氏　　名	サノ　ヒロユキ **佐野　博之**
学位の種類	博士（工学）
学位記番号	博第864号
学位授与の日付	平成25年3月23日
学位授与の条件	学位規則第４条第１項該当　課程博士
学位論文題目	Webインテリジェンスに基づくユーザの知的活動支援に関する研究 (A Study on Structuring Web Contents for Web Intelligence)
論文審査委員	主査　教授　新谷　虎松 教授　内匠　逸 准教授　大囿　忠親

論文内容の要旨

本研究ではWebインテリジェンス技術の応用によって，ユーザの知的活動を支援することを目的としている．ここでのWebインテリジェンス技術とは主に，Webページを Webコンテンツ単位へと分割するためのWebページ分割手法，および，Webコンテンツへのアノテーションなどの，Web情報の構造化技術を差す．構造化されたWeb情報を利用し，Webコンテンツの再利用という観点からユーザの支援を行う．

Web情報の構造化に関連して，Web上にはハイパーリンクと呼ばれるネットワーク構造が存在する．Webにおける既存のハイパーリンク構造はWebページ制作者の観点に基づく構造である．既存のWebコンテンツ間に対してWeb閲覧者が自由にリンクを張ることが可能な機構を実現することによって，Web上に存在する既存のハイパーリンク構造とは異なる，Web閲覧者同士の新たなハイパーリンク構造の実現が期待できる．Web情報を閲覧者の観点から構造化するために，閲覧者がWeb上の情報に対して自由に自身の観点を記述できるようなシステムを試作した．具体的には，Webページ中に存在するWebコンテンツに対して閲覧者が付箋によるアノテーションを行い，Webコンテンツ間に対してハイパーリンクを作成することが可能なシステムとなっている．本システムの実現において，任意のWebコンテンツに対してアノテーションを行うための技術や，アノテーション間の双

方向リンクモデルを提案した．本システムについては 3 章で詳しく述べる．

付箋アノテーションの対象となった Web コンテンツを特定するために，Web ページを Web コンテンツ単位へと分割するための手法に関する研究を行った．Web ページは一般的に複数のコンテンツから構成される．Web ページを記述するための HTML は半構造化文書である．HTML 文書中には各コンテンツの明確な区切りは記述されていない．高精度な Web ページ分割により検索エンジンの精度向上など多くの利点が指摘されており，研究の余地がある．Web ページ分割に関する研究は既に多数存在するが，面積や子ノード数など，コンテンツ量に依存する情報を用いる．その結果，同一 Web サイト内の同じレイアウトの Web ページから異なる分割結果が得られるという問題が存在した．本研究では Web コンテンツの見出し部分を Web コンテンツ間のセパレータとして Web ページ分割を行う手法を提案した．見出し部分の抽出のために，J4.8 アルゴリズムによる決定木学習によって分類器を生成した．評価実験により，分類器の性能は F 値 77.8%，89.3%であることを確認した．得られた見出し部分を用いて Web ページ分割を行った結果，ニュースサイトのニュース記事部分に着目した場合，96.1%の精度でコンテンツ量に依存しない同一の分割結果が得られることを確認した．複数の Web ページ上で提案手法を用いた Web ページ分割を行い，実験対象とした全ての Web ページで 1000 ミリ秒以内に処理が完了することを示した．

3 章，および 4 章で提案した手法を組み合わせることで，Web 情報を閲覧者の観点から構造化することが可能となる．提案手法によって新たに構造化された Web の有用性を実証するために，ユーザの知的活動を支援するためのアプリケーションを試作した．Web ページを Web ブロックへと分割し，Web ブロックをクラウド環境上へと保存することによって Web 情報の再利用性を向上させるためのシステムを試作した．また，Web 情報を元に議論を進めるための議論支援システムや，タブレット端末向けの会議支援システムなどを試作した．提案システムを通じて，本研究で確立した技術が既存の Web コンテンツの再利用性を向上させることを示した．

以下に，本論文の構成を述べる．1 章ではまず，本研究の学術的背景について述べる．研究項目を設定し，それぞれの研究項目の目的について述べる．2 章では本研究を進める上で必要となる基盤技術について述べる．また，既存研究について言及し，本研究の位置づけを行う．3 章では付箋アノテーションシステムの実装について述べる．システムの構成図やシステム実行例のスクリーンショットを用いて全体像についての説明を行った後，DOM ツリーに基づく Web コンテンツの同定手法，および Web コンテンツ間へのハイパーリンク作成について詳しく述べる．4 章では Web ページ分割手法について述べる．分割手法を 3 ステップに分けて詳しく説明を行う．評価実験によって本手法の有効性を示す．5 章では Web 情報を利用してユーザの知的活動を支援するためのアプリケーションについて言及する．最後に 6 章で今後の課題を述べるとともに，本研究をまとめる．

論文審査結果の要旨

Webをベースとした知的な情報基盤を構築するための研究分野はWebインテリジェンスと呼ばれており，その研究は学術的にも社会的にも急務である。本研究はWebインテリジェンスに関する研究であり，Webコンテンツの新たな構造化手法の開発や，構造化されたWebコンテンツを用いてユーザを支援するシステムの開発を目的としている。本論文はそれらの研究成果についてまとめたものである。

Web 上に存在する情報は大多数が半構造化データである。計算機を用いて情報の統合や検索を行うためには，Web情報の機械可読性を向上させるための構造化が不可欠である。一般的に1つのWebページは複数のコンテンツから構成されるが，Webページを記述するためのHTMLでは，コンテンツ間の区切りや各コンテンツの役割などは明確に記述されていない。よって，細かいコンテンツ単位での構造化技術は，注力すべき重要な基盤技術である。本論文ではWeb情報の構造化に関して，WebページをWebコンテンツ単位へと分割するための技術や，Webコンテンツに対するアノテーション技術についての研究といった，Webコンテンツ単位での構造化に着眼している。

本論文は全6章から構成される。

1章では本研究の学術的背景を述べ，研究項目を3つ設定している。また，それぞれの研究項目の目的について言及している。研究項目の詳細は，それぞれ3・4・5章で詳細が述べられている。

2章では本論文を読み進める上で重要となる基盤技術について解説を行っている。次に，関連研究について言及し，本研究の位置付けを行っている。

3章ではWebコンテンツ間の双方向リンクモデルが提案されている。同モデルに基づくアノテーションシステムの実装・評価実験を行うことで有用性を示している。本システムはWebコンテンツの手動構造化システムとして機能する。汎用性の高いアノテーション付与の仕組みによって，任意のWebコンテンツに対してメタデータを与えることや，Webコンテンツ間の関連性の記述を可能としている。また，他の利用者が作成したアノテーションとの間に双方向リンクが自動的に作成されることで，利用者に新たなWebコンテンツの発見を支援する仕組みとなっている。本システムの長期的な運用により，潜在的にはトラックバックとは異なるWebページ閲覧者同士の新しいリンク構造実現に繋がる。

4章ではWebページ分割手法が提案されている。Webページ分割とは，多様なコンテンツを含む半構造化文書であるWebページを，閲覧者にとって意味的にまとまりのある単位へと分割することである。本手法は所望の情報を簡便に把握したいとするWeb閲覧者のニーズを満たす手法として大変興味深い。従来のWebページ分割方法では，同じWebサイト内に存在する同一レイアウトのWebページでも，コンテンツ量によっては異なる分割結果が得られるという問題があった。本論文ではコンテンツ量に非依存な分割を可能とする，新たなWebページ分割手法が提案されている。評価実験によって提案手法の有効性や適用範囲が明確に示されており，有用性の高い技術である。

5章では応用として，3つのアプリケーションが実装されている。1つ目は，閲覧者がWebコンテンツを再利用するためのアプリケーションである。2つ目は，Webコンテンツを基に議論支援を行うためのアプリケーションである。3つ目は，Webコンテンツを会議資料として配布可能な電子会議支援システムである。これらのアプリケーションによって，既存のWeb情報の再利用性が向上する。

6章では本研究についてまとめ，本研究の特色や独創的な点について考察を行うと共に，本研究の貢献について言及している。最後に，今後の課題について議論を行っている。

本研究成果は査読付きのジャーナルとして3編，国際会議に2編の論文が採録済みである。これらの成果は，情報工学分野の中でもとりわけWebインテリジェンス分野の発展に寄与するところが大きい。以上より，本論文は博士（工学）の学位論文に値するものと認める。