個人の選好に基づく複数ニュースサイトの記事収集・閲覧システム

河合 由起子[†] 官上 大輔[†] 田中 克己[†,††]

近年，RSS リーダにみられるように複数の Web サイトにまたがって存在している同じテーマのコンテンツをまとめて閲覧できるシステムが求められている．しかし，既存の Web の情報統合法では，収集した大量の情報をシステムの持つ固定の分類体系を基にカテゴライズするため，そのシステムの分類体系を把握していない利用者にとっては，欲しい情報を速やかに獲得することが困難である．本研究では，利用者がすでに分類体系を把握している Web サイトのポータルページのレイアウトを通して，収集・分類し統合した情報を提示できる My Portal Viewer（MPV）を提案する．また，MPV は収集した情報を利用者の閲覧履歴に基づき動的に分類するという特徴を持つ．これにより，利用者は興味に基づき分類された情報を使い慣れているポータルページを通して閲覧でき，大量の情報を効率的にブラウジングできる．本稿では，ニュースサイトを具体例としてあげ，使い慣れているニュースサイトのポータルページを指定することで，既知の分類体系を通して興味に基づき分類され統合された記事をまとめて提示できる MPV のプロトタイプを構築し，検証する．

Personal Viewer for News Contents Integration Based on User's Behavior

Yukiko Kawai,[†] Daisuke Kanjo[†] and Katsumi Tanaka[†,††]

We propose a novel web information integration system "My Portal Viewer" (MPV) to realize more efficient news articles browsing. Although a variety of systems such as RSS readers have been developed to provide integrated content from multiple web sites, the existing systems present another difficulty to the users, i.e., they still have to search the integrated content pages for the information they want because they are often unfamiliar with the page layout and the content categorization rules implemented in the systems. To solve this problem, MPV provides integrated content to a user with the appearance of the user's favorite web site. Because the user has enough knowledge about the content layout, it is easier for the user to obtain information from the integrated content page. Furthermore, MPV categorizes web content dynamically based on the user's access history which reflects the user's interest. In this paper, we describe the basic algorithms used in MPV and a prototype implementation specialized for news content.

1. はじめに

近年，膨大な Web ページから，信頼度の高い情報を効率的に利用者へ提供できるような Web のサービスが求められている．このようなサービスの実現を目指し，本研究では，複数サイトの Web ページを収集して分類し統合することで，利用者の欲しい情報をまとめて提供できる新たな閲覧システムを提案する．

これまで，複数の Web サイトに存在している同じテーマの情報を収集して統合する，情報統合に関する研究がさかんに行われている[1)~4)]．Google News[3)] では，複数のニュースサイトから Web ページを収集し，収集したページの記事を社会，国際，スポーツなどのカテゴリごとに分類し，統合して提示する．このような情報統合により，利用者は複数のニュースサイトにアクセスすることなく，統合されたページを提供しているサイトにアクセスするだけで，複数のニュースサイトの記事をまとめて閲覧することができる．

しかし，収集した大量の記事をカテゴリごとに分類する際，カテゴリの設定は統合サービスを提供している管理者によって決められており，利用者は目的の情報がどのカテゴリに含まれているかを予測しなければ

† 独立行政法人情報通信研究機構けいはんな情報通信融合研究センターメディアインタラクショングループ
Interactive Communication Media and Contents Group, Keihanna Human Info-Communication Research Center, National Institute of Information and Communications Technology
†† 京都大学大学院情報学研究科社会情報学専攻
Department of Social Informatics, Graduate School of Informatics, Kyoto University

ならない．さらに分類された数項目のカテゴリ内の大量の記事の中から，再度検索する必要がある．たとえば,「社会」,「スポーツ」,「国際」などの 6〜9 項目程度のカテゴリが管理者によって設定されている．そのため，利用者は欲しい記事がどのカテゴリに含まれているかを判別しなければならない．さらに，そのカテゴリへアクセスし，カテゴリ内を探索する必要もある．

本稿では，複数のニュースサイトから Web ページの記事を収集し，収集した記事を利用者の興味に基づき分類し，使い慣れているサイトのポータルページを通して統合した情報を提示できる，新たな情報統合法を提案する．提案する統合法は，次の 2 つの特徴を持つ．

- 複数サイトから収集した Web ページを個人の興味に基づき分類して統合すること
- 好みの分類体系を可視化しているポータルページのレイアウトを利用すること

1 つ目の特徴である興味に基づく情報統合では，利用者の閲覧履歴を基に収集した情報を自律的に分類して統合する．従来の情報統合では複数サイトにアクセスする負荷はなくなったが，利用者は統合された膨大な情報の中から，知りたい情報を探す必要があった．提案手法により，利用者は知りたい情報に関して分類されまとめられた統合ページを閲覧できるため，容易に目的の情報を獲得することができる．たとえば,「スポーツ」のカテゴリから「野球」→「MLB」→「NY」→「松井」と探索することなく,「松井」という新たなカテゴリを提供することで,「松井」に関する記事をまとめて閲覧できる．

2 つ目の特徴であるニュースサイトのポータルページの利用では，統合サイトが作成した新たな統合ページを用いず，利用者の好みのニュースサイトのポータルページを統合ページのインタフェースとして代用する．従来の情報統合では，統合しているサイトの分類体系と，その分類体系を反映しているポータルページ内の情報の配置やリンクといった構成（レイアウト）が独自に決められている．そのため，利用者は提示される統合ページの構成を把握し，収集された膨大な情報の分類体系を理解する必要があった．MPV では，ポータルページは分類体系のルートノードを可視化しているページと考え，ポータルページを通して，利用者へ効果的な情報提示を可能とする．提案手法では，ポータルページのレイアウトはそのままで，ポータルページ内の情報を利用者の興味のある情報へと置き換える．使い慣れているポータルページのレイアウトを利用することで，ポータルページ内の情報の分類・配置状況を潜在的に理解しているため，ページを開いた瞬間に知りたい情報へ迅速にアクセスでき，膨大な情報を容易にブラウジングできるという特徴を持つ．

以下，2 章では，提案する MPV の基本概念と基本構成を示す．次に，3 章では，レイアウトの解析および抽出法を示し，4 章では，利用者の興味に基づいて記事を分類して統合する方法について述べる．5 章では，構築した MPV のプロトタイプの検証を行い，6 章で関連研究について述べる．最後に，7 章でまとめと今後の課題を述べる．

2. 基本概念とシステム設計

本研究は，複数のニュースサイトの記事を収集し，個人の選好に基づいて分類して統合することで，利用者に多量の情報を効果的に提示することを目的としている．本稿では，目的を達成するために 2 つのアプローチを提案する．1 つは，利用者の興味に基づいて収集したページを分類し，統合する手法である．もう 1 つは，利用者の使い慣れたポータルページのレイアウトを通して，統合した情報を提供する手法である．本稿では，特にニュースサイトを対象としており，以下の項目を前提とする．

- 収集されるニュース記事のページは,「タイトル」,「記事」および「画像」で構成される☆.
- 収集されるニュース記事のページには，メタデータとして「書かれた日付」,「記事のタイトル」,「記事の概要」が含まれている．

2.1 MPV の基本概念

MPV の基本概念を**図 1** に示す．利用者は，Web ブラウザのツールバーにある MPV ツールバーのブランク部分に，自身の使い慣れているニュースサイトのポータルページの URL を指定する．MPV ツールバーは，事前に利用者がインストールしているものとする．図中では，利用者は MPV ツールバーに CNN サイトのポータルページの URL を入力している．次に Enter キーを入力すると，複数のニュースサイトの統合された情報が，CNN のレイアウトを通して表示される．この統合結果のページが MPV である．

MPV に表示される内容は，利用者の興味や知識に基づいて統合された情報に一部変換される．変換される部分は，(A) ニュースを分類しているカテゴリごとのキーワード，(B) 画像付きトップニュース，(C) カテゴリごとのニュース記事のタイトル集の 3 つである．

☆「画像」のないニュース記事もあるため，本システムでは統合方法に応じて，画像付きニュース記事と，画像なしニュース記事との利用は異なる．

図 1 基本概念
Fig. 1 Concept of MPV system.

(A) のカテゴリごとのキーワードは，利用者の興味のあるキーワードへ変換される．たとえば，CNN のオリジナルの固定のキーワードは，“World”，“Worlds Business”，“Technology” などであるが，MPV では，“Iraq”，“Matsui”，“Koizumi” などへ変換される．

(B) の画像付きトップニュースは，(A) の利用者の興味に基づいて作成されたカテゴリを利用して，カテゴリ内の未読の画像付き記事へと置換される．図中では，変換された “Iraq” のカテゴリに基づいて関連する記事と画像とがトップニュースの画像付き記事として変換され，提示されている．

さらに，MPV サイトで収集した大量の Web ページの記事は，利用者ごとのカテゴリに基づき分類されて，(C) の記事のタイトル集として統合される．利用者は記事のタイトルのリンクを選択すると，選択した記事のオリジナルの内容を閲覧できる．図中では，“Matsui” に関する記事のタイトルを選択することで，Newsweek サイトのオリジナルの記事がそのまま表示されている．

また，利用者が記事を閲覧するということはその記事の内容について興味があったと仮定し，オリジナルページの記事閲覧後に MPV へ再アクセスすると，(A)〜(C) の内容は書き換えられて提示される．よって，利用者が MPV からオリジナルページの記事を閲覧するたびに，(A)〜(C) は再統合されて提供される．

なお，MPV ツールバーをインストールし，最初に MPV を利用した場合は利用者の閲覧履歴がないため，(A) と (C) の内容の変換は行われず，利用者が MPV ツールバーで指定したオリジナルのポータルページの内容がそのまま表示される．(B) に関しては，MPV サイトで収集した各サイトの記事のうち，最新の画像付きの記事を表示するものとする．

2.2 システムの基本設計

本システムは，図 1 で示したように，ポータルページを指定する MPV ツールバー，統合結果のページである MPV，および MPV を提供する MPV サイトからなる．以下では，MPV ツールバーの機能と，MPV サイトの処理の流れを示す．

MPV ツールバーは，好みのポータルページの URL を指定すると，利用者 ID として cookies ファイルを作成し，入力された URL と cookies ファイルを MPV サイトへ送信する．その後，MPV サイトより，統合された MPV を受信し，利用者へ提示する．

MPV サイトでは，MPV ツールバーより受信した URL と利用者 ID から，蓄積している複数のニュースサイトのページの情報を，利用者の履歴情報を基

に分類し統合する．統合した情報は，指定されたポータルページの内容と置換されて利用者へ送られる．統合結果を返信するまでの手順を以下に示す．ここで，MPV サイトが蓄積しているページの情報とは，収集したページの URL とメタデータ（Title，Description，Keyword，Time），画像データの URL とする．

(1) MPV ツールバーよりポータルページの URL と利用者 ID を受信する．
(2) 受信した URL のページのレイアウトを解析し，置換する部分を検出する．
(3) 受信した利用者 ID に対応する閲覧履歴を基に興味のある記事を推定し（詳細は 4 章），メタデータをデータベースより抽出する．
(4) 抽出した記事に関するメタデータを，レイアウトに合わせて分類し統合する．
(5) (2) の検出情報を (4) の統合情報へと置換する．
(6) 置換した統合結果を MPV へ返信する．

2.3 統合情報へ置換する項目の選定

統合情報をオリジナルのポータルページを通して提供するために，我々は 12 個のニュースサイトのポータルページ☆から置換する項目を選定した．各ポータルページを分析した結果，ニュースサイトのポータルページは主に以下の 5 つの内容に基づく領域で構成されていることが明らかとなった．

(1) ニュースサイトのロゴの画像
(2) カテゴリの項目
(3) 画像とタイトルで構成されるトップ記事
(4) 記事のリンク付きタイトル集（カテゴリごと）
(5) 広告

(1)〜(5) のうち，本システムでは (2)〜(4) を，利用者の興味に基づいた内容に置換する．(1) と (5) に関しては，ニュースの記事との関連性が低いため今回は置換しないものとした．

3. ページ解析および情報抽出法

MPV サイトでは，利用者が指定した任意のポータルページのレイアウト構造を解析し，置換する 3 項目の領域を抽出する．抽出対象となるニュースサイトのポータルページは必ず 3 項目を含むものとする．

提案する抽出法は，ページのレイアウトを xy 座標に変換して領域を算出し，領域間の関連と領域内の特徴とを利用して，領域内で変換する必要のある情報を検出する手法である．

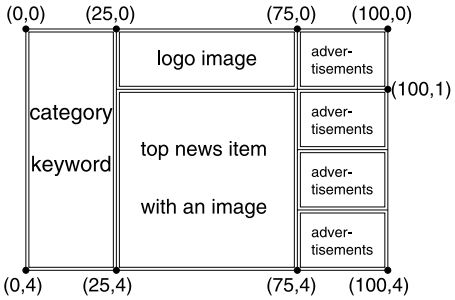

図 2 TABLE 構造を用いて xy 座標に変換しセルを抽出した結果
Fig. 2 HTML layout based on sample TABLE model.

3.1 座標変換によるセルの抽出

多くのポータルページのレイアウトの形成には，HTML の TABLE 構造が利用されており，我々が分析した 12 個のニュースサイトのポータルページでも，TABLE 構造がレイアウト形成に用いられていた．MPV ではこの TABLE 構造を解析して，レイアウトの座標を算出する．

HTML の TABLE 構造は，1 つ以上の行で構成され，各行は 1 つ以上のセルで構成されており，行と列に配列した多次元のデータの表を構成できる[5]．TABLE の行全体は TR で，セルの指定は TD，TH で指定される．TABLE の幅は WIDTH 属性により指定され，ROWSPAN 属性により行の連結，COLSPAN 属性により列の連結が各々指定される．以上の定義より，WIDTH で全体の幅を決定し，TD，TH の出現回数と COLSPAN の値により各行のセルの幅を算出し，x 座標へ変換する．y 座標は，TR の出現回数により全体の高さを決定し，ROWSPAN の値によりセルの高さを算出し，y 座標へ変換する．**図 2** に，以下の簡素化した HTML の TABLE 構造を xy 座標値へ変換し，セルを抽出した結果を示す．

```
<TABLE width=100>
 <TR>
  <TH rowspan="4"><br>category
                  <br>keyword<br></TH>
  <TH colspan="2">logo image</TH>
  <TH>advertisment<br></TH></TR>
 <TR>
  <TH rowspan="3" colspan="2">
                <br>top news item
                <br>with an image</TH>
  <TH>advertisement</TH></TR>
 <TR><TH>advertisement</TH></TR>
 <TR><TH>advertisement</TH></TR>
</TABLE>
```

☆ http://www.cnn.com, http://www.time.com/time/, http://news.yahoo.com/, http://www.asahi.com/english/, http://www.cbsnews.com/, http://www.nytimes.com/, http://www.asahi.com/, http://headlines.yahoo.co.jp/hl/, http://www.yomiuri.co.jp/, http://www.nikkei.co.jp/, http://www.mainichi-msn.co.jp/, http://www.sankei.co.jp/

3.2 セル内の特徴と座標を利用した情報抽出

算出した各セルの xy 座標値と，置換される 3 項目の特徴を基に，オリジナルのポータルページから情報を検出する．検出する 3 項目の特徴を以下に示す．

(A) カテゴリごとのキーワード

（特徴 A-1）セル内の構造がパターン化．

セル内ではカテゴリとなるキーワードの部分だけが異なり，他のタグや条件の配列が同じ構造．

(B) 画像とタイトルで構成されるトップ記事

（特徴 B-1）「(A) のカテゴリ」のセルより後に出現．(B) のセルの x または y は (A) のそれより大きい．

（特徴 B-2）画像とタイトルが同一記事へリンク．

(C) カテゴリごとの記事のタイトル集

（特徴 C-1）(A) のカテゴリと同じキーワードが存在．

（特徴 C-2）カテゴリごとにリンク付きタイトルが 1 つ以上存在．

（特徴 C-3）(C) のセルの y は (B) のセルの y より大きい．

以上の特徴を条件とし，ポータルページの 3 項目の内容を検出する．

4. 記 事 分 類

MPV サイトでは，利用者が指定したポータルページの 3 項目の内容を置換する．置換する内容は収集した記事のうち，利用者の興味に合わせて分類されて選択されたものである．

記事分類では，蓄積している記事のメタデータより「単語情報のテーブル」，利用者の興味情報である閲覧履歴より「興味語および興味木」を各々作成して用いる．分類の概要は，まず，利用者の興味のあるキーワードとなる興味語を閲覧履歴に基づき決定する．この興味語が，カテゴリごとのキーワードへと置換される．次に，興味語を含む画像付き記事のうち，最新かつ未読のものを選択し，画像付きトップ記事へと置換する．最後に，興味語をルートとする興味木を作成し，興味木と各記事の単語の重みを用いて記事を選出する．選択された記事は，カテゴリごとの記事集として置換される．

4.1 単語情報のテーブル

MPV サイトでは，収集したページのメタデータの日付とタイトルおよび概要から，単語とその重みに関する単語情報のテーブルを作成する（**表 1**）．

各ページの単語は，タイトルと概要を形態素解析して抽出した固有名詞，一般名詞の各単語とする．単語の重みは，出現頻度（Term-Frequency）と記事頻度の逆数（inverse Document Frequency）の $tf \times idf$ を用いて，以下の式より算出する．

表 1 単語情報のテーブル
Table 1 Table of pages.

ページの ID（日付）	単語	重み
P_i（04/01/09/12:00）	A	w_{ia}
	B	w_{ib}
	C	w_{ic}
P_{i+1}（04/01/09/12:15）	A	$w_{(i+1)a}$
	B	$w_{(i+1)b}$
	F	$w_{(i+1)f}$

$$w_{ij} = \frac{\log(P_i \text{ 中の単語 } j \text{ の出現頻度} + 1)}{\log(P_i \text{ 中の総単語種数})} \times \log \frac{\text{記事の総数}}{\text{単語 } j \text{ が出現する記事の総数}} \quad (1)$$

4.2 興味語選出と置換

興味語および興味木は，利用者の閲覧履歴を基に作成される．興味語は，オリジナルのポータルページのカテゴリのキーワードと置換される単語である．興味語には重要度があり，利用者が記事を閲覧することで重みが変わる．

興味語 j の重要度は，利用者が閲覧したページ P_1〜P_n に出現する単語 j の重みの総和とする．総和値が閾値以上の単語が興味語となる．閲覧したページを $P_i(i = 1, \cdots, n)$，ページ P_i に出現する単語を j，単語 j の重みを w_{ij} とすると，その総和 I_j は $\sum_{i=1}^{n} w_{ij}$ となる．この I_j 値が閾値以上の場合，j は興味語として選択され，I_j 値の大きい順に，興味語はカテゴリの先頭のキーワードから順に置換される．

興味語をカテゴリのキーワードと置換する場合，ポータルページのオリジナルのレイアウトを変えずに置換しなければならない．そのため，置換可能な興味語の数はオリジナルのキーワードの数に基づき制限される．そこで，MPV では “others” というキーワードのカテゴリを新たに作成し，表示できないキーワードや興味語を格納し，そのカテゴリにマウスを合わせるとプルダウンで格納されたキーワードを表示させる．

図 3 に置換の流れを示す．まず，利用者が MPV を閲覧して記事へアクセスすることで，履歴情報が作成される．履歴情報を基に閲覧したページの単語情報を用いて，興味語を作成する．興味語が初めて作成された場合，最下位となるオリジナルのキーワードが “others” というキーワードへと置換される（図 3 の (b)）．この，“others” にオリジナルのキーワードが格納される．さらに閲覧が続くと履歴情報も更新され，選出される興味語も増加する．選出される興味語の数が，オリジナルのカテゴリのキーワード数 m より多

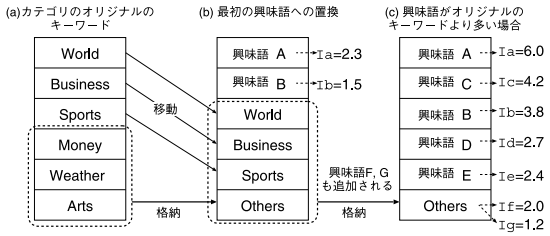

図 3 興味語とカテゴリのキーワードの置換
Fig. 3 Replaced category keywords.

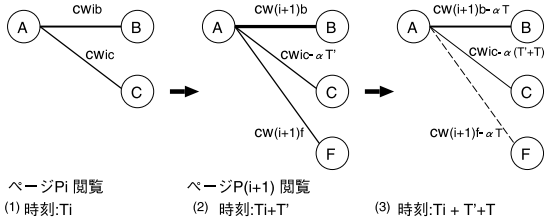

図 4 興味語 A をルートノードとする興味木の再構築
Fig. 4 A keyword tree for user's keyword A is constructed based on the user's behavior.

くなった場合，$m-1$ 個目以降の興味語は，m 個目の “others” に追加される（図 3 の (c)）．

4.3 興 味 木

興味木は，カテゴリの興味語に関する記事を収集したページから選択する際に用いられる．選択された記事のタイトルは，カテゴリごとのタイトル集として置換される．

興味木は，興味語ごとに作成され，各興味語をルートノードとする．子ノードは，興味語を含む同じページに出現する単語となる．ルートノードと子ノードとのリンクの重みは，閲覧したすべてのページから単語間の共起度を算出し，さらに単語の閲覧時刻を抽出し，それらの情報を基に決定される．ここで，共起度とは，各記事のメタデータのタイトルと概要から形態素解析より抽出した単語のうち，同じ記事内にともに出現している割合である．単語 A，B の共起度は，(すべての記事で単語 A と B がともに出現している回数+1)/(すべての記事で単語 A が出現している総数+すべての記事で単語 B が出現している総数) で算出される．

図 4 に利用者がページ P_i を閲覧した後に，A が興味語として選択された場合の，A に対する興味木を示す．まず，A をルートノードとし，その他の単語 B，C がノードとしてリンクが形成される（図 4 の (1)）．各ノードとのリンクには，単語 A と B，A と C の共起度が重み cw_{ib}，cw_{ic} として付加される．図のリンクの線の太さは，重みに比例する．次に，単語 A が出現するページ $P_{(i+1)}$ を閲覧した場合，A のツリーが再構築される（図 4 の (2)）．再構築は，共起語がツリーに存在していない場合は，新たにノードとして追加され，重みとともにリンクが形成される（単語 F の追加）．共起語がすでにツリーに存在している場合は，その重みがリンク値として更新される（単語 B の重みの更新）．また，A と A に共起する子ノードが同時に出現するページを，閲覧しない時間が T 以上たった場合，リンクの重みは αT の割合で減衰する（図 4 の (2)，(3)）．リンクの重みに時間情報を用いることで，利用者の最近の興味を反映できる．

4.4 記 事 選 択

これまで生成した単語情報のテーブルと興味木を用いて，画像付きトップ記事と，各カテゴリの興味語に関連する記事を選択する．

画像付きトップ記事は，単語情報のテーブルから，重みの大きい興味語を含み，かつ日時が最新で未読の画像付きページが選択される．

次に，興味語に関する記事の選択の流れを示す．まず (1) 興味語の出現する記事を選択し，次に (2) 選択した記事から興味木を基にさらに記事を選別する．(2) の記事の選別は，選択された記事に出現する単語の重み w と，興味木のリンクの重み cw とのベクトルの内積値を算出し，内積値が閾値以上の記事とする．さらに，興味語ごとに選別された記事は，類似する内容の記事がある場合はグループ化される．類似記事の判別は，一定時間内に作成されたページをグループ化し，グループ内の記事間での単語の重みのベクトルの内積値を算出し，内積値が閾値以上の記事どうしを類似する記事とする．統合ページで表示する際には，記事の作成時間が新しい未読の記事を順に配列する．レイアウト上に表示できない記事は興味語をアンカとし，リンク先に新たなページを作成し，グループ化した記事のタイトル集として表示される．

5. 実 験

本章では，プロトタイプによる MPV の実行例を示し，レイアウト解析と興味語抽出に関する 3 つの実験とその考察について述べる．実験 1 では，レイアウトの情報抽出について考察する．実験 2 では，閲覧行動による興味語の抽出精度に関して考察し，類似のニュース統合システムとの比較を行う．実験 3 では，興味語を含む記事の分類および提示について考察する．

本稿の実験では，OS に WindowsXP，CPU に Pentium4 2.8 GHz，主メモリに 1.5 GB のデスクトップ PC を MPV サーバとし，Web サーバには apache2,

データベースサーバには PostgreSQL を用いた．開発には，Perl，Microsoft Visual Studio .Net C#，Mecab[6] を利用した．**表 2** に，実験に用いた 5 つのニュースサイトを示す．5 つのニュースサイトから収集した記事は，2004 年 12 月 6 日の 19 時から 20 時までの 225 個の記事である．

記事の単語抽出の閾値を 0.09，興味木の親ノードである興味語の閾値を 0.025，子ノード決定の閾値を 0.01 とした．各閾値は，225 個の各記事に対して約 20 単語が抽出され，そのうち少なくとも 1 単語以上が親となるように決定した．

表 2 統合に利用したニュースサイト
Table 2 News sites for prototype.

ドメイン名	ポータルページの URL	記事ページ数
朝日新聞	http://www.asahi.com	49
読売新聞	http://www.yomiuri.co.jp	49
毎日新聞	http://www.mainichi-msn.co.jp	53
日経新聞	http://www.nikkei.co.jp	34
Yahoo!News	http://headlines.yahoo.co.jp	40

5.1 MPV 実行例

図 5 に，MPV のプロトタイプによる実行例を示し，MPV の有効性について検証する．

(1) は任意のニュースサイトのオリジナルのポータルページと，初めて MPV を利用した結果の MPV とを比較した例である．最初は利用者の閲覧履歴の情報がなく，興味語および興味木は作成されていないため，画像付きトップ記事以外の変化はない．提示された画像付きトップ記事は，収集した記事のうち最新の記事で，別のニュースサイトの記事が提示されている．

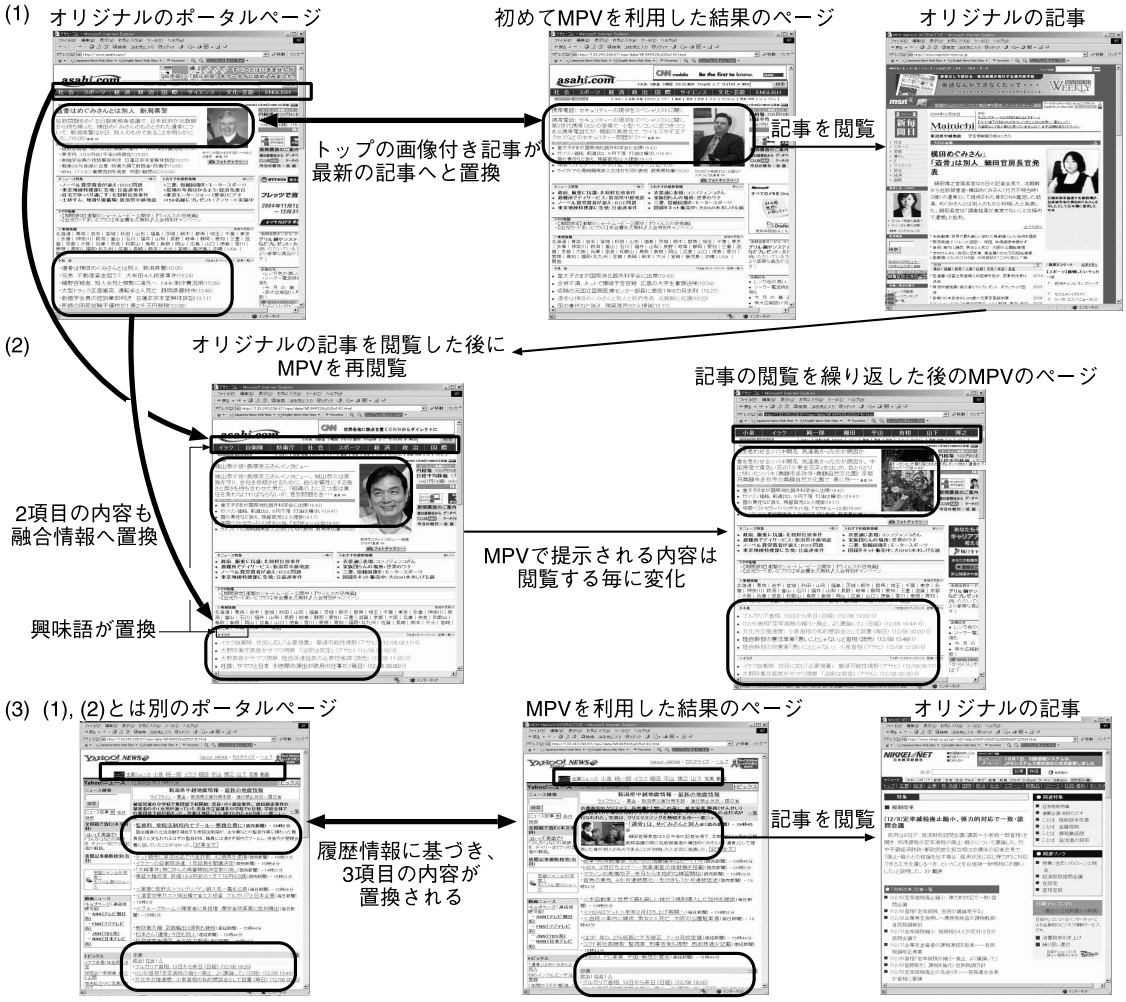

図 5 MPV の閲覧・統合結果表示の例
Fig. 5 The results of MPV page obtained by using prototype system.

(2) は MPV から記事を閲覧することで興味語および興味木が作成された後に，MPV を再度閲覧した例である．閲覧するごとに，3 項目の内容が動的に置換されることが確認できた．閲覧するごとに，新たな興味語とそれに関する記事のタイトルがまとめられて提示されるため，複数サイトから利用者の興味にあわせて記事をまとめて提示できた．

(3) では，(1)，(2) とは別のニュースサイトのオリジナルのポータルページと，閲覧履歴のある利用者の MPV とを比較した例である．閲覧履歴があるため，別のニュースサイトのポータルページでも (2) と同様に 3 項目の抽出および統合が可能であることを確認できた．

以上より，MPV により複数サイトの記事を既存のポータルページでまとめて提示できることが確認され，さらに，収集した情報を興味に基づき分類して提示できることを確認できた．

5.2 HTML 構造による情報検出の考察

本稿では，ポータルページのレイアウトの解析手法を提案した．ここでは，表 2 の 5 つのニュースサイトのポータルページについて，カテゴリのキーワード，トップ記事と画像，カテゴリごとのタイトル集の 3 項目について自動抽出を行った．

抽出結果のうち，カテゴリごとのキーワードの抽出では，3.2 節で述べたように，「セル内の構造がパターン化している」という特徴を利用しており，そのパターンの繰返し回数が 7〜11 回となることが明らかとなった．よって，5 つのニュースサイトでのカテゴリ抽出では，7〜11 回以内で繰り返されている構造に対して自動抽出が可能であると考えられる．

3 項目の自動抽出では，カテゴリのキーワードと，カテゴリごとのタイトル集については，5 つのサイトすべてにおいて良好であった．しかし，トップ記事と画像については，毎日新聞，Yahoo!News の 2 つの自動抽出が行えなかった．これは，3.2 節で示した「トップ記事と画像はカテゴリの後に配置される」という特徴を用いていたが，カテゴリとの間に広告の画像と文章が入っている場合があり，誤認識してしまうことが原因であった．

実験結果より，利用者の指定する任意のポータルページを獲得し，そのレイアウトを自動抽出する現在の提案手法では，すべてのポータルページに対応することは困難と考えられる．そこで，MPV サイトで事前にポータルページを収集し，そのレイアウトを自動解析し，誤認識した領域に関してはそのページ特有の条件を追加して修正することで，実運用性の向上を目指す．

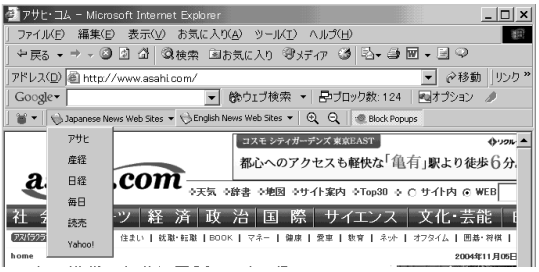

図 6 改良された MPV ツールバー
Fig. 6 Improved MPV toolbar.

現在は，人手により条件を追加することで，正しく抽出され置換されたポータルページを MPV ツールバーに随時追加中である（**図 6**）．利用者は，MPV ツールバーで提示されるポータルページを選択することで，選択したニュースサイトのポータルページを通して統合された記事を閲覧できる．

5.3 閲覧行動にともなう興味語抽出の考察

次に，閲覧行動による興味語の抽出に関して考察し，類似のニュース統合システムとの比較を行う．

MPV では，利用者の閲覧履歴に基づき興味語を決定する．興味語はオリジナルのカテゴリのキーワードと置き換えられる．

表 3 に，利用者が閲覧した記事と閲覧にともない抽出された興味語を示す．興味語は重みの高い順に左側より配列した．なお，MPV には朝日のポータルページを利用した．朝日のポータルページはカテゴリが 8 個あり，1，2 回目の閲覧では，興味語の数がオリジナルのカテゴリ数より少なかったので，オリジナルの「社会」「スポーツ」のカテゴリを表示した．また，4.2 節の図 3 に示したように，8 個目のカテゴリには「Others」を作成し，「Others」ではオリジナルのカテゴリを提供した．「Others」を選択すると，「社会」「スポーツ」「国際」などの各カテゴリに関連する記事を閲覧できる．

閲覧行動の概要は，「イラク」（1，2，3，9 回目），「サッカー」（5，6 回目），「同一の人物の事件」（7，8 回目），「新潟中越地震」（10，11 回目）に関して 2 度以上の閲覧を繰り返し，「任意の人物の事件」に関しては「イラク」と「サッカー」の合間（4 回目）に閲覧している．

実験結果より，同一の内容に関する記事を繰り返し閲覧すると，頻繁に閲覧した記事と関連するカテゴリが優先的に提示されることが確認できた．その結果，利用者は閲覧頻度の高い記事に関するカテゴリをもとに，複数ニュースサイトから関連する記事をまとめて

表 3 実験 2：記事の閲覧履歴と抽出された興味語（7 個）
Table 3 Results of evalution 2.

1 回目	イラク国民議会選，延期に反対，暫定政府のヤワル大統領（朝日）							
	イラク	ブッシュ	大統領	国民	政府	「社会」	「スポーツ」	「Others」
2 回目	大野長官がサマワ視察．陸自派遣延長の必要性強調（読売）							
	イラク	大野	ブッシュ	自衛隊	陸自	大統領	アル	「Others」
3 回目	アジア調査会：元陸自北部方面総監の志方氏が講演（毎日）							
	イラク	自衛隊	東京	大野	ブッシュ	志方	陸自	「Others」
4 回目	K-1 元社長の I 被告，控訴審も実刑．東京高裁（朝日）							
	I 被告	イラク	東京	和義	村上	光	自衛隊	「Others」
5 回目	稲本，半年ぶりに代表復帰．サッカー対ドイツ戦メンバー発表（朝日）							
	稲本	I 被告	イラク	ジーコ	川口	横浜	東京	「Others」
6 回目	サッカー：稲本と西が日本代表復帰 ドイツ戦（毎日）							
	稲本	I 被告	イラク	ジーコ	横浜	川口	ドイツ	「Others」
7 回目	曽我ひとみさん一家 4 人，キャンプ座間を出発（朝日）							
	稲本	曽我	I 被告	ひとみ	ジェンキンス	イラク	ジーコ	「Others」
8 回目	「娘，スパイにされかけた」ジェンキンスさん米誌に語る（Yahoo!）							
	稲本	曽我	ジェンキンス	I 被告	ひとみ	イラク	ジーコ	「Others」
9 回目	テクリートなどで武装勢力の攻撃相次ぐ 25 人死亡（朝日）							
	稲本	曽我	イラク	ジェンキンス	I 被告	ひとみ	ジーコ	「Others」
10 回目	新潟中越地震：被害者にチェックシート配布 仮設住いのケア対策（毎日）							
	稲本	曽我	イラク	ジェンキンス	I 被告	ひとみ	新潟	「Others」
11 回目	子供に息長い「心のケア」必要 中越地震（朝日）							
	稲本	曽我	イラク	ジェンキンス	石井	新潟	中越	「Others」

閲覧できた．たとえば，「イラク」に関するニュースの閲覧の場合，GoogleNews[3] では，「国際」のカテゴリを選択し，そのカテゴリの約 20 個の記事のタイトルの中から「イラク」に関するニュースを探さなければならない．しかし，MPV では利用者の閲覧履歴より「イラク」のカテゴリが作成されるため，「イラク」に関するニュースを再検索することなく，まとめて閲覧できる．さらに，「イラク」に関連する「ブッシュ大統領」や「自衛隊」などのカテゴリも作成されるため，多様な観点に基づいてニュースをまとめて閲覧できるといえる．以上より，閲覧行動による興味語生成の有効性が確認できた．

問題点と解決策

閲覧行動による興味語生成の有効性は確認できたが，一方で履歴に基づく記事分類に関する問題点があげられる．内容の類似する記事を頻繁に閲覧することで，各興味語が類似するため，異なる内容の記事を閲覧できなくなるという問題点が考えられる．たとえば，表 3 の 1 回目から 3 回目に示すように「イラク」関する記事を続けて閲覧すると，すべて「イラク」に関係する興味語で置換されてしまう．

そこで，MPV では 2 つの解決策をとっている．1 つはオリジナルのポータルページのすべての記事を置換せずに，速報や特集などの記事はオリジナルのままで表示させるという点である．たとえば，表 3 の 4 回目の閲覧は，オリジナルのポータルページの速報の記事を選択しており，これにより「イラク」とは無関係の，新たなカテゴリが作成されていることが分かる．もう 1 つの解決法は，オリジナルのカテゴリも選択できる点である．MPV では，新たに「Others」というカテゴリを作成し，このカテゴリから「社会」「スポーツ」などのカテゴリを選択できる．たとえば，表 3 の 5 回目は「Others」の「スポーツ」のカテゴリから選択しており，7 回目は「Others」の「社会」のカテゴリから選択している．この「Others」を利用して新たな記事を閲覧する場合，オリジナルのカテゴリから探すため，GoogleNews と同じ手間がかかる．しかし，その後の閲覧は関連する記事をまとめて閲覧できるため，GoogleNews より効率的な閲覧ができるといえる．

5.4 抽出された興味語に関する考察

本節では，5.3 節で抽出された興味語に関して，興味語を基に分類され提示された記事の再現率について考察する．ここで，収集し抽出したメタデータ（記事のタイトルと概要文）に興味語を含む記事を適合記事とする．再現率は，収集した記事の中で提示された各興味語に対する適合記事の割合であり，(提示された適合記事総数)/(データベース中の適合記事総数) である．

表 4 に，閲覧を繰り返しながら出現した各興味語に対して選択された記事の再現率を示す．閲覧回数が 1〜6 回目までは再現率の平均は 48〜97%と低い値が示されている．これは閲覧回数が少ないため，興味木

表 4 実験 3：実験 2 の興味語に対して分類され提示された記事の再現率
Table 4 Results of evalution 3.

1 回目	イラク	ブッシュ	大統領	国民	政府	「社会」	「スポーツ」	平均
	0.7(7/10)	1.0(5/5)	0.45(5/11)	0.14(1/7)	0.13(2/15)	-	-	0.48
2 回目	イラク	大野	ブッシュ	自衛隊	陸自	大統領	アル	平均
	0.8(8/10)	1.0(3/3)	1.0(5/5)	0.57(4/7)	0.7(4/6)	0.27(3/11)	0.5(2/4)	0.69
3 回目	イラク	自衛隊	東京	大野	ブッシュ	志方	陸自	平均
	1.0(10/10)	0.57(4/7)	0.24(11/46)	1.0(3/3)	1.0(5/5)	1.0(1/1)	0.67(4/6)	0.78
4 回目	I 被告	イラク	東京	和義	村上	光	自衛隊	平均
	1.0(5/5)	1.0(10/10)	0.54(25/46)	1.0(5/5)	1.0(5/5)	0.6(6/10)	0.57(4/7)	0.82
5 回目	稲本	I 被告	イラク	ジーコ	川口	横浜	東京	平均
	1.0(3/3)	1.0(5/5)	1.0(10/10)	1.0(2/2)	1.0(1/1)	0.8(4/5)	0.57(25/46)	0.91
6 回目	稲本	I 被告	イラク	ジーコ	横浜	川口	ドイツ	平均
	1.0(3/3)	1.0(5/5)	1.0(10/10)	1.0(2/2)	0.8(4/5)	1.0(1/1)	1.0(4/4)	0.97
7 回目	稲本	曽我	I 被告	ひとみ	ジェンキンス	イラク	ジーコ	平均
	1.0(3/3)	1.0(4/4)	1.0(5/5)	1.0(4/4)	1.0(4/4)	1.0(10/10)	1.0(2/2)	1.0
8 回目	稲本	曽我	ジェンキンス	I 被告	ひとみ	イラク	ジーコ	平均
	1.0(3/3)	1.0(4/4)	1.0(4/4)	1.0(5/5)	1.0(4/4)	1.0(10/10)	1.0(2/2)	1.0
9 回目	稲本	曽我	イラク	ジェンキンス	I 被告	ひとみ	ジーコ	平均
	1.0(4/4)	1.0(4/4)	1.0(10/10)	1.0(4/4)	1.0(5/5)	1.0(4/4)	1.0(4/4)	1.0
10 回目	稲本	曽我	イラク	ジェンキンス	I 被告	ひとみ	新潟	平均
	1.0(4/4)	1.0(4/4)	1.0(10/10)	1.0(4/4)	1.0(5/5)	1.0(4/4)	0.9(10/11)	0.98
11 回目	稲本	曽我	イラク	ジェンキンス	石井	新潟	中越	平均
	1.0(4/4)	1.0(4/4)	1.0(10/10)	1.0(4/4)	1.0(5/5)	0.9(10/11)	1.0(4/4)	0.98

の興味語と子ノードの重みが小さく，記事の単語との内積値が閾値以下となる記事が多く算出されたためと考えられる．閲覧回数が 7 回目以降では再現率の平均は 98〜100%と高くなっていることが確認できる．

各興味語に対して選択された記事の再現率は，興味語が固有名詞の人名に関しては 100%であった．しかし，「イラク」，「大統領」，「自衛隊」，「東京」といった地名や一般名詞に関しては平均 57%と低かった．これは，地名や一般名詞が記事には含まれているが，含まれている記事の中の重みは低いため，興味木との内積値が閾値より低くなったためと考えられる．

なお，1〜11 回の再現率の平均は 83%で，興味語に対して良好な記事の選別ができたといえる．

6. 関 連 研 究

複数のニュースサイトの記事を利用者の閲覧履歴を基に分類して統合し，既存のポータルページに内容だけを置換するために必要なページの情報抽出技術，ニュース記事統合技術について紹介する．

6.1 情報抽出技術

Web ページから特定の領域を自動抽出するための情報抽出技術は，Web の情報統合技術にとって必要不可欠なものである．そのため，多くの自動抽出に関する研究が行われている[7)〜11)]．

ページ情報の自動抽出は大きく 2 つに分けられる．1 つは，抽出のための例を人手で与えることで，その例を基にマッチするデータを抽出する手法[9),12),13)] で，もう 1 つは，例を与えずに HTML 構造のパターンを発見してデータを抽出する手法[7),8),14)〜16)] である．

STALKER[9)] では，HTML のタグの木構造に着目し，与えられた例にマッチする列の集合を生成し，集合列から単語や数字などのデータを抽出する．さらに，マッチしない列に関しても細分化して新たな列を作ることで，データの抽出を可能にする．MPV でも，抽出するデータの HTML 構造の特徴を例として用いているが，ニュースのポータルページの TABEL タグに注目し，実際に TABLE のパターン構造を基に各 TABLE の領域抽出を行っている点が異なる．

IEPAD[8)] や Arasu ら[7)] は，例を用いることなくタグの繰返しパターンを発見することで，列を切り出しデータを抽出する手法を提案している．しかし，ポータルページでは複雑な TABLE 構造でデータも多種多様であるため，パターンの自動抽出は精密さと時間コストのトレードオフの問題があり，MPV では用いていない．

6.2 記事統合技術

情報統合に関する研究[1),2),17)] は広く行われており，ここでは特にニュースに関する記事の統合技術に注目する．

NewsCrawler[18)] や FeedDemon[19)] は，RSS（RDF Site Summery）で記述された要約を収集し，提示するリーダである．リーダは登録してあるニュースサイ

トを自動的に巡回し，RSS を収集して，そのタイトルをまとめて提示する．利用者は新着順に提示されるタイトルから，キーワードを用いて記事を検索することも可能である．GoogleNews[3] では，約 610 のニュースサイトから収集した記事を 7 つのカテゴリに分け，統合して提供している．トップ記事の決定や類似記事の分類はページランクを用いている．GoogleNews，RSS リーダともに，分類やインタフェースは統合技術を提供している側の設定に依存している点が MPV とは異なる．

Newsbot[4] では，複数のニュースサイトから収集した記事を統合するだけでなく，閲覧履歴を基に関連するニュースを提示する．これにより，利用者は好みのトピックに関する記事を優先的に閲覧できる．しかし，好みの記事に対する分類をしていないため，RSS リーダと同様に一列に並べられたタイトルの中から再検索する必要がある．また，インタフェースも独自のもので，利用者の好みのレイアウトを利用できない．

MyYahoo!News[20] では，ページのレイアウトを利用者が設定できる．また，用意された数十項目のカテゴリのうち，利用者が選択したカテゴリに基づいて収集した記事を分類し，統合して提示できる．レイアウトやカテゴリを利用者が選択できるため，利用者の好みを反映した統合が可能といえる．しかし，カテゴリの選択範囲がシステム側に依存している点が，利用者の閲覧履歴を基に動的に分類している MPV とは異なる．また，インタフェースとして使い慣れているポータルページを指定するだけの MPV と比較すると，レイアウトの設定の負荷は高く，ユーザビリティが低い．

NewsBlaster[1] は，収集した各記事の名詞を抽出してシステムが定義する 6 つのカテゴリに分類し，分類された記事の類似を検出して重複を削除することで複数の記事から 1 つの要約文を作成し提示する．MPV とは，カテゴリ生成と記事分類に利用者の選好を用いている点が異なる．

7. おわりに

本稿では，複数サイトの大量のページを収集し，個人の選好に基づき分類して統合することで，利用者の欲しい情報を提供できる新たな閲覧システム MPV について提案し，プロトタイプによる実験を行った．

利用者の好みのページへ統合情報を置換するためのページのレイアウト抽出の実験では，各カテゴリのキーワードとそれらのタイトル集の抽出は良好であることが確認できた．画像付きトップ記事については自動抽出が行えなかったページもあったが，現在は人手により条件を追加することで実運用性の向上を目指している．

閲覧にともなうカテゴリ生成のための興味語の抽出実験では，同一の内容に関する記事を繰り返し閲覧することで，頻繁に閲覧した記事と関連するカテゴリを生成することが確認できた．これにより，利用者は複数のニュースサイトにアクセスすることなく，閲覧頻度の高い記事に関する記事のタイトルをまとめてブラウジングでき，効果的に記事を閲覧できることを確認できた．また，閲覧した記事から複数の興味語が抽出されてカテゴリが生成されることで，多様な観点に基づきまとめて記事を閲覧できることが分かった．

抽出された興味語に対する適合記事の選択に関しては，再現率は平均 83%であり，良好な記事の選別を確認できた．興味語が固有名詞の場合は再現率が 100%であったが，一般名詞の場合は平均 57%と低かった．これは，記事に出現する一般名詞の重みが低く算出されるため，興味木との類似性が低くなったためと考えられる．しかしながら，選出されなかった記事と一般名詞の興味語との関連性は低く，選出されなかった記事が適合記事とは考えられ難く，提案手法の有効性は示されたと考えられる．

現在はニュースサイトという特定のテーマを同一テーマのポータルページで提供しているが，たとえば旅行や学会サイトなどの情報にも対応できるような，異種のテーマでの情報統合が可能な MPV を検討中である．

参考文献

1) McKeown, K.R., Barzilay, R., Evans, D., Hatzivassiloglou, V., Klavans, J.L., Nenkova, A., Sable, C., Schiffman, B. and Sigelman, S.: Tracking and Summarizing News on a Daily Basis with Columbia's Newsblaster, *Proc. Human Language Technology Conference, 2002*, San Diego, USA, ACM (2002).
2) Radev, D.R., Blair-Goldensohn, S., Zhang, Z. and Raghavan, R.S.: Interactive, Domain-Independent Identification and Summarization of Topically Related News Articles, *5th European Conference, ECDL*, Germany, pp.225–238 (2001).
3) GoogleNews. http://news.google.co.jp/
4) Newsbot. http://uk.newsbot.msn.com/
5) W3C Recommendation: HTML 4.01 Specification (1999). http://www.w3.org/TR/1999/REC-html401-19991224
6) MeCab (2004).

http://chasen.org/~taku/software/mecab/
7) Arasu, A. and Garcia-Molina, H.: Extracting Structured Data from Web Pages, *ACM SIGMOD*, San Diego, CA, pp.337–348 (2003).
8) Chang, C.-H. and Lui, S.-C.: IEPAD: Information Extraction based on Pattern Discovery, *Proc. 10th International World Wide Web Conference*, Hong Kong, China, pp.681–688 (2001).
9) Muslea, I., Minton, S. and Knoblock, C.A.: Active Learning for Hierarchical Wrapper Induction, *Proc. 16th National Conference on Artificial Intelligence*, Orlando, FR (1999).
10) 宮原哲浩，鈴木祐介，正代隆義，内田智之，高橋健一，上田祐彰：極大頻出な縮約可能変数つきタグ木パターンを用いた半構造文書からの情報抽出，第 18 回人工知能学会全国大会 JSAI2004 (2004).
11) Cai, D., Yu, S., Wen, J.-R. and Ma, W.-Y.: Extracting Content Structure for Web Pages Based on Visual Representation, *APWeb*, Vol.2642, Xian, China, Springer, Lecture Notes in Computer Science, pp.406–417 (2003).
12) Kushmerick, N.: Wrapper Verification, *World Wide Web*, Vol.3, No.2, pp.79–94 (2000).
13) Hsu, C.-N. and Dung, M.-T.: Generating Finite-State Transducers for Semi-Structured Data Extraction from the Web, *Information Systems*, Vol.23, No.8, pp.521–538 (1998).
14) Embley, D.W., Jiang, Y.S. and Ng, Y.-K.: Record-Boundary Discovery in Web Documents, *ACM SIGMOD*, Philadelphia, PV, pp.467–478 (1999).
15) Buttler, D., Liu, L. and Pu, C.: A Fully Automated Object Extraction System for the World Wide Web, *Proc. 2001 International Conference on Distrubuted Computing* (2001).
16) Asai, T., Abe, K., Kawasoe, S., Arimura, H., Sakamoto, H. and Arikawa, S.: Efficient Substructure Discovery from Large Semi-structured Data, *Proc. 2nd SIAM International Conference on Data Mining*, Arlington, VA (2002).
17) Lawrence, S., Giles, C.L. and Bollacker, K.D.: Digital Libraries and Autonomous Citation Indexing, *IEEE Computer*, Vol.32, No.6, pp.67–71 (1999).
18) NewsCrawler. http://www.newzcrawler.com/
19) Feed Demon. http://www.bradsoft.com/feeddemon/index.asp
20) MyYahoo!. http://my.yahoo.co.jp/?myHome

(平成 16 年 9 月 14 日受付)
(平成 17 年 1 月 29 日採録)

(担当編集委員 石川 博，原 隆浩，片山 薫，佐藤 聡，土田 正士)

河合由起子（正会員）

1997 年九州工業大学情報工学部電子情報工学科卒業．2001 年奈良先端科学技術大学院大学情報科学研究科情報システム学専攻博士後期課程修了．同年独立行政法人情報通信研究機構（旧，独立行政法人通信総合研究所）入所，現在に至る．博士（工学）．主に個人適応，情報検索と情報統合の研究に従事．日本データベース学会等会員．

官上 大輔

独立行政法人情報通信研究機構専攻研究員．博士（工学）．1998 年立命館大学理工学研究科博士課程後期課程単位取得退学．ユーザ適応，セマンティックウェブ等の研究に従事．人工知能学会，日本データベース学会等会員．

田中 克己（正会員）

1974 年京都大学工学部情報工学科卒業．1976 年京都大学大学院修士課程修了．1979 年神戸大学教養部助手．1986 年同大学工学部助教授．1994 年同大学工学部教授（情報知能工学科）．1995 年同大学大学院自然科学研究科情報メディア科学専攻専任教授．2001 年京都大学大学院情報学研究科社会情報学専攻教授，現在に至る．2003 年から独立行政法人情報通信研究機構（旧，独立行政法人通信総合研究所）メディアインタラクショングループリーダ兼任．京都大学工学博士．主にデータベースとマルチメディア情報システムの研究に従事．人工知能学会，日本ソフトウェア科学会，IEEE Computer Society，ACM 等各会員．