断片化Web情報の構造化に基づくコンテンツ閲覧支援環境について

大囿 忠親† 伊藤 太樹† 平田 紀史† 柿元 宏晃† 白松 俊† 新谷 虎松†
†名古屋工業大学大学院情報工学専攻

1 はじめに

本研究では，断片化されたWebページなどの情報をユーザに構造化させることで，そのユーザの観点によるWebを再構成することを支援し，その結果として，Webコンテンツの閲覧を支援するためのシステムの実現を目指している．ここでは，Webページを適切な意味単位に基づき分割し，ユーザに分割されたWebページの断片を収集し再構成させることで，従来とは異なる方法でユーザの意見や好みを収集することを目指している．ここでは，断片化されたコンテンツを構造化するためのエディタを提供し，さらに，構造化されたコンテンツからモバイル用など新たなコンテンツとして再利用するためのシステムを提供する．これにより，ユーザがどのようなコンテンツを収集，構造化，そして，再利用するかを追跡することが可能になる．これらの情報を用いることで，ユーザの観点から構造化された，新たな情報を得ることが可能になる．

本稿では，断片化Web情報の構造化について議論し，断片化Web情報の構造化に基づくコンテンツ閲覧支援環境について述べる．

2 断片化Web情報

本研究では，Webページ内に含まれる複数のコンテンツをそれぞれ断片化Web情報と呼ぶ．本研究では，Webページを断片化する方法として次の2つの方法を用いた．1つ目の方法は，Webページをブロックと呼ぶ意味的なまとまりのある単位に分割する方法である．Webページをブロックに分割し，それぞれのブロックを断片化されたコンテンツとする[2]．図1は，ニュース記事のページを主となるコンテンツや広告などのブロックに分割した例である．

もう一つの方法は，Webページ内のコンテンツに付箋を付与する方法である．ここでの付箋とは，紙の付箋のようにWebページ内の一部分を指定可能なアノテーションである[3]．付箋が付与された周辺のコンテンツを断片化Web情報とする．

A Web Browsing Support System based on Structured Fragments of Web Contents
†Tadachika OZONO †Taiki ITO, Norifumi HIRATA, Hiroaki KAKIMOTO, Shun SHIRAMATSU, Toramatsu SHINTANI
†Department of Computer Science and Engineering, Graduate School of Engineering, Nagoya Institute of Technology

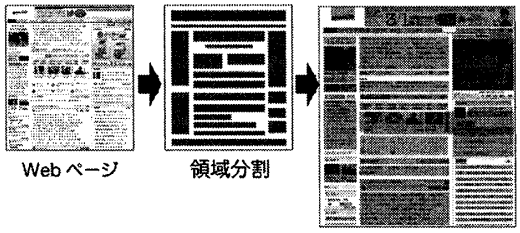

図1: ブロックへ分割されたWebページ

本研究では，複数の断片化Web情報間の関係を定義することを断片化Web情報の構造化と定義する．例えば，Webページは，Webページ作成者の観点で断片化Web情報を構造化したものと見なすことができる．Webページを構造化された断片化Web情報として解析することで，Webページのレイアウトに内在するWebページ作成者の意図を抽出し，検索エンジンの精度向上などに応用できる．

3 断片化Web情報の構造化

断片化Web情報の組み合わせ方には，ユーザのなんらかの好みや意見などの関する情報が含まれると期待される．Web上のマッシュアップと同様に，ユーザの好みや意見を文章として陽に記述したコンテンツを作成するよりも，既存のコンテンツの断片を組み合わせることは簡単である[1]．断片化Web情報の組み合わせからユーザの意見や好みを抽出することが可能になれば，意見より広いユーザを対象とした意見抽出なども可能になる．

本研究において，断片化Web情報間の関係とは，2つの断片化Web情報間のラベル付きのリンク，および，複数の断片化Web情報のグループ，の2つである．ラベル付きのリンクにより，断片化Web情報をノードとする有向グラフを形成する．ラベルは，リンクの両端の断片化Web情報間の関係を表す．例えば，修辞関係のように意味を表すリンク，上下左右などの隣接関係を表すリンク，もしくは，HTMLにおけるリンクのように参照を表すリンクなどが考えられる．

ここで重要な点は，閲覧者の観点からのリンクを実現することである．従来のHTMLにおけるaタグに基づくリンクは，コンテンツ作成者の観点によるリンク

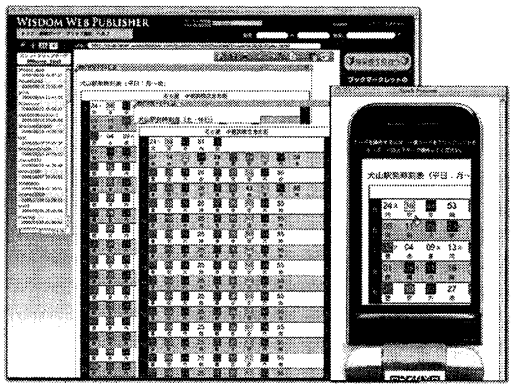

図2: 断片化Web情報の構造化エディタ

であるといえる．閲覧者の観点によるリンクを自由に作成できる環境を実現することで，より利便性の高いコンテンツ処理が可能になる．ブックマーク，アノテーション，もしくは，タグなども閲覧者の観点によるリンクであると見なすことができるが，対象となるコンテンツの粒度が，基本的にはページ単位である点が課題であった．異なる複数のWebページに含まれているが，類似の内容を表す断片化Web情報間のリンクは，Webページを解析する上で重要な情報になる．

複数の断片化Web情報のグループとは，同じWebページに属する断片化Web情報のグループのように，リンク先が複数の断片化Web情報のグループを指す場合などに用いる．

4 コンテンツ閲覧支援システム

断片化Web情報を構造化するシステムの実装例として，Web上やPC上の情報をシームレスに組み合わせて再利用することで携帯電話上でのコンテンツ閲覧を支援するシステムを試作した．ここでの断片化Web情報は，Web上の情報にとどまらず，PC上のワープロ書類やプレゼンテーション資料などを含んでいる．

本システムでは，断片化Web情報の収集，断片化Web情報の構造化，そして，断片化Web情報の出力，の3段階でコンテンツの閲覧を支援する．

断片化Web情報の収集では，Webページを自動的に断片化するシステム，Webページを手動で断片化するシステム，および，Webページへの付箋付与システム，を実装した．Webページを自動的に断片化するシステムでは，WebページをBlockNetと呼ぶ構造に変換することで解析精度を高めた．Webページを手動で断片化するシステムは，Webブラウザ上やPC上の画面でで断片化したい範囲を指定するシステムを実装した．Webページへの付箋付与システムでは，貼り付け時と解析時における付箋の貼り付け位置が保存されるようにした．

断片化Web情報の構造化では，断片化Web情報を組み合わせることで構造化するためのエディタを実装した．図2は，本研究で試作した断片化Web情報の構造化エディタである．図2では，PC用のWebページから収集した電車の時刻表を断片化Web情報として構造化している．ここでは，2つの時刻表の断片が構造化されている．

断片化Web情報の出力では，断片化Web情報の構造を利用することで，プリンタ，携帯電話，および，スマートフォンなどの出力先の種類に対して出力形式を調整することが可能である．構造化された断片化Web情報がどのように出力されるかを調べることには意味があり，例えば，モバイル機器ではどのようなコンテンツが求められているかを調べることが可能になる．図2の右側の携帯電話の画面は，PC用のWebページから生成された携帯電話用のコンテンツが表示されている例を表している．

5 おわりに

本稿では，断片化Web情報の構造化と，それに基づくコンテンツ閲覧支援環境について述べた．ユーザの好みや意見を文章として陽に記述するよりも，既存のコンテンツの断片を組み合わせることは簡単である．断片化Web情報の組み合わせからユーザの意見や好みを抽出することが可能になれば，意見より広いユーザを対象とした意見抽出なども可能になる．今後は，断片化Web情報の構造からユーザの意見や好みを解析するための手法を目指す．本システムはコンテンツの再利用に関する情報を得るためのプラットフォームとしても利用可能であり，断片化Web情報の構造解析手法を導入することで，閲覧だけにとどまらないユーザのコンテンツ利用に関する知見を得ることが期待される．

参考文献

[1] 新谷 虎松, 大囿 忠親, “知的Webのためのマッシュアッププログラミング,” 情報処理, Vol. 50, No. 5, pp. 444-453, 2009.

[2] Taiki Ito, Hiroyuki Sano, Tadachika Ozono, Toramatsu Shintani, ”A Hierarchical Web Page Segmentation Algorithm using Machine Learning”, In the Proc. of The Eleventh International Conference on Intelligent Systems and Control 2008, 2008.

[3] 佐野 博之, 浅見 昌平, 大囿 忠親, 新谷 虎松, “Webエージェントを用いたWebコンテンツへの付箋アノテーションの実現”, コンピュータソフトウェア, Vol. 26, No. 3, pp. 69-77, 2009.