類似構造を有する内容類似ページの収集・統合方式の提案

Proposal of Web Pages Integration Method based on Page Layout and Page Content

河合 由起子[1] 宮上 大輔[2] 田中克己[3]

Yukiko KAWAI Daisuke KANJO
Katsumi TANAKA

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[1][2]．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．しかしながら，これまでの MPV では，利用者が指定したオリジナルページの構成にのみ注目していた．そこで，本研究では，ページの構成だけでなく，そのページの内容に基づいて情報を分類して提示する新たな統合手法を提案する．具体的には，指定されたページの内容を解析するために，まず，大量の収集したページを基に自動作成した属性辞書を用いて，収集したページの属性とインスタンスを抽出する．次に，利用者の指定したページから同様に属性辞書を用いて属性とインスタンスを抽出し，それらを比較して類似性を検出する．本方式により，利用者は指定したページのレイアウトを通して，指定したページの内容と類似するページを，ページ内の各属性を視点にまとめて閲覧できる．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．

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 [1][2]. 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.

1. はじめに

近年，膨大なWebページから類似あるいは相違な「内容」や「構造」を有するページを集約できるような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ページを収集してページの構造や内容を基に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][2]．MPVは(1)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，(2) 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，という2つの特徴を持つ．これらの特徴より，利用者は知りたい情報に関して分類された情報を，使い慣れているページの構造を通してまとめ閲覧できる．しかしながら，同一のトピックに関する情報をまとめて閲覧できるが，ページの内容の類似性や相違性を反映した分類や統合が困難であった．

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，新たな統合方式を提案する．本稿では，特に会議ごとでページの内容と構成が変わるCall for Paper (CFP)ページを対象とする．本方式では，新たに次の2つの特徴を持つ．

・ 利用者が指定したページ全体の内容と類似するページを分類すること．
・ ページ内で利用者が指定した属性（単語）を抽出し，その属性のインスタンス（内容）を基に分類し統合すること．

上記の特徴を実現するために，我々は，ページ内の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．これら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ページを収集し，収集した大量のページより属性辞書を自動作成する．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．

以下，2.では，新たな提案方式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，4.で実装したシステムを基に検証を行う．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

2.1 基本概念

基本概念を図 1 に示す．利用者は，興味のある学会の CFP

[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

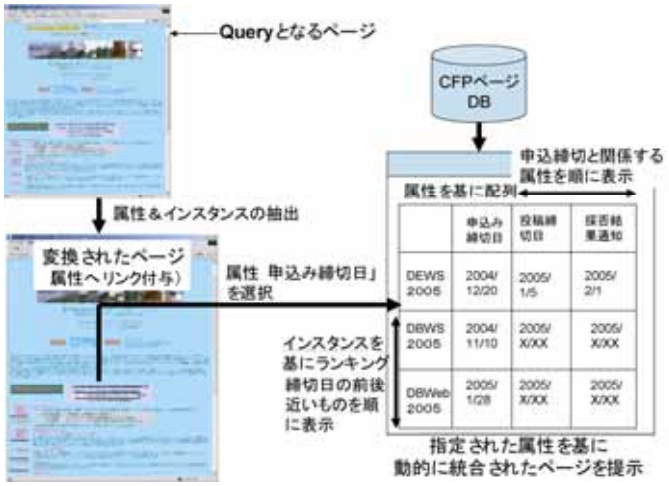

図 1　統合システムの基本概念
Fig.1 System Concept

をサンプルとして指定する．図中では，利用者は MPV ツールバーに 2005 年 2 月に開催された国内会議 DEWS の CFP の URL を例として入力している．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トップページが表示される．このトップページは 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，利用者の指定した特定の会議（DEWS2005）のトップ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，DEWS2005 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利用者へ提示する．内容として変更されているのは，抽出した属性の属性名に対してリンクが付与されている点である．ここで属性名とは，DEWS2005 の CFP ページの場合，「論文締切」「採否結果通知」等にあたる．

2.2　属性とインスタンスに基づいた統合ページ

本方式では，多種の CFP ページを収集しており，CFP ページ間で同じ属性でもインスタンスが異なる．そのため，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 1 ページに統合すると閲覧する手間は O_{nm} 必要となる．そこで，類似する属性ごとに統合した別ページを用意する．利用者は指定したページ内の任意の属性のリンクアンカーを選択すると，その属性のインスタンスが類似する CFP について統合したページを閲覧できる．

提案システムでは利用者の指定したページと類似する構造と内容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「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」「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「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」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．

2.3 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（CFP)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)抽出した属性名が出現している箇所にリンクを付与して利用者へ提示.

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名と近い位置の属性名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名をもつCFPページを選択.
(4)(3)で選択したページから，利用者の指定した属性名に対するインスタンスを抽出.
(5)(4)で抽出したインスタンスとMPVトップページのインスタンスを(3)で選択された全てのページで比較
(6)(5)の比較結果より，類似性の高いインスタンスをもつCFPページを順にランキング.

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

提案手法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，収集した大量のWebページを利用することで自動作成される．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．この文章出現頻度(document frequency)の高いキーワードを属性名として辞書に登録する．

本抽出方式では，属性辞書の作成に単語の出現頻度以外に文字の表現形式（サイズや太さ）も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目指す．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CFPページのテキストを収集.
(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サイズと太さを n より検出.
(6)文章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.

$$WS_i = df \times S_i$$

df = log(単語iの出現回数+1)/log(単語の種類ｓ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.

$$WB_i = df \times B_i$$

B_i = (単語iの太さの総和)/(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，ページ内の属性間に出現する文字列の特徴を基にインスタンスを検出する．本稿では文字列の抽出法について述べる．

(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，本方式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)提示しているトップページの各属性の位置情報を検出.
 (a)トップページの各属性の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)(b)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．

・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[3]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.

4.2 属性辞書

表 1 自動作成された属性辞書
Table 1 Attribute dictionary

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	環境	シンポジウム	締切	科学
工学	電子	テーマ	分野	問題
会員	支部	セッション	協会	ホームページ
社会	言語	大会	方法	下記
頻度と太さの重みWB_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	方法	エネルギー	日程	カテゴリ
論文	大会	一般	対象	言語
国際	システム	申込	詳細	要綱
技術	経済	科学	内容	結果

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となるページ内のレイアウトの情報も含めた重み付けを検討する.

4.3 指定ページの属性とインスタンス抽出

利用者が指定した CFP ページをトップページへ変換した結果について考察する．任意の CFP ページの統合結果のページの 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．抽出されなかった属性名としては「目的」「その他」などであった．「目的」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」に関しては，98 番目に検出されていた．属性抽出に関しては，一度作成した属性辞書の単語を辞書作成の際のキーワードとして再利用することで，再収集・再構築を行い精度の向上を目指す予定である．また，属性辞書を半自動で作成することで，属性辞書を編集して実運用性を高めることも必要であると考えている．インスタンスの抽出に関しては，属性間の文字列全てをインスタンスとして抽出しているが，各ページで共通する属性の文字列の相異を検出することで，インスタンスの抽出精度の向上を検討中である．

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[4][5][6]．また，収集した情報の分類も行われている[7][8][9].

STALKER[4]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．本方式でも，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[5]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，実用性が低く適用が困難である．

FeedDemon[7]や NewsCrawler [8]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![9]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．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． 本方式では，利用者はカテゴリなどの設定をする必要がなく，好みのページを指定するだけで，そのページの構造と内容，さらにページ内の単語に基づいて統合され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情報統合方式について提案した．提案方式では，利用者は指定したページの構成を Viewer として利用できるだけでなく，指定したページの内容と類似したページを分類して統合できる．本方式では，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利用者が指定した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となるページの構造情報も含めて，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 (ICME2004)*, 2004.

[2] 河合由起子, 官上大輔, 田中克己: 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, 日本データベース学会Letters (DBSJ Letters), Vol.3, No.1, pp.17-20, 2004

[3] MeCab: http://chasen.org/~taku/software/mecab/

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

[5] Chang Chia-Hui and Lui Shao-Chen.: IEPAD: Information Extraction based on Pattern Discovery. *In Proceedings of the Tenth International World Wide Web Conference*, pp.681-688, 2001.

[6] Arvind Arasu and Hector Garcia-Molina.: Extracting Structured Data from Web Pages, *ACM SIGMOD*, pp.337-348, 2003.

[7] FeedDemon: http://www.bradsoft.com/feeddemon/index.asp/

[8] NewsCrawler: http://www.newzcrawler.com/

[9] MyYahoo!: http://my.yahoo.co.jp/

河合 由起子 Yukiko KAWAI

独立行政法人情報通信研究機構勤務．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．1976 年京都大学大学院前期博士課程修了，工学博士．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, 人工知能学会，日本ソフトウェア科学会，情報処理学会，日本データベース学会会員．