DEWS2005 2C-i5

類似カテゴリと類似構造を有するページ集合の統合・提示方式の提案

河合由起子† 官上 大輔† 田中 克己†,††

† 独立行政法人情報通信研究機構 〒619–0289 京都府相楽郡精華町光台 3–5
†† 京都大学大学院情報学研究科社会情報学専攻 〒606–8317 京都市左京区吉田本町
E-mail: †{yukiko,kanjo}@crl.go.jp, ††tanaka@dl.kuis.kyoto-u.ac.jp

あらまし 我々はこれまで，複数ニュースサイトの記事を収集し利用者の閲覧履歴に基づき分類し，好みのトップページを通して統合した情報へアクセスできる My Portal Viewer (MPV) を提案してきた．利用者は MPV により，使い慣れているページのレイアウトを通して興味に基づき分類された情報へアクセスでき，大量の情報を効率的に閲覧できる．これまでの MPV は利用者が選択したオリジナルページのレイアウトにのみ注目していた．そこで，本研究では，ページのレイアウトだけでなく，そのページの内容にも基づいて情報を分類し統合する方式を提案する．具体的には，利用者が指定した好みのページの内容に基づいてページを収集して分類し，そのページ内で利用者が指定する属性とそのインスタンスのキーワードを基に分類したページの情報をダイナミックに統合して提示する．ページの内容理解は，動的に構築される属性辞書を基にページの属性とインスタンスを抽出することで類似するカテゴリを検出する手法である．利用者は指定したページの内容に合わせて，ページ内の各属性を視点にまとめられた情報を閲覧できる．本稿では，学術会議の案内ページを具体例として挙げ，複数の案内ページの統合・提示方式について検討する．
キーワード 情報統合，MPV，ページ構造，ビュアー, 属性

My Portal Viewer for Information Integration based on Page Layout and Content

Yukiko KAWAI†, Daisuke KANJO†, and Katsumi TANAKA†,††

† National Institute of Information and Communications Technology 3–5 Hikaridai, Seika-cho, Soraku-gun, Kyoto, 619–0289 Japan
†† Department of Social Informatics, Graduate School of Informatics, Kyoto University Yoshida-honmachi, Sakyo-ku, 606–8317, Japan
E-mail: †{yukiko,kanjo}@crl.go.jp, ††tanaka@dl.kuis.kyoto-u.ac.jp

Abstract A novel web application called "My Portal Viewer (MPV)" has been developed to provide web users with high quality content, which is needed due to rapidly growing amount of content on the web. It provides the fused news to a user based on two viewpoints through a user friendly interface and the user's preferences. MPV automatically selects and merges content from many news pages based on the user's interest and knowledge. In this study, we propose improved MPV. Improved MPV provides the user with an integrated web page through a specified page for a query based on two viewpoints through a user-friendly interface using the specified page and also the content of that page. The approach of improved MPV is that similar content to the favorite page is collected, categorized and integrated based on the attribute of a specified keyword. Whenever a user accesses an attribute keyword on the specified page using an attribute's link, he/she can acquire the desired content efficiently because improved MPV presents not only his/her favorite interface but also the links to the integrated pages based on content similar to the favorite page and specified attribute. In this paper, we apply the new MPV to CFP page and describe the approach.
Key words Information Integration, MPV, Layout, Viewer, Attribution

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，我々は，複数のニュースサイトの大量の記事を収集し，利用者の好みに基づき分類して統合・提示する My Portal Viewer (MPV) を提案し，検証してきた [1] [3]．MPV 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トップページのレイアウト（構成）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て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．

2 つ目の特徴であるニュースサイトのトップページの利用では，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トップページ内の情報の配置やリンクといった構成が独自で決められている．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，利用者が普段閲覧しているトップページの構成はそのままで，その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．利用者は使い慣れている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ズできるという特徴をもつ．

これまで 2 つの特徴をもつ MPV のプロトタイプを実装し，利用者の閲覧行動から生成される MPV ページの評価実験を行った [2]．その結果，利用者は既存のニュースのトップページを通して，自身の閲覧履歴に基づいて分類され統合された記事へアクセスでき， 効率的な閲覧ができることを確認できた．しかし，ニュースサイトのトップページではページの構成が定常的なため有効であったと考えられる．そのため，ページの内容の変化に伴い構成が変わるトップページに対しては適応が難しい．

そこで，本研究では，ページの構成だけでなくページの内容も反映した情報の分類および統合を可能にする，新たな MPV を提案する．本稿では，特に会議ごとでページの案内の内容と構成が変わる Call for Paper (CFP) のページを具体例とした効果的な情報統合を目指す．

MPV では新たに次の 2 つの特徴を持つ．

- 利用者が指定したページの内容と類似するページを収集して分類すること．
- 利用者が指定したページ内のテキスト（属性）に基づいて情報を配列して統合すること．

これらの特徴を実現するためには，ページの内容を理解する技術が必要である．そこで，我々はページ内容を理解するために，テキスト情報を属性とインスタンスに分類して取り扱う．例えば，CFP ページでは，属性は「会議名」,「トピック」,「口頭発表論文締切」などであり，各属性のインスタンスは「DEWS」,「XML，半構造データ」,「2005 年 1 月 5 日」などにあたる．提案する新たな MPV では，ページの属性とインスタンスを抽出し，抽出した属性とインスタンスを基にページを収集し分類して統合する．具体的には，まず Web から任意のテーマに関する属性辞書を作成し，次に属性辞書を用いて利用者が指定したページから属性を抽出し，抽出した属性とインスタンスを基にページを収集および分類する．さらに，利用者がページ内の属性を指定すると，(1) 指定された属性とページ内で配置関係の近い属性と，(2) 指定されたページのインスタンスと類似性の高い収集したページのインスタンスを基に動的に配列して統合する．提案方式により，利用者は指定したページの内容と類似したページの情報を，指定した属性を基にまとめて閲覧できる．

以下，2. では，新たに提案する MPV の基本概念と基本構成を示す．次に，3. で，属性に基づくページの収集と収集したページの統合法について述べる．5. では，関連研究について述べ，最後に，6. 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

本研究では，複数の Web サイトの大量のページを統合し，利用者に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 2 つのアプローチを提案している．一つは，利用者の使い慣れたページの構成を用いることで，利用者の好みを反映したページ構成を通して，融合した情報を提示する手法である．もう一つは，その好みのページの内容に基づいてページを収集して分類し，ページの任意の属性とインスタンスを基に統合して提示する手法である．

2.1 基本概念

MPV の基本概念を図 1 に示す．

2.1.1 類似ページの収集

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 CFP 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 2005 年 2 月に開催される DEWS の CFP の URL を入力している．次に，Enter キーを入力すると，複数の CFP の融合された情報へアクセスできる CFP のポータルページとなる MPV のトップページが表示される．この MPV は DEWS のトップページと同じ構成で表示される．

MPV サイトでは，まず，利用者の指定した DEWS のトップ

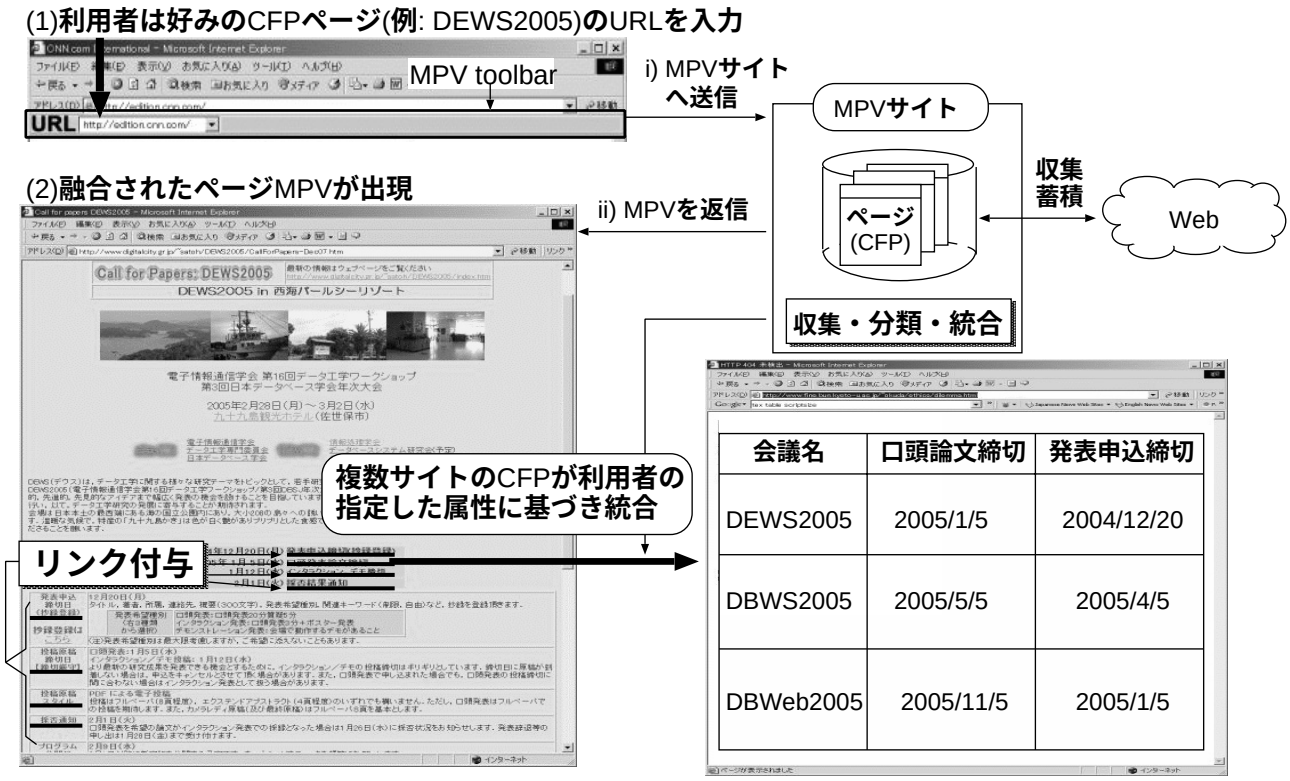

図 1　基 本 概 念

ページから属性辞書（詳細は 3.）を用いて属性とインスタンスを抽出し，属性のインスタンスを基にページを分類する．ページの収集と同時に，DEWS のページの構成をそのまま利用して内容だけを一部変換したものを MPV として利用者へ提示する．内容の変換は，抽出した属性の属性名に対するリンクの付与である．ここで属性名とは，DEWS の CFP ページの場合，「発表申込締切（抄録登録）」「口頭発表論文締切」「採否結果通知」「トピック」等にあたる．

2.1.2　属性とインスタンスに基づく統合

各 CFP ページには属性が複数あるため，全ての属性に関する情報を統合して 1 ページで提示すると，その中から再検索する必要があり効率的な閲覧が困難になる．指定したページに n 個の属性があり，収集したページが m 個ある場合，それらを統合すると再検索の手間は $O(nm)$ 必要となる．そこで，MPV では，利用者が各属性名のリンクを選択すると，その属性名に基づいて複数の CFP をまとめて統合ページとして提供する．

利用者が属性名を選択すると，MPV サイトでは利用者の指定した属性に対するインスタンスの類似性を基に収集したページを選択する．さらに，利用者が指定した属性と他属性との位置関係を検出し，指定された属性と近い位置に配置されている属性を順に配列する．図中では，利用者が「口頭発表論文締切」の属性を選択し，その属性と近い配置にある「発表申込締切（抄録登録）」「インタラクション，デモ締切」「採否結果通知」が順に横に配列される．

また，配列された各属性はそのインスタンスに基づいて，ランキングされる．ランキングは，利用者が指定した属性のインスタンスと収集したページで指定された同一の属性に対するインスタンスとを比較して，類似性の高いインスタンスを順に配列するものとする．図中では利用者が指定した「口頭発表論文締切」の属性に対するインスタンスが「2005 年 1 月 5 日（水）」となっており，この日付と近いインスタンスをもつ他の CFP が順にランキングされ，統合結果として提示されている．

2.2　システムの基本設計

本システムは，利用者の Web インタフェースとなる MPV と，MPV を提供する MPV サイトからなる．Web のインタフェースは，利用者が Viewer として利用したいページの URL を入力するツールバーと，融合された結果が表示されるページ部分とで形成される．

利用者が MPV のトップページとして指定した URL は，MPV サイトへ送られる．MPV サイトでは以下の手順で MPV のトップページを作成する．

（1） テーマに対する属性辞書を作成．

（2） 受信した URL の CFP ページを獲得．

（3） 獲得したページを属性辞書を基に解析し，属性およびインスタンスを抽出．

（4） 抽出した属性名にリンクを付与して，MPV のトップページとして利用者へ提示．

次に，利用者が MPV のトップページ内の属性名を選択すると，選択された属性名が MPV サイトへ送られ，MPV サイトでは以下の手順で収集した CFP ページを統合する．

（1） MPV トップページの全ての属性のページ内の位置情報を獲得．

（2） 位置情報より，利用者の指定した属性と近い位置の属性を順にテーブルの横に配列し提示．

（3） 利用者の指定した属性をもつ CFP ページのインスタンスを選択.

（4） 利用者の指定した属性に対するインスタンスを抽出.

（5） （4）のインスタンスと（3）のインスタンスを比較し，類似性の高い（3）のインスタンスから順にテーブルの縦にランキングする.

3. 分類および統合手法

MPV サイトでは，利用者が指定したページの構成と内容に基づき収集したページを分類して統合し，提示する.

提案手法では，まず，利用者が MPV のトップページとして指定したページから属性辞書を基に属性およびインスタンスを抽出する．抽出した属性にはリンクが付与される．次に，利用者が MPV のトップページのリンクが付与されている属性を選択すると，その指定された属性とそのインスタンスを基にテーブルに配列することで統合する.

3.1 属性辞書の作成と属性抽出

MPV ではページの内容を判別するために，属性辞書に基づきページ内の情報から属性とインスタンスを検出する．属性辞書は，特定のテーマに関する属性のキーワード集である．この属性辞書は，Web を基に自動作成される．辞書作成法は，特定のテーマに関するページを収集して，収集したページのキーワードの出現頻度を基に決定する．例えば，CFP の場合,「投稿締切」「採否結果」「トピック」「運営組織」などが各 CFP ページで頻繁に出現している．この出現頻度の高いキーワードを属性名として辞書に登録する.

提案する属性辞書の作成は，単語の出現頻度以外に文字の表現形式（サイズや太さ）を特徴量として用いる．これにより，CFP ページの作者の意図を反映した単語の抽出が可能となり，属性の効果的な抽出を行う．以下に，属性辞書の作成手順を示す.

（1） テーマとなるキーワードを決定

テーマが CFP の場合,「論文募集」「会議」「ワークショップ」等を利用.

（2） キーワードを基に検索して結果を収集し，テキスト情報を獲得

（3） テキストを形態素解析して普通名詞の単語を検出

（4） 検出した単語の出現回数を全てのページ n より検出

（5） 検出した単語のフォントのサイズと太さを全てのページ n より検出

（6） 出現回数とフォントサイズを基に下記の式で単語 i の重み ws_i を算出して配列

$$ws_i = tf \times S_i \tag{1}$$

$tf = log$(単語 i の出現回数＋１)/log(単語の種類の総数)

$S_i =$ (単語 i のフォントサイズの総和)/(単語 i の出現回数)

（7） トップ 30 の単語 i を属性辞書として登録

（8） トップ 60 の単語 i から出現回数とフォントの太さを基に下記の式で単語 i の重み wb_i を算出して配列

$$wb_i = tf \times B_i \tag{2}$$

$B_i =$ (単語 i の太さの総和＋１)/(単語 i の出現回数)

（9） トップ 30 の単語 i を属性辞書として追加登録

以上より，作成した属性辞書を用いて，(2) の収集した CFP ページや利用者の指定したページから属性辞書に出現する属性名を抽出する．利用者が指定したページから抽出された属性名に対しては，リンクが付与される.

3.2 インスタンスの抽出

次に，インスタンスの抽出手法を述べる．提案方式では，ページ内の各属性間に出現する文字数を基にインスタンスを検出する.

（1） 属性辞書を用いて，収集した CFP ページおよび利用者の指定した CFP ページから属性を抽出

（2） 各ページの属性を出現する順にランキング

（3） 属性 i 番目と i+1 番目の間の文字数をカウント（HTML タグは除く)

（4） (3) の結果を属性 i のインスタンスとして検出

（5） (3) が NULL の場合属性 $i+1$ の属性名とインスタンスを属性 i のインスタンスとして検出

3.3 動的な配列による統合

本節では，属性とインスタンスを基に収集した CFP ページの情報を統合する手法を述べる．ページには属性が複数あるため，全ての属性に関する情報を統合したページを提示すると情報量が多く，その中から再検索する必要がでてくる．そこで，MPV では利用者が知りたい属性に関する情報のみをまとめて提示することで，効率的な閲覧を提供する．統合手法は，利用者が指定したページの属性の位置情報とインスタンスを用いる．指定されたページの各属性のページ内の位置情報を用いることで，利用者は指定したページの構成を通して各属性に基づいて統合された情報を閲覧できる．以下では，まず属性に基づく配列手順を述べ，次にインスタンスに基づく配列手順を述べる.

（1） 利用者へ提示している MPV のトップページの各属性の位置情報を検出

（a） MPV のトップページの各属性が出現する順番を検出

（b） 属性 i 番目と i+1 番目の間の文字数をカウント (HTML タグは除く)

（c） (2) の結果を属性 i と $i+1$ 番目の距離として決定

（2） 指定された属性との距離が近い属性ほど関連性が高いものとして配列.

次に，インスタンスを基に配列された属性をもつ CFP ページをランキングする．ランキングは利用者が指定した属性のインスタンスと類似するインスタンスをもつ CFP ページを順に配列する．類似度の検出は，指定した属性のインスタンスとのベクトルの内積値を用いる．ただし，時間情報や場所情報の場合は，内積値ではなく下記の条件を基に配列する.

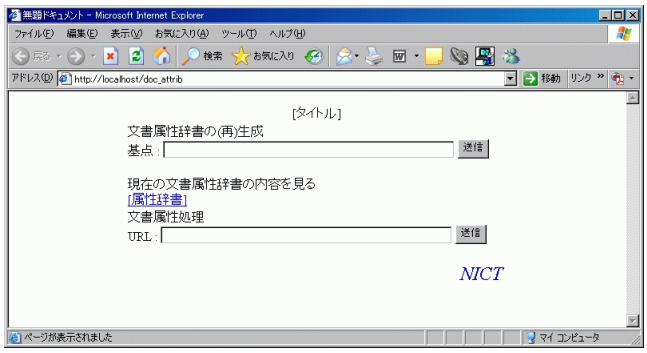

図 2　辞書作成および MPV トップページ作成のための入力画面

表 1　単語情報のテーブル

Table 1　Table of pages

頻度とフォントサイズの重み ws_i より抽出				
論文	学会	情報	技術	委員
国際	システム	ソフトウェア	申込	ワークショップ
原稿	シンポジウム	環境	締切	科学
工学	電子	テーマ	分野	問題
セッション	会員	支部	協会	ホームページ
社会	言語	大会	方法	下記
頻度とフォントの太さの重み ws_i より抽出				
日時	場所	資格	技能	著者
国際	セミナー	セッション	世界	会場
タスク	エネルギー	方法	日程	カテゴリ
論文	大会	一般	対象	言語
システム	全国	申込	詳細	要領
技術	経済	科学	内容	結果

- 時間情報の場合は数値の近いものを配列
- 場所情報の場合は地理的位置の近いものを配列

4. 実　　験

提案方式を基に，プロトタイプを作成し属性辞書の評価実験を行った．

4.1 システム構成

実験では，OS に WindowsXP，CPU に PentiumM 1.7GHz，主メモリに 2.0GB の PC を MPV サーバとし，Web アプリケーションサーバには tomcat5.5.7 を用いた．検索には GoogleAPI を，形態素解析開発には，Mecab [4] を利用した．

検索結果として上位 496 件を収集した．フォントサイズの重みを用いた属性抽出では上位 30 単語を登録し，フォントの太さの重みを用いた場合は出現頻度の結果上位 60 件を再度ランキングして上位 30 単語を登録した．今回は，CFP のテーマとして “（論文募集 OR CFP）（会議 OR ワークショップ）”を入力のキーワードとした．検索結果の CFP ページから形態素解析にて抽出した普通名詞の総単語数は 5970 語であった．

図 2 に属性辞書のテーマとなるキーワードと MPV のトップページとなる URL とを入力する Web ページを示す．

4.2 属 性 辞 書

表 1 に属性辞書の作成結果を示す．属性辞書は，5790 語から式 (1) を用いて頻度とフォントサイズの特徴量を算出して上位 30 単語を抽出した結果と，式 (2) を用いて頻度とフォントの太さの特徴量を算出して上位 30 単語を抽出した結果を登録した．表より，5790 語から選出された 60 語には，属性として取り扱える語が多く含まれており，良好な辞書が作成できたと言える．再現率を（属性辞書の適合属性数）/（属性辞書に登録されている単語総数）とすると，55 %であった．今後は属性辞書の作成のために抽出する単語の数の閾値を検討し，属性抽出の精度を高める．また，フォントサイズや太さだけでなく配置位置の情報も含めた重み付けを検討する．

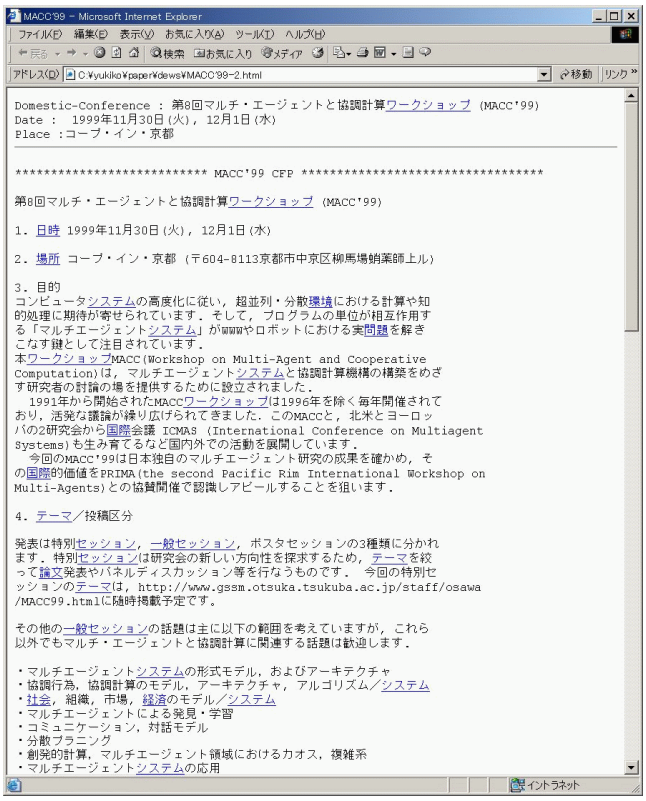

図 3　MPV のトップページの一例

4.3 指定ページの属性抽出

図 3 に，利用者が指定した CFP ページを MPV のトップページへ変換した結果を示す．結果より，10 項目の属性名に関して 8 項目の属性名が抽出されていることが確認できた．抽出されなかった「目的」の属性に関しては式 (1) の式より 71 番目に検出されており，上位 60 番目までのみ取り扱っていたため，辞書として登録されなかった．また,「その他」の属性に関しても，98 番目に検出されていた．

属性精度を高めるためには，属性辞書の登録数を増やすことで解決できると考えられるが，現在の提案方式では登録数を増加させるとインスタンスも多く含まれてしまうため，インスタンスの抽出精度が低くなると考えられる．今後はインスタンスの抽出に関しては，HTML の構造を踏まえた抽出を導入することで属性およびインスタンスの両方の抽出精度を高める検討を行う予定である．

5. 関 連 研 究

Web ページから情報を抽出する技術は，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 [5] [6]．また，収集した情報の分類も行われている [7].

STALKER [5] 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，情報抽出を可能にする．MPV でも，インスタンスの抽出では HTML 構造の特徴を用いているが，属性の抽出では Web 全体から例となる属性を抽出して属性辞書を自動作成し，その辞書を用いて任意のページの属性を抽出している点が異なる．

IEPAD [8] や Arasu ら [6] は，例を用いることなくタグの繰り返しパターンを発見することで，列を切り出す情報抽出法を提案している．しかし，CFP ページの構造は単純なものは少なく，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，実用性が低く MPV では適用していない．

NewsBlaster [7] は，複数のニュースサイトから収集した記事を 6 つのカテゴリに分け，カテゴリ毎の記事を品詞に基づき類似する記事へと分類する．分類された記事集から，自然言語処理技術を用いて一つの要約文が作成され，融合結果として提示される．複数記事の要約は有効だが，カテゴリ分類は静的で，ページの内容を表す属性やインスタンスに基づいた動的な分類および融合はされていない点が MPV と異なる．

また，収集した情報を効果的に提示する技術も開発されている．FeedDemon [9] や NewsCrawler [10] は，RSS（RDF Site Summery）で記述された要約を収集し，閲覧できるリーダである．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，記事のタイトルをまとめて利用者へ提示できる．しかし，ページの内容に基づいた記事の動的な統合および提示はされていない．そのため利用者はキーワードを明示的に入力する必要があり，またキーワードと関連する情報を閲覧することができない．

MyYahoo! [11] では，利用者が設定したカテゴリに基づいて，収集した複数のニュースサイトの記事を分類し，タイトルをカテゴリ毎にまとめて提示する．レイアウトも利用者が設定できる．しかし，利用者が選択できるカテゴリは MyYahoo!サイトが提供しているカテゴリに限られている．また，レイアウトは，明示的に利用者がカテゴリの表示する順序や場所を設定する必要があり，設定自体を使い慣れるための負荷も高い．MPV では，利用者の好みのページを通して統合した情報へアクセスできる点が異なる．

6. まとめと今後の課題

本研究では，複数ページを統合し，ユーザの欲しい情報を提供できる新たな MPV について提案した．新たな MPV では，利用者は指定したページの構成を基に Viewer として利用するだけでなく，そのページの内容と類似したページを収集して統合し提供する．提案方式は，Web から収集した数百件分のページを基に属性辞書を作成し，収集したページと利用者が指定したページから属性とインスタンスを抽出し，それらを基にページの内容を解釈して分類および統合する．属性辞書の作成では，出現頻度以外にページの構成である見た目を反映してフォントのサイズおよび太さを特徴量として用いて属性を抽出し登録した．また，ページの統合では利用者が指定したページ内の各属性の配置状況を反映して，属性間の文字数を位置情報としてカウントすることで距離を算出し，その距離の近い属性を順に配列した．さらに，属性に対するインスタンスの類似性を検出し，その類似性を基にランキングして統合を行った．

本稿では，特に会議ごとでページの案内の内容が変わる Call for Paper (CFP) のページを具体例として実験を行った．実験結果より，再現率が 55 %となる属性辞書の作成を確認できた．また，作成した属性辞書を基に MPV のトップページの属性を効果的に抽出できたことを確認した．今後はフォントサイズや太さだけでなく配置位置の情報も含めて，属性およびインスタンスの抽出精度を高める予定である．また，別テーマの料理についても実験を行い検証する予定である．

文　献

[1] Yukiko Kawai, Daisuke Kanjo, and Katsumi Tanaka. My portal viewer for content fusion based on user's preferences. In *IEEE International Conference on Multimedia & Expo(ICME)*, 2004.

[2] 河合由起子, 官上大輔, 田中克己. 個人の選好に基づく複数 web サイトの情報収集・閲覧システム. データベースと Web 情報システムに関するシンポジウム (DBWeb), 2004.

[3] 河合由起子, 官上大輔, 田中克己. 興味と好みに基づく複数 web ページの情報融合・提示システムの検討. データ工学ワークショップ (DEWS2004), 2004.

[4] MeCab (2004). `http://chasen.org/~taku/software/mecab/`.

[5] Ion Muslea, Steven Minton, and Craig A. Knoblock. Active Learning for Hierarchical Wrapper Induction. In *Proceedings of the Sixteenth National Conference on Artificial Intelligence*, Orlando, FR, 1999.

[6] Arvind Arasu and Hector Garcia-Molina. Extracting structured data from Web pages. In *ACM SIGMOD*, pp. 337–348, San Diego, CA, 2003.

[7] Kathleen R. McKeown, Regina Barzilay, David Evans, Vasileios Hatzivassiloglou, Judith L. Klavans, Ani Nenkova, carl Sable, Barry Schiffman, and Sergey Sigelman. Tracking and Summarizing News on a Daily Basis with Columbia's Newsblaster. In *Proceedings of the Human Language Technology Conference, 2002*, San Diego, USA, 2002. ACM.

[8] Chia-Hui Chang and Shao-Chen Lui. IEPAD: information extraction based on pattern discovery. In *Proceedings of the Tenth International World Wide Web Conference*, pp. 681–688, Hong Kong, China, 2001.

[9] Feed Demo. `http://www.bradsoft.com/feeddemon/index.asp`.

[10] NewsCrawler. `http://www.newzcrawler.com/`.

[11] MyYahoo! `http://my.yahoo.co.jp/?myhome`.