3G7-2

コンテンツエージェントに基づくWebClipシステムの試作

Implementing a Web Clip System based on a Content Agent

浅見昌平*1 Shohei Asami　驛昌弥*1 Masaya Eki　大囿忠親*1 Tadachika Ozono　新谷虎松*1 Toramatsu Shintani

*1名古屋工業大学大学院 工学研究科 情報工学専攻
Dept. of Computer Science and Engineering, Graduate School of Engineering, Nagoya Institute of Technology

To clip a desired part of a content on a web page, a web page is divided into a needful area and a needless area. In this paper, We present the WebClip System that storage parts of web archives. The system is used for making supporting an archive for a needful area on a web page. We describe a DOM tree analysis and agent's interactions with a system user.

1. はじめに

Web 上の様々なデータは，出版物に比べ，安定して参照できない問題点がある．そこで，世界各国の国立図書館を中心に，Web 上の情報を文化資産として収集し，長期保存する Web アーカイブが行われている [Abiteboul 02, 廣瀬 03]．Web アーカイブは，クローラによって自動的に収集され，一定周期ごとの更新を時系列順に保存される．収集された Web ページは，収集時点での状態のまま保存されるため，元の Web ページが書き変わったとしても，過去の状態を参照することができる．

ユーザは，閲覧中の Web ページに有用な情報を発見した際，その情報を再び閲覧するために，ブックマークへの登録を行う．ブックマークに登録された Web ページは，将来にわたって閲覧時の状態を保っているとは限らないため，ユーザにとって Web アーカイブは有効な手段である．多くの Web アーカイブ機関では，保存した Web アーカイブを公開しており，ユーザは保存された過去の Web ページを閲覧することができる．しかし，Web アーカイブ機関では，クローラが自動的に収集するため，ユーザが求める状態の Web ページを保存しているとは限らない．また，Web アーカイブは，1 つの Web ページ全体を保存するため，ユーザによっては，必要な情報以外の部分はノイズになると考えられる．

本稿では，ユーザが閲覧中の Web ページから，保存したい領域を選択し，その選択領域をアーカイブする手法を提案する．Web ページの一部分だけを保存した Web アーカイブを WebClip と定義する．ユーザ自身が保存する対象を決定することで，ユーザのニーズに適応した WebClip を作成できる．2 章では，保存したい領域を選択するための手法の説明，および関連研究の紹介を行い，3 章では本研究で実装した WebClip システムについて述べる．4 章では，本システムの実験結果を示し，最後に 5 章で本稿をまとめる．

2. 保存領域の選択

Web ページにおいて，ユーザが保存したい情報は，一部分であることが多い．例えば，ブログに投稿されたプログラミングメモ，あるいはニュースサイトに配信された事件の記事である．図 1 の例に示すように，Web ページには，保存したい

連絡先: 浅見昌平，名古屋工業大学大学院 情報工学専攻 新谷研究室，〒 466-8555 名古屋市 昭和区 御器所町 名古屋工業大学，TEL:(052)735-7968，FAX:(052)735-5477，asami@toralab.ics.nitech.ac.jp

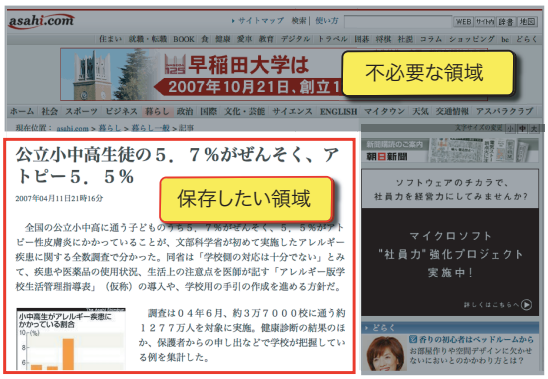

図 1: Web ページの必要な部分と不必要な部分の例

記事以外に，メニュー，広告，および空白の領域が含まれる．ユーザが作成する Web アーカイブには，図 1 の不必要な領域を含めないように，保存領域を選択することが求められる．

2.1 Document Object Model

DOM(Document Object Model) は，HTML 文書および XML 文書のための API として定義されている．アプリケーションは，DOM を操作することによって HTML 文書の構造を取得でき，Web ページにおける HTML タグの描画位置まで取得できる．しかしながら，DOM ツリーは，Web ブラウザを介して見た通りの構造をしていない．同じ Web ページのように見えたとしても，全く異なった HTML 文書であることがある．例えば，図 2 に示す (a)，(b) は，それぞれ Web ブラウザを介してみたスクリーンショットである．(a) の DOM ツリーは (c) で示され，(b) の DOM ツリーは (d) で示される．(c) と (d) を比較すると，全く異なった HTML で記述されている．

見た目が類似した Web ページのコンテンツでも，DOM ノードの構造が異なるため，DOM ツリーの解析も異なった結果が出力される．例えば，図 2 中の「HTML タグ構造の例です.」というテキストノードは，(c) では<BODY>の子ノードであるが，(d) では<BODY>から 4 階層下の子孫ノードである．Web ブラウザのウィンドウの大きさが各ユーザで異なるため，同じ DOM ツリーの構造でも，描画されたときの配置は異なる．

単に DOM ノードの階層関係を辿るだけでは，描画される DOM ノードの位置を考えた抽出が難である．そこで，本研究では，ユーザの Web ブラウザ上で，描画される DOM ノード

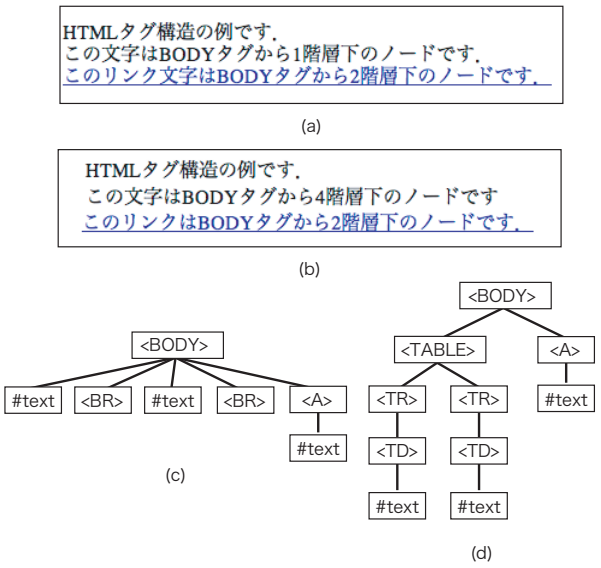

図 2: 異なる HTML 構造をした Web ページの例

図 3: Clipmarks の実行例

の位置情報を取得し，保存領域に描画されている DOM ノードを抽出する．

2.2 DOM ノードの抽出

保存領域に描画されている DOM ノードを抽出する手法について述べる．本研究では，ユーザの保存したい領域を，長方形の枠で囲まれた領域と仮定する．すなわち，保存領域に描画されている DOM ノードは，長方形に含まれる DOM ノードの位置を調べることで抽出できる．

Web ページに描画される DOM ノードの位置情報は，DOM ノードのプロパティから取得できる．DOM ノードは長方形の形で描画されるため，取得する位置情報は，長方形の左上の頂点 X 座標 (X)，Y 座標 (Y)，幅 (W)，および高さ (H) である．保存領域の長方形 (X_u, Y_u, W_u, H_u)，i 個目の DOM ノードの位置情報 (X_i, Y_i, W_i, H_i) に関してユークリッド距離 (U_i) を求め，U_i が最も小さい DOM ノードを抽出する．

2.3 関連研究

HTML の抽出に関して，レイアウトを考慮した Web ページを分割する研究が行われている．Web ページを分割することによって，描画位置が類似した DOM ノードをまとめることができる．[Zou 06] らは，Web ページをいくつかの区分に分割するために，Zone ツリーアルゴリズムを提案した．W3C によって勧告されている DOM ノードの分類を用いて，1 つ以

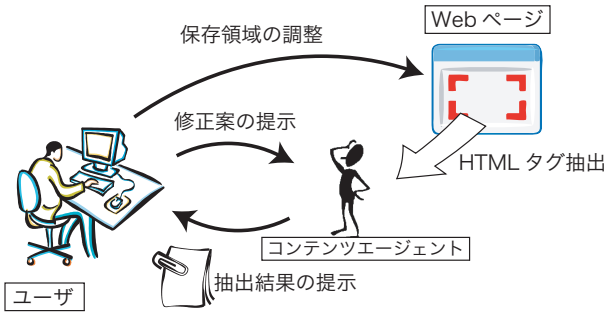

図 4: コンテンツエージェントによる仲介

上の DOM ノードを 1 つの Zone にまとめる．Web ページを分割することによって，その後の検索ステップにおける高速化，および精度の向上に繋がると主張している．本研究において，ユーザが作成する保存領域は，Web ページの中で分割された 1 つ以上の区画に該当すると考えられる．描画位置から DOM ツリーを解析する点で，本研究の抽出処理と関連性がある．

Web ページの一部分を保存するための Web サービスに，Clipmarks[*1]がある．ユーザは，閲覧中の Web ページから，保存したい DOM ノードを直接選択する．選択された DOM ノードは，結合されて保存される．図 3 に Clipmarks の実行例を示す．Clipped と記された領域は，その領域の DOM ノードを選択したことを表す．ユーザは直感的な操作によって，保存する情報を決定できる．

3. WebClip システム

本研究では，ユーザが Web ページを閲覧中に，必要な領域だけを保存可能な WebClip システムを実装した．必要な領域だけを保存した Web アーカイブを，WebClip と呼ぶ．WebClip を作成する際，エージェントの仲介によって，ユーザが必要な情報を選択することを支援する．

3.1 コンテンツエージェントの提案

DOM ノードの抽出における問題として，DOM ノードが Web ページに描画されている通りの構造をしていないことを述べた．そのため，ユーザが描画位置から直感的に情報を選択する際，意図した情報を含む DOM ノードを取得することは困難である．本研究で用いる DOM ノードの抽出手法では，抽出結果がユーザの意図した情報であるのか知ることができない．そこで，ユーザに抽出結果を提示し，意図した情報でなければ再度抽出を行う，仲介役のコンテンツエージェントを提案する．

図 4 に，コンテンツエージェントによる，ユーザと抽出結果の仲介を示す．コンテンツエージェントは，ユーザが指定した保存領域から DOM ノードの抽出を行い，ユーザに結果を提示する．ユーザは提示された結果に対し，修正案をコンテンツエージェントに与えることができる．修正案は，「より大きな領域」,「次の候補」,「前の候補」，および「より小さな領域」である．また，保存領域の大きさ，または位置を調整することで，コンテンツエージェントに再び，DOM ノードを抽出させることができる．ユーザは，意図した DOM ノードが抽出するまで，保存領域の調整，および修正案の提示を繰り返す．

*1 http://clipmarks.com/

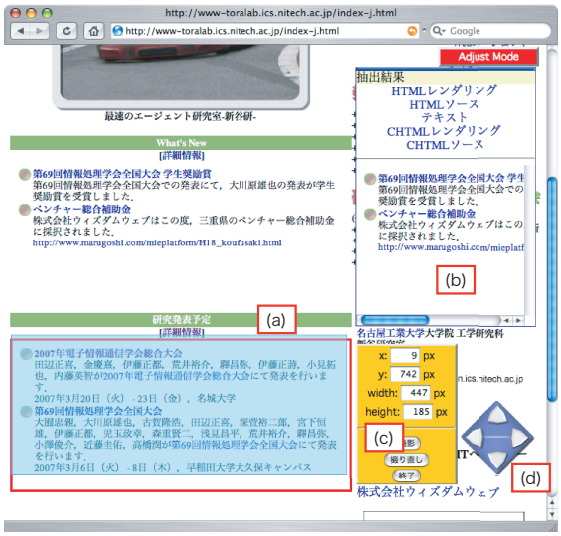

図 5: WebClip システムの実行例

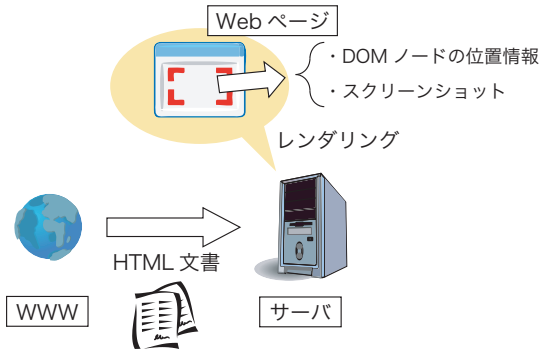

図 6: サーバ・レンダリング方式

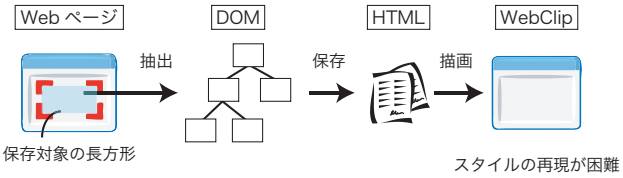

図 7: WebClip システムにおける処理の流れ

3.2 DOM ノードの抽出結果の修正

ユーザが修正案を提示すると，コンテンツエージェントは抽出結果の修正を行う．抽出結果は，DOM ツリーの操作を用いて修正され，修正案と DOM ツリーの操作は次のように対応する．

「より大きな領域」 抽出結果の親ノード (parentNode)
「次の候補」 抽出結果の兄弟ノード (nextSibling)
「前の候補」 抽出結果の兄弟ノード (previousSibling)
「より小さな領域」 抽出結果の 1 番目の子ノード (firstChild)

DOM 操作によって抽出結果を修正し，修正結果を再びユーザに提示する．また，ユーザが保存領域を調整した際，再び全ての DOM ノードと保存領域のユークリッド距離を求め，新たな抽出結果を提示する．

3.3 保存機能

本システムでは，WebClip をサーバ上のデータベースに格納し，サーバ上で管理する．図 5 に，WebClip システムの実行例を示す．(a) の青い長方形は保存領域を表している．(b) は抽出結果を，(c) は保存領域の位置情報，およびメニューを，(d) は抽出結果の修正を行うインタフェースを，それぞれ表している．ユーザは，(c) のメニューから保存ボタンを押すことによって，提示された抽出結果を保存することができる．

コンテンツエージェントは，抽出結果，Web ページの URL，保存領域の情報 (X_u, Y_u, W_u, H_u)，およびユーザの Web ブラウザ情報を，サーバ上のシステムに送信する．ユーザの Web ブラウザ情報は，ウィンドウの幅 (W_b)，ウィンドウの高さ (H_b)，およびユーザエージェントである．サーバ上のシステムは，受信した WebClip をデータベースに格納する．ユーザは，保存した WebClip を，サーバ上の Web ページを介して閲覧することができる．

3.4 更新機能

Web アーカイブは，保存元の Web ページが更新されたとき，新たに保存元の Web アーカイブを作成することが求められる [小城 05]．Web アーカイブでは，クローラが保存元の Web ページの更新を調べ，更新が確認できれば新たに Web アーカイブを作成する．しかし，本システムはユーザの Web ブラウザ上で動作するため，WebClip を自動的に更新することができない．DOM ノードを抽出する処理では，Web ページに描画される位置情報を必要とするため，Web ページをレンダリングしなければならない．WebClip を自動的に更新するためには，ユーザの閲覧状態をサーバ上で再現することが必要である．Web ページをサーバ上でレンダリングする手法として，置田らはサーバ・レンダリング方式を提案している [置田 06]．

図 6 にサーバ・レンダリング方式の概要を示す．サーバ・レンダリング方式は，対象の URL から HTML 文書を取得し，サーバ上で仮想的に Web ページをレンダリングする手法である．サーバ上で Web ページをレンダリングすることによって，DOM ノードが描画される位置を取得し，WebClip の作成を自動化することができる．レンダリングの際，ユーザの Web ブラウザ情報を与えることで，ユーザの閲覧状態が再現可能である．本システムでは，最初の WebClip をユーザが Web ブラウザ上で作成し，更新の作成をサーバ上のプログラムが自動的に行う．

3.5 実装

図 7 に本システムにおける処理の流れを示す．ユーザは閲覧中の Web ページにおいて，ブックマークレットから WebClip システムを起動する．ユーザは WebClip システムを用いて，保存領域の長方形を作成する．本システムは，DOM ツリーを解析することによって，ユーザが指定した保存領域の DOM ノードを抽出し，サーバに HTML 文書として抽出結果を保存する．ユーザが作成した WebClip を閲覧する際，保存された HTML 文書は Web ページとしてレンダリングされる．

WebClip を閲覧する際の問題点として，ユーザが WebClip を作成したときの状態が再現できないことが挙げられる．ユーザ閲覧時では，抽出した DOM ノード以外にも DOM ノードが存在し，Web ページでのレイアウトに影響を与えている．例えば，抽出した DOM ノードの親ノードにスタイルシートが指定されていたとき，スタイルシートが継承され，レイアウトに影響される．そこで，サーバ・レンダリング方式を利用して，閲覧時の Web ページのスクリーンショットを作成する．サー

表 1: 実験結果

評価項目	Web ページ数
修正することなく抽出できた	23
修正することで抽出できた	14
抽出できなかった	9
システムが実行できなかった	4

バ・レンダリング方式では，閲覧時の Web ページを仮想的にレンダリングするため，ユーザが閲覧した状態を再現できる．また，Web ブラウザ上で動作するプログラムは，セキュリティの関係上，スクリーンショットを撮ることができないが，サーバ上では API を用いることでスクリーンショットを撮ることが可能である．

4. 評価

本研究で実装した WebClip システムを用いて，内容が異なる 50 の Web ページから，任意の領域を保存する実験を行った．まず，対象の Web ページにおいて，本システムを起動する．次に，閲覧した際の位置関係から，保存対象の文章のまとまりを選ぶ．例えば，図 5 の (a) のように，文章の段落を保存対象とした．保存対象から得た抽出結果が，意図した文章のまとまりであるかを判断する．

評価方法は，「修正することなく抽出できた」，「修正することで抽出できた」，「抽出できなかった」，および「システムが実行できなかった」の 4 段階で評価する．現在，<FRAME>が使用されている Web ページでは，本システムを実行することができないため，「システムが実行できなかった」という項目を含めた．

表 1 に実験結果を示す．37 の Web ページにおいて，保存領域から意図した DOM ノードを抽出できた．修正が必要だった Web ページにおいて，ほとんどが 3 回以内の修正によって，意図した DOM ノードが抽出できた．

抽出できなかった Web ページの主な原因は，<TABLE>内の DOM ノードであった．抽出したい情報が，<TABLE>内の複数の<TR>，<TD>に分割されている場合に，修正しても思った通りの DOM ノードが抽出できなかった．

図 8 に抽出に失敗した Web ページの構造を示す．(a) は<TABLE>によって構成される表組みのレイアウトである．この表組みから，(a) に示すように 2 列目にある情報を抽出するとき，(b) の結果が得られる．抽出したい DOM ノードが DOM ツリー内で離れて存在するため，2 つの DOM ノードを含む<TABLE>が抽出された．この<TABLE>に対して修正を行っても，2 つの DOM ノードのみを抽出することはできない．実際の Web ページでは，レイアウトのために用いられている (b) のような<TABLE>が存在するため，意図した情報だけを抽出できない領域が存在した．

5. まとめ

本稿では，Web ページから必要な情報だけを Web アーカイブするための，WebClip システムについて述べた．WebClip システムでは，ユーザが Web ページの保存領域を指定し，その領域から該当する DOM ノードを抽出する．DOM ノード抽出の際，Web ページで閲覧した状態と，DOM ツリーの構造が異なる問題があり，位置情報を考慮した抽出処理が必要である．また，抽出結果を，ユーザが意図した保存対象と一致させるため，エージェントによる仲介を提案した．ユーザは抽出結果に対し，修正案をエージェントに提示することができる．エージェントは，修正案を抽出結果に反映させ，結果をユーザに提示する．エージェントの仲介により，ユーザが必要な情報を選択するための支援を実現した．

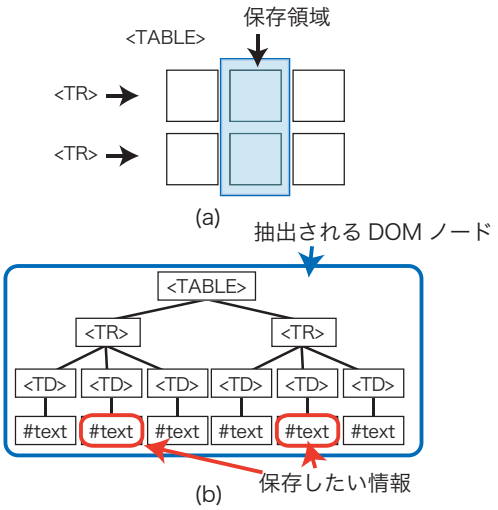

図 8: TABLE タグによる抽出失敗例

本システムを用いた抽出実験を行い，50 の Web ページから WebClip を作成した．37 の Web ページの領域に対しては，修正を含め上手く抽出できたが，9 の Web ページの領域からは抽出に失敗し，4 の Web ページではシステムを実行できなかった．今後の課題として，抽出に失敗した原因である<TABLE>への対応，<FRAME>を用いた Web ページへの対応が挙げられる．

参考文献

[Abiteboul 02] S. Abiteboul, G. Cobena, J. Masanes and G. Sedrati: “A First Experience in Archiving the French Web,” Proc. of the Sixth European Conference on Research and Advanced Technology for Digital Libraries (ECDL ’02), pp. 1-15, 2002.

[廣瀬 03] 廣瀬信己: “国立図書館におけるウェブ・アーカイビングの実践と課題,” 情報処理学会研究報告, Vol. 2003, No. 51, pp. 95-111, 2003.

[小城 05] 小城正士, 廣瀬信己, 河野浩之: “Web アーカイブにおける時系列閲覧: 単一コレクションへの適用,” DBSJ Letters, Vol. 4, No. 1, pp. 153-156, 2005.

[置田 06] 置田誠, 山口典男, 重松隆之, 高橋修, 宮本衛市: “携帯電話機用 WEB ブラウザのサーバ・レンダリング方式の提案と実装評価,” 情報処理学会論文誌, Vol. 47, No. 7, pp. 2107-2116, 2006.

[Zou 06] J. Zou, D. Le and G. R. Thoma: “Combining DOM Tree and Geometric Layout Analysis for Online Medical Journal Article Segmentation,” Proc. of the Sixth ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL ’06), pp. 119-128, 2006.