンス数} \times \text{取得情報種類数} \qquad (2)$$

に近い値となる．この際，取得する情報種類数は，名称，場所，開始日，終了日の4種類であるため，ウェブページ中に出現するインスタンス数が15個であれば，全情報総数は，インスタンス数の4倍の60件となる．しかし，地域情報の中には，一部の情報が記載されず，人が閲覧しても得られない情報がある．その場合，取得すべき情報とはできないため，情報総数には含めず，全ての取得すべき情報のうち，どれだけの情報を取得できたかを評価の基準とする．

4.2 実験結果

地域情報が記述されているウェブページを入力としたときの地域情報取得結果を表3に示す．

表3について，実全情報総数は，正確度を求める際に定義した bの値であり，ウェブページ中に存在している全ての情報が総数が表されている．取得情報総数については，bと同様に，正確度を求めるために定義した aの値であり，評価実験において，正しい情報を取得できた総数である．正確度は，前項で定義した式1を用いて計算した値である．

表 3　実験結果 (一部)

実験番号	対象情報	全情報総数 (b)	取得情報総数 (a)	正確度 [%]
1	習志野市イベント	56	45	80
2	船橋市イベント	74	67	91
6	江戸川区イベント	52	39	75
10	台東区イベント	60	48	80
17	船橋市道路工事	56	56	100

また，各情報の種類ごとにおける正確度にまとめたものを表 4 に示す.

表 4　情報種類ごとの正確度 (一部)

実験番号	実験対象	正確度 [%]			
		開始日	終了日	名称	場所 (会場)
1	習志野市イベント	79	79	86	79
2	船橋市イベント	100	80	89	94
6	江戸川区イベント	100	100	46	54
10	台東区イベント	100	100	100	20
17	船橋市道路工事	100	100	100	100

5　おわりに

利用者のページ遷移を前提として情報提供を行うウェブサイトを対象として，機械的処理によって，情報抽出を行う手法を提案した．提案手法は関東の 20 市区のウェブサイトを対象として評価実験を行った．これらの対象は，提案手法に適したウェブであり，この報告で触れられていない幾つかの自治体のウェブページからは有意な情報を得ることが出来なかった．これは，HTML タグの使われ方と，表現される情報に関連が規則的に発生しないウェブページで発生する．例えば，表を構成するタグである (td) の中に複数の情報がかかれている場合などである．今後，タグに依存しない情報抽出の方式と組み合わせて目的の情報を得ることを目指している．

また，本稿で提案した手法では，依然として地域固有の情報を人手によって保守する必要があり，目的とした機械処理を完全には達成していてないという課題が残されている．

謝辞

本研究は，渡邊悠介 (現: NTT コムウェア)，“共生コンピューティングのための地域情報取得手法”，千葉工業大学修士論文,2008 年 3 月を基に開発が進められた.

参考文献

1) 野口龍太郎, 山田泰寛, 池田大輔, 廣川佐千男. 頻度情報を用いた web 文書群からのテンプレート抽出. *DEWS2004*, 2004.

2) 村上義継, 坂本比呂志, 有村博紀, 有川節夫. Html からのテキストの自動切り出しアルゴリズムと実装. 情報処理学会論文誌. 数理モデル化と応用, Vol. 42, No. 14, pp. 39–49, 20011215.

3) 梅原雅之, 岩沼宏治, 鍋島英知. 事例に基づくシリーズ型 html 文書から xml 文書への半自動変換. 人工知能学会論文誌, Vol. 17, No. 5, pp. 408–416, 2002.

4) 南野朋之, 齋藤豪, 奥村学. 繰返し構造に基づいた web ページの構造化. 情報処理学会論文誌, Vol. 45, No. 9, pp. 2157–2167, 9 2004.

5) 中野雄介, 山登庸次, 武本充治, 須永宏. Web アプリケーションの結果ページからの結果部分抽出法. *DEWS2007*, 2007.

6) Chia-Hui Chang and Shao-Chen Lui. Iepad: information extraction based on pattern discovery. *WWW '01: Proceedings of the 10th international conference on World Wide Web*, pp. 681–688, 2001.

7) Takuo Suganuma, Hideyuki Takahashi, and Norio Shiratori. Agent-based middleware for advanced ubiquitous communication services based on symbiotic computing. *7th IEEE Int. Conf. on Cognitive Informatics*, pp. 300–309, 2008.

8) 梅原雅之, 岩沼宏治, 鍋島英知. 事例に基づくシリーズ型 html 文書の意味論理構造の自動認識. 人工知能学会論文誌, Vol. 17, No. 6, pp. 690–698, 2002.

9) 吉永直樹, 鳥澤健太郎. Web からの属性情報記述ページの発見. 人工知能学会論文誌, Vol. 21, pp. 493–501, 2006.

10) M YOSHIDA. Extracting ontologies from world wide web via html tables. *Proc. PACLING 2001*, pp. 332–341, 2001.