【ショート論文】

複数 Web ページの注目領域を対象とした情報探索と集約手法の提案

田崎　雄一郎[†1]　福原　知宏[†2]　佐藤　哲司[†1]

新聞からスクラップブックを作成するというように，Web 探索においても複数の情報源から情報を集約することは多い．それぞれの Web ページ中には利用者にとって必要な情報と不必要な情報が混在しており，利用者は要不要の判断を繰り返しながら情報を収集する．収集を行いながら情報を把握することは難しく，探索中に収集した一連の情報の把握を支援するための情報集約手法が必要である．本論文では Web ページ中の部分毎の情報（部分領域）を対象とした情報探索・集約手法を提案する．提案法は，ページ中の部分領域を単位とした探索を行いながら，利用者自身が情報を集約する．提案手法を実装したシステム評価実験を行い，部分領域を対象として情報の探索と集約を行う本手法により，利用者が十分有効に情報を収集できることが確認できた．今後は従来手法との比較を行い，従来手法に比べた提案手法の有効性を確認する．

Method of information search and aggregation targeting the attention area of multiple Web pages

Yuichiro Tasaki,[†1] Tomohiro Fukuhara[†2] and Tetsuji Satoh[†1]

When we search on the Web, to organize and to understand information which is collected from multiple Web-page is difficult because key information is distributed over Web pages and within a page. So, aggregation method of key information is needed to find key information gathered during the search. In this paper, we suggest information aggregation method that extracts key information from Web pages. Proposed method aggregates partial areas of Web pages as key information. We had an experiment for evaluating the effectiveness of proposed method. We confirmed that users were able to aggregate key information well.

1. はじめに

情報を収集する際，複数の情報源から集約することは多い．例えば日々配送される新聞から，興味を持った記事を切り抜いてスクラップブックを作成するというように，様々な情報・話題が含まれる媒体中から，特定の話題に対して利用者自身が情報の選択と集約を行う場面は多くある．このような場面では，目的の全体像となる一連の情報を自身なりに把握するために，閲覧者自身が情報の集約とその保存を行っていると考えられる．

Web 探索においてはどうだろうか．Web 探索において，情報の収集は多くの場合検索エンジンを利用して，Web ページを単位として行われている．Web ページも新聞や雑誌と同様に，利用者が必要とする情報，利用者にとって不必要な情報が混在している．利用者はこのような情報の混在した Web ページを，求める情報があるかどうかの判断も含めて繰り返し閲覧を行う必要がある．ページの閲覧を繰り返す中で，最終的に辿り着く情報のみが目的とする情報である場合もあるが，一連の探索過程中で得られた複数の情報が，最終的な目的とする情報として必要な場合もある．探索を行いながら一連の情報を集約・把握するのは難しく，探索中に収集した一連の情報の把握を支援するための情報集約手法が必要である．

本論文では，Web ページ中の部分情報を利用して情報を探索・収集することで，利用者が必要だと判断する情報への容易なアクセスと，探索中に必要と判断した情報を集約する手法を提案する．提案法は，Web ページを単位として利用者に提示される検索結果を，部分毎の情報（部分領域）に分解し提示することで，必要な情報への容易なアクセスを支援する．また部分領域をインタラクティブに保持し集約する機能を実装することで，利用者の必要となる情報を集約することを可能する．集約した領域の情報を基に部分領域の生成と提示にフィードバックを行うことで，利用者が必要とする情報の選択を支援する．

本論文の構成は以下の通りである．2 章で関連研究を紹介し研究の位置づけを明確にする．3 章で提案する注目領域を対象とした探索・集約手法を詳述し，4 章で試作したシステムについて述べる．5 章は試作システムを用いて利用者による評価した結果を述べ考察し，6 章で今後の課題を示す．最後に 7 章でまとめを示す．

†1 筑波大学大学院 図書館情報メディア研究科
Graduate School of Library Information and Media Studies, University of Tsukuba
†2 産業技術総合研究所 サービス工学研究センター
Center for Service Research, National Institute of Advanced Industrial Science and Technology

2. 関 連 研 究

2.1 Web ページの部分情報に注目した研究

Web ページ中の部分情報に注目した研究は盛んに行われており，特に情報検索の精度向上や情報要約を目的とする，ページ中の部分情報であるパッセージや重要文などを抽出する手法が知られている．典型的な重要文抽出の手法として，出現頻度や共起，タイトル語や見出し語に注目した単語の重要度や，文や単語間のつながり，文の位置情報などを利用する手法がある [1]．重要な文だけでなく，文書中の連続した一部分をパッセージとして抽出する手法に Salton ら [2]，望月ら [3] がある．Salton らは段落を単位としたパッセージの抽出を，望月らは文書中の語彙的連鎖を利用して，検索クエリに応じたパッセージの抽出を試みている．

ページ中の主要な部分をコンテンツと定義し，コンテンツを推定・抽出する手法に吉田ら [4] がある．そこでは，特定の Web ページにしか出現しない情報に注目し，コンテンツを自動的に推定し抽出する手法が提案されている．

これらの研究は，情報検索や要約の精度向上を目的としていることから，抽出される情報はプレーンテキストであり，抽出した情報のサイズが非常に大きいなど，そのまま利用者への提示を考慮した情報の抽出になっておらず，そのまま適用することは難しい．

2.2 複数 Web ページ中の情報を集約する研究

複数 Web ページ中からの情報を集約し，利用者に対して提示する研究も盛んに行われている．祖父江ら [5] は，ニュース記事を対象に，記事の専門性・難易度・新規性を推定し，利用者が注目する重要度に応じて新聞紙面レイアウトで記事の提示を行っている．Parapar ら [6] も複数ニュース記事の一覧提示を行っている．利用者が入力した検索語からニュース記事群を取得し，コンテンツ部分の抽出と関連語を収集する．収集した関連語を用いて要約した記事の一覧を利用者に提示するシステムを提案している．

マッシュアップと呼ばれる複数 Web ページからの情報集約手法について，Han ら [7] が研究を行っている．Han らは Web Contents Description Language(WCDL) と呼ばれる形式で記述されている Web サービスを利用し，複数 Web ページ中のコンテンツを利用者に抽出・提示している．

これらの研究は，複数ページ中の情報を横断的に収集し利用者に提示するという，情報推薦を主な目的としている．一方，本研究は部分情報を用いた一覧的な閲覧と，利用者自身による情報の集約から，探索した目的の全体像となる一連の情報の把握を支援することを目的としている．従来，検索精度の向上や集約した情報の推薦に用いられていた Web ページ中の部分情報を，複数の Web ページから自動的に抽出して利用者に提示し，利用者自身による探索と集約からの情報把握に用いた点が本研究の新規性である．

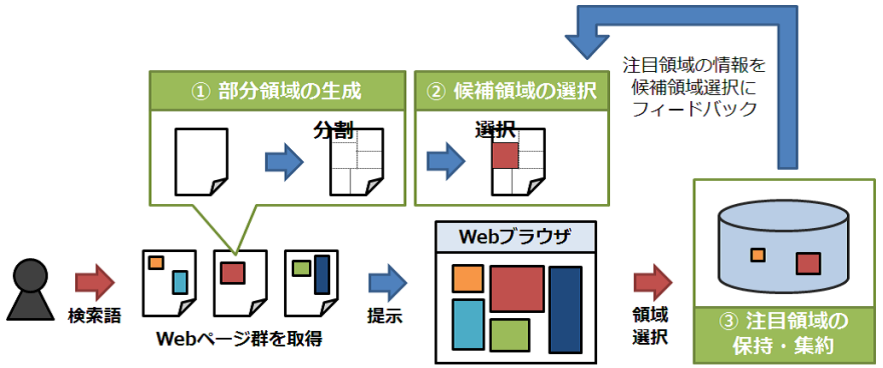

図 1 提案手法の全体像

3. 部分領域を対象とした情報の探索と集約

本論文では，Web ページの部分情報を対象とした情報の探索・集約手法を提案する．

Web ページ中には部分毎に様々な情報が混在しており，利用者に必要な情報はページ全体ではなく，特定の一部分にのみ記述される場合も多い．本手法は，Web ページ中の部分毎の情報を部分領域という単位に分割し，利用者に提示する．利用者は部分領域を単位として情報を探索することで，情報の要・不要の判断，情報の選択をしやすくなる．またページを単位とする閲覧では，それぞれのページを独立して閲覧した後に別ページを閲覧する必要があるが，本手法では複数 Web ページ中の情報を領域単位で一覧的に配置する．

ページ中の部分情報を表す用語を，以下のように定義する．

- **部分領域**： Web ページ中の情報を部分毎に分割した領域群
- **候補領域**： 部分領域の中から，実際に利用者に候補として提示を行う領域群
- **注目領域**： 候補領域の中から，利用者自身が注目・選択した領域群

本手法の全体の流れを図 1 に示す．Web ページ分割による (1) 部分領域の生成，(2) 候補領域の選択と提示，提示された候補中から利用者自身による (3) 注目領域の保持・集約，という一連の流れからなる．以下，(1)〜(3) の各段階について処理の内容を詳細に述べる．

3.1 Web ページ分割による部分領域の生成

提案手法では Web ページの部分領域を用いた情報の探索と集約を行う．そのため，まず探索対象となる Web ページを分割することで，部分領域を生成する．まとまりを持つ単位で部分領域を生成するため，HTML で記述された Web ページを対象に，World Wide Web Consortium（W3C）で規定されたブロック要素[*1]の HTML タグを主に用いてページを分割する．ブロック要素のタグには，例えば ⟨div⟩ や ⟨p⟩ などが挙げられ，構造のまとまりを持つ単位での分割が可能になる．

本研究では利用者へ部分領域を提示する際に，特に領域の大きさが重要となってくると考え，ブロック要素タグから生成した各領域のテキスト長に注目した．テキスト長の閾値を設け，HTML 階層構造の近接する領域の結合と，閾値を超える領域を細分化することで領域の大きさを決定する．領域のサイズ決定に用いる適切なテキスト長の閾値を利用者実験により調査した．実験の結果から，300〜500 字程度のテキスト長を用いることで，提示に適切な部分領域を生成することが可能であるとの結果が得られた [8]．ブロック要素のタグとテキスト長を用いることで，提示に適切なサイズの領域が生成できることを確認している．

3.2 候補領域の決定

分割した Web ページから得られた部分領域群を，利用者に対して注目領域の候補として提示する．情報集約のためのシステムとしては，候補領域の段階で検索語やスニペットを含む領域という絞り込みをせずに，網羅的に提示するのが適当であると考えた．そこで，ページ分割により生成された部分領域を，網羅的に候補領域として提示する手法とした．部分領域生成の際にテキスト長の閾値に満たなかった領域や極端に縦横比が大きい領域など明らかに注目領域となりえない領域を除き，生成された領域のほぼ全てを候補として提示する．

本研究では探索の対象として，検索クエリに対応した Web ページ群を検索エンジンから取得する．検索エンジンから得られる結果には，Web ページ中で検索クエリが含まれている部分をプレーンテキストで抜き出した，スニペットと呼ばれる短文のテキストが付随しており，筆者らはこのスニペットを利用して候補領域を選定する方法 [9] も検討した．このようにスニペットの文字列が含まれる部分領域，すなわち Web ページ中からスニペットを抜き出した部分領域のみを提示した場合には，当然ながら検索クエリが含まれない領域が利用者に対して提示されない．利用者が目的とする情報に検索クエリが含まれないという場合も

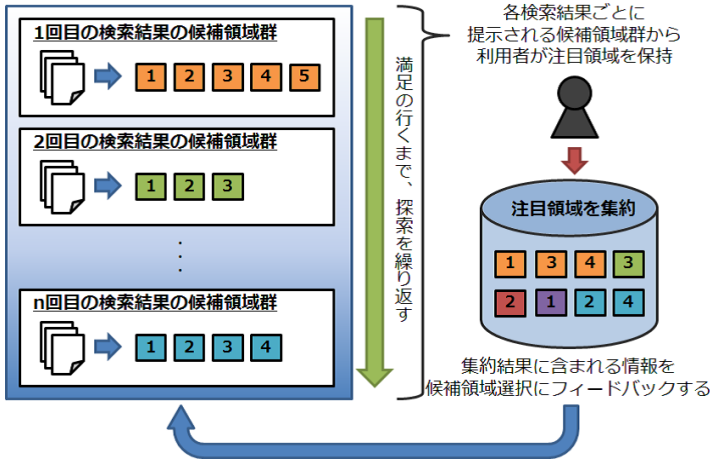

図 2 探索過程における情報の集約

経験上多く[*2]，本手法では利用者自身による注目領域の決定を支援することを目的としているため，検索語やスニペットを用いて候補領域を絞り込むことで，本来は注目領域となる部分領域を候補領域から除外してしまうことがないようにした．

3.3 注目領域の決定・保持

本手法が提示する候補領域群から，さらに利用者自身による注目領域の決定を支援する手法が必要となる．注目領域の決定はシステムと利用者とのインタラクションにより，利用者自身が探索の中で容易に行えることを目指す．図 2 に示すように，利用者が必要であると判断した注目領域を次々と保持することで，特定の話題について探索した一連の情報が集約されていく．集約された情報は探索の途中，終了後のいつでもアクセスすることができ，利用者が興味を持った一連の情報の関係性や全体像などの確認が可能である．

集約したそれぞれの領域内には利用者自身が必要とした情報が含まれており，探索を繰り返すたびに必要な情報が蓄積されていく．今回は未実装であるが，将来的には蓄積された情報から，例えば単語の出現頻度や共起関係を用いてフィードバックを行い，候補領域の選択や注目領域決定を支援する．

⋆1 HTML の厳密な記法では，body タグの直下にはブロック要素のタグのみしか記述できないとされる．ブロック要素ではないタグ（⟨a⟩, ⟨img⟩ など）はインライン要素として扱われる．

⋆2 例えば「カレー 作り方」で検索した場合，求める情報は「じゃがいも」などの材料であったり，「30 分煮込む」などの方法を求めている），スニペット領域のみの提示では適切な情報の探索・収集を行うことが難しい．

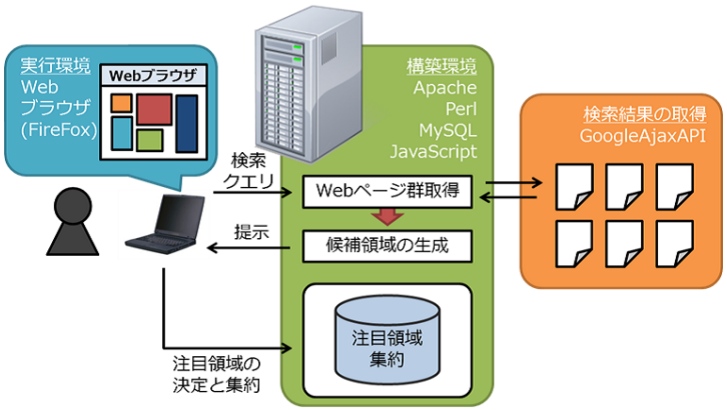

図 3 システム構成図

4. 注目領域を探索・集約するシステムの実装

提案する情報探索・集約手法を，Web ブラウザ上で実行できるシステムを実装した．システムの構成図を図 3 に示す．本システムは利用者が入力した検索クエリに対する結果ページ群を，検索エンジン[*1]を利用して取得する．取得した結果群のそれぞれの Web ページに対して前項で述べた手法で，ページ分割による部分領域の生成と候補領域の選択を行い，利用者に対して一覧的に領域を提示する．なお，今回は集約された注目領域からのフィードバックによる選択支援は行っていない．

検索クエリに「土下座」を入力したときのシステム実行結果を図 4 に示す．図 4 の左側（候補領域）は複数の検索結果ページからシステムが抽出した候補領域群が提示されていることを表す．提示された領域は，テキストとその周辺にある画像のみを対象として構成されており，文中のアンカータグなどは除去している．図 4 の右側（注目領域）は利用者が候補領域を探索し，自身が目的とする情報が含まれている注目領域を選択・保持した領域を表す．

図 4 の赤線で囲まれた範囲に存在する 3 つの候補領域は，同じ Web ページから生成されたものである．テキスト長を基準に領域の生成を行っているため，ページ中のテキスト量が

図 4 システム実行例

多い場合，それぞれのページから複数の候補領域が生成される．逆に，Web ページ中にまとまった量のテキストを持つ領域が存在しない場合（例えば YouTube のようなテキスト以外を主なコンテンツとしたページ）は，候補領域が一つも提示されない場合も存在する．

Web ブラウザ上には複数 Web ページ中の部分領域が一覧的表示されており，横断的に情報を閲覧することが可能である．提示される領域は利用するディスプレイのサイズに基づいて，画面中に候補領域を検索エンジンから得られた結果のランキング順に敷き詰めている．

各候補領域の右上（図 4 の青丸部）にはチェックボックスが設置されており，利用者はこのチェックボックスを用いて，候補領域中から自身の検索意図に応じた注目領域を選択する．例えば今回の検索クエリである「土下座」の場合でも，土下座の歴史について調べる場合や，土下座をした有名人に関するニュースを収集するなど様々な検索意図が存在し，利用者それぞれの意図に応じた注目領域を選択する．選択した注目領域は，システムからのリンクでいつでもアクセスできる別の画面（図 4 の右側，注目領域）に保持される．この保持された領域は探索の途中，検索クエリを変更した後なども常に保持し続けており，探索の途中・終了後のいつでも探索過程の中で注目した一連の情報を確認することができる．集約した注目領域は，興味の遷移などから利用者自身が必要ないと感じた場合，注目領域を表示した画

⋆1 実装では GoogleAjaxAPI を使用したが，API 部分を書きなおすことで Yahoo!検索 Web API など，他の検索エンジンを利用することも容易である．

面のチェックボックスからいつでも保持を解除することができる．

このように本システムを利用することにより，探索を行いながら注目する情報を集約でき，探索した目的の全体像となる一連の情報を把握することができる．

5. 評価実験

5.1 概　　要

実装システムを用いて，被験者による本手法の評価実験を行う．評価実験では本論文で提案する，部分領域を用いた情報探索と，注目領域を用いた情報の保持・集約の評価を目的とする．実装システムを用いて課題を探索することで，部分領域により十分な情報を収集できるか，保持機能を用いて一連の情報の集約と把握が可能になるかを明らかにする．

実験では，被験者に実装システムを用いて検索課題の調査を行ってもらった．検索課題は，新燃岳の噴火に関するニュース記事を利用者に渡し，それに関連して被験者自身が興味を持った話題（例えば，噴火の概要や経済への影響などを想定）の調査を 15 分間実施した．なお，被験者は課題探索前にシステムの利用方法を把握するため，被験者は別の課題を 10 分間で簡単に調査した．

評価実験は 2 名の協力者にシステム利用を依頼し，システムの機能が利用された回数を示すクリックログの取得と，課題探索後のアンケート調査を行った．被験者は探索しながら注目領域の保持を行い，探索終了後には，保持した領域の情報だけを用いて課題のまとめ資料を作成した．集約した注目領域から改めて情報をまとめることで，探索しながら十分な情報を収集できたかどうかを確認する．

5.2 実験の結果

システムの利用結果を表 1 に示す．表 1 から，探索中に入力したクエリ数や候補領域中から元の Web ページを閲覧した回数などが把握できる．また保持した領域の数からシステムの機能が利用された回数が分かり，画像が含まれていた領域が多く保持される傾向にあることが分かった．

アンケート結果から，部分領域を提示する手法について以下のように閲覧性の高さに関して肯定的な意見が得られた．

- 複数の領域が一画面中で閲覧でき全体像が把握しやすい
- ページ全体を閲覧する必要性が減って早く検索ができる

また部分領域の提示に関して，検索エンジンが提示する結果と比較した意見や，部分領域による情報アクセスについて以下のような意見が得られた．

表 1　システム利用結果

評価項目	利用者 A	利用者 B
入力クエリ数	9	9
元ページ閲覧回数	1	5
候補領域を生成した Web ページ数	208	192
保持領域数	18	15
画像を含む保持領域数	11	10

- 領域内に画像が表示される場合があることで判断がしやすかった
- 情報判断のために読まなければならないテキスト量が多い
- 抽出された部分の周辺の情報が欲しい
- 念のため見ておきたいという理由から元のページを閲覧したい

探索をしながら領域を保持する機能については，機能に対する要望も含め，以下のような意見が得られた．

- ワンクリックで簡単に保持ができた
- ある程度のまとまりを持っている領域を単位として保持をすることで，どうしてその領域を保持したのかという振り返りが行いやすい
- 保持領域はメモとして用いたのでもう少し小さい領域にしたい
- 保持した領域の配置を変えたい

5.3 考　　察

システムの利用結果から，利用者が分割前の Web ページを閲覧した回数はそれぞれ 1 回，5 回である．これは分割の対象となった Web ページ数が約 200 件ずつであることを考えると非常に少ない．候補領域を網羅的に提示しただけでなく，それぞれの領域がまとまりを持った状態で生成され，元の Web ページから独立した状態で必要な情報へアクセスすることができたからだと考えられる．また，単に情報の閲覧をするだけではなく，複数の領域が一画面内で閲覧でき全体像が把握しやすいなどの意見が得られたことからも，部分領域を提示することで，閲覧性の高い情報探索が可能となったと考えられる．

情報の保持についてもシステムの利用結果から，それぞれ 18 個，15 個の領域が保持されている．保持された領域の半数以上は画像が含まれた領域であり，テキストではない直観的な情報が利用者の判断に有効に作用したと考えられる．アンケートの結果からも容易な操作で注目領域を保持できることや，領域がある程度まとまりを持った情報で構成されていることから，保持した情報を振り返る際に有効であるとの意見が得られた．

以上のことから，注目領域を探索・集約する本手法が，利用者が目的とする情報の取捨選

択と，探索した目的の全体像となる一連の情報の把握に十分有効だと考える．

一方，情報の集約支援に関して，領域のサイズや配置の変更といった編集機能への意見がいくつか得られた．特に部分領域のサイズについて，提案した Web ページ分割手法により，提示に適したサイズで部分領域を生成できることは確認したものの，提示に適した領域のサイズと，集約に適した領域のサイズが同一であるかどうかは確認できていない．今後利用者による評価からの確認と，その結果から領域の編集を支援する機能の提案が必要であると考えている．

6. 今後の課題

評価実験では実際にシステムを利用した検索課題の調査・収集とそのまとめを行ってもらい，提案手法が Web 上からの情報探索において有効であるとの見通しが得られた．本手法そのものの有効性は見通せたが，従来手法に比べた優位性の確認が必要であり，他システムとの比較実験を計画している．実験により従来の Web ページを単位とした閲覧・探索に比べ，より閲覧性の高い情報の探索が可能となること，部分領域を集約することにより一連の探索で得た情報の把握ができるという有効性が確認できるものと考える．

実施する実験計画の概要について述べる．実験は主に提案手法を実装した本論文で詳述したシステムと，他システムとを比較する．比較するシステムは以下の 3 つを想定している．

(1) 検索エンジン（Google 検索，Yahoo!検索など）
検索クエリのサジェスト機能やプレビュー機能を除去したものを用いる

(2) 検索結果を領域化して提示・保持するシステム
検索結果は検索エンジンのものと変わらないが，各々の結果が領域化されており，必要な領域を利用者が保持できるシステム

(3) 部分領域を提示・保持するシステム
本論文で提案した，部分領域を探索・保持するシステム

3 つのシステムを比較することで，1) 部分領域を単位として情報を探索することの有効性，2) 探索を行いながら領域を収集する手法の有効性の 2 つを確認する．探索課題の分類としては，大きく以下の 3 つのパターンを考えている．その中でも特に複数の答えが存在する課題，利用者の興味によって答えが異なる課題の 2 つに対して本手法・システムが有効に作用すると考えている．

- ひとつの答えが明確に存在する課題： 例）有名人 A さんの誕生日を調べる
- 複数の答えが存在する課題： 例）カレーの隠し味として入れられるもの一覧
- 利用者の興味によって答えが異なる課題： 例）旅行計画，興味のある話題の収集

以上のような課題とシステムを用いて，利用者による比較実験を実施する予定であり，結果から従来手法に比べた本手法の有効性が確認できると期待される．

7. おわりに

本論文では，利用者が必要とする話題に関する情報の探索とその集約を目的として，Web ページ中の部分領域に着目した手法を提案した．提案手法は部分領域を対象として探索を行い，利用者自身が注目領域の取捨選択を行いながら情報を集約する．提案手法を実装したシステムで利用者に探索と集約を行う評価実験を行った．評価実験の結果から，部分領域による目的とする情報の収集と，探索した目的の全体像となる一連の情報の把握に十分有効に作用する見通しが得られた．今後は従来手法との比較による定量的な有効性の検証を行い，より充実した支援を行うためのシステムの改善に取り組む．

謝辞 本研究の一部は科研費（21500091）の助成を受けたものである．ここに記して謝意を示す．

参 考 文 献

1) 奥村学, 難波英嗣: テキスト自動要約, オーム社, 2005.
2) G. Salton, J. Allan, C. Buckley: Approaches to Passage Retrieval in Full Text Information Systems, SIGIR '93, pp.49-58, 1993.
3) 望月源, 岩山真, 奥村学: 語彙的連鎖に基づくパッセージ検索, 自然言語処理, vol.6, No.3, pp.101-126, 1999.
4) 吉田光男, 山本幹雄: 教師情報を必要としないニュースページ群からのコンテンツ自動抽出, DBSJ Journal, Vol.8, No.1, 2009.
5) 祖父江翔, 瀬合将士, 山本孝二, 田村哲嗣, 速水悟: 検索新聞：新聞形式による検索情報要約システムの提案, 第 25 回人工知能学会全国大会 (JSAI2010), 3D1-2, 2010.
6) J. Parapar, A. Barreiro: NowOnWeb:a NewsIR System, Procesamiento del lenguaje natural, N.39, pp.287-288, 2007.
7) H. Han, J. Guo, T. Tokuda: Deep Mashup: A Description-Based Framework for Lightweight Integration of Web Contents, The 19th International Conference on World Wide Web (WWW 2010), pp.1109-1110, 2010.
8) 田崎雄一郎, 佐藤哲司: Web ページの階層的な分割と提示に関する一検討, 第 2 回データ工学と情報マネジメントに関するフォーラム (DEIM2010), A4-2, 2010.
9) 田崎雄一郎, 島田諭, 福原知宏, 佐藤哲司: Web ページからの注目領域抽出に基づく横断型情報閲覧システム, 第 19 回 Web インテリジェンスとインタラクション研究会, 2011.