

地方自治体ウェブからのイベント情報抽出手法

潮 田 達 也^{†1} 藤 田 茂^{†2}

地方自治体ウェブサイトが存在するイベント情報ページから機械利用可能な形式で情報を取得することを目的とした研究を行っている。Web ページからの情報抽出の既存手法として HTML タグに注目したウェブラッパー手法があるが、タグの構造に情報抽出の性能が左右されるという課題がある。そこで本論文では、HTML 文書からタグを除去してテキスト情報へ変換し、そのテキスト情報とあらかじめ抽出すべき対象に関する単語を収録した単語集を比較することで目的の文字列を抽出する手法を提案する。東京都 23 区及び千葉県 56 市町村に対して評価実験を行った。提案手法は全体で 73% のイベント情報を抽出することが出来た。既存手法は LR ラッパー手法が 52%、Tree ラッパー手法が 55%、PLR ラッパー手法が 32% であった。提案手法は単純なアルゴリズムと単語集を組合せる事で既存手法よりも高い割合でイベント情報が抽出可能である事を確認した。

Event information extraction technique from local government web

TATSUYA USHIODA^{†1} and SHIGERU FUJITA^{†2}

The research to aim to acquire information on the event information page that exists in the local government web in the form for which the machine can be used is done. There is a problem that the performance of the information extraction is controlled by the structure of tag though there is a web rapper technique for paying attention to the HTML tag as an existing technique of the information extraction on the Web page. Then, it proposes the technique for extracting a target character string by removing tag from the HTML document, converting into text, and comparing the text with the collection of words that collects the word concerning the object that should be extracted beforehand in this thesis. The assessment experiment was done to Tokyo 23 district and Chiba Prefecture 56 municipality. The proposal technique was able to extract event information on as a whole 73%. LR-Wrapper was 52%. Tree-Wrapper was 55%. PLR-Wrapper was 32%. The proposal technique confirmed event information was rating higher than an existing technique extractive by the combination of a simple algorithm and the collection of words.

1. はじめに

各地方自治体や一般の企業では、ウェブサイトを利用した情報提供が行われている。各自治体ウェブサイトを紹介して提供する情報の例として、イベント情報や道路工事情報という情報と住民に直接関わることの分別や収集方法などがある。これら地方自治体のウェブサイトは人が利用する事を目的としている。そのため、ウェブページは HTML によって記述されているが、それらは人が閲覧しやすいデザインや位置を考慮した構成を実現するために利用され、文書を構造化するために利用していない。また、本来の HTML タグが持つ意味とは異なる用途で用いられており、機械的な処理による情報取得は考慮されていない。機械的な処理による情報取得が可能になれば、取得した情報を再利用する事が可能であると考えられる。例えば、イベントや工事を行う場所付近の交通案内等に役立てる事が可能である。しかし、現状ではまだセマンティック Web は普及していない。

機械的な情報の取得を可能にするために、これまでウェブページを対象とした情報取得手法が提案されている。ひとつはウェブページの特徴である HTML タグに注目して情報の取得を行うウェブラッパー手法²⁾⁻⁵⁾が複数提案されている。しかし、現状の HTML では、どのように HTML タグを使用するかはウェブページの作成者により決定されるため、各ウェブページによってラッパーを生成する必要がある。そのため、情報を取得するウェブサイトごとやウェブサイトのデザインが変更されるたびにラッパーを新たに生成することは作業負担が大きいの。

また、文字列の類似度による情報取得手法⁶⁾が提案されているが、イベント情報においてイベントや施設の名称などの地域特有の固有名詞は多様となる傾向がある。そのため、類似度を用いた場合に十分な情報取得が行えずに間違った情報を取得してしまうと考えられる。本論文では、各地方自治体のウェブサイトにて公開されている情報から、単語比較を用いたイベント情報の抽出を行う手法を提案し、評価実験を行ったのでこれを報告する。

^{†1} 千葉工業大学大学院 情報科学研究科

^{†2} Graduate School of Information and computer science, Chiba Insutitute of Technology

^{†2} 千葉工業大学 情報科学部

Faculty of Computer and Information Science, Chiba Insititute of Technology

2. イベント 情報

イベント情報ページには、各地方自治体ごとにその地域で行われている行事・催し物についての情報が掲載されている。しかし、各地方自治体によりウェブページのデザイン、書式、記載されている内容、情報の件数や説明文の量や質は様々である。

本論文ではイベント情報ページを対象としてHTML 文書からイベント 情報を抽出する。イベント情報の中には、日時や場所だけでなくイベント 内容や問い合わせ先などの様々な情報が記述されている。その中で、各地方自治体ウェブサイトのイベント 情報ページにおいて記述されていると期待できる情報として、

- イベント 名称
- 日時 (期間の場合は開始日、終了日を含む)
- 会場

の 3 種類の情報が挙げられる。本論文ではこれらの情報に関して抽出を行うこととする。

3. イベント 情報ページの種類

地方自治体より提供されているウェブページを扱う場合、考慮すべきであると考えられるウェブページの形式は以下の 2 形式¹⁾ である。

- シングル・インスタンス型¹⁾
ウェブページ内にイベント 情報が一件だけ掲載されている形態のウェブページ
- マルチブル・インスタンス型¹⁾

ウェブページ内の同一のカテゴリについて、複数件のイベント 情報が掲載されている形態のウェブページ。なお、この形式のページには表を用いたページと表を用いずにイベント 情報を掲載したページが存在する。

これらシングル・インスタンス型とマルチブル・インスタンス型は、野口らの研究¹⁾ による定義である。この中で、イベント の名称、日時や会場など、一つの事柄を表す情報の単位をインスタンスと称している。

4. ウェブからの情報取得に対する関連研究

4.1 ウェブラッパーによる情報取得²⁾⁻⁵⁾

ウェブラッパーによる情報取得手法は、文書中に存在する HTML タグに着目し、同じ意味を持つテキストは、同じ HTML タグに囲まれているとした手法である。この手法では、あ

らかじめ HTML タグと情報を取得すべき項目を関連付けさせておくことで、関連付けられた HTML タグを判断基準とし、そこに記述された文字列を取得する情報としている。

この手法に関する研究は多く²⁾⁻⁵⁾、さまざまなラッパーが開発されてきた。以下にその例となる研究について述べる。

LR ラッパー²⁾

LR ラッパーは、取得したい情報において、左側に共通する最長接尾語と右側に共通する最長接頭語を求めることで、該当する文字列の間にあるテキストを取得する手法である。切り出すための文字列が長ければ長いほど、目的の情報以外を取得してしまう確率は少なくなる。しかし、そのような切り出し文字列を機械的に決めることが出来ない場合がある。

Tree ラッパー³⁾

Tree ラッパーは、HTML 文書を木構造として扱い、取得した情報と同じ意味を持つ情報は、同じ木構造に記述されている HTML タグに囲まれていると想定する。そして、該当する木構造で記述されている HTML タグに囲まれたテキストを取得する手法である。

PLR ラッパー⁴⁾

PLR ラッパー (Path-Left-Right ラッパー) は、Tree ラッパーと LR ラッパーを組み合わせたラッパーである。Tree ラッパーでは、HTML タグの木構造から特定した HTML タグ内のテキストを取得しているため、その中に含まれる不要な文字列まで取得していた。これを LR ラッパーの概念を取り入れることで、前後の不要な文字列を省略する手法である。

複数の WebWrapper の利用⁵⁾

複数のウェブページに対応するために、取得すべき情報の一部から同様の表現がされている正解データを学習し、複数のパターンにより構成された LR ラッパーを複合的に組み合わせる手法が提案されている⁵⁾。

5. 既存手法²⁾⁻⁵⁾ の問題

これらの手法に共通した問題として、ウェブページのデザインが変更されてしまうと、HTML タグの利用法や構造が大きく変わるため、同じ HTML タグでも大きく意味の異なる可能性がある。そのため、ウェブページのデザイン変更や、対象のウェブページを変更するたびにラッパーを生成する必要がある。これらのラッパー手法は HTML タグを用いて情報を抽出している。そのため、一つの HTML タグ内に複数の異なる情報が記述されている場合はタグを用いただけでは対象とする情報のみを抽出することが出来ない。

5.1 文字列の類似度による情報取得⁶⁾

文字列に着目した手法として、梅原らの事例に基づくシリーズ型 HTML 文書からの半自動変換がある。ウェブラッパーを用いた手法とは異なり、HTML タグ自体に大きな意味を持たせず、文書中に現れるテキストの区切りとするために利用しているだけである。この文書はシリーズ型 HTML 文書と呼ばれる、記述されている情報の意味や文書の構造が相互に類似している文書に注目しており、事例となる文書を与えることで、同じシリーズの文書から情報の取得を行うものである。そのため、地域情報を提供するウェブページがシリーズ型文書るとき、効果的な情報取得が行えたと期待される手法である。ウェブラッパー手法と異なり、文字列の類似度という観点から取得を行っているため、ウェブページのデザインが変更された場合でも対応できる可能性があり、ウェブページごとにラッパーを生成するような手間は必要ない。

しかし、イベント情報においてはイベントや施設の名称などの地域特有の固有名詞は多様となる傾向がある。具体的な例を挙げると、“体力測定へ 15 (いこう)！& PLAY SPORTS！”、“富士吉田 WINTER CAMP”、“スポーツフュススタ習志野”などである。そのため、文字列の類似度では、十分な情報の取得が出来ずに間違った情報を取得してしまう可能性がある。

6. イベント情報抽出手法

本論文で目的とするのは各地方自治体に存在するイベント情報ページからのイベント情報抽出である。このイベント情報にはイベント名称・日時・場所・その他の情報が記載されている。

提案手法の流れを図 1 に示す。提案手法は、イベント情報が記述された HTML 文書から単語集を用いた情報抽出を行う。本手法はタグを用いて情報を抽出することは無く、既存手法では抽出することが出来ないウェブページでも情報抽出が可能であると考えられる。提案手法を利用するためには、イベント情報抽出に用いる単語をあらかじめ単語集にして与える必要がある。与える単語集はイベント名称・日付・場所・除外の 4 種類である。なお、除外については今回抽出しないその他の情報に関する単語を収録している。また、この処理で用いる形態素解析器は Sen⁷⁾ である。

次に今回用いた単語集の構築手法を述べる。各単語集は事前に訓練例として各自治体イベント情報ページを取得、そのページ内でイベント名称・日付・場所が記述されている文の中で出現頻度の高い単語を収集して登録している。除外と呼ばれる単語集も同様に今回抽出するイベント名称・日時・場所以外の内容が記述されている文の中で出現頻度の高い単語を取

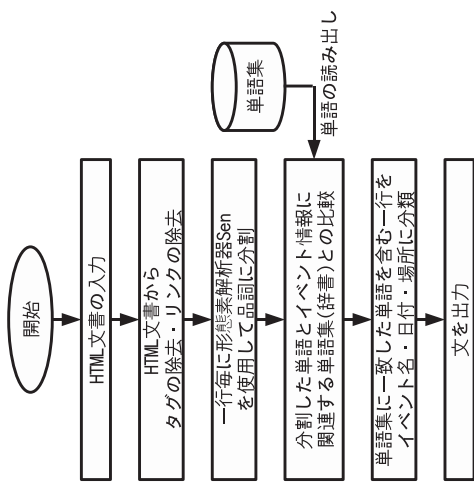


図 1 イベント情報抽出手法

集して各単語集に登録している。

以下にそれぞれの単語集の内容、単語の例と単語数を記載する。

- (1) 単語集 (イベント名)
イベント名に関する単語を集めた単語集 (計 80 単語)
 - 単語の例
大会 講座 祭 展 講演 マーケット
- (2) 単語集 (日付)
日付に関する単語を集めた単語集 (計 18 単語)
 - 単語の例
期日 期間 正午 開場
- (3) 単語集 (場所)
場所に関する単語を集めた単語集 (計 46 単語)
 - 単語の例
会場 センター 図書館 会館
- (4) 単語集 (除外)

抽出しない文に関する単語を集めた単語集 (計 15 単語)

- 単語の例

参加 問合せ 申込み

これら単語集は、イベント情報に関係する単語のみを登録している。

7. テキストからのイベント 情報抽出

7.1 前 処 理

イベント 情報抽出の前処理として、入力とした HTML 文書に対して HTML タグ及びハイパーリンクの除去を行う。この処理により、HTML 文書をテキスト情報のみにすることと本処理にて不要な情報の抽出が行われることを抑える。また、イベント 情報ページには別の イベント 情報ページへのリンクがあり、必ずそれぞれの イベント 名称が記載されている。そのため、別の イベント 名称を除去せずに残してしまうと、本処理にて誤った情報が抽出されてしまうので別ページへのハイパーリンクを取り除く。

7.2 単語比較による情報抽出

HTML 文書の前処理により HTML タグ及びハイパーリンクの除去を行い、テキスト情報のみにする。

この手法は、まず単語集にある単語を元に イベント 情報から目的の情報を抽出するために、イベント 情報に記載された文を一行ずつ読み込む。次に、読み込んだ一行を日本語形態素解析器 Sen により品詞に分割し、イベント情報に関連する単語を集めた単語集を用いて比較を行う。単語集の単語と一致した一行については単語集の種類により、イベント 名称・日付・場所・除外に分類する。イベント 名称・日付・場所・除外の単語集には 6 節に述べた通りの単語が登録されている。イベント 名称・日付・場所に分類された一行はそのまま出力される。このとき一行の品詞数がイベント 名称・日時・場所に分類された文の中で最も長い文の品詞数よりも多い場合は出力しない。これは分類された文の中に説明文などの長文が含まれていた場合、それらの文を出力しないようにするためである。

一致した単語を含む文にはその単語に関する内容が記述されている可能性が高いと考えられ、目的の情報の抽出が可能であると考えられる。しかし、除外に分類された文については除外すべき単語に一致したということでも出力しない。この方法で、イベント 情報に記載された複数日時の取得と HTML タグに依存しない情報の取得を行う。

8. 評 価 実 験

8.1 実験環境と評価用データ

表 1 評価用システム

Table 1 System for evaluation

実装言語	JDK 1.6
HTML Parser	Jericho HTML Parser
形態素解析器	Sen Version 1.2.2.1
登録した単語数	計 159 個

実験用の評価システムを JDK1.6 で作成し、表 1 の実験環境で実験を行った。評価実験の対象とするウェブページは、平成 21 年 1 月 10 日時点の東京都 23 区及び千葉県 35 市町のイベント情報の記載されたウェブページ計 230 ページである。なお、東京都 23 区についてはそのすべての自治体ウェブページにイベント情報が記載されたページが存在したが、千葉県全 56 市町村には イベント情報を記載したページを持たない自治体ウェブページが 21 市町村存在する。そのため、今回は イベント情報のページが存在しない自治体は除いた 35 市町で実験を行った。

イベント 情報ページについては各地方自治体ごとに一つのページに一つの イベント 情報が記載されているページを採用しているウェブサイトでは 5 ページずつ、一つのページに複数の イベント 情報が記載されているページを採用している場合は 1 ページずつ イベント 情報ページを収集している。もし、各地方自治体ウェブページに存在する イベント 情報ページが 5 ページに満たないページ数の場合には存在するページを全て収集している。しかし、これらは イベント 情報が記載されたページ数である。そのため、各地方自治体 イベント 情報ページ数とそのページに記載された イベント 情報数が同じであるとは限らない。

8.2 評 価 手 法

提案手法を実装した評価用システムに、前項にて収集しておいた各地方自治体 イベント情報ページを入力として与え、個々のウェブページの情報抽出結果について、8.3 節にて述べる正確度の定義を用いて評価する。また、同じデータを用いた既存手法との比較実験を行う。この中で Tree ラッパー手法と PLR ラッパー手法については各地方自治体毎に一ページを訓練例としてラッパーを生成している。

8.3 評価指標

提案手法を用いて取得した情報と実際にウェブページに記述されているイベント情報とを比較し、正しく取得出来ている情報数が幾つであるかを評価する。

評価の基準として式 (1) に示す正精度 (Accuracy)(%) の定義を用いる。

$$Accuracy = \frac{\sum_{j=1}^n C(Eventname_j, date_j, Place_j, Others_j)}{(Total_of_event_information \times 4) - X} \times 100 \quad (1)$$

X: Total where information that in each event information should doesn't exist

$$However, C = \begin{cases} The_event_name_exists + 1 \\ The_date_exists + 1 \\ The_place_exists + 1 \\ Others_exist + 0 \\ Others_don't_exist + 1 \end{cases}$$

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する。そのため、式 (1) の分子にある C は各イベント情報毎に抽出されたイベント名称 (Eventname) ・日時 (date) ・場所 (Place) ・その他除外すべきもの (Others) の情報総数である。なお、本手法ではテキストからの情報抽出を行うため、必要な情報を抽出してしまいう可能性がある。そのため、抽出すべきイベント名称・日付・場所の情報と同様に unnecessary 情報をまとめて一つの情報として扱う。 unnecessary 情報は抽出すべきではない情報である。そのため、各インスタンスごとに抽出されたかった場合のみ他の情報と同様に C の総数に 1 を加える。また Total_of_event_information は全イベント情報の総数である。X は各イベント情報に始めから存在しない取得すべき情報の総数である。一方で分母の総数は、入力したイベント情報が記述されたウェブページにおける、全ての情報についての総数である。本論文の提案手法は本来ある情報の数に対して、正しく抽出した情報の数がどの程度であるかを評価の基準とした。

9. 実験結果

イベント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イベント情報抽出結果を表 2 に示す。

表 2 について、東京都 23 区、千葉県 31 市と千葉県 4 町の正精度を示している。正精度は

前項で定義した式 (1) を用いて計算した値である。全情報総数は、式 (1) の分母の値であり、ウェブページ中に存在している全ての情報の総数が表されている。取得情報総数については、式 (1) の分子の値であり、評価実験において人手により抽出結果と元の HTML 文書と比較した場合に今回対象としたイベント情報が取得出来た総数である。また、既存手法についても同じデータで行った実験結果を示す。

表 2 実験結果
Table 2 Outcome of an experiment

対象地域	全情報総数	取得情報総数				正精度 [%]	
		提案手法	LR	Tree	PLR	提案手法	Tree
東京都 23 区	1008	791	556	550	284	79	55
千葉県 31 市	920	626	430	507	288	68	48
千葉県 4 町	80	55	51	51	21	69	63
						64	27

10. 考察

表 2 の中で、PLR ラッパー手法⁴⁾ を用いた結果が低い理由として、ページ内に出現する文字列の頻度によって目的の文字列を抽出するパスを生成している。この PLR ラッパー手法でも LR ラッパー手法や Tree ラッパー手法と同様に訓練例を与えている。しかし、訓練例と一緒に目的の文字列がどの位置に存在するかの情報は与えていないためにイベント情報が正確に抽出出来ないと考えられる。

今回の実験では誤って unnecessary 情報が抽出されたかどうかについても実験結果と元の HTML 文書と比較して行っている。そのため、実験結果から本手法は既存手法よりも unnecessary 情報が抽出されていないと考えられる。これは、イベント情報ページにて見出しの単語を用いている場合が多いため、「問い合わせ」や「申込」といった今回は抽出しない unnecessary 情報を表す言葉と単語集の単語が一致したために抽出されなかったと考えられる。

提案手法の問題として「問い合わせ」や「申込」のあとにイベントの場所の住所と同じ住所が記載されている場合はイベント情報の場所についての情報が抽出されてしまいう可能性がある。実験中にイベント名称・日付・場所が抽出された上で場所の住所だけが 2 度抽出されてしまいうページが評価用データの 4 割に存在した。このような事が起こらないようにする 1 つの方法として、一度場所に関する単語が認識されて 1 行が抽出されたらそれ以降は場所に関する単語の比較は行わないという方法がある。この方法では 1 つのページに 1 つの

イベント情報が書かれたシングル・インスタンス型のウェブページに対しては有効であり、そのようなイベントページのみを対象としている場合には効果が期待できる。しかし、今回対象とした地方自治体の中には 1 つのページに複数のイベント情報が書かれたマルチプル・インスタンス型のウェブページも存在する。そのため、1 つの場所に関する単語が認識されて 1 行が抽出された場合、それ以降を抽出しないようにしてしまうとマルチプル・インスタンス型のウェブページではほぼ情報の抽出が行えなくなるという問題がある。

また、本手法は HTML タグを使用しない手法であるために、ウェブラッパー手法のようにページのデザインが変更されるたびに新しいラッパーを生成する事は無い。しかし、比較に単語を用いるという点で単語集に登録されている内容が問題になる。特にイベント名称や場所に関しては各地域特有の単語が使用される場合があり、単語が登録されていない場合は抽出できないことがある。これは今回対象としなかった他の各地方自治体イベント情報ページに対して評価実験を行う場合でも同様の問題が起きると考えられる。この問題の解決方法として、イベント名称・日付・場所ごとに出現頻度の高い単語を収録した単語集を用意する。この単語集に各地方自治体ごとに対応する地域限定の単語を登録した単語集を用意して一緒に使用するというものがある。この場合は新たな単語を単語集に登録する必要があるが、各地域特有の単語を登録する事で新たなイベント情報ページに対応出来ると考えられる。

今回は各地方自治体イベント情報ページを対象として実験を行った。対象としたページはイベント情報が記載されていると分っているページであり事前知識が無い状態での実験は行っていない。しかし、このイベント情報ページ群は各自自治体ウェブサイトごとに一箇所に集められていることが多い。そのため、それらページが存在する場所を把握すればイベント情報ページのみの取得が可能だと考えられる。

11. 結 論

地方自治体が公開しているウェブページには有用な情報が多い。しかし、現状では人が閲覧することを前提としているため、HTML タグによる文書の意味把握ができず、機械処理による情報抽出が困難である。

この問題に対して、既存のウェブラッパー手法²⁾⁻⁵⁾では HTML タグを用いてタグで囲まれた情報を抽出している。しかし、これらの手法では一つの HTML タグで囲まれた日時・会場に関する情報が改行タグを用いて複数行記述されている場合と抽出すべき情報自体が直接 HTML タグで囲まれていない場合に目的の情報を抽出する事ができないという問題があった。

この問題に対して本論文では、既存手法で適用できなかったイベント情報の抽出を行うために、HTML 文書から HTML タグを取り除いたテキスト情報からの単語比較を用いた抽出手法を提案した。

本提案手法の有効性を確認するために、東京都 23 区と千葉県 35 市町の地方自治体イベント情報ページを対象として既存手法との比較実験を行った。評価指標として式 (1) の正確度を用いた実験結果により、本提案手法では全体で 73% のイベント情報を抽出が行えた事を確認した。また、既存手法では LR ラッパー手法が 52%、Tree ラッパー手法が 55%、PLR ラッパー手法が 32% のイベント情報を抽出することが出来た。

本提案手法により、既存手法で用いられたアルゴリズムよりも単純なアルゴリズムと事前で作成しておいた単語集を用いることで、イベント情報が抽出出来ることを確認した。

参 考 文 献

- 1) 野口 龍太郎, 山田 泰寛, 池田 大輔, 廣川 左千男: 頻度情報を用いた Web 文書群からのテンプレート抽出, DEWS, (2004).
- 2) N. Kushmerick: Wrapper induction: Efficiency and Expressiveness Artificial Intelligence, Artificial Intelligence, Vol.118, No.1-2, pp.15-68 (2000).
- 3) 村上 義継, 坂本 比呂志, 有村 博樹, 有川 節夫: Html 文書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, Vol.42, No.14, pp.39-49 (2001).
- 4) 山田 泰寛, 池田 大輔, 廣川 左千男: 半構造化文書に対する木構造と文字列を組み合わせたラッパーの自動生成手法, 情報処理学会論文誌研究報告, Vol.2003, No.98, pp.115-122 (2003).
- 5) 植松 幸生, 内山 俊郎, 片岡 良治, 松井 藤五郎, 大和田 勇人: 複数の Web Wrapper による高精度な情報抽出, 情報処理学会研究報告, Vol.2007, No.6, pp.117-123 (2007).
- 6) 梅原 雅之, 岩沼 宏治, 鍋島 英和: 事例に基づく 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.16, No.5, pp.408-416 (2001).
- 7) sen home, <https://sen.dev.java.net/>.