Web ページの注目領域を対象とした情報探索・集約に関する研究*

田崎 雄一郎(学籍番号 201021749)
研究指導教員:佐藤 哲司

1. はじめに

新聞や雑誌から必要な情報を抽出し，スクラップブックを作成するというように，利用者自身による情報の集約は広く行われている．

Web からの情報探索ではどうだろうか．Web 探索において，情報収集は多くの場合 Web ページを単位として行われる．Web ページも新聞や雑誌と同様に，利用者自身が必要とする情報，利用者にとって不必要な情報が混在している．また，複数ページの情報が目的とする情報に必要となる場合も多い．探索を行いながら一連の情報を集約・把握するのは困難であり，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．

本研究では 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の集約を支援する手法を提案する．

2. 提案手法

本研究では，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．本手法は，様々な情報が混在した Web ページ中の情報を，部分領域を単位として分割し利用者に提示し，利用者自身による情報の集約を支援する．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ごとの領域に分割した領域群．
- **候補領域**
  各部分領域を利用者に対して提案するために，提示の優先度をスコア付けされた領域群．
- **注目領域**
  候補領域として提示された中から，その探索において必要な情報が含まれる領域．提案手法では，利用者自身が必要な注目領域を選択する．

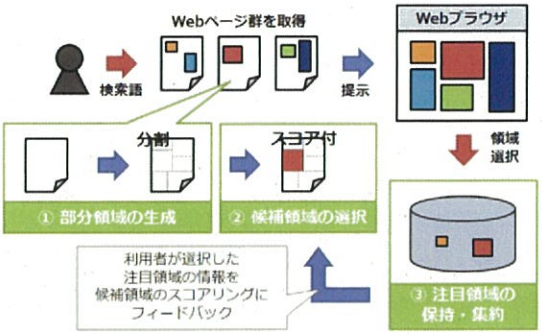

図 1 提案手法の概要

本手法の概要を図 1 に示す．本手法は 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) の各処理の内容を詳細に述べる．

2.1. 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．ページ中からの領域抽出には吉田ら[1]や Salton ら[2]が挙げられるが，利用者への提示を目的とした領域抽出に関する研究は少ない．

本研究では利用者への適切な提示が可能な分割を目指し，HTML 要素とテキストの分量に注目して分割を行った．利用する HTML 要素には や<p>などのブロックレベル要素として規定された要素を，テキストの分量は分割と結合の閾値を定め利用した．閾値は利用者実験から適切に分割できるよう決定した．

2.2. 候補領域の選択

複数の Web ページ中から抽出した部分領域群に対しスコアリングすることで，部分領域を提示する優先度を決定する．付与された優先度順に，部分領域を候補領域として利用者に提示する．

領域のスコアリングには，領域中に含まれるテキストと画像を利用した．テキストによるスコアは各行の文字数に注目し，以下の式から算出する．また領域内に画像が含まれる場合は利用者の情報判断に有効だと考えられるため，スコアを加算した．

---

* "A Study of Information Seek and Aggregation Method Based on Partial Contents from Multiple Web Pages" by Yuichiro TASAKI

図2 注目領域を集約するシステム

$$textScore = \frac{\sum_{i-startNo}^{endNo} \exp(-\frac{(\chi i - \mu)^2}{2\sigma^2})}{endNo - startNo}$$

2.3. 注目領域の保持・集約

複数の Web ページから部分的な情報を集約する研究として，Parapar ら[3]や mash-up 研究などが挙げられる．これらは自動的に情報を提示する情報推薦の研究が多く，利用者自身による集約の支援は少ない．

本手法では提示された候補領域群から，利用者が注目領域の決定を繰り返すことで情報の集約を支援する．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から支援した．

図 2 は実装した提案システムの実行例を示す．図左側には候補領域が一覧的に提示されており，利用者はその中から注目領域を選択する．検索クエリを変更しながら選択・保持を繰り返すことで情報を集約する．集約した領域の情報を元に，本手法では候補領域のスコアリングにフィードバックを行った．

3. 評価実験

3.1. 実験の概要

既存手法との比較実験を行うことで，提案システムを用いた情報探索・集約の有効性と，既存手法に対する優位性を示す．比較する既存手法には，検索エンジンを利用した探索と，Web ブラウザのブックマークによるページ全体の保持を対象とした．

実験は 8 名の協力者に，提案手法・既存手法それぞれを用いた情報の探索と集約，さらに集約の結果を元に Microsoft Power Point での発表資料作成を依頼した．実験中にはシステムの利用ログを記録し，協力者から実験や提案手法に関するアンケートを収集した．

3.2. 実験の結果と考察

利用者実験により得られた，既存手法・提案手法それぞれにおけるシステムの利用ログを表 1 に示す．利用者アンケートの結果とあわせ，比較実験の結果を検討した．

表1 システムの利用ログ

	既存手法	提案手法
検索クエリ数	10.8	10.9
リンククリック数	17.9	19.3
領域生成数	—	833.5
領域保持 or ブックマーク数	11.5	19.1
資料作成時間(分)	19.9	20.3

提案手法と既存手法に，検索のクエリ数などについて，有意な差は得られなかった．アンケートからはどちらのシステムにも肯定・否定的な意見が得られ，部分領域を探索する提案手法が，Web ページ全体の探索と同程度の情報アクセスが可能であると考えられる．

情報集約については概ね肯定的な結果だった．Web ページのブックマークが平均 11 件されたのに対し，領域は 19 件弱保持されていた．t 検定を利用したところ，両者の間に 10% 有意水準で有意な差があることも分かり，アンケートからも注目領域保持の容易さに関して肯定的な意見が得られた．注目領域を対象とした情報集約が，ブックマークを用いて Web ページ全体を集約する手法よりも有効であると考えられる．

4. おわりに

本研究では，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．提案手法をシステム実装し，Web 検索エンジンとブックマーク機能を利用した探索・集約との比較実験を行い，提案手法の有効性を検証した．その結果，注目領域を対象とした提案手法が，ユーザの情報集約のプロセスを支援できることを明らかにした．

文献

[1] 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.

[2] G. Salton, J. Allan, C. Buckley: Approaches to Passage Retrieval in Full Text Information Systems, SIGIR'93, pp.49-58, 1993.

[3] J. Parapar, A. Barreiro: NowOnWeb:a NewsIR System, Procesamiento del lenguaje natural, N.39, pp.287-288, 2007.

[4] Y.Tasaki, T.Fukuhara, T.Satoh: Aggnel: An information aggregation system of partial contents from multiple Web pages, The Fifth International Symposium on Mining and Web (MAW-12), S2-2, 2012.