Webデータベースにおける入力フォーム情報の自動抽出(セッション4: メタデータとWebデータベース応用)

NAKATOH TETSUYA
Computing and Communications Center, Kyushu University

大森 敬介
九州大学大学院システム情報科学府

廣川 佐千男
九州大学情報基盤センター

出版情報：情報処理学会研究報告. 情報学基礎研究会報告. 2005 (42), pp.87-94, 2005-05-19. 一般社団法人情報処理学会
バージョン：

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

2005－DBS－136（12）
2005－FI－79（12）
2005／5／20

Webデータベースにおける入力フォーム情報の自動抽出

中藤哲也[†]　大森敬介[††]　廣川佐千男[†]

ブラウザに表示される入力フォームにおいて，属性ごとにキーワードを指定して検索が可能なWebデータベースが増えている．さらに，このようなサービスをアプリケーションから直接利用する枠組としてWebサービスがある．多数のWebサービスのプールから必要なものを選択し，組み合わせることにより新たなサービスを構築する研究が多くの注目を集めている．しかし，一般に公開されているWebサービスはWebデータベースと較べごく少数である．本稿では，Webデータベースを Webサービスとして利用できるようにするために，入力データの属性をWebデータベース・サイトのフォーム・インターフェースから自動的に抽出する方式を提案した．また，国内の2,800件のWebデータベースから無作為に選んだ134件のサイトについて抽出実験を行ない，精度，再現率，F値の３つの観点から評価した．

Automatic Extraction of Input Form Information from Web Databases

Tetsuya Nakatoh,[†] Keisuke Ohmori[†] and Sachio Hirokawa[†]

There are increasing number of Web sites which dynamically generate web pages from their data bases according to users request specified with attributes and keywords. On the other hand, Web services are programmable components to provide services via the Web and are gaining much attention due to its composition mechanism. However, the number of available Web services is very small compared to Web databases. In this report, the authors propose a method which transforms a Web database to a Web service by extracting the set of input attributes to the site. An empirical evaluation is conducted by assessing precision, recall and F-measure of extracted attributes for 134 sites randomly chosen from 2,800 Web databases.

1. はじめに

Web上で利用できる情報には，静的なページから得られる情報だけでなく，検索に対して動的に生成されたWebページから得られる情報も非常に多い．そのような検索機能を提供するページは，一般に**検索サイト**と呼ばれている．検索サイトには，Googleなどのような一般のWebページ群検索サイトの他に，自サイト内のデータベースに対する検索機能を提供するサイトも多い．それらの情報は一般に直接参照する事ができず，検索によって動的に生成されるWebページによってのみ参照可能である．そのため，それらのページはInvisible Web[10),11)]，Deep Web[1)]，Hidden Web[3),4)]などとも呼ばれている．

そのような検索サイトは特定のテーマに限定した質の高い情報やサービスを提供している事が多く，またその情報量は直接参照可能なWebページの情報量よりも多いと言われている．このため，それらのデータの自動的な取り扱いは，情報抽出の重要な研究テーマの一つである．

我々は，そのような検索サイトを自動的に解析する事で，情報の入出力を自動化し，いわゆるメタサーチシステムを動的に構築するシステム**DAISEn**[16)]を提案してきた．本システムでは，特定の分野に関する検索サイトを選び，選択されたサイトへのキーワード検索を自動的に行ない，結果を統合してユーザに提示する事が可能である．

近年検索サイトには，これまでとは異なる新しい方向性がみられるようになって来た．複数の項目を用いた複雑な質問が行なわれ，そしてURLの単純なリストの代わりに，幾つかの項目から構成された情報の集まりのリストを返す検索サイトが増えている．例えば，

† 九州大学 情報基盤センター
Computing and Communications Center, Kyushu University
†† 九州大学大学院 システム情報科学府
Graduate School of Information Science and Electrical Engineering, Kyushu University

Amazon.com[15] は本のリストを返す．kakaku.com[18] は PC のリストと共にそれらの価格を返す．Travelocity[20] は指定されたエリアのホテルのリストを返す．これらの専門的な検索サイトの入力形式は，一般的な検索エンジンのものより複雑である．入力としていくつかのキーワードを組み合せて指定することを必要とし，それぞれのキーワードが異なった属性を表す．乗り換え検索の Jorudan[17] では，出発と到着駅，日時が必要である．また，ホテルのための検索サイト Mytrip[19] では，チェックインの日付，チェックアウトの日付，人数，部屋数，価格の上限と地域が必要である．これらの検索サイトを，単純なキーワードを用いた一般的な検索エンジンと区別して，Web のインターフェイスを持つデータベースという意味で **Web データベース**と呼ぶ．

我々は現在，これらの Web データベースの情報を連携，統合することを目的に研究を進めている．Web データベースの統合は，すなわち利用者にとってより使いやすいシステムを構築することに他ならない．例えば，複数の PC パーツの Web データベースを統合することで最も安い PC パーツを扱う店を探すことができる．ホテル予約と航空機予約を組み合わせれば，出張の準備を手早く行なうことが可能となる．

一方，ネットワーク上の情報サービスの新しい形として近年 Web サービスが注目されており，Web サービスの連携として Web Composition に関する研究が行なわれている[13]．しかしながら，今のところイントラネット内での運用が主であり，公開され利用可能な Web サービスは限定的である．今後，公開される Web サービスの増加には期待が持てるが，それを上回る多数のサイトで人間に対するユーザインターフェースを用いたサービス，すなわち Web データベースが提供され続けると考えられる．

本研究は，これらの複雑な Web データベースが持つ機能，サービスを動的に結合，連携し，新たなサービスを構築するという長期的なプロジェクトの一環である．その目的のためには，以下の 5 つの機能を実装する必要がある．

(A) 入力項目を持つ Web データベース URL の取得
(B) Web データベースからのフォーム情報取得
(C) 入力項目の分類
(D) 入力項目の統合
(E) 検索結果から個別データの抽出

本稿では，(B) のフォーム情報抽出を扱う．従来のデータベース，あるいは Web サービスであれば，データスキーマが明示的に与えられている．しかしながら，Web データベースにおいてはブラウザ経由での利用しか想定されていない．従って，各 Web データベースが扱うデータスキーマは, 入出力のページのフォーム情報や検索結果の出力情報から抽出，推定する必要がある．

関連研究として，検索結果の HTML ファイルに現れる反復パターンを発見し，個別データを自動的に抽出するための研究 (E) が数多くなされている[16]．それらは，狭い意味でラッパーと呼ばれている．一方，本稿の主要テーマは入力データスキーマの自動抽出であり，まだ多くの研究はない．Zhang ら[14] は，クエリーフォーム全般に存在する隠された共通の文法を想定し，その文法に沿った構文解析によりフォーム情報抽出を行なっている．しかしながら，入力項目の近くにそのラベルが存在することを想定しており，我々が想定する `TABLE` タグを用いた構造を想定していない．このため，我々の想定する構造がより現実に即していると考えている．

Web 上のサービス連携に関する従来の研究[2),12] では，各データベースの詳細情報が開発元から提供されること，あるいは共通形式のデータへの変換プログラムが提供されることを想定している．本稿で提案する手法は，各 Web データベースの Web インターフェースだけから必要な情報を得るものであり，各サイトの開発，運用システムとは完全に独立に実現できる．

北村ら[5] は，WWW より情報を抽出し統合するスクリプト言語 MetaCommander を実装している．HTML ページから目的のデータを抽出する為の手順をスクリプトとして記述するシステムであるが，タグや文字列として表された HTML 文書にどのようなデータ構造が含まれているかをスクリプトを書くユーザーが考え，そのデータ構造の表現形式をタグや文字列として記述する必要がある．すなわちデータスキーマが自動的に抽出される訳ではない．

情報融合のエージェントについての関連研究としては，Knoblock らによる ARIADNE[6] がある．これは，学習に基づいた情報抽出エージェントを容易に構築するための枠組みと，それらを組み合わせるための枠組みを与えている．しかし，対象は一般の Web(Visible Web) であり，本研究で扱う「入力情報」は対象になっていない．

フォーム情報について，これまでに具体的な調査を行なってきた[9]．本稿では，フォーム情報抽出，及び入力フィールドのスキーマ抽出を自動的に行なうアルゴリズムを提示し，それを実装したツールについて述べる．加えて，国内の 2,800 件の Web データベース

から無作為に選んだ 134 件の Web データベース・サイトについて，本ツールを，精度，再現率，F 値の 3 つの観点から評価する．

2. 入力項目とフォーム情報

ユーザに Web データベースを提供しているサイト（検索サイト）は一般にブラウザ経由の利用しか想定されていない．統合システムの構築時に利用できる情報は入力ページや検索結果の HTML ファイルのみである．従って，入力項目の抽出に用いる事ができる情報は，入力ページの HTML ファイルにおいて FORM タグ (<FORM>, </FORM>) で囲まれる部分である☆．この情報のことを特に**フォーム情報**と呼ぶ．本稿では，複数の入力項目を持つ検索サイトの HTML ファイルから統合に必要なフォーム情報を抽出する手法を提案する．本章では，統合対象となる入力ページの構造と入力項目の属性名について説明する．

図 1 のような複数の入力項目を持つ検索サイトの入力ページから，検索の統合に必要なフォーム情報を取得する事を考える．一般に，各入力項目の直前には入力項目の属性名を示す文字列がある．たとえば，図 1 では「タイトル」,「著者名」,「出版者」,「出版年」,「件名」,「キーワード」,「分類」の文字列である．本稿では，これらを各入力項目のラベルと呼ぶ．ラベルはその検索サイトの機能的意味を示している．

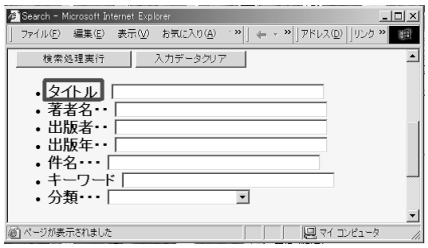

図 1 入力項目とそのラベル

従来の研究[14] においては，ラベルとして各入力項目の直前の文字列が想定されていた．しかしながら，我々の調査の結果，複数の入力項目を持つサイトでは，多くの場合 TABLE タグが用いられていることが明らかになった．このため本研究では，TABLE タグで表される入力項目群からラベルを抽出するために，ラベルは左端，上端，直前に現われるとするヒューリスティックスを提案する．そのような位置にラベルが現われている検索サイトの例を下記に示す．

☆ 出力結果のページからの情報を用いることも可能である．[8]

図 2 では，ラベルは入力項目の左端に現われている．この図では，プルダウンメニュー中の,「全て」,「標題」,「著者名」,「出版者」,「件名」,「フルタイトル」がラベル候補となる．

図 2 入力項目の左端に現われるラベル

図 3 では，ラベルは入力項目の上端に現われている．この図では，プルダウンメニュー中の,「フリーワード」,「タイトル」,「フルタイトル」,「著者」,「出版者」,「件名」,「分類」,「ISBN」がラベル候補となる．

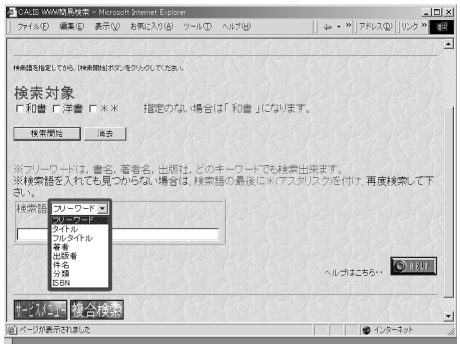

図 3 入力項目の上端に現われるラベル

図 4 では，ラベルは入力項目の直前に現われていることがわかる．この図では，プルダウンメニュー中の,「書名／タイトル」,「著者名／制作」,「出版社／発売者」がラベル候補となる．

3. フォーム情報の定式化

フォーム情報を抽出するアルゴリズムの説明の前に，抽出すべき情報を整理し，フォーム情報の定式化を行なう．この定式化された表現に従って，ブラウザへの表示のために構成された HTML 文書から，可能な限り論理的な構造を持った HTML 形式の中間表現へと変換する．特に，ラベルは一般に文字列である場合が多い．そのようなラベルの情報を他の要素と同様に扱

図 4　入力項目の直前に現われるラベル

うために，type 値が “word” である INPUT タグと同様の構造に変換する．

図 5 の例のような具体的におけるフォーム情報は，FORM タグの action 値 “./input.cgi” と method 値 “GET”，INPUT タグの type 値 “text” と name 値 “te”，OPTION タグの value 値 “opt1” と “opt2” と “opt3”，及び OPTION タグ直後の文字列「属性名 1」，「属性名 2」,「属性名 3」である．

```
<FORM action="./input.cgi" method="GET">
   <SELECT name="select">
      <OPTION value="opt1"> 属性名 1 </OPTION>
      <OPTION value="opt2"> 属性名 2 </OPTION>
      <OPTION value="opt3"> 属性名 3 </OPTION>
   </SELECT>
   <INPUT type="text" name="te">
</FORM>
```

図 5　HTML ファイルにおける FORM タグ

```
フォーム情報 := (form*);
form := (method, action, input*);
method := GET | POST ;
input := (type, name, value*, term*, pointer*, initial*);
type := text | radio | checkbox | select | word | etc ;
pointer := 整数;
initial := 整数
```

図 6　フォーム情報の BNF 表記

BNF 表記で表したフォーム情報の構造を図 6 に示し，フォーム情報の構成要素について下記に示す．

フォーム情報　複数の form から構成される．

form　FORM タグ 1 つ分の情報を持ち，action, method と複数の input から構成される．

action　FORM タグにおける action 値であり，一般に cgi プログラムが指定される．

method　FORM タグにおける method 値であり，“GET” か “POST” が指定される．

input　INPUT タグや SELECT タグ 1 つ分の情報を持ち，type, name 及び複数の value, term, pointer, initial で構成される．

type　INPUT タグにおける type 値のことで，“text”, “radio” や “checkbox” である．SELECT タグは “select” という type 値を持った INPUT タグに変換する．またラベル候補の文字列は，“word” という type 値を持った INPUT タグに変換する．

name　INPUT タグや SELECT タグの name 値である．

value　INPUT タグにおける value 値であり，SELECT タグの value 値は OPTION タグの value 値を用いることとし，OPTION タグが複数の場合 value 値も複数と定義する．

term　INPUT の type 値が “radio” か “checkbox” の場合は INPUT タグ直後の文字列とする．type 値が “select” の場合は OPTION タグ直後の文字列とし，OPTION タグが複数の場合 term 値も複数と定義する．また，type 値が “word” の場合は「ラベル候補の文字列」である．

pointer　この input のラベル候補が何番目の input であるかを示す．

initial　input の type 値が “radio” か “checkbox” の場合は何番目に “checked” が付いていたかを示す数字である．type 値が “select” の場合は何番目の OPTION タグに “checked” が付いていたかを示す．

4. フォーム情報抽出アルゴリズム

図 7 は，複数の入力項目を持つ検索サイトの HTML ファイルからフォーム情報を取得する手順である．まず，(1)〜(4) に示した手順で HTML ファイルを前処理した後に，(5)〜(9) に示した手順でフォーム情報の抽出を行なう．

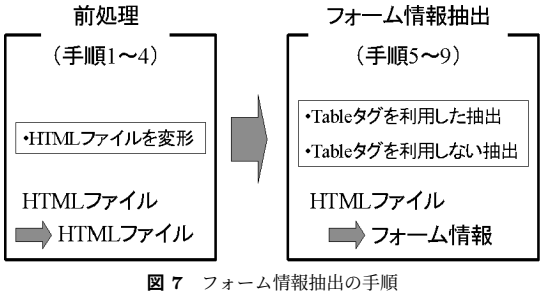

図 7　フォーム情報抽出の手順

このアルゴリズムによって，各テキスト入力フィールドに関して最大 3 つのラベルが取得される．

(1) 無視するタグの除外

HTML ファイルから，<ADDRESS>，<SCRIPT>，<!>，<FONT>，<LABEL>，<B>，<I>，<U>，<S>，<TT>，<SUP>，<SUB>，<NOBR>，<CENTER>，<A> の各タグ（終了タグを含む）を除去する．

(2) OPTION タグの集約

SELECT タグで囲まれた OPTION タグ部分を集約し，INPUT タグへと変換する．OPTION タグの value 値を INPUT タグの value 値とし，INPUT タグの type 値を新たに “select” とする．更に OPTION タグ直後の文字列を INPUT タグの直後に配置する．OPTION タグが複数ある場合には，value 値と直後の文字列をそれぞれコンマで区切り，配置する（図 8，9）．

```
<SELECT name=na>
  <OPTION value=val1>セレクト 1
  <OPTION value=val2>セレクト 2
</SELECT>
```

図 8 SELECT タグから INPUT タグへの変換（変換前）

```
<INPUT type="select" name="na"
  value="val1,val2"> セレクト 1, セレクト 2
```

図 9 SELECT タグから INPUT タグへの変換（変換後）

(3) “radio” 及び “checkbox” タイプの集約

INPUT タグの type 値が “radio” あるいは “checkbox” であり，name 値が同じものが連続して現われる場合，それら連続する INPUT タグを 1 つに統合する．統合方法は，それら連続する INPUT タグ中の value 値と INPUT タグ直後の文字列をそれぞれコンマで区切り，INPUT タグ中の value と INPUT タグ直後の位置にそれぞれ配置する（図 10，11）．

```
<INPUT type=radio name=na value=val1>ラジオ 1
<INPUT type=radio name=na value=val2>ラジオ 2
```

図 10 連続する INPUT タグの統合（統合前）

```
<INPUT type=radio name=na value="val1,val2">
ラジオ 1, ラジオ 2
```

図 11 連続する INPUT タグの統合（統合後）

(4) INPUT，SELECT タグへの番号付加

INPUT タグそれぞれに対し，何番目の FORM タグの INPUT タグかを示す “form-num”，及び，その FORM タグ中の何番目の INPUT タグかを表す “input-num” をそれぞれ INPUT タグ中に新たに付加する（図 12，13）．

```
<INPUT type=text name=na>
```

図 12 INPUT，SELECT タグへの番号付加（付加前）

```
<INPUT type=text name=nam
   form-num=1 input-num=1>
```

図 13 INPUT，SELECT タグへの番号付加（付加後）

(5) FORM タグ中の action 値と method 値の取得

各 FORM タグ中から，method 値として “GET” か “POST” を取得し，また，action 値も取得する．もし，action 値が “./input.cgi” のような相対 URL の場合は，絶対 URL へと変換する（図 14）．

```
<FORM method="GET" action="./input.cgi">
```

図 14 FORM タグ中の action 値と method 値の取得

(6) TABLE タグの内容の 2 次元配列化

TABLE タグで囲まれる部分を<TR>や<TH>と<TD>を考慮して 2 次元配列へと格納する．<TH>と<TD>に “colspan” や “rowspan” のように複数の行と列にまたがることを表す指示がある場合はこのことも考慮する（図 15，16）．

```
<TABLE>
 <TR><TD>A1</TD><TD>B1</TD><TD>C1</TD></TR>
 <TR><TD>A2</TD><TD>B2</TD><TD>C2</TD></TR>
 <TR><TD>A3</TD><TD>B3</TD><TD>C3</TD></TR>
</TABLE>
```

図 15 TABLE タグの内容の 2 次元配列化（2 次元配列化前）

A1	B1	C1
A2	B2	C2
A3	B3	C3

図 16 TABLE タグの内容の 2 次元配列化（2 次元配列化後）

(7) 2 次元データの整形

TABLE タグで囲まれた部分を格納した 2 次元配列において，一列，または一行全てのデータが空の場合は 2 次元データの整形を行なう（図 17，18）．

(8) 一般入力項目のラベル取得

TABLE タグで囲まれていない部分に入力項目がある場合，各入力項目の直前の文字列をラベルとして取得する．

空	空	空
空	A1	B1
空	A2	B2

図 17 2 次元データの整形（整形前）

A1	B1
A2	B2

図 18 2 次元データの整形（整形後）

文字列 1<*input1*> **文字列** 2<*input2*><*input3*>

上図における入力項目のラベルとして，入力項目 <*input1*> は直前のラベルとして文字列 *1* を取得する．入力項目 <*input2*> は直前のラベルとして文字列 *2* を取得する．入力項目 <*input3*> は直前のラベルとして <*input2*> を取得する．

(9) 2 次元データからの入力項目ラベル取得

TABLE タグで囲まれた部分の 2 次元データを解析することで，データ中に含まれる入力項目のラベルを取得する．入力項目からみて左端，上端，直前の三種類の文字列または入力項目をラベル候補として取得する．ラベル候補として入力項目を取得した場合は，その入力項目のラベルを再取得する．

文字列11	···	**文字列**12 <*input1*> **文字列**13
···	···	···
文字列41	···	**文字列**42 <*input2*> **文字列**43 <*input3*>

上図における入力項目のラベル候補として，入力項目 <*input1*> は左端のラベルとして**文字列**11，直前のラベルとして**文字列**12 を取得する．入力項目 <*input2*> は左端のラベルとして**文字列**41，上端のラベルとして <*input1*>，直前のラベルとして**文字列**42 を取得する．入力項目 <*input3*> は左端のラベルとして**文字列**41，上端のラベルとして <*input1*>，直前のラベルとして**文字列**43 を取得する．

5. 抽出アルゴリズムの評価

5.1 評 価 実 験

前章で述べたフォーム情報抽出アルゴリズムの評価を行なうために，このアルゴリズムを実装したツールを作成し実験を行なった．

評価実験の対象として，以前収集した Web データベース 2,800 件[7] から，テキスト入力フィールドを持つ Web データベース 150 件を無作為に選び，wget☆を用いて HTML ファイルを取得した．150 件のうち，有効な HTML を取得できた Web データベース 134 件を今回の評価実験の対象とした．

評価の手順を次に示す．

(1) 正解例の作成 上記の Web データベース 134 件それぞれにテキスト入力フィールドを識別する ID，プルダウンメニューを識別する ID を付加する．作業員 1 名は Web データベースを閲覧，各テキスト入力フィールドのラベルにあたる文字列または ID を判断し，このラベルをそのテキスト入力フィールドの正解例として記録する．

(2) 本アルゴリズムによる抽出 上記の Web データベース 134 件に対し，本アルゴリズムを実装したツールを用いて各テキスト入力フィールドのラベルを自動で抽出する．

(3) 定量的評価 人手で準備した正解例の中の何割をツールで抽出できたかを表す再現率（Recall），ツールで抽出したものの中で人手で準備した正解例に含まれるものが何割だったかを表す精度（Precision），および F 値（F-measure）の 3 つの値を求める．

(4) 従来の手法との比較 従来の手法，即ち，テキスト入力フィールドの直前の文字列や ID をラベルとして抽出する手法の F 値を求め，本ツールの F 値と比較する．

テキスト入力フィールドのラベル抽出における再現率 R，精度 P，F 値 F は，一般的な情報検索における定義にならい，それぞれ以下のように定義した．

ある Web データベースにおいて n 個のテキスト入力フィールドがあるとする．各 $i = 1, 2, ..., n$ について，人手で準備した正解例のラベルの集合を H_i，本アルゴリズムにより抽出されたラベルの集合を A_i とする．このとき，その Web データベースにおける R，P，F は以下の式で表される．

$$R = \frac{1}{n} \sum_{i=1}^{n} \frac{|H_i \cap A_i|}{|H_i|} \tag{1}$$

$$P = \frac{1}{n} \sum_{i=1}^{n} \frac{|H_i \cap A_i|}{|A_i|} \tag{2}$$

$$F = \frac{2}{\frac{1}{R} + \frac{1}{P}} = \frac{RP}{2(R+P)} \tag{3}$$

☆ wget: Web 上のデータを一括取得するためのソフトウェア

5.2 実験結果と考察

評価実験で得られたF値をグラフ化したものを図19に示す．横軸はF値を降順にソートしたサイトをとり，縦軸はF値を表している．

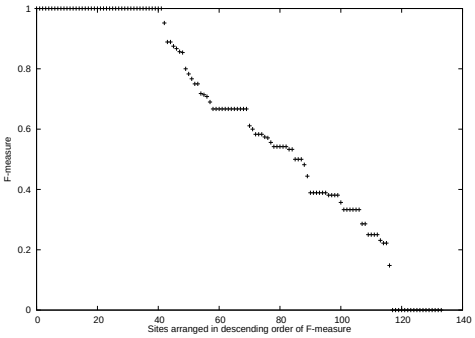

図 19 F値のグラフ

この結果より次のことが分かる．

（A）F値が1のサイトは134件中42件（31%）

（B）F値が0のサイトは134件中17件（13%）

（C）それ以外のサイトは134件中75件（56%）

（A）は，42件のサイトについて抽出が完全に成功していることを示す．

（B）は，17件について全く抽出できなかったかあるいは抽出したものが全て間違っていたことを意味する．17件を個別に確認したところ，抽出に失敗した理由は以下の4つであった．

- ツールにプログラム上のバグがあったためラベルの取得に失敗した．
- 人手により正解例と判断したラベルが不適切であった．
- 本アルゴリズムで想定していない位置にテキスト入力フィールドのラベルが存在した．
- TABLEタグでなく，<dt>，<li>，<dd>などのリストを構成するタグを用いて表の構造を表していた．

（C）は，ツールを用いて正解例の一部を取得できたことを示す．この75件について，正解例以外が取得された原因を確認するため，再現率と精度の関係を調べた．図20は再現率，精度，F値をまとめたグラフであり，横軸にはF値を降順にソートしたサイトをとり，縦軸は再現率，精度，F値の各値である．図21は再現率と精度の相関関係を表したグラフ（Recall, Precision値の小数点第2位を四捨五入した値でgrid上に配置したグラフ）であり，x軸は精度，y軸は再現率，z軸はサイト数である．

これらのグラフから，再現率が高いが精度が低いサイト，逆に精度が高いが再現率が低いサイトに着目した．

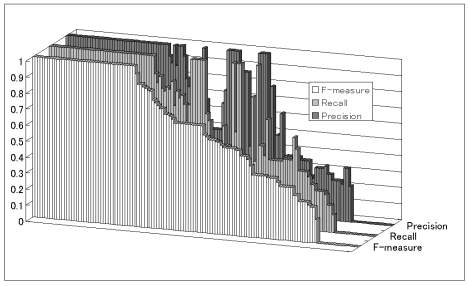

図 20 再現率，精度，F値のグラフ

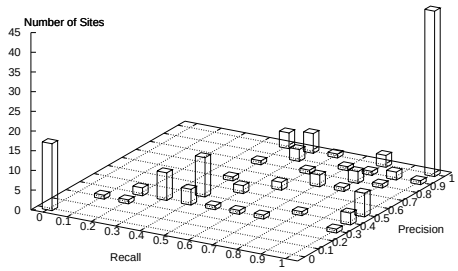

図 21 再現率と精度の相関関係

再現率が高いが精度が低い，即ちツールで取得したラベルに正解例のラベル以外のものが含まれるサイトの理由は下記であった．

- ツールで取得した上端，左端，直前の3つのラベルのうち1つだけが正解例と一致したため精度が低い．

一方，精度が高いが再現率が低い，即ちツールで取得できたラベルは正解例のラベルと一致したが，ツールでは正解例のラベル全てを取得できなかったサイトの理由は，以下の2種類に分類できた．

- 人手により正解例と判断したラベルが不適切であった．
- 正解例のラベルはテキスト入力フィールドの直後にあり，想定外の位置にラベルが現れたためツールで取得できなかった．

以上から，再現率と精度の値を向上するために，ツールのバグ修正のほかに以下の3つのような対応が考えられる．

- <dt>，<li>，<dd>などのリストを構成するタグを考慮したアルゴリズムの改良．

- 複数取得されたラベル候補のうち，選択すべきラベルを判断するための重み付けや選択手法[8)]．
- 想定されるラベル位置についての再検討．

次に，テキスト入力フィールドの直前の文字列やプルダウンメニューをラベルとして抽出する従来の手法の F 値を求め，本ツールの F 値と比較した．図 22 は本ツールの F 値と従来の手法の F 値を比較した図であり，縦軸は F 値，横軸はサイト数である．図 22 から，本ツールの F 値が従来の手法の F 値よりも良い結果であり，本アルゴリズムが有用であることが分かる．

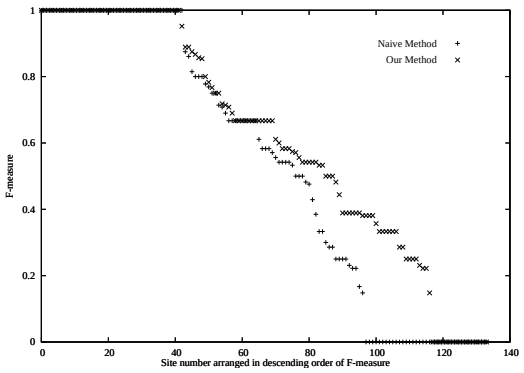

図 22 従来の手法との比較

6. ま と め

本稿では，Web データベースを統合することの有用性を示し，統合の手順を示すとともに統合に必要な技術に触れた．このうち，フォーム情報取得アルゴリズムについて提案を行ない，実装したツールの評価を行なった．

今後，アルゴリズムの改良を行なう事により，抽出精度を高めると共に，他の技術と組み合わせることで，Web データベースを自動的に変換し，Web サービスとして提供する仕組みを構築する予定である．

参 考 文 献

1) BrightPlanet, The Deep Web: Surfacing Hidden Value, BrightPlanet White Paper, 2000.
2) S. Chawathe, H. Garcia-Molina, J. Hammer, K. Ireland, Y. Papakonstantinou, J. Ullman, and J. Widom. "The TSIMMIS Project: Integration of Heterogeneous Information Sources". In Proceedings of IPSJ Conference, pp. 7-18, Tokyo, Japan, October 1994.
3) P. Ipeirotis, L. Gravano and M. Sahami, PERSIVAL Demo: Categorizing Hidden-Web Resources, JCDL2001, 2001.
4) P.Ipeirotis, L.Gravano and M.Sahami, Probe, Count, and Classify: Categorizing Hidden-Web Databases, ACM SIGMOD 2001, 2001.
5) Yasuhiko Kitamura, Tomoya Noda, and Shoji Tatsumi, Single-agent and Multi-agent Approaches to WWW Information Integration, Multiagent Platforms, Lecture Notes in Artificial Intelligence, Vol. 1599, Berlin et al.: Springer-Verlag, 133-147, 1999.
6) Knoblock, C. A., S. Minton, J. L. Ambite, N. Ashish, I. Muslea, A. G. Philpot, and S. Tejada, The Ariadne Approach to Web-Based Information Integration, International Journal of Cooperative Information Systems, vol.10, no.1-2, pp.145-169, 2001.
7) T. Nakatoh, K. Ohmori, Y. Yamada and S. Hirokawa, *COMPLEX QUERY AND METADATA*, Proc. ISEE2003, pp. 291-294, 2003.
8) 大森 敬介, 中藤 哲也, 原 由加里, 廣川 佐千男. 検索サイトにおける入力項目と検索結果のフィールド名の対応調査 FIT2004, pp. 89-90, 2004.
9) 大森敬介, 中藤哲也, 山田泰寛, 原由加里, 廣川佐千男, 複雑な検索機能を持つ検索サイトの動向調査 DEWS2004, I-1-05, 2004.
10) P. Pedley, The invisible web, ASLIB, 2001.
11) C. Sherman and G. Pric, The Invisible Web, Infomation Today, Inc., Medfore, New Jersey, 2001.
12) 菅坂 玉美, 益岡 竜介, 佐藤 陽, 北島 弘伸, 丸山 文宏. 知的エージェント環境 SAGE の EC への適用, 取引フェーズへの適用. 第 6 回マルチ・エージェントと協調計算ワークショップ (MACC), 日本ソフトウェア科学会, 1997 年 12 月.
13) S. Thakkar, C. A. Knoblock, J. Ambite and C. Shahabi, Dynamically Composing Web Services from On-line Sources, Proc. of 2002 AAAI Workshop on Intelligent Service Integration, Edmonton, Alberta, Canada.
14) Zhen Zhang, Bin He, Kevin ChenChuan Chang, Understanding Web Query Interfaces: BestEffort Parsing with Hidden Syntax, SIGMOD2004.
15) Amazon.com, `http://www.amazon.com/`
16) 専門検索サイトの動的統合による次世代検索システム DAISEn, Directory Architecture for Integrated Search Engines, `http://daisen.cc.kyushu-u.ac.jp/`
17) Jorudan, `http://www.jorudan.co.jp/`
18) kakaku.com, `http://www.kakaku.com/`
19) Mytrip, `http://www.mytrip.net/`
20) Travelocity, `http://www.travelocity.com/`