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

2007－ＦＩ－86 (11)
2007－ＤＤ－60 (11)
2007／3／27

Webサイトの適応型サイトマップの作成とそのオンラインショッピング支援への応用

伊藤慎一† 西口直樹† 松本章代† 小西達裕†
高木朗‡ 小山照夫¶ 三宅芳雄∫ 伊東幸宏†

†静岡大学 〒432-8011 静岡県浜松市城北3-5-1
‡株式会社CSKシステムズ 〒107-0062 東京都港区南青山2-26-1
¶国立情報学研究所 〒101-8430 東京都千代田区一ツ橋2-1-2
∫中京大学 〒470-0393 愛知県豊田市貝津町床立101
E-mail: riir@inf.shizuoka.ac.jp

Webユーザがオンラインショッピングや就職支援サイトなどを利用する場合，同一目的のサイトを巡回，比較して最適なサイトを選択している．その際，その目的のサイトに共通に存在する典型的な情報の記述を，各々のサイト中から見つけ出す必要がある．しかし，各サイトの構造は一様ではなく，それらの情報の記述の位置もサイトによって異なるため，サイトの比較に大きな労力を要している．そこで本研究では，語の出現頻度・タグ情報を用いた機械学習により求めたフィルタを用いて，サイト内から典型的情報の記述を抽出して提示する方法を提案する．また，オンラインショッピングサイトに対して指定された情報の記述部分を抽出するプロトタイプシステムの試作とそこで作成したフィルタの性能評価について報告する．

A method to make adaptive site map of Web site and its application to on-line shopping

Norikazu ITOH† Naoki NISHIGUCHI† Akiyo MATSUMOTO†
Tatsuhiro KONISHI† Akira TAKAGI‡
Teruo KOYAMA¶ Yoshio MIYAKE∫ and Yukihiro ITOH†

†Shizuoka University 3-5-1 Johoku, Hamamatsu-shi, Shizuoka, 432-8011 Japan
‡CSK SYSTEMS Corporation 2-26-1 Minamiaoyama, Minatoku, Tokyo, 107-0062 Japan
¶National Institute of Informatics 2-1-2 Hitotsubashi, Chiyoda-ku, Tokyo, 101-8430 Japan
∫Chukyo University 101 Tokodachi, Kaizu-cho, Toyota -shi, Aichi , 470-0393 Japan
E-mail: riir@inf.shizuoka.ac.jp

When users of internet intend to get service and cruise around multiple web sites which serve the common services, they usually compare each site and choose the best site. For example, we usually check price, delivery date, the way of payment, and so on before we order to get the service. However, location of descriptions representing such typical information in each site is different from each other, and the user pay a lot of cost to find the information in every site. In this pager, we propose a method to specify descriptions representing typical information in web sites and help users in finding typical information. In order to specify a description for typical information, we make filters by using a machine learning technique, C4.5. We also introduce our prototype system for on-line shopping and report accuracy and recall of each filter.

1 はじめに

現在，Web検索エンジンの品質向上のために様々なアプローチが存在するが，本研究では，オンラインショッピングや就職支援などの同一目的のサイトを検索・巡回する，という側面からWeb検索エンジンの品質向上を試みる．検索結果から目的のサイトを探し，サービスを受ける手順の中で，ユーザは利用規約・料金等，この種のサイトに共通して存在する典型的な情

報を得てサービスを受けると考えられる．しかしこれらの情報の記述は検索結果のページには含まれず，サイトの中までたどる必要があり，探す労力がかかる．つまり，検索結果と同時にサービスを受ける上で必要な情報を得ることはユーザにとって魅力であり，便利であるといえる．そこで，本研究では同一目的のサイトを検索し，利用する際にそのサイトに共通に存在する情報を提示する手法を提案する．ユーザによって欲しい情報は異なると考えられるため，ユーザが自由に選択し，作成できるツリー型のサイトマップを提示する．そしてどのサイトでも同じ形式のサイトマップが形成される．本研究ではこのサイトマップを適応型サイトマップと呼ぶ．また，オンライン上のサービスのうち，インターネット利用者の9割が経験しているというオンラインショッピングサイトを対象にする．

2　システム構成

本論文で提案する適応型サイトマップを作成するためのシステムの構成（Fig.1）について述べる．本システムは既存の検索エンジンを利用し，その検索結果のサイトごとに適応型サイトマップを作成する．検索結果にはサイトごとに適応型サイトマップを表示するためのリンクを追加することで，ユーザが閲覧できるようにする．

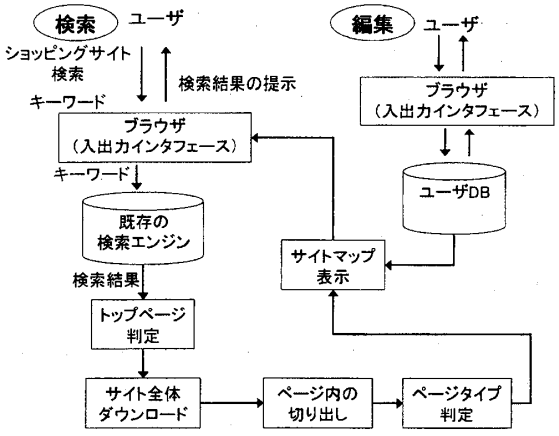

Fig.1 システム構成図

サイト全体を解析するため検索結果を受け取ると，トップページを特定し，トップページから内部リンクをたどることでサイト全体をダウンロードする．そして，指定した情報が存在するページを特定し，ユーザが作成したサイトマップを表示する．現在はFig.1のシステムで実装しているが，この流れでは処理時間がかかってしまうため，実用化に向け，あらかじめ各サイト解析を行っておくシステムにする必要がある．

本論文ではこのシステムにおける，サイトに共通に存在する情報のページの特定手法について3章で，試作システムについて4章で述べる．そして5章でフィルタの評価実験について説明する．高精度のフィルタができれば，サイトマップを作成できるといえる．

3　サイトに共通に存在する情報のページの特定手法

3.1　サイトに共通に存在する情報とは

Web上のオンラインショッピングサイト，オンラインショッピングに関する書籍，ユーザのアンケートから，ユーザがオンラインショッピングを行う際に必要な情報を決定する．

支払い方法，送料，返品，掲示板，注文の流れ，商品納期，ラッピング，保証，領収書，注文フォーム，Eメール注文，FAX注文シート，FAQ，お客様の声，問い合わせ，商取引法，ログイン，会員登録，ポイント，プライバシーポリシー，メールマガジン，サイトマップ，セキュリティ

これらの情報が存在するページをオンラインショッピングサイト内から特定する．ユーザがシステムを利用する際は，この情報の中から自由に情報を選択できるようにする．

3.2　指定した情報のページの特定手法

指定した情報のページを特定することはページ分類に似ている．従来のページ分類の研究では，以下の研究が行われている．

Haasらは，(1) 目次，索引(organizational)，(2) 参照，支援(documentation)，(3) 記事，論文(text)，(4) ホームページ(homepage)，(5) マルチメディア(multimedia)，(6) 入力フォーム(tool)，(7) 検索画面(database entry)の7種類のページタイプに分けている[1]．ページタイプそれぞれに出現している語に着目し，その特徴的な語による分類を試みているが，調査対象の文書数が少ない上，分類区分もWebの実態を反映しているとは言い難い．

一方，松田らはオンラインショップやリンク集といったWebの特徴的なページタイプに該当するもののみを取り出すものであり，実用性を考慮したものとなっている[2]．ページタイプの判定には，ページ内の特徴的なキーワードに加えて，HTMLタグ構造，リンク数，画像サイズなど，スタイル的，構造的なファク

タも用いている．

以上で述べたようなすべての Web 空間では，ページタイプの種類が多すぎて，その多様性に対応することが難しいこともあり，限定した空間の中で指定したタイプに分類する研究もなされている．岡田らは，ニュースサイトに含まれるページを，あらかじめカテゴリ化されている「経済・国際・社会・政治・スポーツ」の 5 つのタイプに分類している．語の出現頻度を用いた決定木によって実現している[3]．

以上のようなページ分類における従来の研究では，ページ全体を対象にし，ページごとに指定したタイプのページであるかを判別している．しかし，本研究で特定する 3.1 節で挙げた情報はページ内の一部に存在し，ページ内には他の情報も含まれている．そのため，従来のようにページ全体を対象にすると余分な情報も含まれるため，そのページを特定することは難しいと考えられる．実際に予備実験として，ページ全体を対象にした判別の精度を検証する．ただし，1 ページ中に複数の情報が存在するため，1 ページごとにどの情報が存在しているかを判別する．

3.2.1 従来の手法：ページ全体を対象にした判別

従来のページ分類の研究で行われているページ全体を対象にした判別の精度を検証するために，判別対象ページを 3.1 節で挙げた情報のうち，「プライバシーポリシー」「商取引法」「支払い方法」の 3 つとし，ショッピングサイト内からこれら 3 つの情報が存在するページを判別できるかを確かめる．まず，3 つの情報が存在するページとそれら以外のページを収集する．判別するために C4.5[4]による決定木を利用する．決定木の独立変数には，指定した情報が存在するページ集合（α）と指定した情報が存在しないページ集合（β），それぞれにおける特徴的な語を用いる．特徴的な語は以下の手順で決定する．

1) ページごとに形態素解析を行う．
2) α，βそれぞれにおいて，形態素ごとに tf・idf（出現頻度の比）を計算する．
3) α，βそれぞれ上位 50 語ずつ，計 100 語を独立変数とする．

2)の各語の tf・idf（出現頻度の比）を計算する際，サ変動詞・形容動詞は語幹が一致する場合，残りの動詞・形容詞は原形が一致する場合は同じ語としてカウントする．また，助詞・助動詞は除去する．そして，tf・idf と出現頻度の比はそれぞれ次の式で求める．

$tf \cdot idf = tf \times idf$

tf：語の出現頻度　　idf：$\log(n/df)+1$

n：すべての文書数　　df：語の出現文書数

$$\text{出現頻度の比} = \frac{f_1}{s_1} \Big/ \frac{f_2}{s_2} \times \log f_1$$

f_1：該当情報のページでの語の出現頻度

f_2：該当情報以外のページでの語の出現頻度

s_1：該当情報のページ全体での語数

s_2：該当情報以外のページ全体での語数

tf・idf では該当情報のページ集合における出現頻度が高い語が上位に含まれ，出現頻度の比では出現頻度よりも該当情報のページ集合とそうでないページ集合の出現頻度の比が高い語が上位に含まれる．この tf・idf と出現頻度の比のどちらの上位語を利用した方が判別に適しているかを比較する．

実験を行うための教師データ数（Table 1）とテストデータ数（Table 2）は以下のとおりである．実験結果を Table 3 に示す．

Table 1 教師データ数

情報のタイプ	該当	該当以外	合計
プライバシーポリシー	519	3627	4146
商取引法	542	3672	4214
支払い方法	600	3334	3934

（単位：ページ）

Table 2 テストデータ数

情報のタイプ	該当	該当以外	合計
プライバシーポリシー	45	85	130
商取引法	26	86	112
支払い方法	68	87	155

（単位：ページ）

Table 3 指定した情報のページの判別精度

対象：ページ全体

X	Y	HIT	FA	MISS	CR	適合率	再現率	F 値
A	a	35	4	10	81	89.7%	77.8%	83.3%
	b	28	7	17	78	80.0%	62.2%	70.0%
B	a	25	25	1	62	50.0%	96.2%	65.8%
	b	26	22	0	65	54.2%	100.0%	70.3%
C	a	57	29	10	58	66.3%	85.1%	74.5%
	b	63	32	4	55	66.3%	94.0%	77.8%

（単位：ページ）

X：情報のタイプ　A：プライバシーポリシー　B：商取引法
C：支払い方法　Y：上位語の求め方　a：tf・idf　b：出現頻度の比
HIT：ユーザ判定○プログラム判定○，FA：ユーザ判定○プログラム判定×，MISS：ユーザ判定×プログラム判定×，CR：ユーザ判定×プログラム判定×
適合率(P)：HIT / (HIT+FA)　　再現率(R)：HIT / (HIT+MISS)
F 値(適合率と再現率の調和平均)：2 / (1 / P + 1 / R)

「プライバシーポリシー」に関しては F 値が約 80%

以上得られたが，まだ MISS が多い．「支払い方法」に関しては F 値が約 70%であるものの，「商取引法」に関しては F 値が 70%を切る結果となってしまい，FA が多い．このようにある程度の精度で判別できるが，まだ正しく判別できていないページが多い．やはり，ページ内には様々な情報が存在し，該当情報以外の余分な情報が含まれてしまうことが原因であり，ページ全体を対象にすると，高精度では判別できない．

3.2.2 提案手法：指定した情報が記述されている部分を対象にした判別

3.2.1 項で述べたとおり，従来と同様の手法ではページ内に存在する情報を高精度で判別することは難しい．そこで，原因となったページ内に含まれる余分な情報に対処するために，ページ内の切り出しを行い，該当情報の範囲を抽出し，切り出した範囲を判別する手法を提案する．ページ内の切り出しとは，ページ中で意味的に構造が区切られる範囲を切り出すことである．このページ内の切り出しが正しくできれば，作成したルールで該当情報のページであるかを判別できると考えられる．そのためにはページ内の意味的な構造を分析する必要がある．我々は検索精度を向上させるために，ページ内の見出し構造と表構造の解析を行ってきた[5][6]．この構造解析を利用することでページ内の指定した情報が記述されている部分の切り出しを行う．

見出し構造を利用した切り出し方法

Web ページ内の見出し構造の解析を利用することでページ内の切り出しを行う．ここでいう見出しとは，

(A) 一行の短い文で書かれており，他の見出しや文，図，表に対し一目で内容が分かるように付けられた標題．

(B) 事柄をいくつかに分けて書き並べている一つ一つ，他の見出しや文，図，表の標題にならないものもある．箇条書き．

と定義されている．Fig.2 では①～⑥が見出しである．これらの見出し間には階層構造が存在する．Fig.2 では①～④が同じ階層であり，⑤，⑥は①の下位階層である（逆に言えば①は⑤，⑥の上位階層）．そして，見出しの支配範囲とはある見出しから同じ階層もしくは上位の階層の見出しまでのことであり，Fig.2 では，見出し①～⑤の支配範囲はそれぞれ①′～⑤′である．この階層構造を利用し，ページ内の切り出しを行う．

しかし，この見出し構造の解析では（B）の「箇条書き」を見出しとしているが，「箇条書き」には支配範囲が存在しない．見出しだけに指定した情報が記述されているケースはほとんどないと考えられるため，本研究ではこの「箇条書き」のような支配範囲が存在しない見出しを対象外とする．また，画像で表示されている見出しを扱っていないが，画像を利用して表示されている見出しも多く存在する．そこで，画像の見出しを扱うために以下の戦略を追加する．

- 画像が表示されない場合のテキストを指定している．（例．<img alt="テキスト"）
- テキストの回り込みを指定していない．
- リンクになっていない．

この 3 点の条件を満たす画像を見出しとして抽出する．また画像の見出しの支配範囲を判定するために，(ア)画像の見出し同士，(イ)画像の見出しとテキストの見出し，における階層構造を判定する必要がある．このために，以下の戦略を追加した．(ア)における階層構造を判定するために，以下の条件を比較する．

- 画像の幅
- 画像の高さ
- 行内での画像の位置の指定（左・中・右）

(イ)においては画像の見出しを上位階層とする．ただし，インデントに差が存在する場合のみ画像の見出しを下位階層とする．

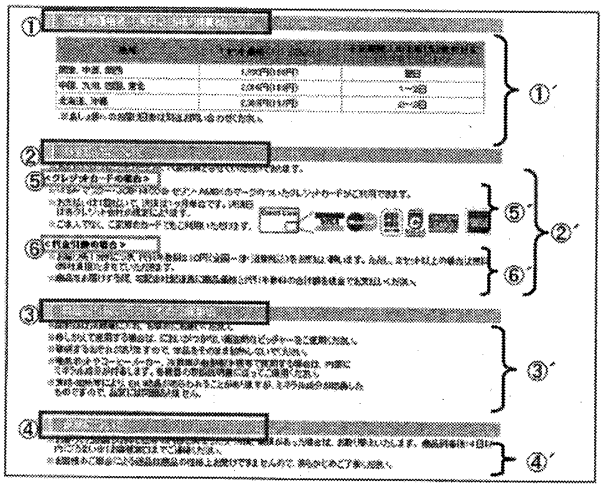

Fig.2 Web ページ内の見出し構造

表構造を利用した切り出し方法

Web ページ内の＜TABLE＞タグで記述されている表構造の解析を利用することでページ内の切り出しを行う．Fig.3 では表内の見出しが①～⑤であり，その見出しに対応するセル，つまり支配範囲がそれぞれ①′～⑤′である．この関係を利用してページ内の切り出しを行う．

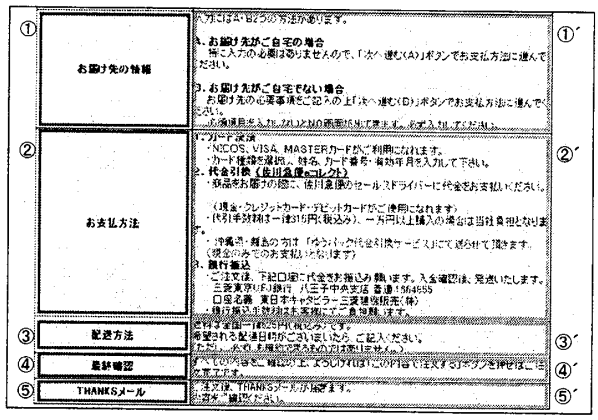

Fig.3 Web ページ内の表構造

上記の見出し構造・表構造の解析を利用してページ内の切り出しを行い，切り出した範囲の判別を行うが，指定した情報の見出し自体にも特徴的な語が存在する傾向が見られる．そのため，見出しに含まれる語を用いて見出しを限定する．また，この見出しに含まれる特徴的な語を用いて見出しを限定することで，取り除いてしまうページがないようにするため，50 箇所すべての見出しを取り出せる語を限定する語として採用する．

この 2 つの手法を用いてページ内の切り出しを行った際，指定した情報の範囲を正しく切り出している範囲があるかを検証する．この精度を図るために，対象の情報タイプを「プライバシーポリシー」の記述があるページとし，「プライバシーポリシー」の記述範囲を正しく切り出すことができているかを確認する．

まず，見出しを限定する語を決定するために，「プライバシーポリシー」を含むページ 50 ページに対し，見出しに含まれる語を人手で抽出し，その見出しすべてを含む語を決定する．その結果「個人情報」「プライバシー」「Privacy」のいずれかが見出しに含まれていることがわかった．

次に見出しの語を調査したページとは別の「プライバシーポリシー」の記述がある 63 ページに対し，見出し構造判定プログラムと表構造判定プログラムを適用し，精度を確認する．以下にアルゴリズムを示す．

1) 「プライバシーポリシー」が存在するページの見出しの特徴的な語「個人情報」「プライバシー」「Privacy」が Web ページ中の見出しに含まれるか判定する．
2) 1) で含まれていると判定された見出しの支配範囲に含まれる部分を抽出する．
3) 抽出した部分を目視で確認する．

この実験では 46 ページが正しく抽出でき（73.0%）た．このうち表構造の解析を利用して抽出できたページは 3／46 ページである．10 ページは見出しが取り出せず，7 ページは支配範囲が間違っている，という結果になった．

以上で述べた見出し構造・表構造の解析の他に，入力フォームを示すタグを利用してページ内の切り出しも行う．Web ページには入力フォームの範囲を示すタグ（<form>）が存在する．「注文フォーム・会員登録・ログイン・メールマガジン・問い合わせ・掲示板」にはユーザが入力するフォームが存在し，このタグを用いて範囲を区切っているケースが多い．このため，この入力フォームの範囲を示すタグを利用したページ内の切り出しも行う．

決定木を作成する際の教師データは，人手で指定した情報の記述部分を抽出したデータとする．対象ファイルを判別する際は，上記アルゴリズムを実現したプログラムによってページの切り出しを行い，切り出した範囲を作成したルールによって判別する．

4 試作システム

ユーザにより必要な情報が異なるという点に対応すべく，自由に適応型サイトマップを作成できるインタフェース（Fig.4）を構築する．このインタフェースでは，3.1 節で挙げた情報からユーザが欲しい情報を選択できる．また，自由に中間ノードを作成し，カテゴリ化することができる．

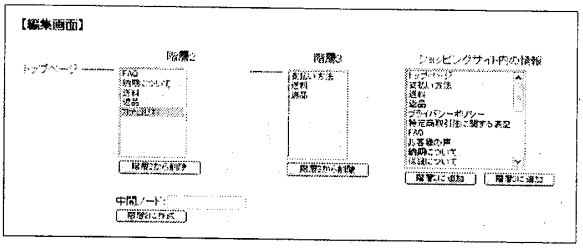

Fig.4 ユーザインタフェース

このインタフェースでユーザが作成したサイトマップをそれぞれのサイトで作成する．そして Web 検索エンジンの検索結果における各サイトのサマリーの下に作成したサイトマップへのリンクを提示する．そのリンクをたどることでサイトマップが表示される．オンラインショッピングサイトに適用した結果を Fig.5 に示す．このサイトマップでは，ユーザが指定した情報

を特定できると，そのページへのリンクとなる．

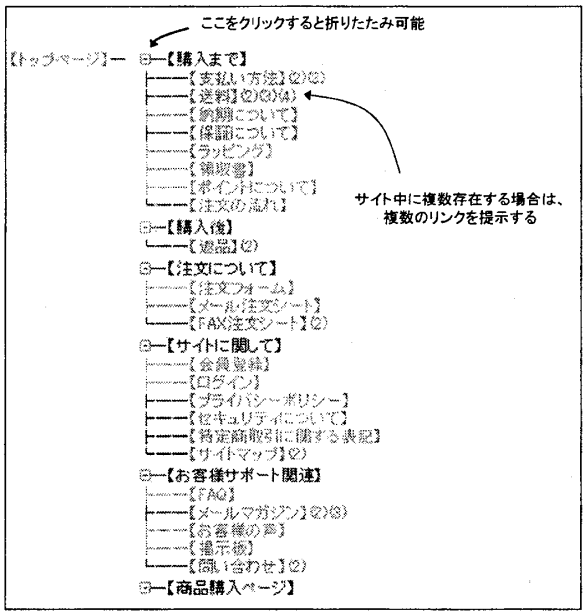

Fig.5 作成したサイトマップ

5 フィルタの評価

5.1 評価実験

提案手法の有効性を検証するために，3.2.2 項のテストで使用したデータと同様のデータを用いて比較する．ただし，該当情報の学習データは目視で確認し，該当情報が記述されている部分を抽出する．テストでは，見出し構造・表構造の解析と入力フォームの範囲を示すタグを利用しページ内の切り出しを行う．切り出した範囲を決定木によって判別する．実験結果を Table 4 に示す．

Table 4 指定した情報のページの判別精度

対象：情報が記述されている部分

X	Y	HIT	FA	MISS	CR	適合率	再現率	F 値
A	a	41	9	4	66	82.0%	91.1%	86.3
	b	28	1	17	74	96.6%	62.2%	75.7
B	a	20	0	6	78	100.0%	76.9%	87.0
	b	22	0	4	78	100.0%	84.6%	91.7
C	a	48	3	20	85	94.1%	70.6%	80.7
	b	49	4	19	84	92.5%	72.1%	81.0

（単位：ページ）

X：情報のタイプ　A：プライバシーポリシー　B：商取引法
C：支払い方法　Y：上位語の求め方　a：tf・idf　b：出現頻度の比

Table 4 のように対象を情報が記述されている範囲にすると，3 つの情報のタイプについては，ページ全体を対象にした判別よりも高精度で判別することができた．その他の情報のタイプでも提案手法を用いて高精度で判別できるか否かを検証する．

「プライバシーポリシー」「商取引法」「支払い方法」以外の情報のタイプも同様の手法で判別できるかを検証するために，教師データを収集する必要がある．C4.5 を利用し，ニュースサイトのページの見出しと本文だけを取り出す研究[7]では，分類するタイプの教師データ数は，タイプごとに均等ではなく，存在した数を採用している．本研究でもそれぞれの情報のタイプの教師データ数を均等ではなく，数千のショッピングサイトから収集できた数を採用する．

ただし，「注文フォーム・会員登録・ログイン・メールマガジン・問い合わせ・掲示板」については，入力フォームが存在するという特徴を持つ．そのため，独立変数として，tf・idf（出現頻度の比）の上位語だけでなく，入力フォームを示すタグとその属性を加える．また「サイトマップ」では内部リンクの集合であるため，内部リンク数を独立変数に追加する．その他の情報については，特徴的な語だけを用いるが，「会員登録」が存在する部分の上位 50 語は，「静岡県」「愛知県」などの都道府県名がほとんどを占めていたため，上位 100 語を採用した．これらを独立変数とし，判別精度を計算した（Table 5，Fig.6）．

Table 5 では，tf・idf（出現頻度の比）を利用した判別の適合率・再現率・F 値を示しているが，適合率は平均で 94.4%（94.4%）以上と高い値であるが，再現率は平均で 73.8%（63.0%）であった．適合率と再現率の調和平均である F 値の平均は 81.5%（74.1%）となった．適合率は同じ値であるが，再現率に差があり，tf・idf の F 値の方が 9.2%高くなった．Fig.6 は，それぞれの情報のタイプの F 値のグラフである．tf・idf の上位語を用いた判別と出現頻度の比を用いた判別での精度を比較すると，tf・idf の方が安定して高い精度で判別できたことがわかる．

Table 5 の tf・idf による判別の精度を情報のタイプごとに確認すると，「E メール注文」「お客様の声」については再現率が 50%を切り，「メールマガジン」については再現率が 52.0%となり，この 3 つでは F 値が 70%以下となってしまった．それ以外については F 値 70%以上と，高い精度で特定することができた．

5.2 評価実験の考察

5.2.1 tf・idf と出現頻度の比の差

本手法で指定した情報のタイプ，23 タイプのうち，14 タイプでは tf・idf を利用した判別の精度の方が高

Table 5 適合率・再現率・精度

情報のタイプ	tf・idf			出現頻度の比		
	適合率	再現率	F 値	適合率	再現率	F 値
D	92.9%	96.3%	94.5%	100.0%	85.2%	92.0%
E	100.0%	73.1%	84.4%	100.0%	73.1%	84.4%
F	84.6%	91.7%	88.0%	75.0%	25.0%	37.5%
G	91.3%	87.5%	89.4%	91.7%	91.7%	91.7%
H	100.0%	87.0%	93.0%	100.0%	91.3%	95.5%
I	100.0%	88.2%	93.8%	100.0%	47.1%	64.0%
J	88.9%	36.4%	51.6%	88.9%	36.4%	51.6%
K	100.0%	58.3%	73.7%	100.0%	41.7%	58.8%
L	100.0%	62.9%	77.2%	100.0%	65.7%	79.3%
M	100.0%	48.0%	64.9%	100.0%	64.0%	78.0%
N	100.0%	95.1%	97.5%	100.0%	78.0%	87.7%
O	100.0%	80.0%	88.9%	100.0%	44.0%	61.1%
P	92.9%	52.0%	66.7%	88.2%	60.0%	71.4%
Q	93.3%	66.7%	77.8%	100.0%	61.9%	76.5%
R	90.6%	70.7%	79.5%	92.6%	61.0%	73.5%
S	76.6%	87.8%	81.8%	93.1%	65.9%	77.1%
T	94.9%	84.8%	89.6%	95.0%	86.4%	90.5%
U	100.0%	57.1%	72.7%	82.1%	54.8%	65.7%
V	90.5%	61.3%	73.1%	85.7%	38.7%	53.3%
W	92.0%	92.0%	92.0%	95.7%	88.0%	91.7%
平均	94.4%	73.8%	81.5%	94.4%	63.0%	74.1%

Table 6 対応表

記号	情報のタイプ	記号	情報のタイプ	記号	情報のタイプ
D	ポイント	K	FAX 注文シート	R	セキュリティ
E	掲示板	L	FAQ	S	送料
F	ラッピング	M	お客様の声	T	返品
G	保証	N	ログイン	U	納期
H	領収書	O	会員登録	V	注文の流れ
I	注文フォーム	P	メールマガジン	W	問い合わせ
J	E メール注文	Q	サイトマップ		

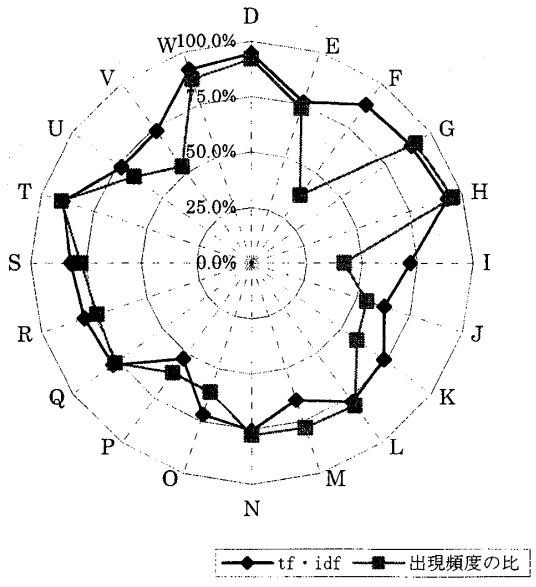

Fig.6 情報のタイプ別の F 値

くなった．残り 9 のタイプの情報では出現頻度の比を利用した判別の方が高い精度を示したものの，tf・idf を利用した判別精度と大きな差はない．つまり，該当情報のページにおける頻度が高い語を利用した方が判別には適しているということが確認できた．しかし，「お客様の声」の判別に関しては，出現頻度の比を利用した判別の精度の方が明らかに高い．「お客様の声」のページでは，頻度は少ないが，そのページにだけ存在する語が「お客様の声」のページの判別には有効である，ということになる．そこで，tf・idf と出現頻度の比の 2 つの上位語を組み合わせて判別を試みたが，F 値 68.4%と 2 つの間の精度となった．このように，独立変数の語によって精度が変化するため，どの語を独立変数として使用すると精度が高くなるかという点で他の手法も検討したい．

5.2.2 取りこぼしの要因

本手法で適合ページを取りこぼす主な要因を調査したところ，以下に挙げるケースに該当する事例が多く見受けられた．

これはページ内を正しく切り出すことができていないことが要因である．本手法では，見出し構造・表構造・入力フォームの範囲を示すタグ（<form>）の 3 つでページの切り出しを行っているが，切り出す際に余分な範囲を加えてしまう，また切り出す範囲が少ない，このために決定木で取りこぼす要因となっている．表構造では，ページ内の切り出しはほとんど正しくできているが，他 2 つの手法では正しく切り出すことができていないケースが存在した．

見出し構造の解析では，見出しと見出しの階層構造を判定しているが，正しく階層構造を判定できていないために，範囲の超過・不足となっていた．現在，ページ全体に対して見出しの階層構造を解析していて，サイトメニューなど，本文とは意味的な関係がない部分も解析している．また，HTML 文書の記述方法には多様性があるため，対応できていないページも存在した．これらを改善することで正しく切り出すことができる．

入力フォームの範囲を示すタグ（<form>）では，ページ製作者によって正しい範囲をこのタグで区切っていない場合があるため，余分な範囲を含めてしまっているケースがあった．

また，見出し構造・表構造では対応できないページも存在した．これは画像が見出しとなっていて，alt タグによる代わりのテキストを指定していないケース

である．テキスト情報が得られないために，意味的な構造を解析することができない．今後，画像解析に踏み込んだ解析を行う必要がある．

5.2.3 精度向上の検証

教師データ数を変化させることで精度を向上させることができるか検証した．

高須らは教師データ数を変化させたときの決定木による判別精度を検証し，教師データ数を増やすほど判別精度は向上することを実験から確認している[8]．そこで，本手法においても，教師データ数の変化によって判別精度が変化するかを検証した．対象の情報のタイプを「支払い方法」とし，教師データの「支払い方法」のページ数を 200, 400, 600 と変化させたところ，精度向上がみられた（Table 7）．

以上のように，教師データ数を増加させることで精度向上を確認することができた．再現率が低い情報のタイプでは，該当情報ページの教師データ数が少ないため，このように増加させることで精度が改善することができると考えられる．

Table 7 教師データ数を変化させたときの判別精度

Z	Y	HIT	FA	MISS	CR	適合率	再現率	F 値
200	a	40	3	28	85	93.0%	58.8%	72.1%
	b	29	1	39	87	96.7%	42.6%	59.2%
400	a	47	3	21	85	94.0%	69.1%	79.7%
	b	44	3	24	85	93.6%	64.7%	76.5%
600	a	48	3	20	85	94.1%	70.6%	80.7%
	b	49	4	19	84	92.5%	72.1%	81.0%

Y：上位語の求め方　a：tf・idf　b：出現頻度の比
Z：教師データの該当情報のページ数

6 おわりに

本稿では，適応型サイトマップを作成するための Web ページ内の一部に存在する情報のページを特定する手法について議論を行った．

Web ページ全体を対象に判別を行うと，ページの一部に存在する意味的な情報のページを特定することはできない．我々の先行研究である見出し構造・表構造の解析と入力フォームの範囲を示すタグを利用し，ページの切り出しを行い，切り出した部分を判別するという手法を提案した．さらに，これを評価するために，ページ全体を対象にした判別と精度を比較したところ，精度向上がみられ，本手法の有効性を示すことができた．

今後は 5.2 節，5.3 節で述べた点を見直し，更なる精度向上を目指したいと考えている．そして，本稿で提案したシステム全体を，ユーザ満足度など，ユーザの視点での評価を行いたいと考えている．

参考文献

[1] Haas, S.W., Grams, E.S. Readers, Authors, and Page Structure:A Discussion of Four Questions Arising from a Content Analysis of Web Pages. JASIS. Vol.51, No.2, p.181-192(2000)

[2] 松田勝志，福島俊一：文書タイプ分類による問題解決向き WWW 検索システムの開発と評価，情報処理学会研究報告，1998-FI-53，vol.99，No.20，pp.9-16（1999）

[3] 岡田望，西園敏弘，古畑裕介：文書分類における決定木アルゴリズム適用法の検討，電子情報通信学会オフィスインフォメーションシステム研究会，OIS2004-80，pp.97-102（2005）

[4] J. R. Quinlan. C4.5： Programs for Machine Learning. Morgan Kaufmann, CA（1993）

[5] 西口直樹，松本章代，小西達裕，高木朗，小山照夫，三宅芳雄，伊東幸宏：見出しの階層関係を利用した WWW 検索精度の改善，電子情報通信学会技術研究報告，NLC2005-114～125, pp.1-6（2006）

[6] 松本章代，西口直樹，小西達裕，高木朗，小山照夫，三宅芳雄，伊東幸宏：表構造における意味的関係に基づく WWW 検索精度の向上，情報処理学会研究報告，2006-DD-55，Vol.2006，No.58，pp.5-11（2006）

[7] 新納浩幸，佐々木稔：Web ページ内の目的部分の自動抽出，情報処理学会研究報告，2004-NL-162，Vol.2004，No.73，pp.33-40（2004）

[8] 高須淳宏，相原健郎：テキスト分類における訓練データと性能の実験的考察，Nii Journal No.6(2003)