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

河合由起子† 官上 大輔† 田中 克己†,††

† 独立行政法人通信総合研究所 〒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 近年，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 Web ブラウザが求められている．しかし，現在の Web の情報融合システムでは，収集した情報をシステムの仕様に基づき分類し統合して表示するため，利用者はそのシステムの分類体系やページのレイアウトにすぐに順応できず，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，収集した情報を個人の興味および知識を基に分類して統合し，さらに統合した情報を利用者の好みのページのレイアウトを通して提示できる My Portal Viewer (MPV) を提案する．MPV は，利用者の使い慣れている Web サイトのポータルページを利用することで，利用者に伏在する好みのレイアウトを通して，融合した内容を提示するという特徴をもつ．また，利用者の閲覧履歴に基づき興味の分類体系を動的に構築し，利用者の興味に基づき自律的に情報を分類し融合して，提示するという特徴ももつ．本稿では，ニュースを具体例として挙げ，利用者が好みのニュースサイトのポータルページを指定することで，そのページのレイアウトを通して，利用者の興味を基に分類され融合された記事を閲覧できる MPV について検討する．
キーワード 情報融合，個人適応，ビュアー，興味，ポータルページ

My Portal Viewer for Content Fusion based on User's Preferences

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

† Communications Research Laboratory 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 applications called "My Portal Viewer (MPV)" has been developed to provide web users with higher quality content, which is needed due to rapidly growing amount of content on the web. It provides the fused news to the 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 after gathering these pages from various news web sites. Our unique approach is that the layout of the MPV page is applied to the users' favorite news portal page, and part of the original content is replaced by the fused content. Whenever a user accesses an MPV page after browsing other news pages, the user can acquire the desired content efficiently because MPV presents a refreshed page based on the user's behavior, which refrects his/her interests and knowledge. In addition to the MPV framework, methods for replacing and selecting have been develped that are based on the user preference's using HTML table model and semantics technology is described.
Key words Information Fusion, Personalization, Viewer, Preference, Portal Page

1. はじめに

近年，膨大な Web ページから，より信頼度の高い情報をより効率的にユーザ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 Web のサービスの実現を目指し，本研究では，複数の Web ページを融合し，ユーザの欲しい情報を提供できる新たな Web ブラウザについて検討する．

今日の情報融合では，複数の Web サイトから大量の Web ページを収集し，収集した Web ページをカテゴリに基づいて分類して融合することで，カテゴリごとのまとまった情報を提供できる．これにより利用者は，各 Web サイトにアクセスすることなく，融合されたページを提供している特定のサイトに

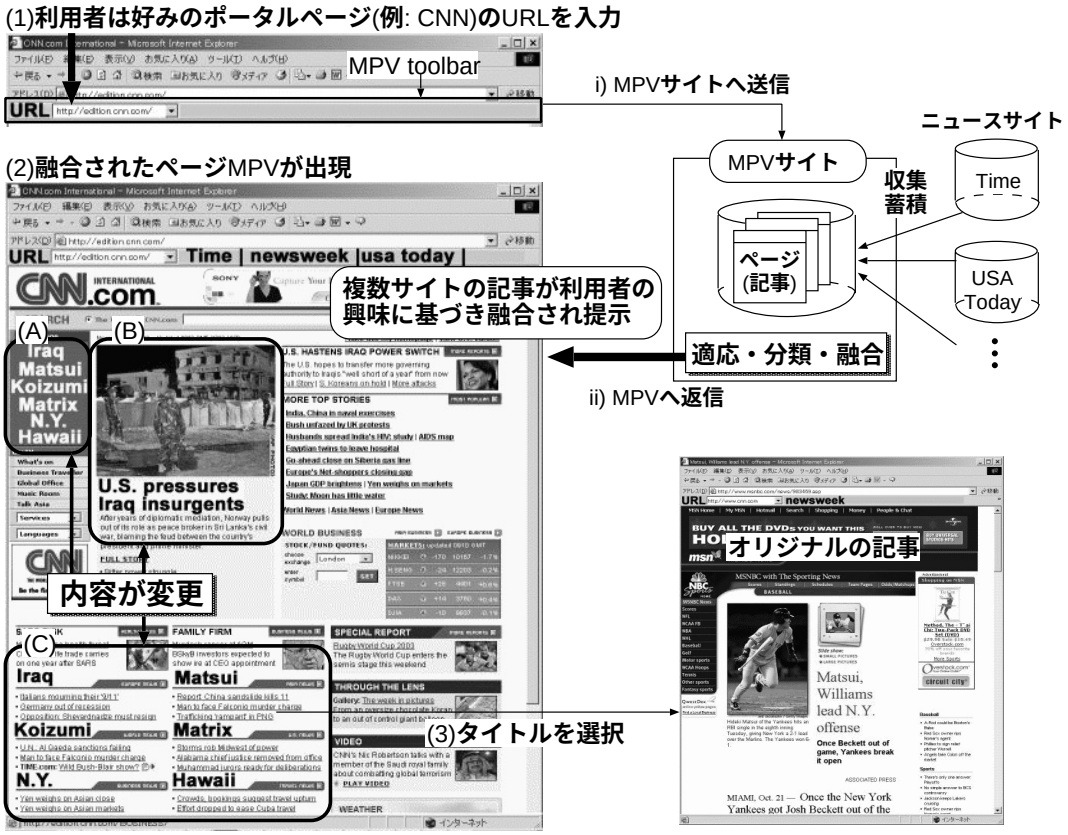

図 1　基 本 概 念

アクセスするだけで，カテゴリごとにまとめられた複数のページの情報を閲覧することができる．

しかし，カテゴリの設定は融合サービスを提供している管理者によって決められているため，利用者は融合されたページを閲覧する際，あらかじめ設定されているカテゴリを予想する必要がある．また，自身の知りたい情報がどのカテゴリに分類されているかを推測する必要もある．例えば，複数のニュースサイトの情報を融合した場合,「政治」,「スポーツ」,「国際」などの複数のカテゴリを管理者が設定しており，利用者は設定されているカテゴリの種類を把握し，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，融合されたページのレイアウトも融合サービスを提供している管理者によって決められているため，欲しい記事を迅速に見つけるためにも，融合されたページのレイアウトを使い慣れる必要がある．

本稿では，情報融合の際に必要なカテゴリの分類体系を，利用者が容易に把握できる新たな情報融合の構成法を提案する．提案する情報融合の構成法は，2 つの特徴をもつ．一つは，情報融合サイトが作成した新たな融合ページを用いず，利用者の使い慣れている Web サイトのポータルページを融合ページのインタフェースとして代用する．これにより，利用者に伏在する好みの分類体系を可視化した一つの「使い慣れているページのレイアウト」を通して，融合した内容を提示できるという特徴をもつ．もう一つの特徴は，従来の静的なカテゴリの分類ではなく，利用者の閲覧履歴に基づき興味の分類体系を動的に構築することで，利用者の興味に基づき自律的に分類し融合して，提供できるという特徴である．

本稿では，具体的にニュースサイトを例に挙げ，複数のニュースサイトの情報を利用者の興味に基づき分類し融合して，利用者の指定したニュースサイトのポータルページのインタフェースを用いて，融合された情報を閲覧できる My Portal Viewer (MPV) について検討する．

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，利用者が Viewer として代用するページのレイアウトを分析し，融合した情報を置換する内容を抽出する手法を述べる．4. 章では，ユーザの興味と知識に基づいて融合する情報を選出し，融合する手法を提案する．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

本研究では，複数の Web サイトの大量の Web ページを融合し，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の 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 Web ページのレイアウトを用いることで，利用者の好みを反映した Viewer 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，利用者の興味や知識に基づいて収集したページを分類して融合することで，個人の嗜好に適応して情報を融合する手法である．本稿では，特にニュースサイトを対象とし，以下の項目を前提とする．

- 収集されるページはニュース記事とし，ニュース記事は

「タイトル」,「記事」および「画像」で構成される(注1).

- 収集されるページには，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(注2).

2.1 好みを反映した Viewer

一般的に，利用者が Web ブラウザのブックマークやホームの機能を利用する場合，自身の興味のある情報を閲覧する目的だけではなく，普段から使い慣れて見慣れている，好みのページを利用する目的もあると考えられる．後者の使い慣れているページの利用では，利用者はページ内のどの辺りに情報が配置されているか，という空間的な情報分類ができていると考えられ，さらに，ページのリンク先の情報をある程度予測できるという特徴をもつ．そのため，使い慣れていないページと比べて，利用者は使い慣れているページの分類体系を把握できていると考えられるため，サイト内の目的の情報へ少ないクリック数で辿り着けるという利点がある．

例えば，複数のニュースサイトがある場合，欲しい情報を閲覧するために最初にアクセスするサイトは，普段から見慣れている同じニュースサイトであることが多く，さらに知識を深めたい場合に他のニュースサイトへアクセスする傾向がある．MPV では，利用者にとって使い慣れたページは，利用者の好みの分類体系を可視化したページの一つと考え，融合した情報を効果的に提示できる Viewer として利用した．

2.2 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融合された情報が，CNN のレイアウトを通して表示される．この融合結果のページが 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融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毎のキーワード，(B) 画像付きトップニュース，(C) カテゴリ毎のニュース記事のタイトル集の 3 つである．

(A) のカテゴリ毎のキーワードは，利用者の興味のあるキーワードへ変換される．例えば，CNN のオリジナルでは，(A) カテゴリ毎のキーワードは，”World”，”Worlds Business”，”Technology” などが固定で提示されるが，MPV では，”Iraq”，”Matsui”，”Koizumi” などへ変換され，提示されている．

(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，それに関する記事と画像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．

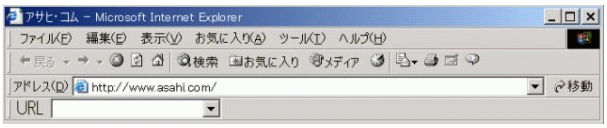

図 2　MPV ツールバー: 利用者はブランク部分に URL を入力

さらに，MPV サイトで収集した大量の Web ページは，利用者ごとのカテゴリに基づき分類され選別されて，(C) のニュース記事のタイトル集として融合される．利用者は提示された記事のタイトルを選択すると，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，MPV サイトで収集したニュースサイトの一つである，Newsweek サイトのオリジナルの記事の内容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，利用者が閲覧した記事の内容について新たな知識を得たと考えられるため，オリジナルのページの閲覧後，MPV へ再度アクセスすると (A)〜(C) の内容は書き換えられて，提示される．よって，利用者は MPV からオリジナルのページの記事を閲覧する度に，新たな融合された情報を MPV から獲得できる．

なお，MPV ツールバーをインストールし，最初に MPV 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 ツールバーに入力した URL のオリジナルのポータルページのレイアウトで，内容がそのまま表示される．ただし，(B) に関しては，MPV サイトで収集した記事のうち，最新の画像付きのニュース記事を表示するものとする．

2.3 システムの基本設計

本システムは，図 1 で示したように，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．以下では，Web のインタフェースの機能と，MPV サイトの処理の流れを示す．

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（IE5.0 以上で動作）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．図 2 に，MPV ツールバーを示す．MPV ツールバーのブランク部分に，利用者は使い慣れた好みの Web ページのポータルページの URL を入力する．MPV ツールバーは，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，融合された MPV を受信し，利用者へ提示する．MPV の表示と同時に，MPV ツールバーの URL の入力部分の横には，MPV サイトが収集し蓄積した，各ニュースサイトのドメイン名が表示される．ドメイン名の配列は，MPV で提示されているニュース記事数の多いニュースサイトの順に，左から順に配列される．

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，利用者の興味情報をもとに分類し，好みのレイアウ

(注1)：「画像」のないニュース記事もあるため，本システムでは融合方法に応じて，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が異なる．
(注2)：調査した 5 つの主要な英語のニュースサイトのうち，3 つのニュースサイトのニュース記事にメタデータが付与されていた．

トへと変換して，融合結果を返信する．融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，収集したページのうち，URL とメタデータ（Title，Description，keyword，Time）のみとする．

（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 のポータルページのソースを HTTP/1.1 GET する．

（3） 受信した利用者 ID に対応する興味情報 (詳細は 4.章) をデータベースより select する．

（4） GET したポータルページのレイアウトを解析し，変換する部分を検出する．

（5） 蓄積しているページの情報から，利用者の興味情報を用いて，融合する情報を選出する．

（6） (4) で検出した情報を，(5) の融合情報へと置換する．

（7） 置換した融合結果を MPV へ返信する．

2.4 融合情報へ置換する項目の選定

オリジナルのポータルページのレイアウトを通して，融合された情報を提供するために，我々は 5 つのニュースサイトのポータルページを分析し，置換するべき項目を選定した．分析結果より，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） カテゴリ毎のキーワード

（3） 画像とタイトルで構成されるトップ記事

（4） カテゴリ毎に分類される記事のタイトル集

（5） 広告

MPV サイトでは，利用者の興味に基づきページを分類して融合する特徴をもつ．本システムでは，従来の情報を提供するサイト側で静的に設定されている (2) と (4) を，利用者の興味に基づいたキーワードとそのキーワードに関するタイトル集へ置換する．また，(3) の画像付きトップ記事は，利用者の注目度も高いと考え，利用者の最も興味のある未読の記事へと置換する．(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．以上より，MPV では，(1)，(5) の内容はオリジナルのままとし，(2)〜(4) までの 3 項目の内容を置換するものとした．

3. 情報検出手法

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を解析し，変換すべき 3 項目の内容部分を検出する．検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．

提案する検出方法は，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．

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 5 つのニュースサイトのポータルページでも，全てに TABLE 構造がレイアウト形成として用いられていた．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．

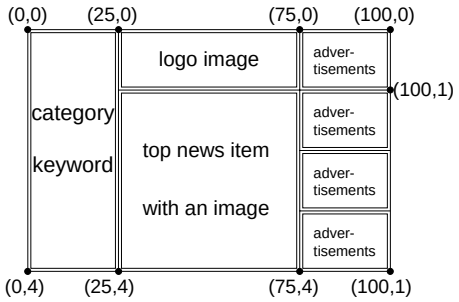

図 3 TABLE 構造を用いて xy 座標に変換しセルを抽出した結果

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 [1]．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 3 に，次の簡素化した 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"><br>top news item
                  <br>with an image</TH>
  <TH>advertisement</TH></TR>
 <TR><TH>advertisement</TH></TR>
 <TR><TH>advertisement</TH></TR>
</TABLE>
```

3.2 セルの座標値とセル内の特徴を利用した情報検出

算出した各セルの xy 座標値と，置換される 3 項目の内容の各々の特徴を基に，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 項目の情報の特徴を以下に示す．

(A) カテゴリ毎のキーワード

- キーワードに基づいてセル内の構造がパターン化．

(B) 画像とタイトルで構成されるトップ記事

- 「カテゴリ毎のキーワード」の x 座標値より大きい．
- 画像 1 枚とタイトル 1 つが同一ニュース記事をリンク．

(C) カテゴリ毎に分類される記事のタイトル集

- 「画像とタイトルのトップ記事」の y 座標値より小さい．
- 「カテゴリ毎のキーワード」と同じキーワードが存在．
- カテゴリ毎に 1 つ以上のリンク付きタイトルが存在．

以上を条件とし，ポータルページの 3 項目の内容を検出する．

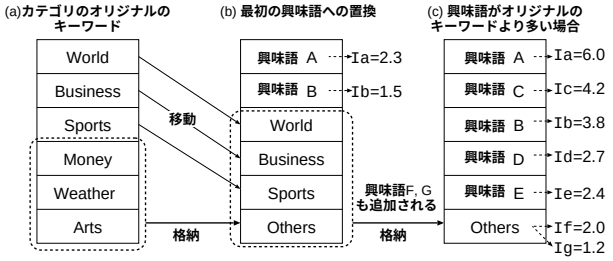

図 4　興味語とカテゴリのキーワードの置換

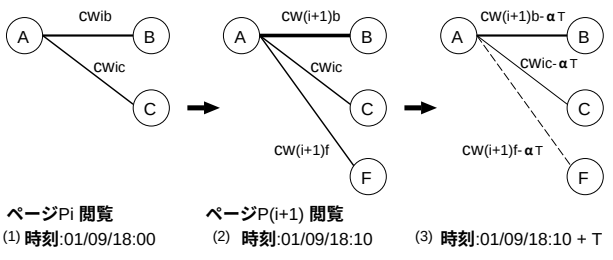

図 5　興味語 A をルートノードとする興味木の再構築

4. 情報選択および融合法

MPV サイトでは，利用者が指定したポータルページから，変換すべき 3 項目の内容を検出し，その後，収集したページの情報を利用者の興味に合わせて融合し，融合した内容へ置換する．選択および融合は，収集した各ニュース記事のページのメタデータより抽出した「単語情報のテーブル」と，利用者の興味情報である「興味語および興味木」を利用する．

4.1 ページのテーブル

MPV サイトでは，収集したページのメタデータの日付と概要から，単語とその重みに関するテーブルを作成する（表 1）．

表 1　ページのテーブル

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	$w_{(i+1)f}$

各ページの単語の抽出は，概要を形態素解析し，固有名詞，一般名詞，動詞の各単語を抽出する．単語の重みは，出現頻度 (Term-Frequency) の tf と，品詞の種類に対応した重み $W_c(c = 1 \cdots 3)$ を用いて，以下の式より算出する．

$$w_{ij} = tf \cdot W_c = \frac{log(P_i\text{中の単語 } j \text{ の出現頻度} + 1)}{log(P_i\text{中の総単語種類数})} \cdot W_c \qquad (1)$$

4.2 カテゴリのキーワード抽出と置換

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、利用者が閲覧することで各興味語の重みが変わるため，興味語は動的に選出され，MPV の内容はダイナミックに融合される．

興味語 j の選出方法は，利用者が閲覧したページ P_i〜P_n に出現する単語 j の重みの総和を算出し，総和値が閾値以上の単語とする．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，$I_j = \sum_{i=0}^{n} w_{ij}$ 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，提案システムではポータルページのオリジナルのレイアウトを変えずに内容だけを変更するため，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制限される．そこで，提案システムでは”others”というキーワードのフォルダを新たに作成し，表示できないキーワードや興味語を格納し，そのフォルダ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 4 に置換の流れを示す．まず，利用者が初めて MPV を閲覧して記事を閲覧することで，履歴情報が作成される．履歴情報には，利用者 ID と閲覧した記事の URL が追加される．この利用者の履歴情報とページの単語情報をもとに，興味語が作成される．この興味語が初めて作成された場合，最下位のキーワードが”others”というキーワードへと置換される（図 4 の (b)）．”others”には，置換された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の数 m よりより多くなった場合，$m-1$ 個目以上の興味語はまとめられて，m 個目の”others”へ追加される（図 4 の (c)）．

4.3 興 味 木

興味木は，大量のページから、各カテゴリである興味語に関連する記事を選択する際に用いる．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，選択された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は，閲覧した全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の情報を抽出し，それらの情報を基にノード間の重要度を決定し，形成される．

図 5 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 5 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 5 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 が追加）．共起語が既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 が更新）．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が，閲覧されない時間が T 以上たった場合，リンクの重

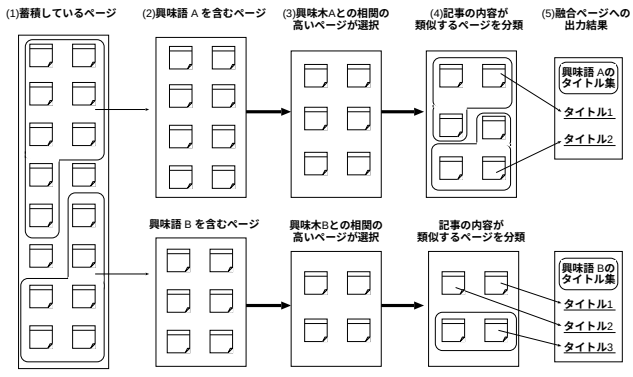

図 6　ニュース記事のページ選択の流れ

みが $\alpha\ T$ の割合で削減される（図 5 の (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　ページの選択および融合

単語情報のテーブルと興味木とを用いて，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，興味語の重みの大きい順に，単語情報のテーブルから，興味語を含む単語で日時が最新の画像付きページを探索することで，選択される．

各カテゴリの興味語に関する記事の選択の流れを，図 6 に示す．まず興味語のルートノードである興味語の出現するページを選択する（図 6 の (2)）．次に,，選択した興味語ごとのページから，興味木を基にさらに選別する（図 6 の (3)）．興味木を用いたページの選別は，選択されたページ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を利用者の興味を反映した記事として選別する．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（図 6 の (4)）．類似記事の判別は，一定時間 (プロトタイプでは 24 時間）内に作成されたページをグループ化し，グループ内のページ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ページどうしを類似ページとし，再グループ化する．グループ化されたページのうち，一つをランダムに選択し，選択したページのタイトルをタイトル集として置換する（図 6 の (5)）．

ランダムに選択されたページ以外のページのタイトルは，プルダウン式により表示される．提示されている代表のタイトルにマウスを合わせると，類似する他の記事のタイトルが出現するという方法とする．また，類似するページのタイトル間で重複する単語がある場合，それらの単語を薄く表示する．これにより，タイトルの内容の違いを強調して提示できる．

5.　MPV の検討

以下では，提案した MPV サイトと MPV ツールバーのプロトタイプによる，MPV のレイアウトの解析と興味語の抽出について検討する．

5.1　HTML 構造による情報検出手法

本稿では，利用者が指定した任意のポータルページのレイアウトの解析手法を提案した．プロトタイプでは，A〜E の 5 つのニュースサイトのポータルページについて,（１）カテゴリのキーワード,（２）トップ記事と画像の 2 項目について自動抽出を行った．表 2 に抽出結果を示す．

表 2　ページのテーブル

ポータルページ	カテゴリのキーワード	トップ記事と画像	繰り返し回数
A	○	○	11
B	○	○	11
C	○	○	8
D	○	○	7
E	○	○	8

繰り返し回数とは，3.2 節で述べたように，カテゴリ毎のキーワードを抽出する際に「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，そのパターンが繰り返されている回数を示す．抽出結果より，ポータルページ A〜E 全てに関して,（１）,（２）とも自動抽出が可能であることが確認できた．また，繰り返し回数は 7〜11 回となることがことが確認され，A〜E 以外のサイトでのカテゴリ抽出には，7〜11 回以内で繰り返されている構造に適応できると考えられる．

しかし，A〜E 以外のポータルページのトップ記事と画像に関しては，カテゴリのキーワードの隣の列にレイアウトされておらず，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，自動で抽出されなかった．

現在は利用者が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しているが，全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．解決方法としては，一つには，MPV サイトで事前にポータルページのレイアウトを解析し，自動抽出が行えなかったポータルページについてはそのページ特有の条件を追加することで抽出する方法が考えられる．もう一つは，利用者が任意に指定したポータルページを基に，利用者自身が変換したい 3 つの内容の部分を指定することで抽出する方法が考えられる．前者の解決方法は，ポータルページ全てに対して，各ポータルページごとに条件を追加することは負荷が高いと考えられる．しかし，ニュースサイトの場合レイアウトが類似しているため，追加された条件を随時適応させることで，自動抽出の適合性を向上できると考えられる．後者の場合は，置換される部分を指定するという利用者側の負荷が生じるため，負荷を軽減する容易な指定方法を検討する必要がある．

5.2　興味語の検出

MPV では，利用者の閲覧履歴に基づき，個人の興味のある単語を興味語として検出する．この興味語は収集したページの中から，個人の興味や知識に合った内容のページに分類し，選択し統合して提示する際に利用される．

プロトタイプでの興味語の検出結果では，興味語間の類似性が高くなる傾向が明らかとなった．例えば，選挙に関する記事のニュースを続けて閲覧すると，著名な立候補者や現大統領の固有名詞が選出され，さらに閲覧を続けても，興味語の各固有名詞の重みが同様に増加するため，検出される興味語の類似性が高くなるという結果であった．

現在は，興味語間の類似性を検出し，グループ化する手法を検討中である．類似性の検出には，興味語間の共起度から類似性の高い興味語を選出し，選出した興味語の重みの最大値の単

語のみを興味語として再検出する手法が考えられる．

6. 関連研究

NewsBlaster [2] [3] 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，利用者の興味に基づいた分類および融合はされておらず，利用者は欲しい情報の要約文を探索する必要がある．

FeedDemon [4] や NewsCrawler [5] 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．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が，キーワードを明示的に入力する必要があり，利用者の興味に基づいた自律的な記事の分類および融合はされていない．

MyYahoo! [6] 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．提示されるレイアウトも利用者が設定できる．しかし，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．また，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．さらに，興味が変わった場合に，設定したカテゴリを利用者が変更する必要もある．

7. まとめと今後の課題

本稿では，複数のページを融合し，ユーザの欲しい情報を提供できる新たな Web ブラウザ MPV について検討した．MPV では，利用者は使い慣れているポータルページを Viewer として利用することで，自身の好みのレイアウトを通して，融合された情報の閲覧が可能になることを示した．また，閲覧履歴に基づき興味の分類体系を動的に構築することで，利用者の知りたい情報に基づき情報を分類して融合が可能な手法を提案した．MPV の新たな情報融合の構成法により，利用者は情報融合の際に必要なカテゴリの分類体系を容易に把握できると考えられる．

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閲覧しているが，多種のテーマを異種のテーマのポータルページで閲覧可能な Viewer を検討中である．そのため，オリジナルのポータルページの情報抽出法を異種テーマへ応用する必要があり，今後の課題となる．また，情報選択法としてセマンティック Web 技術を導入し，ページと利用者の興味および知識の 3 つのオントロジーを，閲覧履歴から動的に構築する手法も検討中である．

文　献

[1] W3C Recommendation. *HTML 4.01 Specification, http://www.w3.org/TR/1999/REC-html401-19991224*. 1999.

[2] 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.

[3] R. Barzilay, K. McKeown, and M. Elhadad. Information fusion in the context of multi-document summarization, 1999.

[4] `http://www.bradsoft.com/feeddemon/index.asp`.

[5] `http://www.newzcrawler.com/`.

[6] MyYahoo!: http://my.yahoo.co.jp/?myHome.