興味に基づく複数 Web ページの情報統合・提示システムの提案

Personal Viewer for Web Page Integration based on User's Preferences

河合 由起子[1] 宮上 大輔[2] 田中克己[3]

Yukiko KAWAI Daisuke KANJO
Katsumi TANAKA

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．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，個人の興味を基に分類され統合された記事を閲覧できる MPV について検討する．

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 integrated 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. In addition to the MPV framework, methods for replacing and selecting have been developed that are based on the user preference's is described.

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Webのサービスの実現を目指し，本研究では，複数サイトのWebページを収集して統合し，利用者の欲しい情報を提供できる新たな提示システムを提案する．これまで，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統合する情報統合に関する研究が盛んに行われている[1][2][3]．情報統合により利用者は，各Web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特定のサイトに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まとめて閲覧することができる．

しかし，収集したWebページの情報をカテゴリごとに分類する際，カテゴリの設定が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いるため，利用者は統合されたページを閲覧する場合，設定されているカテゴリの項目やその項目の内容を予想する必要がある．例えば，複数のニュースサイトの情報を統合した場合，「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．

本稿では，情報統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統合の構成法を提案する．提案する統合法は，次の2つの特徴をもつ．

- 複数サイトから収集した情報を興味に基づき分類し統合
- 好みの分類体系を可視化しているViewerを利用

1つ目の特徴では，利用者の閲覧履歴を基に個々の興味や知識に基づく分類体系を動的に構築することで，複数のサイトから収集したWebページの情報を自律的に分類し統合する．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，利用者は複数サイトへアクセスし探索することなく，自身の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，容易に目的の情報を獲得することができる．例えば，「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が作成されることで，「松井」に関してまとめられた情報へ容易にアクセスでき閲覧できる．

2つ目の特徴では，統合サイトが作成した新たな統合ページをViewerとして用いず，利用者の使い慣れているWebサイトのポータルページを統合ページのViewerとして利用する．従来の統合されたページのViewerは，統合サービスを提供しているサイトの管理者によって決められており，統合ページ内から欲しい情報を迅速に見つけるため，利用者はViewの構成を把握する必要があった．Viewerの構成を把握することは，ページ内で情報がどのように分類され配置されているかという，可視化された分類体系の理解につながることであり，サイト内の情報の分類体系の把握にもつながる．そのため，統合されたページを閲覧する利用者にとって，Viewerの構成の把握はブラウジングする際の重要な知識であると言える．提案手法では，個人に潜在する好みの分類体系を可視化していると考えられる「使い慣れているページのViewer」を通して統合した情報を提示する．これにより，利用者はページ内の情報の分類体系を容易に把握でき，リンク先の情報の内容をも予測できる．

本稿では，具体的にニュースを例に挙げ，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統合して，利用者の指定したニュースサイトのポータルページのViewerを用いて，自身の知りたい情報をまとめて閲覧できるMy Portal Viewer (MPV)について検討する．

2. 基本概念とシステム設計

本研究では，複数 Web サイトの大量の Web 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，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．

[1] 正会員 独立行政法人情報通信研究機構 yukiko@nict.go.jp
[2] 正会員 独立行政法人情報通信研究機構 kanjo@nict.go.jp
[3] 正会員 京都大学大学院 情報学研究科社会情報学専攻
独立行政法人情報通信研究機構
tanaka@dl.kuis.kyoto-u.ac.jp

・ 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」,「画像」および「記事」で構成される．
・ 収集されるページには，メタデータとして「書かれた日付」,「概要」が含まれる．

2.1 好みを反映した Viewer

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，普段より興味のある情報を閲覧する目的だけではなく，使い慣れているページを利用する目的もある．使い慣れているということは，利用者はページ内のどの辺りにどういった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，その結果，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．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，統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

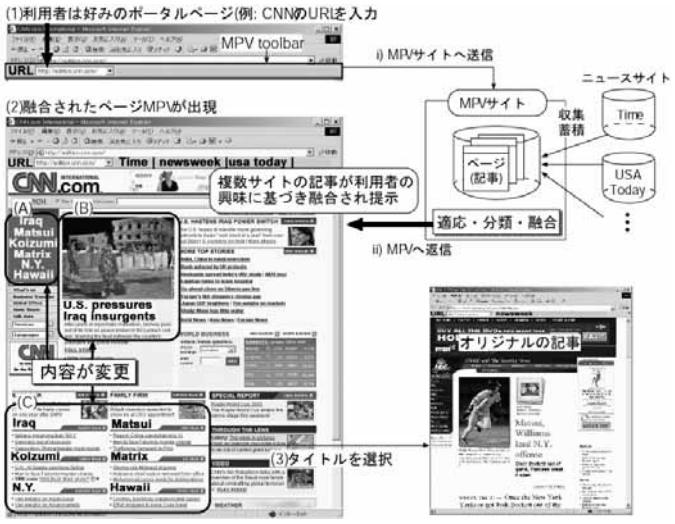

図 1 統合システム MPV の基本概念
Fig.1 Concept of MPV

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のインタフェースのブランク部分に，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．図では，利用者は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．

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．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", "Sports" などであるが，MPVでは，"Iraq", "Matsui", "Koizumi" などに変換されている．また，この新たなカテゴリに基づいて，興味のあるカテゴリ内の未読の記事がトップニュースの(B)として置換される．さらに，収集した大量のWebページは，新たな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統合され提示される．利用者はMPVの記事のタイトルを選択しクリックすると，オリジナルのニュース記事のページを閲覧できる．図では，提示されたタイトルをクリックすると，Newsweekサイトのオリジナル記事が表示されている．

また，利用者が記事を閲覧するということは，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を閲覧するたびに，新たに統合された情報を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ツールバーで指定したオリジナルのポータルページのViewerで，内容がそのまま表示される．ただし，(B)に関しては，MPVサイトで収集した各サイトの記事のうち，最新の画像付き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，Viewerを指定するMPVツールバー，統合結果のページMPV，およびMPVを提供するMPVサイトからなる．MPVのツールバーのブランク部分に利用者は好み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．MPVサイトでは，MPVツールバーから情報を受信すると，以下の手順でページを統合する．

(1) MPVツールバーよりURLと利用者IDを受信する．
(2) 受信したURLのページのソースを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5章)をデータベースよりselectする．
(4) GETしたページの構造を解析し，置換部分を検出する．
(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，統合する情報を選出する．
(6) (4)の検出情報を，(5)の統合情報へと写像する．
(7) 写像した統合結果をMPVページとして返信する．

2.4 統合情報へ写像する項目の選定

統合された情報を，オリジナルのポータルページのViewerを通して提供するMPVを実現するため，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5つの内容に基づく領域，(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)画像付きトップ記事，(4)カテゴリ毎に分類される記事のタイトル集，(5)広告，で構成されていることが明らかとなった．

MPVサイトでは，利用者の興味に基づきページを分類して統合する特徴をもつ．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．

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 Viewer の構造を解析し，変換すべき 3 項目の内容を検出する．提案手法は，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．

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[4]．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，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>
```

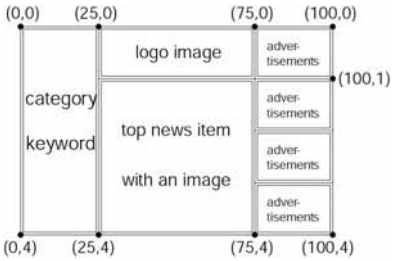

図 2　TABLE 構造を基に変換しセルを抽出した結果
Fig.2　Extracted cells based on sample HTML table model

3.2　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)カテゴリ毎のキーワード：カテゴリとなるキーワードに基づいてセル内の構造がパターン化．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)画像付きトップ記事：「(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．画像とタイトル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：「(A)のカテゴリ」と同じキーワードが存在．カテゴリ毎にリンク付きタイトルが 1 つ以上存在．(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て利用し，3 項目の内容を検出する．

4. 情報分類および統合法

MPV サイトでは，利用者が指定したポータルページから検出した 3 項目のオリジナルの内容を，蓄積したページの情報から利用者の興味に合わせて分類，選択および統合した内容へと置換する．分類および融合は，各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」を作成し，利用する．

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）．

表 1　単語情報のテーブル
Table 1　Table of word information

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} (04/01/09/12:10)	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 w_{ij} は，出現頻度(Term-Frequency)の tf と，品詞の種類に対応した重みWc (c=1,…,3)を用いて，$w_{ij}=tf$・Wcより算出する．

4.2　興味語および興味木

興味語と興味木は，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 の内容は閲覧のたびに新たに統合される．

興味語 j の選出方法は，利用者が閲覧したページ P_i～$P_{(i+1)}$ 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,…,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，$I_j=\sum_{i=1}^{n} w_{ij}$ 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．

興味木は，収集したページから，カテゴリである興味語と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．選択された興味語ごとに興味木は作成され，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．

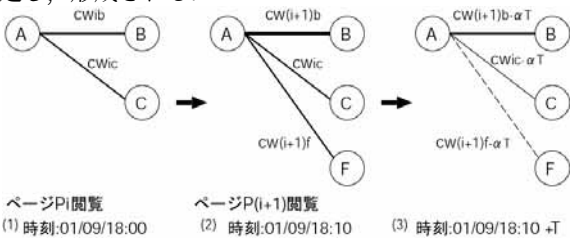

図 3　興味語 A をルートノードとする興味木の再構築
Fig.3　Reconstruction of the keyword tree rooted by the interest keyword A

図 3 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 3 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 3 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 3 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．

4.3 ページの選択および統合

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は，まず興味語のルートノードである興味語の出現する記事を選択する．次に，選択した各興味語の記事から，興味木を基にさらに選別する．興味木を用いた記事の選別は，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を利用者の興味を反映した記事として選別する．選択された記事は，さらに類似する内容の記事がある場合は統合される．類似する記事の判別は，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で，ページ内の単語の共起頻度が類似しているページとする．類似する記事の統合法は，単語数が最大のタイトルのみを表示させ，そ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提示法とした．

5. プロトタイプ

図 4 MPV の統合結果表示例
Fig.4 Example of MPV output with information integration

図 4 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討する．図は，過去に記事を閲覧し，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．閲覧する毎に，3 項目の内容は動的に置換され，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ことが確認できた．これにより，利用者は複数サイトから知りたい情報をまとめて閲覧でき，また，使い慣れている Viewer で閲覧できるため, MPV 内のカテゴリやトップ記事といった項目を容易に発見でき，さらに知りたい記事のオリジナルのページへ 1 クリックでアクセスでき, MPV の有効性が示されたと言える．また，別のニュースサイトのポータルページでも 3 項目の抽出および統合情報の提示が可能であることが確認できた．

6. おわりに

本稿では，複数の Web 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 MPV について提案し，プロトタイプによる検討を行った．MPV により，収集した情報を利用者の閲覧履歴を基に個人の興味のある情報の分類体系を動的に構築し，自律的に分類して統合し，さらに，利用者が使い慣れているポータルページの Viewer を通して統合情報を提示できた．プロトタイプシステムでの統合結果では，好みの分類体系を通して統合された情報の閲覧ができ，容易に欲しい情報の獲得が可能になることを示した．

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，多種のテーマを異種のテーマのポータルページで提供可能な Viewer を検討中である．また，情報の分類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] K. McKeown, R. Barzilay, D. Evans, V. Hatzivassiloglou, J. Klavans, C. Sable, B. Schiffman, and S. SigelmanMeehan.: “Tracking and summarizing news on a daily basis with columbia's newsblaster”, Proceedings of the Human Language Technology Conference (2002).
[2] NewscCawler: http://www.newzcrawler.com/
[3] MyYahoo!: http://my.yahoo.co.jp/
[4] W3C Recommendation.: “HTML 4.01 Specification”, http://www.w3.org/TR/1999/REC-html401-19991224

河合 由起子 Yukiko KAWAI
独立行政法人情報通信研究機構専攻研究員．2001 年奈良先端科学技術大学院大学博士後期課程修了，博士（工学）．個人適応化・セマンティック Web に関する研究・開発に従事．情報処理学会，日本データベース学会など正会員．

官上 大輔 Daisuke KANJO
独立行政法人情報通信研究機構専攻研究員．博士（工学）．1998 立命館大学理工学研究科博士課程後期過程単位取得退学．インタラクション，ユーザ適応などの研究に従事．人工知能学会，日本データベース学会など正会員．

田中 克己 Katsumi Tanaka
京都大学大学院情報学研究科社会情報学専攻教授，および，独立行政法人情報通信研究機構グループリーダ（専攻研究員）. 1976 年京都大学大学院博士前期課程修了, 京大・工博．主にデータベース，マルチメディアコンテンツ処理の研究に従事．IEEE Computer Society，ACM，情報処理学会，日本データベース学会など正会員．