データベースシステム 128－5
（２００２．７．１７）

# ＨＴＭＬ文書からのイベント情報抽出

三宅 新二[i iii] 岡部 一光[i iii] 鳥越 秀知[ii iii] 横田 一正[iv]

i (株)両備システムズ ソフトウェアカンパニー 〒700-8504 岡山市豊成 2-7-16
ii 国立詫間電波工業高等専門学校 〒769-1192 香川県三豊郡詫間町香田 551
iii 岡山県立大学 情報系工学研究科システム工学専攻 〒719-1197 総社市窪木 111
iv 岡山県立大学 情報工学部 〒719-1197 総社市窪木 111
E-mail: i {shinji,okabe}@ryobi.co.jp, ii torigoe@dg.takuma-ct.ac.jp, iv yokota@c.oka-pu.ac.jp

**あらまし** HTML 文書からイベントに関する情報を抽出するための実現方式と問題点を議論する。イベント情報を抽出するためには、イベント情報の特定、必要項目の特定と抽出、不足項目の補足が必要となる。このために、HTML の構造情報の解析、タグと値のパターンマッチング、値の変換などを行う。これによって異なった形式で作成された複数の HTML 文書からイベント情報を抽出することが可能になる。複数の HTML 文書から同一形式でレコードを抽出することにより、HTML 情報の利用の可能性を広げる。

**キーワード** HTML，情報抽出，イベント情報，情報統合

# Extraction of Event Information from HTML Documents

Shinji MIYAKE[i iii] Kazumitsu OKABE[i iii] Hidetomo TORIGOE[ii iii] Kazumasa YOKOTA[iv]

i Ryobi Systems Corporation, Software Company, 2-7-16 Toyonari, Okayama-shi, Okayama, 700-8504 Japan
ii Takuma National College of Technology, 551 Takuma-cho, Mitoyo-gun, Kagawa, 769-1192 Japan
iii Okayama Prefectural University, Graduate course of Information Science and System Engineering, 111 Kuboki, Soja-shi, Okayama, 719-1197 Japan
iv Okayama Prefectural University, Faculty of Information Science and System Engineering, 111 Kuboki, Soja-shi, Okayama, 719-1197 Japan
E-mail: i {shinji,okabe}@ryobi.co.jp, ii torigoe@dg.takuma-ct.ac.jp, iv yokota@c.oka-pu.ac.jp

**Abstract** In this paper, we describe a method of extracting event information from HTML documents, and discuss some problems involved in the method. In order to extract event information, specification of event information area, specification of required terms and extraction, and supplement of insufficient terms are required. For this reason, analysis of structure information of a HTML document, pattern matching for tag and data area, and conversion of values are performed. Various kinds of event information are extracted from HTML documents. This method increases the availability of HTML information by extracting records in the same form from HTML documents.

**Keyword** HTML, Information Extraction, Event Information, Information Integration

## 1. はじめに

ネットワーク環境の急速な拡大に伴って、分散している情報源を統合することが注目されている。しかし、表現、構造、意味が同種のもの同士であれば、問題は生じないが、現実には、さまざまな異種性が存在しているため、その異種性の解消が大きな問題となっている。そこで、必要な情報だけに着目して、表現における異種性を解消して抽出し、情報統合することを考える。

インターネット上には多くの情報があふれており、かつ、個々のホームページは多種多様である。しかし、これらの情報を有効に利用するには、機械的な処理が不可欠である。本論文では、特定範囲のイベント情報に絞って、HTML 文書から情報を抽出し、一覧表にまとめることを提案する。

イベント情報を比較検討する場合、いろいろなホームページを検索して、イベントに関する情報を比較検討することとなる。しかし、ホームページごとに記述形式は異なっている。このような場合に、ホームページごとに抽出パターンを定義し、必要な情報だけを収集し、一覧表にまとめる。これにより、複数のホームページの異種性を解消し、情報統合した一覧表にまとめることができ、ソートや抽出など、情報の比較検討が容易になる。

情報抽出は、ラフなマッチングパターンの記述表現で、精度の高い情報抽出を目指している。このため、項の抽出は、正則表現を利用したマッチング方式とする。さらに、項の組合せは、HTML 文書のタグの階層構造と、抽出項の出現順序から、最適な組合せを求める方式で行う。この二段階方式により、項の出現順序における自由度が向上し、省略値の補足や、ノイズの排除も可能となる。マッチングパターン方式では、有効な抽出パターンを登録する必要がある。このため、定期的に再利用できる方が有効利用できると判断し、処理対象を「イベント情報の抽出」に絞る。

まず 2 節では、HTML 文書におけるイベント情報について議論する。3 節ではイベント情報をレコードとして抽出する方法を考え、4 節では抽出の洗練化について議論し、5 節で抽出処理の概要を述べる。

## 2. HTML 文書におけるイベント情報

イベント情報の記述パターンは多種多様である。以下に個々のイベント情報の例(抜粋)を示す。

・保健関連の情報

```
<P><B><FONT COLOR="#ff00ff" SIZE="+1">■</FONT>
        <FONT COLOR="#009900" SIZE="+1">3歳児健康診査</FONT>
        <FONT COLOR="#ff00ff" SIZE="+1">■■</FONT></B><BR>
    <B>●対象</B>10 年 12 月生まれ</P>
<P><B>●持参</B>母子健康手帳・健康診査アンケート </P>
<P><B>●内容</B>身体測定・内科診察・歯科健診・保健指導ほか</P>
<P><B>●受付時間・場所</B> 各所とも 13 時～14 時</P>
<P><B><FONT COLOR="#000099">〔倉敷〕</FONT></B>12(水)・26(水)市保健所</P>
<P><B><FONT COLOR="#000099">〔玉島〕</FONT></B>18(火)玉島保健福祉センター</P>
<P></P>
```

・企業イベントの情報

```
<!--社員をやる気にする給与制度セミナー--><A NAME="100"><HR><BR>
<CENTER><FONT SIZE="4"><STRONG>
```

```
<A HREF="http://www/">社員をやる気にする給与制度セミナー</A></STRONG></FONT>
<BR><BR></CENTER>
<TABLE WIDTH="100%" CELLPADDING=5>
<TR><TH WIDTH="20%"><FONT size=2>開催日＆会場</FONT></TH>
<TD BGCOLOR="#DDDDFF"><font size=2>・<strong>2002 年 6 月 13 日(木)
</strong><dir>XXセンター　3階会議室<br>XX 区;</dir></font></TD></TR>
<TR><TH><FONT size=2>主催者</FONT></TH>
<TD><font size=2>コンサルティンググループ</font></TD></TR>
<TR><TD COLSPAN=2>
<font size=2>企業活性化を目標においた給与制度<br><br>
<br>定員:10 名<br>参加費:2,000 円</font></TD></TR></TABLE>
<!--社員をやる気にする給与制度セミナー-->
```

実際の HTML 文書はもっと多くの情報を含んでいる。しかし、イベント名、場所、日時などの限られた情報を抽出できれば、イベント情報の概要は把握可能であり、利用の可能性は広がる。

そこで、HTML 文書の中から、イベント名、場所、日時などを抽出するため、以下のアプローチを考える。

1)構造からのアプローチ

HTML 文書の構造(タグの出現順序)に着目して、以下の範囲を特定する。

－個々のイベント情報の範囲。

－補足すべき値の含まれる範囲。

補足すべき値の範囲も特定しておくのは、次に示すように特定の情報(７月)が個々のイベント情報とは別の場所に記述されている場合があるためである。

| ７月のイベント | | 補足すべき値の範囲 |
|---|---|---|
| 特別企画展 | １７日　オープニング（市民センター） | 個々のイベントの範囲 |
| | １８日　勉強会（市民ギャラリー） | |
| 文化セミナー | ２５日 | 個々のイベントの範囲 |
| | 県立大学 | |

2)項の出現順序からのアプローチ

個々のイベント情報の範囲から、必要な情報(イベント名、場所など)を抽出し、出現順序に着目して、情報を構造化する。この方法により、日付が先にくる場合にも対応でき、自由度が高くなる。

・特別企画展 ―┬―１７日――市民センター
　　└―１８日――市民ギャラリー

・文化セミナー――――２５日――県立大学

補足すべき値の範囲からも、必要な情報(７月)を抽出し、個々のイベント情報をレコードとして出力するときに補足する。

## 3. レコードの抽出方法

以下の手順でレコードを抽出する。

－事前処理

－抽出範囲の特定

－項の抽出

－レコードの出力

### 3.1. 事前処理

HTML 文書には、レイアウトに関する多くのタグがあり、データ部分の記述とあわせると、さまざまな表記方法がある。このため、マッチング処理の負担が大きい。事前に以下の処理を行い、不要な情報を削除し、表記方法を揃えておくことで、マッチング処理の効率を高める。

－文字コード（SJIS に変換する）
－半角文字と全角文字（区別するかどうか、タグ部分とデータ部分に分けて指定可能とする）
－英小文字と英大文字（区別するかどうか、タグ部分とデータ部分に分けて指定可能とする）
－さまざまな表記方法（曜日／月の英語表記、大晦日などの表現）
－不要情報の削除（コメントタグ部分、BODY タグの範囲外、FORM タグなどの入力用の指定部分など、抽出に関係しない情報を除外）
－タグの整理（CENTER/STRONG タグなど、抽出範囲の選択、項の抽出に関係しない情報を除外）
－終了タグの補足（タグの対応関係のマッチングを行い、終了タグを補足する）

HTML 文書ごとに抽出パターンが異なるため、抽出範囲の選択や項の抽出において必要となるタグも異なる。FONT タグにて特定の色に変換している部分や、STRONG タグにて強調している部分に着目する場合など、必要なタグを残しておく必要がある。

この理由により、整理すべきタグは、HTML 文書ごとに指定(変更)可能とする。

### 3.2. 抽出範囲の特定

HTML 文書の構造(タグの出現順序)に着目して、以下の範囲(項の抽出範囲)を特定する。

－固有部：　個々のイベント情報の範囲。
－共通部：　補足すべき値の含まれる範囲。

抽出範囲の特定は、タグの出現パターンに着目して、前後のマッチングパターンを指定することにより行う。前後のマッチングパターンの指定においては、正則表現を可能とし、タグ部分を指定可能とする。複数のタグの組み合わせを指定することにより、抽出範囲を特定する精度を高める。

一つの抽出パターンに対して、複数の範囲が対応することがあり、同じ識別子となる。

前後のマッチングパターンにより特定した範囲は、範囲の識別子で対応付ける。この識別子によって特定した範囲を区別し、範囲の識別子ごとに項の抽出、レコードの出力を行う。範囲の指定は、固有部、共通部ともに複数パターンを指定可能とする。

また、HTML 文書では、終了タグの対応付けがない場合がある。しかし、事前にタグの対応関係のマッチングを行い、終了タグを補足しておくため、問題はない。

(1)指定方法

抽出範囲の特定は、固定部と共通部を区別して、以下のように指定する。

**<固有部の指定>::=event(<マッチングパターン>, <範囲の識別子>, <マッチングパターン>)**
**<共通部の指定>::=common(<マッチングパターン>, <範囲の識別子>, <マッチングパターン>)**

前後の<マッチングパターン>により範囲を特定し、<範囲の識別子>で対応付ける。
<マッチングパターン>には、タグ部分を指定可能とし、以下に示す正則表現を可能とする。

〔正則表現〕　（ａａ）？　：　ａａの０回または１回の繰り返し
（ｂｂ）＊　：　ｂｂの０回以上の繰り返し
（ｃｃ）＋　：　ｃｃの１回以上の繰り返し
［ｄｄ｜ｅｅ］　：　ｄｄまたはｅｅを選択

タグ部分のマッチングでは、現実的な利用を考えて以下の制約を設ける。
－タグ名は全体指定とする。<TAB*>のような指定には対応しない。
タグの属性部分も指定可能とし、以下のマッチングを可能とする。
－属性指定を省略した場合は、タグ名だけのマッチングを行う。
－複数の属性が指定された場合、属性の指定順序は逆転可能とする。
－指定された属性だけをマッチングし、他の属性は無視する。
－属性名も全体指定とする。属性値には以下の正則表現を可能とする。
〔正則表現〕　＊　：任意の文字列と一致(属性値の指定を省略しても同じ)
［ｄｄ｜ｅｅ］：ｄｄまたはｅｅと一致
－2番目の<TD>タグを指定したい場合など、タグの順番を<TD(2)>と指定可能とする。
ROWSPANの有無によって、順番が変わる場合でも、以下のように対応できる。
<TD ROWSPAN><TD(3)>：先頭のTDがROWSPAN指定なら、さらに3番目
<TD ^ROWSPAN><TD(2)>：先頭のTDがROWSPAN指定でなければ、さらに2番目

(2)指定例
以下のように指定する。
event( (<TABLE>,<TR>), #001, </TR> )
common( <HR>, #002, <HR> )

### 3.3. 項の抽出
3.2 項で特定した抽出範囲から、前後のマッチングパターンにより、項として抽出する部分を特定する。項は、項の識別子と抽出部分をペアにして抽出する。また、レコードの抽出時に順序性を保証する必要があるため、対応するタグと対応付けておく。

項の抽出は、特定のキーワードやタグの出現パターンに着目して、前後のマッチングパターンを指定し、抽出部分を特定することにより行う。前後のマッチングパターンの指定においては、正則表現を可能とし、タグ部分とデータ部分の両方を指定可能とする。
また、HTML文書では、終了タグの対応付けがない場合がある。しかし、事前にタグの対応関係のマッチングを行い、終了タグを補足しておくため、問題はない。

(1)指定方法
項は、以下のように指定する。
**<項の指定>::=term(<項の抽出>,<項の出力>)**

<項の抽出>::=(<マッチングパターン>, @<変数>, <マッチングパターン>, <範囲の識別子>)
（前後の<マッチングパターン>を指定することにより抽出部分を特定する。
<マッチングパターン>には、タグ部分とデータ部分の両方を指定可能とし、正則表現も可能とする。タグ部分は、抽出範囲の特定と同様の指定を可能とする。
<範囲の識別子>で示される範囲に対してマッチングを行い、抽出部分は「@<変数>」で指定し、<項の出力>と対応させる。）

<項の出力>::=(<項の識別子>, @<変数>)
（項を抽出するタグの位置と対応付けて、<項の識別子>と抽出部分(@<変数>)をペアにして出力する。<項の識別子>は、レコードの抽出時に利用する。
年号を西暦に変換する場合などに備えて、@<変数>の編集、計算操作も可能とする）

(2)指定例

項のマッチングパターンの指定例を、表形式、リスト形式、その他の場合に分けて、以下に示す。

・表形式の場合

term( ( <TR>, (<TR>, <TD>)+), @X, </TR>, #001) , (EVENT,@X) )

この指定は、<TR>タグを検出後、(<TR>タグ、<TD>タグ) の組を 1 回以上検出した後のデータを抽出することを示す。

・リスト形式の場合

term( ( <UL>,<LI>), @X, [<LI>|</UL>], #001) , (EVENT,@X) )

この指定は、<UL>タグ、<LI>タグを検出後のデータを抽出することを示す。最初の<LI>タグに対応してデータを抽出する。

・その他の場合

term( ( <TD>,"イベント："), @X, </TD>, #001) , (EVENT,@X) )

この指定は、<TD>タグを検出後のデータ部分で、"イベント名："を検出した直後の文字列からデータを抽出することを示す。

### 3.4. レコードの出力

3.2 項で特定した固定部の範囲で抽出した項を、指定された順番にまとめ、レコードとして出力する。このとき、不足する項がある場合は、共通部から補足する。

このため、レコードの指定では、出力する項を指定するだけでなく、3.2 項で特定した固定部および共通部の範囲の識別子を指定する。

(1)指定方法

レコードは、以下のように指定する。

**<レコードの指定>::=record(<レコードの抽出>,<レコードの出力>)**

<レコードの抽出>::=(<項の識別子, <項の識別子>, ・・)

(レコードを作成する<項の識別子>をすべて指定する。)

<レコードの出力>::=(<出力ファイル名>)

(指定されたファイルに、CSV 形式でレコードを出力する。出力としては、ここでレイアウト指定を可能としたいが、とりあえず CSV 形式としている。)

(2)指定例

以下のように指定する。

record( (EVENT, PLACE MONTH,DAY), (EVENT.CSV) )

レコードとして出力する項の識別子は EVENT、PLACE、MONTH、DAY である。

出力ファイルは、EVENT.CSV である。

## 4. 抽出の洗練化

項の抽出では、マッチングによる抽出だけでは、必要な情報を特定できない場合も多い。このため、抽出部分の正当性確認や、再抽出により、情報の抽出精度を高め、利用可能性を高める。

たとえば、マッチングだけでは、抽出部分が対象の情報かどうか判断できない場合は、次に示す抽出部分の体裁や、文字列によって判断する。この場合は、指定文字列が多くなるだけでなく、指定もれによる抽出もれや、誤った情報の抽出などの異常が発生しやすくなる。しかし、出現パター

ンによるマッチングだけでは特定できない場合があるため、指定可能とする。

－抽出部分の文字数、数値の範囲、文字種別（数字／英字など）の妥当性を確認

－抽出部分に指定された文字列（いずれか）を含むことを確認

－抽出部分に指定された文字列（すべて）を含まないことを確認

また、抽出すべきデータが、範囲指定または列挙されている場合にも対応する。たとえば、以下のような範囲指定や列挙の場合がある。

－範囲：「３月から６ヶ月間」、「３月から６月まで」、「３～６月に」、「３月～６月に」

－列挙：「９月、１２月、・・」、「３月・６月」、「３月と６月に・・」

このため、単独の項として認識するだけでは不十分で、直前の項との関係で、「範囲(終了値)」、「範囲(期間)」、「列挙」などの意味付けを行う。

さらに、正則表現だけでは抽出範囲を特定できない以下の場合にも対応する。

1) マッチングによる抽出部分の全体ではなく、抽出部分の一部分だけを抽出対象とする場合がある。必要な部分を特定するため、以下の工夫を行う。

－括弧、区切り記号に着目して、対応する部分を特定する。

－形態素解析を利用し、対応する名詞部分に特定する

2) TR タグと TD タグにより、以下のような表を提示する場合は、タイトル行にマッチングするデータ部分を項として抽出したい。しかし、正則表現によるマッチングでは表現できないため、別途抽出パターンの指定方法を提供する。このとき、ROWSPAN や COLSPAN 指定(グループ化)も考慮して、対応する位置の情報を抽出する。

| イベント | 場所 | 日時 |
|---|---|---|
| XX 学会 | 鬼怒川 | ７月１９日 |

「<TD TITLE="イベント">」のように属性値でタイトル名を指定させ、合致するデータ部分を抽出する。

3) 括弧の対応付けは正則表現だけでは表現できない。前後のマッチングパターンが括弧の場合は、対応付けを行う。

## 5. 処理の概要

HTML 文書からレコードを抽出するまでの概要を、以下に示す。

1)抽出範囲の指定

ホームページから抽出すべきイベント情報が指定されている部分、補足すべき情報が指定されている部分を、抽出範囲の指定により特定する。

event(マッチングパターン, #001, マッチングパターン)

common(マッチングパターン, #002, マッチングパターン)

2)項の指定

特定した抽出範囲から抽出すべき項を指定する。レコードの出力に必要な項をすべて指定する。

term((マッチングパターン,@X,マッチングパターン,#001),(EVENT,@X))

term((マッチングパターン,@X,マッチングパターン,#001),(PLACE,@X))

term((マッチングパターン,@X,マッチングパターン,#001),(DAY,@X))

term((マッチングパターン,@X,マッチングパターン,#002),(MONTH,@X))

3)レコードの指定

レコードとして抽出する項を指定する。また、レコードを出力するファイルも指定する。

record( (EVENT, PLACE MONTH,DAY), (出力ファイル名) )

4)前処理

以下の処理を行い、マッチング効率をよくする。

－HTML 文書の不要な部分、不要なタグなどを削除。

－記述されていない終了タグを補足。

－半角文字を全角文字に変換、英小文字を英大文字に変換。

6)抽出範囲を特定

1)で指定した<範囲の指定>により、抽出範囲を特定する。

7)項を抽出

6)で特定した抽出範囲に対して、2)で指定した<項の指定>を適応し、項を抽出する。

8)レコードを出力

6)で特定した抽出範囲に対して、3)で指定した<レコードの出力>を適応し、指定ファイルに以下のようなレコードを出力する。

EVENT, PLACE, MONTH, DAY　　←タイトル行（項の識別子を出力）

特別企画展，市民センター，７，１７

特別企画展，市民ギャラリー，７，１８

文化セミナー，県立大学，７，１７

## 6. まとめ

複数の HTML 文書から情報を抽出し、一覧表にまとめる。これにより、表現の異種性を解消できることを確認予定である。

また、抽出パターンの指定方法の改善、情報抽出の有効性の向上など、多くの課題を解決し、ツールの完成度を向上させる予定である。

このツールにより抽出したデータを、有効活用できるようになれば幸いである。

なお、本システムは作成中であるため、評価を行っていない。本システムの作成後は、以下のチェックを行い、評価する予定である。

－実際に抽出したレコードの中に余分なものや、漏れがないこと

－いろいろなホームページに適応し、パターン登録の操作性や有効性

－定期的に更新されるホームページに適応し、継続利用の有効性と可能性


## 謝　辞

さまざまな議論をいただく岡山県立大学の國島助教授、および横田研究室の皆様に感謝します。


## 参　考　文　献


[1] 有限会社マークアップ，図解でわかる正則表現，株式会社ディー・アート，東京，2002.

[2] 横田一正　他，マルチメディア情報学　第３巻　情報の表現，岩波書店，東京，2000.

[3] 横田一正　他，マルチメディア情報学　第７巻　情報の共有と統合，岩波書店，東京，1999.

[4] 北研二　他，情報検索アルゴリズム，共立出版株式会社，東京，2002.

[5] 梅原雅之　他，”事例に基づく HTML 文書からの XML 文書への半自動変換，”人工知能学会論文誌，16 巻 5 号 B，pp408-416，2001.

[6] 関根聡，“テキストからの情報抽出，”情報処理，40 巻 4 号，pp.370-373，Apr，1999.