情 報 学 基 礎 57−19
自 然 言 語 処 理 136−19
（2000. 3. 22）

Webページからのタイプ別情報抽出・分類方式

山田洋志　福島俊一　松田勝志

{h-yamada, fuku, mat}@hml.cl.nec.co.jp

NEC ヒューマンメディア研究所

ユーザの目的に応じた情報検索・情報提供を実現するために,ページタイプ分類を利用した情報抽出・分類方式を提案し,試作システムで精度を評価した. 本方式では,ページタイプ分類を使用することで,それぞれのページタイプに適した抽出・分類処理を行える. そのため,多くの種類の文書が混在するWebページに対しても必要な情報だけを高い精度で抽出することができる. また,分類結果を表や図を用いてユーザに提示することで特定の目的に応じた情報を提供するサービスを実現できる.

試作システムによる求人情報とイベント情報の抽出精度の評価では,記述が比較的一定している勤務地や開催日時などの情報で適合率90%以上を達成した. 記述の自由度が高い,求人の応募資格やイベント名では65–75%にとどまっている. 検索誤りの主な原因は,情報を判別するキーワードのミスマッチと,表や箇条書きの前後からの抽出の誤りであった. 抽出もれの原因としては,表や箇条書きのパターンや情報判別用のキーワードのの不足が主なものであった.

Information Extraction and Classification Method for Web Pages based on Document Type

Hiroshi YAMADA　Toshikazu FUKUSHIMA　Katsushi MATSUDA

NEC Human Media Research Laboratories

This paper describes a novel information extraction method which realizes task oriented information retrieval. This method uses page-type classification method which judges type of Web pages. Introducing the page-type concept, extraction systems can select appropriate algorithm or rules for the target page-type. Hence, extraction performances will be increased.

This paper also demonstrates accuracy of extraction. Extraction precisions of work place at help-wanted advertisement and date at event information are 90% and over, since these information are relatively set. Precisions of requirement at help-wanted advertisement and event name at event information are unfortunately between 65% and 75%, because there are many description styles of these information. The causes of the extraction errors is mismatches of keywords and extracting errors from table captions. The extraction omissions are caused by lack of pattern, namely, table, article, keyword and so on.

1 はじめに

ここ数年WWWが急速に普及し,だれもが利用できる情報源としての地位を固めている. 一方で,Webページの増加によって求める情報を選び出すことが難しくなり,さまざまな検索技術が開発されている.

筆者らは,ユーザがWWWを利用する目的に着目し,目的に役立つ情報を選択するため,ページタイプという概念を提案し,その判別方式を開発した[1, 2, 3]. ページタイプはWebページの内容や分野でなく,ページの種類や形式を示している. 現在,判別できるページタイプとしては,カタログページ,リンク集,プレゼントページなどがある. たとえば,購入を目的とした製品情報の収集では,製品名だけではなくページタイプが「カタログ」であるという検索条件を使うことで検索結果を絞り込める.

ページタイプ判別を利用すると,単語だけを使った検索と比較して不要な検索結果を大幅に削減できる. さらに,検索結果に対して,重要な情報を抽出して一覧表示したり,分類して表示したりすることで検索結

果の一覧性を高め,また,より詳細な絞り込みができる.その際,抽出する情報や分類の基準はページタイプによって異なっている.たとえば,求人情報ならば勤務地や職種,プレゼントならば締め切りや応募方法を抽出あるいは分類したい.ページタイプを利用すると各タイプごとに処理方法を変えられるため,必要な情報だけを正確に抽出・分類できる.

本稿では,各ページタイプに属するWebページ中から,キーとなる情報を抽出・分類する方式について述べる.さらに,求人情報,イベント情報のふたつのページタイプを対象にした試作システムと情報抽出精度について述べる.

2 ページタイプの概要

ユーザがWWWを使用する目的に応じて検索結果を絞り込むことができれば,必要な情報を見つけ出すことが容易になる.単語やキーワードを指定した検索は,ページのテーマや分野をある程度絞り込むことができるが,検索結果の中には商品の仕様,ユーザによる感想,日記の一部などが混在する.たとえば,商品の購入を考えているユーザであれば最初に必要になるのは仕様を記述したカタログであり,その後,類似商品の比較や購入店の選択を行う.そのようなタスクに応じた情報を分けて検索するために,筆者らはページタイプの判別方式を開発した[1, 2, 3].

ページタイプは内容や分野ではなく形式の種類やページの目的を示す.たとえば,同じ単語を含んでいてもカタログやリンク集というタイプの違いが生じる.逆に,同じタイプのWebページであってもテーマや分野は異なっている.ページタイプによる分類・検索は,分野によるディレクトリサービスや単語検索の機能と相補的なもので,両者を組み合わせることで,ユーザの目的とする情報をより的確に絞り込むことができる.

ページタイプの判別には,単語やタグの種類,タグ内の属性値,ファイルサイズ,リンクや画像の有無などを総合的に判定して数値化する.たとえば,カタログページの判定には以下の条件を使用している.

- 特定の単語を含む: 商品,サービス,製品,お客さま,問い合わせ,価格,仕様,特長
- co.jpドメインで,URLに'product'という文字列を含む.
- $<TABLE>$タグを使用している.
- ドメイン内へのリンクが多く,ドメイン外へのリンクが少ない.

これらの条件に重みをつけ,各ページがいくつを満たしているかで「カタログらしさ」を数値化する.

よく利用され実用的な判別精度が達成できそうなものから判別条件を作成し,すでに製品カタログ,リンク集,掲示板・チャット,プレゼント,調査報告,求人情報,イベントの各タイプの判別を実現している.

3 ページタイプを利用した情報抽出・分類方式

ページタイプを利用して検索結果が得られたとき,ユーザにとってキーとなる情報や最適な分類軸はそれぞれのユーザの目的によって更に細分化される.たとえば,求人情報であれば,勤務地や職種,給与,応募資格など,イベントであれば開催日や開催地に注目さらには分類することがユーザの目的にかなっている.

従来,このような形で情報を提供するサービスの実現のために,情報提供する企業や店舗あるいはポータルサイト側が情報の抽出・整理コスト(人手)を負担している.今後の情報量の増大や作業のコストを考えると情報収集・抽出の自動化が望ましいが,従来の情報抽出技術で十分な精度を得ることは難しい.5W1Hのような構文を利用した情報抽出や,地名や日付などの固有情報抽出の技術を未整理のWebページ集合に適用すると,不要な情報が大量に抽出される.

本稿で提案する情報抽出・分類方式では,(1)ページタイプによるWebページの分類・選別,(2)ページタイプに合わせた情報抽出,(3)抽出した情報およびWebページの分類,の3段階で処理を行う.これにより高精度の情報抽出・分類を実現する.

ページタイプ分類 ロボットなどで収集したWebページのタイプを判定し,必要なページを選別する.

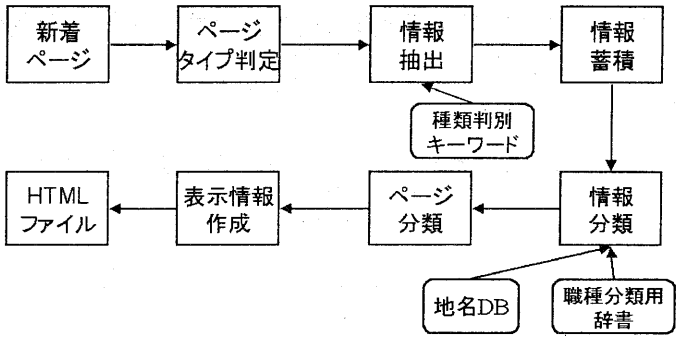


図 1: 処理概要

情報抽出 各ページタイプ別にキー情報を抽出する．すでにページタイプが判別されているため抽出方法はタイプごとに選択することができる．そのため，たとえば，同じ地名であっても求人情報の勤務地とイベント情報の開催地では別の抽出方法を選択することができる．また，同じページ中にそれ以外の地名が記述されていても抽出しないことができる．このように，タイプ別の情報抽出というアプローチを取ることで情報抽出の精度を高めることが可能になる．

情報分類・ページ分類 各ページタイプのキー情報の間の関係に関する知識を利用して分類する．たとえば，日付を分類する場合，プレゼントの締め切りであれば，間際の日付を細かく分類し，先のものは月ごとに分類するなどの処理ができる．次いで，抽出したキー情報の分類結果に応じてWebページを分類する．キー情報の分類をそのままページの分類とする，あるいは，キー情報の組み合わせで分類するなどの方法がある．

4 試作システム

求人情報とイベント情報を対象として，ページタイプ別情報抽出・分類方式を利用したサービスを試作した．ロボットが集めたページから求人情報とイベント情報を取り出し，勤務地や開催月ごとにまとめて表示する．以下，本システムの処理概要を述べる (図1)．

ページタイプ判定 ロボットで収集したページから「求人情報」，「イベント」タイプのページを選別する．

情報抽出 各タイプから以下の情報を抽出する．

- 求人情報：勤務地，職種，資格
- イベント：開催地，日時

抽出はページ内の表あるいは箇条書き箇所を対象とし，情報の種類の判別にはキーワードを利用する．詳細は4.1節で述べる．

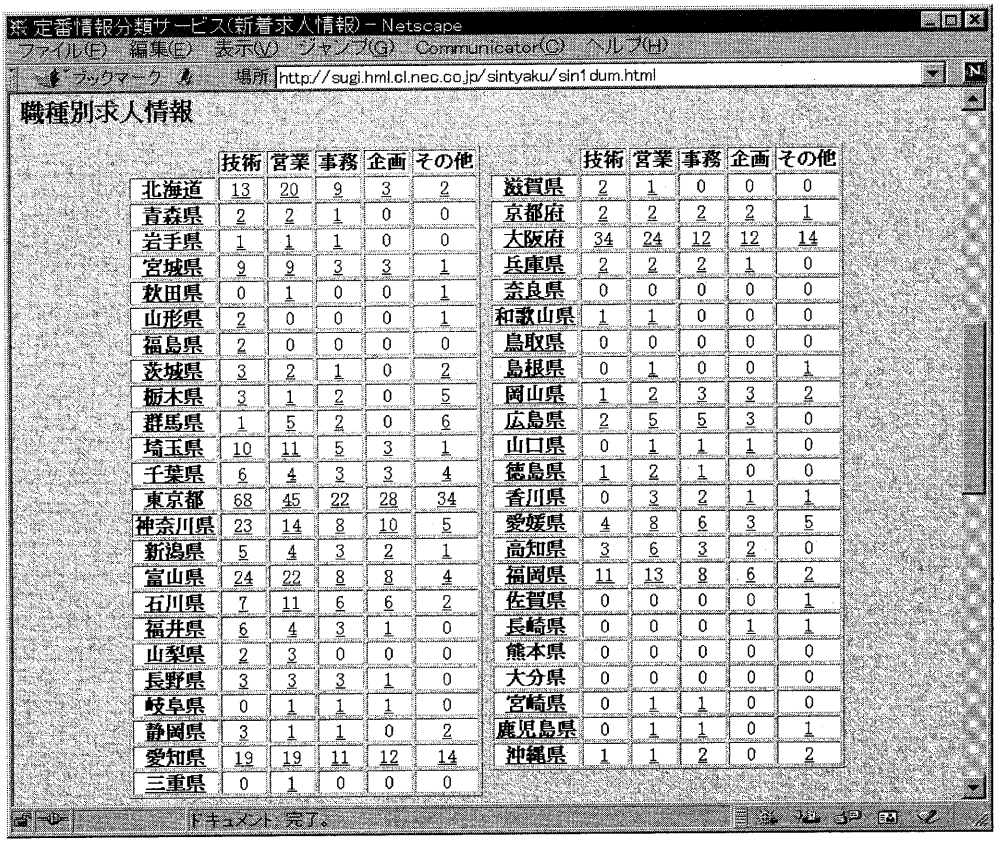
情報蓄積 抽出したデータをデータベースに追加する．また，古い情報や重複を削除する．

情報分類 地名は都道府県別，日付は月ごとに分ける．分類用の知識として，地名を都道府県に結びつけるための地名DBと職種分類用の辞書を使用した．詳細は4.2節で述べる．

ページ分類 抽出した情報の分類結果でWebページを分類する．求人情報は，勤務地，職種，勤務地と職種の組み合わせ別に分類する．イベント情報は，開催地，開催月別に分類する．

表示情報の作成 ページ分類結果から表示用のWebページを作成する．図2は求人情報を勤務地 (都道府県別) と職種の組み合わせで分類したインデックス，図3，図4は，イベント情報を開催月と開催地で分類したインデックスである．都道府県別の求人情報表示は図3と同じ形式で作成した．

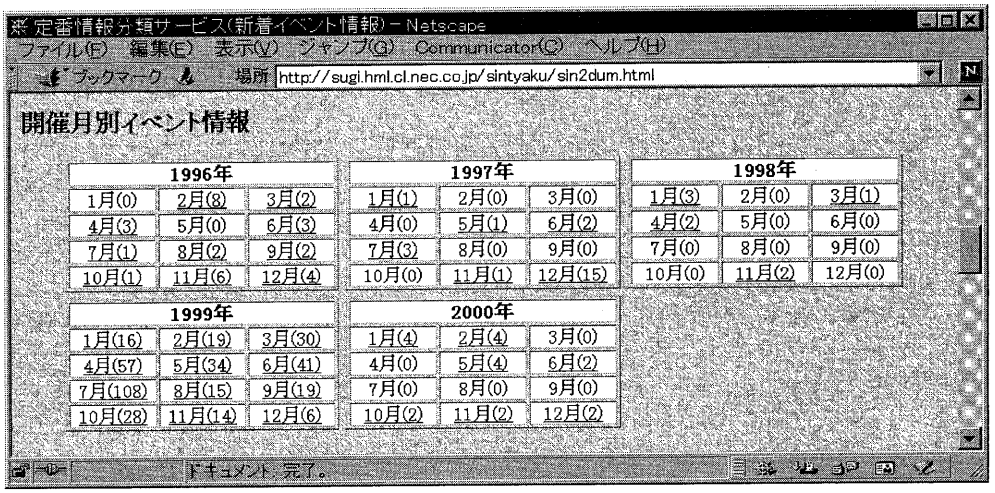
これらのインデックスから見たい分類項目のリンクやマップを選択すると情報の概要を記述した一覧が表示される．さらにそこからオリジナルのページのリンクをたどることができる．



	技術	営業	事務	企画	その他
北海道	13	20	9	3	2
青森県	2	2	1	0	0
岩手県	1	1	1	0	0
宮城県	9	9	3	3	1
秋田県	0	1	0	0	1
山形県	2	0	0	0	1
福島県	2	0	0	0	0
茨城県	3	2	1	0	2
栃木県	3	1	2	0	5
群馬県	1	5	2	0	6
埼玉県	10	11	5	3	1
千葉県	6	4	3	3	4
東京都	68	45	22	28	34
神奈川県	23	14	8	10	5
新潟県	5	4	3	2	1
富山県	24	22	8	8	4
石川県	7	11	6	6	2
福井県	6	4	3	1	0
山梨県	2	3	0	0	0
長野県	3	3	3	1	0
岐阜県	0	1	1	1	0
静岡県	3	1	1	0	2
愛知県	19	19	11	12	14
三重県	0	1	0	0	0

	技術	営業	事務	企画	その他
滋賀県	2	1	0	0	0
京都府	2	2	2	2	1
大阪府	34	24	12	12	14
兵庫県	2	2	2	1	0
奈良県	0	0	0	0	0
和歌山県	1	1	0	0	0
鳥取県	0	0	0	0	0
島根県	0	1	0	0	1
岡山県	1	2	3	3	2
広島県	2	5	5	3	0
山口県	0	1	1	1	0
徳島県	1	2	1	0	0
香川県	0	3	2	1	1
愛媛県	4	8	6	3	5
高知県	3	6	3	2	0
福岡県	11	13	8	6	2
佐賀県	0	0	0	0	1
長崎県	0	0	0	1	1
熊本県	0	0	0	0	0
大分県	0	0	0	0	0
宮崎県	0	1	1	0	0
鹿児島県	0	1	1	0	1
沖縄県	1	1	2	0	2

図 2: 求人情報分類(地域×職種別)



1996年		
1月(0)	2月(8)	3月(2)
4月(3)	5月(0)	6月(3)
7月(1)	8月(2)	9月(2)
10月(1)	11月(6)	12月(4)

1997年		
1月(1)	2月(0)	3月(0)
4月(0)	5月(1)	6月(2)
7月(3)	8月(0)	9月(0)
10月(0)	11月(1)	12月(15)

1998年		
1月(3)	2月(0)	3月(1)
4月(2)	5月(0)	6月(0)
7月(0)	8月(0)	9月(0)
10月(0)	11月(2)	12月(0)

1999年		
1月(16)	2月(19)	3月(30)
4月(57)	5月(34)	6月(41)
7月(108)	8月(15)	9月(19)
10月(28)	11月(14)	12月(6)

2000年		
1月(4)	2月(4)	3月(0)
4月(0)	5月(4)	6月(2)
7月(0)	8月(0)	9月(0)
10月(2)	11月(2)	12月(2)

図 3: イベント情報分類(開催年月別)

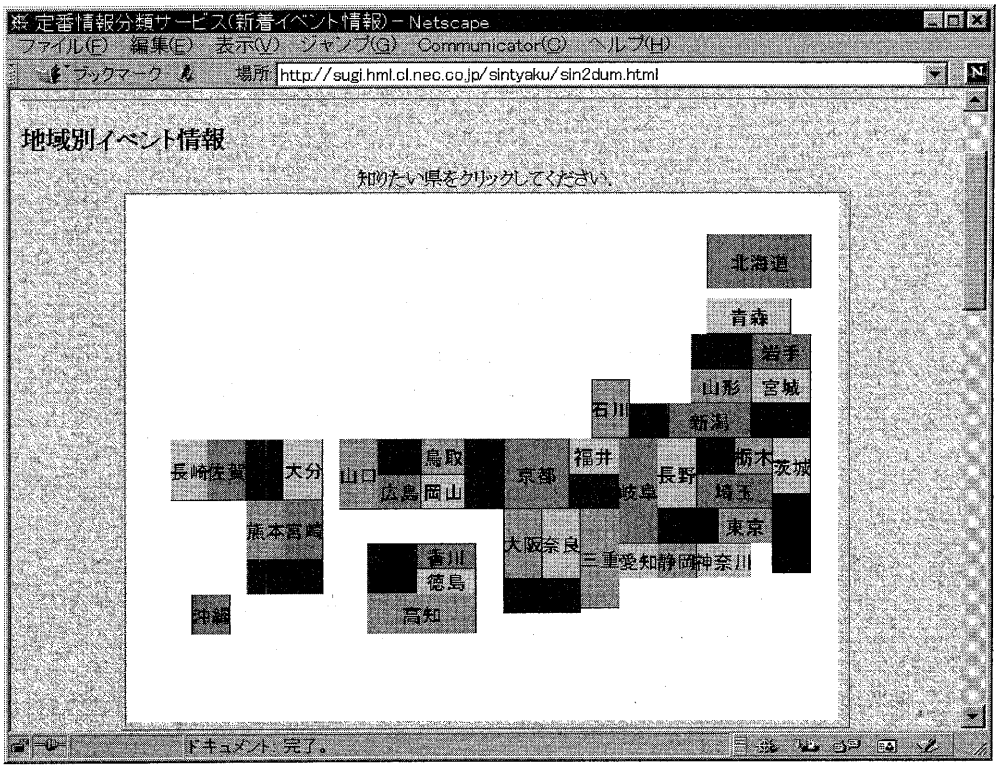


図 4: イベント情報分類(開催地別)

■表形式1

名称	○○大会
開催日	1999年9月20日
場所	川崎市××ホール

■表形式2

名称	開催日	場所
○○大会	1999年9月20日	川崎市××ホール
△△コンサート	1999年9月22日	横浜市☆☆公民館

■箇条書き形式1

- 名称: ○○大会
- 開催日: 1999年9月20日
- 場所: 川崎市××ホール

■箇条書き形式2

【名称】 ○○大会
【開催日】 1999年9月20日
【場所】 川崎市××ホール

図 5: 情報抽出の対象とする記述形式

4.1 情報抽出の詳細

本節では情報抽出の方法について説明する.

情報を正確に抽出するために,ページ中の表と箇条書きを抽出対象とした(図5). 表は,HTMLの<$TABLE$>タグを使った箇所を対象とする. 箇条書きは,<UL>などの箇条書き用のタグのほか,単なるテキストで先頭に記号(「・」など)が付いているものも対象にした.

ページタイプに合わせたキーワード(表1)を見出しとする箇所を抽出する. 「求人情報」の'勤務地'と「イベント」の'開催地'はいずれも地名であるが,キーワードで区別する.

抽出の例外として,「イベント」の'名称'が表あるいは箇条書きの中に書かれていない場合は,直前の強調された文字列(<H>タグ,<B>タグ)を抽出した.

表 1: 情報抽出用キーワード

求人情報			イベント		
情報	数	キーワードの例	情報	数	キーワードの例
勤務地	2	勤務地,……	開催年月日	8	日時, 開催月日,……
資格	5	条件, 資格,……	開催地	4	場所, 開催地,……
職種	3	職種,……	イベント名	6	名前, 名称,……

表 2: 職種分類用キーワード

職種	数	キーワード	職種	数	キーワード
営業	2	営業, 販売	技術	13	技術, 製造,SE,……
事務	9	事務, 人事,……	専門	7	薬剤師, 弁護士,……
企画	8	企画, 広告,……			

4.2 情報分類の詳細

本節では抽出した情報の分類について述べる. 分類のための知識として地名DBと職種分類用辞書を使用した. 地名DBは, 市区町名がどの都道府県に属するかの対応関係を登録している. 同名の地名がある場合は登録していない. 職種分類用辞書には各職種に対応するキーワードを列挙する (表2).

'勤務地' および '開催地' は, 都道府県ごとに分類する. 都道府県名と抽出したテキストとの照合を行う. 都道府県名が含まれていない場合 (市から書いてある場合など) は, テキストの先頭部分を地名DBと照合して都道府県ごとに分類する.

「求人情報」の '職種' は, 分類用辞書 (キーワード集) を参照して5種類に分類する. 分類用辞書は, 求人情報ページの職種部分から手作業で抽出して作成した.

表3に求人情報とイベントページから抽出, 分類したデータの例を示す.

5 情報抽出精度

抽出精度を評価した. ロボットで収集したWebページ約60万件からページタイプが「求人情報」と「イベント」と判定されたページを取りだし, そこから情報を抽出した.

求人情報

種類	抽出したテキスト	分類
勤務地	札幌事業所: JR・地下鉄新札幌駅より徒歩5分	北海道
職種	ネットワークエンジニア	技術
資格	40歳以下の経験者	(分類せず)

イベント

開催日	平成11年7月4日	1999/07
名称	郷土芸能等披露、表彰式	(分類せず)
開催地	瀬戸内町特設会場	鹿児島県

表 3: 情報抽出・分類の例

表 4: 求人情報抽出精度

	情報数	正解	誤り	もれ	再現率	適合率
勤務地	224	129	0	95	57.6	100.0
職種	261	124	10	127	47.5	92.5
資格	231	72	39	120	31.2	64.9

表 5: イベント情報抽出精度

	情報数	正解	誤り	もれ	再現率	適合率
開催地	69	44	4	21	63.8	91.7
日時	69	43	4	22	62.3	91.5
名称	70	33	11	26	47.1	75.0

5.1 求人情報

「求人情報」は929ページ見つかったが,そのうち収集した時間順に300ページを調査した. ページタイプ判定の誤りにより,求人情報以外のページが6ページあったため,294ページを抽出精度の評価対象とした. これら294ページから抽出するべき情報を人手で判定した. '勤務場所','職種','資格'のそれぞれについて,抽出ページ数と正誤数を表4に示す. 表4で,「正解」は正しく情報が抽出できたページ数,「誤り」はページ中に記載があるが抽出する箇所を間違えたページ数,「もれ」は記載があるが抽出できなかったページ数である. いずれの情報についても記載がないのに誤って抽出した例はなかった. すべての情報が全てのページに記載されているわけではないので,情報数が全ページ数(294)より少なくなっている. 特に,勤務地については省略されているページが多い. この場合,企業の所在地が勤務地になると推測されるが,今回の評価では記載無しとして扱った.

5.2 イベント情報

イベント情報は110ページ見つかった. ページタイプ判定の誤りにより,イベント情報以外のページが12ページあったため,イベント情報は98ページであった. さらに同一ページのコピーが異なるURLで含まれていたので重複した26ページを除き,72ページを抽出精度の評価対象とした. 開催場所,開催日時,イベントの名称のそれぞれについて,抽出ページ数と正誤数を表5に示す.

ひとつのページに複数の情報が記載されている場合,ひとつでも正しく抽出できていれば成功とした. ただし,ほとんどの場合は,表形式で記されていて,全部抽出できるか全部できないかのどちらかになった. '開催日'のうち年月だけが抽出でき,日が抽出できなかったページがある. 試作システムでは月別の分類表示をしているので正解とした. '開催地'が施設で書かれていて都道府県名が不明のページがある. 情報は抽出できていると考えて正解に分類したが,県別の分類はできていない.

5.3 考察

全体に適合率が高く,正確な情報抽出という目的が達成されている. 特に,評価したページにおいて比較的一定した記述をされていた'勤務地'や'開催地','職種','開催日'については90%以上の高い適合率が得られた. 一方,「求人情報」の'資格'と「イベント」の'名称'については,ページ内での記述位置や見出しの付け方などが多様なため精度が落ちている.

検索誤りの主な原因は,情報を判別するキーワードが別の場所にマッチしたことと,表や箇条書きの前後

からの抽出の誤りであった.‘資格’の誤りの多くは,「条件」というキーワードが求人条件以外の場所にマッチしたことによる.「イベント」の‘名称’は,表や箇条書きの前に記述されている場合も多く,キーワードが見つからない場合に直前の強調部分を抽出したが,これが抽出誤りの主な原因となった.

抽出もれの原因としては,今回抽出対象としている表や箇条書き以外の形式で記述されているものがもっとも多い.表が入れ子になっているなど形式が複雑であったり,ブラウザで表示すると表に見えるが実際は表とテキストの組み合わせである場合がある.次いで,各情報を見つけるためのキーワードの不足がある.また,キーワードとなる見出しが付いていない表や箇条書きもあった.

「求人情報」の‘勤務地’については,具体的な地名ではなく「本社」,「各営業所」と書かれているページがあり,もれのうち32件を占めている.「イベント」の‘開催日’のもれは,月と日が別の欄に書かれていたり,月が表の見出しやタイトル部分に含まれ日だけが表中に記述されている場合がある.

精度改善のための課題としては以下がある.

- 記述形式の多様化: $<TABLE>$の入れ子など,現在対応している単純な表形式以外にも,対応可能な形式が多い.種類を増やすことで再現率を上げる.
- キーワードの増強と修正: 未登録のキーワードを追加する.また,登録済みのキーワードのうち誤りの多いもの(「条件」など)を見直す.
- 住所以外の地名対応: 建物名や駅名など重要な施設については住所との対応を用意することで場所の分類を行えるようになる.
- 文脈ルールの導入: 勤務地省略時は所在地を用いる,タイトルからの日時や場所の抽出など,表や箇条書き以外から情報を抽出する.
- 他のページの参照: ページに書かれていない情報がリンク先やリンク元ページにある場合がある.たとえば,イベント施設のサイト内の情報で,会場が自明なため書かれていない場合など.抽出精度との兼ね合いになるが,近くにあるページの参照も検討したい.

6 おわりに

ユーザの目的に応じた情報検索・情報提供を実現するために,ページタイプ分類を利用した情報抽出・分類方式を提案した.

本方式では,ページタイプ分類を使用することで,それぞれのページタイプに適した抽出・分類処理を行える.そのため,多くの種類の文書が混在するWebページに対しても必要な情報だけを高い精度で抽出することができる.また,分類結果を表や図を用いてユーザに提示することで特定の目的に応じた情報を提供するサービスを実現できる.

試作システムによる抽出精度の評価では,記述が比較的一定している場合(勤務地,開催日時など)で適合率90%以上を達成した.記述の自由度が高い,応募資格やイベント名では65-75%にとどまっている.

ページタイプ技術に基づく情報抽出技術は,高い適合率を実現しており,再現率についても改善が見込まれる.今後,精度の改善と適応タイプの拡大を行うとともに,分類メニューや結果の表示方法などのインタフェース面も検討し,情報抽出技術を利用した新しいサービスとして実際に運用を行いたい.

参考文献

[1] 山田,松田,竹元,赤峯,福島,“インターネット多角的検索システムOTROS”, 情処57回大会,3L-01〜04,1999

[2] 松田,福島,“文書タイプ分類による問題解決のためのWWW検索システム”, 情処58回大会,4T-02,1999

[3] 松田,福島,“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”, 情処FI研,FI-53-2,1999

[4] 山田,福島,松田,“Webページからのタイプ別情報抽出・分類方式”, 情処60回大会,1N-06,2000