参照ページからの情報を利用したWeb探索支援

板橋 英夫† 望月 源† 白井 清昭† 奥村 学‡
† 北陸先端科学技術大学院大学 情報科学研究科
‡ 東京工業大学 精密工学研究所

1 はじめに

本研究は，ユーザが実際にWebページを閲覧する前に，そのWebページに関する情報を提示することにより，Web探索を支援することを目的とする．Webページに関する情報は，そのWebページそのものから取り出すのではなく，そのWebページにリンクをはっているページから取り出す．以下，情報を取り出す対象となるページを対象ページ，対象ページにリンクをはっているページを参照ページと呼ぶ．また，参照ページの中で，対象ページへのリンクを該当アンカー，対象ページの内容を説明している文章を参照箇所と呼ぶ．対象ページ，参照ページ，参照箇所の関係を図1に示す．参照ページには対象ページに関する客観的な意見や感想などが記述され，対象ページそのものから取り出される情報とは性質の異なる情報が得られると考えられる．本研究では，複数の参照ページの中から参照箇所（図1の斜線部）を抽出し，ユーザに提示する．ユーザは，参照箇所を見ることにより，そのWebページが有用であるかを判断することができる．また，実際に参照箇所を取り出し，参照箇所にどのような情報が含まれるのかについての分析を行う．

参照ページから得られる情報を用いてWeb探索支援を行う研究は過去にも行われている．検索エンジンGoogle[1]とCLEVER[2]は，参照ページの数によって検索結果をランク付けしている．鷲崎らは，参照ページの中から対象ページへのアンカーを含む1文を対象ページに対する注釈と見做し，これを取り出して対象ページの解説を作る研究を行っている[3]．しかし，対象ページについて書かれた記述がアンカーを含む一文だけだとは限らない．Amitayは，HTMLタグを手がかりとしてアンカー周囲の文章を切り出し，対象ページに関する情報が書かれているとみなした．さらに，機械学習を用いて重要文選択を行い，対象ページの要約を作成した[4]．また，Webページと同様のハイパーテキスト構造を持つものとして論文の参照関係がある．難波らは他の論文を参照している周囲の文を取り出し，被参照論文の要約を作る研究を行っている[5]．

本研究のアプローチは，参照箇所をHTMLタグを利用して取り出すという点でAmitayの研究に近い．本研究では，さらに複数の参照ページから取り出された情報を分類，整理することにより，対象ページに関する情報をユーザにわかりやすく提示することを目指す．

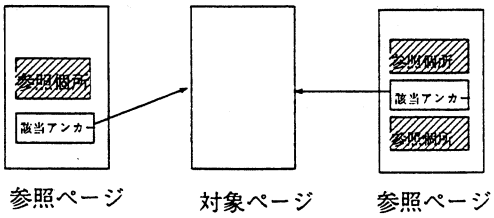

図1: Webページ間での参照関係

2 参照箇所の抽出

参照箇所を抽出する手法を検討するために，インターネットから参照・被参照関にあるWebページを以下のように収集した．まず，検索エンジンにクエリを入れ，検索結果の上位200件（満たない場合は最大数）を対象ページとした．次に，それぞれの対象ページにリンクをはっているページを参照ページとして収集した．ただし，参照ページが10件以下の対象ページは用いていない．今回は，クエリとして「チャット」と「窓の杜」を選んだ．収集した対象ページ，参照ページの数を表1に示す．これらの参照ページを分析し，参照箇所を抽出する手法を考えた．以下，その手法を説明する．

2.1 参照箇所抽出アルゴリズム

まず，参照箇所の抽出を試みる前に，該当アンカーがナビゲーション目的での参照であるかどうかを判定する．ナビゲーション目的での参照とは，例えば，あるサイトのページにおいて，「戻る」と書いてそのサイトのホームページへリンクをはる場合などである．この場合，その該当アンカーの周辺には対象ページに関する情報が存在しないことが多い．そこで，該当アンカーがナビゲーション目的での参照であるかどうか

表1: 実験に用いたWebページの数

	チャット	窓の杜
対象ページ数	14	7
参照ページ数	386	296
平均参照ページ数	27.6	42.3

を判定し，その場合には参照箇所は存在しないとみなす．これは，「戻る」「トップへ」「Back」「Home へ」など，ナビゲーション目的で良く使われる文字列のリストをあらかじめ用意し，それらとアンカー文字列が一致するかどうかによって判定する．

次に，参照箇所の抽出を行う．参照箇所は，参照ページが以下に示す4つのパターンに当てはまるかどうかを順番に調べ，当てはまった時点で参照箇所を抽出する．

(1) リストタグを手がかりとする場合

参照ページの中には，リストタグを用いて，他のページへのリンクとそのページに関する説明を列挙しているものが多く見られた．その例を以下に挙げる[1]．

> ⟨li⟩ Vector — フリーソフト、シェアウェアを中心にした国内最大級のダウンロードサイト。検索機能が優れている。
> ⟨li⟩ 窓の杜 — オンラインで公開されている国内外の優秀な Windows 用ソフトウェアがダウンロードできる。
> ⟨li⟩ Download ASCII — フリーソフトやシェアウェア、メーカー各社の各種デバイスドライバーの無償ダウンロード。

このように，該当アンカーの直前に ⟨li⟩ タグがあれば，その ⟨li⟩ タグから次の ⟨li⟩ タグまでの部分を参照箇所として取り出す．上の例では，下線部が参照箇所として取り出される．

同様に，⟨dl⟩ タグを使って他ページへのリンクを列挙しているページもある．

> ⟨dt⟩ ・DOS/V,WINDOWS フリーウェア、シェアウェアホームページ
> ⟨dd⟩ 窓の杜 Windows Forest

このように，該当アンカーの直前に ⟨dd⟩ タグがあれば，それに対応する ⟨dt⟩ タグから ⟨dd⟩ タグまでの間を参照箇所として取り出す．

(2) br タグを手がかりとする場合

リストタグの代わりに ⟨br⟩ タグを用いて，他のページへのリンクとそのページに関する説明を列挙しているページもある．その例を以下に挙げる．

> 岩波書店 －全出版物案内 ⟨br⟩
> インプレス－書籍・雑誌・ソフト ⟨br⟩
> ＳＳコミュニケーションズ

また，⟨br⟩ を使わずに他のページへのリンクを列挙する場合もある．

> ... imagine iBoard レンタル掲示板 INGNET ...

[1]本論文で挙げる Web ページの例では，⋯は他のページへのリンクを，⋯は対象ページへのリンク（該当アンカー）を，下線部は抽出された参照箇所を表わす．

このように，該当アンカーの前後に他のページへのリンクが3つ以上続く場合，該当アンカー文字列とその直後にある文字列を参照箇所として抽出する．具体的には，参照ページの中から以下のパターンにマッチする部分を見つける．

アンカー$_1$ 文字列$_1$ ⟨br⟩$_1$
アンカー$_2$ 文字列$_2$ ⟨br⟩$_2$
アンカー$_3$ 文字列$_3$ ⟨br⟩$_3$

ここで，文字列$_i$ と ⟨br⟩$_i$ は空でもよいとする．マッチすれば，該当アンカーとその直後にある ⟨br⟩ タグもしくは他ページへのアンカーまでの部分を参照箇所として抽出する．例えば，該当アンカーがアンカー$_2$ のときには，「アンカー$_2$ + 文字列$_2$」を参照箇所として抽出する．

(3) テーブルタグを手がかりとする場合

テーブルタグを使って他のページへのリンクとそのページに関する説明を列挙するページも多く見られた．その例を以下に挙げる．

マグネット	サンリオのキャラクターでチャットができます。会員制（無料）
ゆめみ亭	チャットポータルサイト。初心者でも気軽に参加できます
chat.co.jp	チャットポータルサイト。カテゴリ別にチャットルーム有り

このように，テーブルタグを使ってアンカーを同じ列に並べ，かつ該当アンカーの右のセルにアンカー以外の文字列が存在するときには，その文字列を参照箇所として抽出する．なお，この例ではわかりやすさのために枠線を表示しているが，実際の Web ページでは，テーブルタグをレイアウトのために使用し，枠線を表示していない場合が多い．

また，以下のように，アンカーとリンク先ページの説明が交互に記述されている場合もある．

掲示板
自分の掲示板をつくりたい人はこちら
チャット
夜まで語りあかそう！
無料ホームページ
ホームページスペースを無料で 50MB 提供

そこで，テーブルの同じ列にアンカーとアンカー以外の文字列が交互に並んでいた場合，該当アンカーの下のセルにある文字列を参照箇所として取り出す．

(4) その他

上記のいずれのパターンにも当てはまらない場合には，該当アンカーの近傍を参照箇所として抽出する．参照箇所の境界は HTML タグによって決める．具体的には，該当アンカーの前に存在する HTML タグを

探し，参照箇所の先頭とする．同様に，該当アンカーの後に存在するHTMLタグを探し，参照箇所の末尾とする．ただし，以下のHTMLタグは無視し，参照箇所の境界としない．

- 文字修飾タグ (⟨font⟩,⟨b⟩,⟨i⟩ など)
- ⟨image⟩ タグ
- ⟨a⟩ タグ
- コメント
- ⟨br⟩ タグ (ただし，無視するのは1回のみ．2回目に現われたときは参照箇所の境界とする)

例を以下に挙げる．

> 意味⟨br⟩
> Windows ユーザー定番のフリーウェア、シェアウェア集。ていねいな解説があるので初心者でも安心です。...

この例では，テーブルの1つのセルの中に該当アンカーが存在する．したがって，該当アンカーの前後にあるテーブルタグを検出した時点で参照箇所の境界を決めている．

3 参照箇所の分類

抽出された参照箇所の内容を分析し，どのような情報が参照箇所に含まれるかについて調査した．その結果，参照箇所は大きく分けて以下の3つのタイプに分類できることがわかった．

(1) 説明タイプ

対象ページの内容を説明しているタイプである．その例を以下に挙げる．

> 株式会社インプレスが厳選したWindows用オンラインソフトを紹介するサイト。

この場合，参照箇所として取り出される情報は，対象ページそのものから取り出される情報 (対象ページのタイトルや要約など) と似ている．したがって，説明タイプの参照箇所を提示することは，対象ページから得られる情報を提示することと比べて，あまり差がないといえる．しかし，説明タイプの参照箇所から，対象ページからは得られないような情報が得られる場合もある．以下に例を挙げる．

> Windows Forest. A webzine about Windows online software at Impress Corporation,Tokyo,Japan.

この場合，対象ページは日本語で書かれているが，参照箇所からは英語による対象ページの説明が得られている．このように，対象ページの言語以外での説明が得られることは，説明タイプの参照箇所の大きな特徴である．

(2) 意見タイプ (ページ型)

対象ページに対する意見を述べているタイプである．その例を以下に挙げる．

> チャットでお世話になってるｃｕｅさんのＨＰです☆トップの写真がとってもきれいです☆☆ＢＢＳのアイコンがかわいい～～♪

この他にも，対象ページの雰囲気やレイアウトなど，対象ページに関する様々な意見が得られる．このような他者の客観的な意見は，対象ページそのものからは得られない情報であり，参照ページから情報を収集することの利点である．

(3) 意見タイプ (コンテンツ型)

対象ページそのものではなく，対象ページが紹介しているコンテンツに関する意見を述べているタイプである．例えば，対象ページが商品を紹介するページであり，参照ページでその商品に対する意見を述べている場合がある．以下は，携帯端末P503isのカタログサイトを参照するページ中の記述である．

> 使い勝手はP503iと概ね同じ。ただ、パナソニック端末はいまだにインライン変換でなく、しかも単文節変換だ。連文節変換のさらに一歩先を行く予測入力変換「POBox」搭載のSO503iに比べると、2歩下がっている感じ。ここはちょっと残念。

このような対象ページのコンテンツに関する意見も，対象ページそのものからは得られない情報であり，ユーザに対象ページの有用性を判断させる重要な材料となる．

複数の参照ページから抽出される参照箇所を無秩序に並べて提示しても，ユーザにとってわかりやすいとは言えない．

そこで参照箇所を上記3つのタイプに分類し，整理して提示すればユーザも理解しやすくなる．しかし，自動的に分類する手法は現在検討中である．今のところ，以下のような特徴を手がかりにすることを考えている．

- 参照箇所が短い場合は説明タイプであることが多い．特に，抽出された参照箇所が該当アンカーのアンカー文字列と一致するような場合は，説明タイプであることが多かった．
- 意見タイプの参照箇所は，説明タイプの参照箇所と比べて，「かわいい」「きれい」などの形容詞が使われている場合が多い．

4 評価実験

本節では，2.1節で提案した参照箇所抽出アルゴリズムの評価実験について述べる．ここでは，クローズドテストとオープンテストの2種類の実験を行う．クローズドテストは，表1のWebページの集合について，参照ページから参照箇所を抽出した．一方，オープンテストでは，Webページ作成のための素材を提供

表 2: 実験結果 (クローズドテスト，チャット)

	完全一致	部分一致	適用率
再現率	0.578	0.916	—
精度	0.513	0.863	—
- リストタグ	0.235	1.000	(0.101)
- br タグ	0.784	0.914	(0.346)
- テーブルタグ	0.769	0.872	(0.116)
- その他	0.295	0.787	(0.436)

表 3: 実験結果 (クローズドテスト，窓の杜)

	完全一致	部分一致	適用率
再現率	0.467	0.778	—
精度	0.447	0.772	—
- リストタグ	0.510	0.837	(0.199)
- br タグ	0.520	0.800	(0.203)
- テーブルタグ	0.333	0.556	(0.037)
- その他	0.406	0.754	(0.561)

表 4: 実験結果 (オープンテスト)

	完全一致	部分一致	適用率
再現率	0.315	0.620	—
精度	0.345	0.690	—
- リストタグ	0.250	1.000	(0.095)
- br タグ	0.467	0.600	(0.179)
- テーブルタグ	0.750	0.938	(0.190)
- その他	0.178	0.578	(0.536)

するページ「まゆ工房」と，他者による商品の評価を掲載するページ「リブラ」の 2 つを対象ページとし，それぞれの参照ページから参照箇所を抽出した．参照ページの数はそれぞれ，86, 40 である．これらの参照ページは，2 節の参照箇所抽出アルゴリズムの検討には用いていない．オープンテスト，クローズドテストともに，人手で抽出した参照箇所を正解として評価を行った．

クローズドテスト (チャット)，クローズドテスト (窓の杜)，オープンテストの実験結果をそれぞれ表 2, 表 3，表 4 に示す．これらの表において，「完全一致」は，自動抽出した参照箇所が人手で抽出した参照箇所と完全に一致したときに正解とみなした場合の精度 (適合率) と再現率を表わす．一方，「部分一致」は，自動抽出した参照箇所が，人手で抽出した参照箇所を完全に包含しているときに正解とみなしたときの評価である．これは，人手で抽出した参照箇所と完全に一致しなくても，それを含む記述が抽出できれば，ユーザにとって有益な情報となりうると考えたためである．また，2.1 項で述べた個々の抽出パターンが参照箇所の抽出にどれだけ有効かを評価するために，それぞれのパターンを適用した割合 (適用率) と，そのときの精度も示した．

「リストタグ」と「テーブルタグ」のパターンは，適用率は低いが，比較的良い精度が得られていることがわかる．一方，「br タグ」は，クローズドテストに比べてオープンテストでの精度が低く，アルゴリズムの検討に用いた参照ページに特化したパターンであるといえる．「その他」については，オープンテストにおいても適用率が高く，「部分一致」で評価した精度も 5 割を越えている．しかし，「その他」のパターンで参照箇所を決めたとき，非常に長い記述が抽出される場合も多い．このような場合は，たとえ対象ページに関する情報が含まれていたとしても，ユーザは長い記述を読まなければならない．したがって，提案アルゴリズムで抽出した参照箇所の中から，対象ページに関する情報を選別することが必要となる．我々は，この際，提案アルゴリズムのように HTML タグのみを手がかりとするのではなく，言語的な情報も積極的に用いるべきであると考えている．

5 おわりに

本研究では，Web 探索支援を目的に，参照ページから Web ページに関する情報を取得し，ユーザに提示する方法を検討した．HTML タグを手がかりに参照箇所を抽出すること試み，その手法を予備実験によって評価した．また，対象ページから得られない情報として，他言語による対象ページの説明，対象ページに関する他者の意見，対象ページのコンテンツに関する他者の意見が参照ページから取得できることがわかった．

今後の課題としては，まず参照箇所抽出アルゴリズムの改良が挙げられる．今回は HTML タグのみを手がかりとしたが，それだけでは参照箇所を抽出することはできない．したがって，言語的な情報も用いる必要があろう．また，参照箇所を 3 節で述べた 3 つのタイプに自動的に分類する手法を考え，参照ページから得られた情報をユーザに分かりやすく提示することも課題のひとつである．

謝辞

本研究において，有用なコメントを頂いた日本学術振興会 特別研究員，難波英嗣氏に感謝致します．

参考文献

[1] Sergey Brin, Lawrence Page. "The anatomy of a large-scale hypertextual Web search engine". Computer Networks and ISDN Systems vol.10 pp.107-117.1998

[2] S. Chakrabarti et al.1999 "Mining the Web's Link Structure" IEEE computer Vol. 32, No. 8, pp60-67.1999

[3] 鷲崎 誠司他. "ハイパーリンクの構造を利用した検索結果の選択手法" 情報処理学基礎 55-10

[4] E.Amitay. "InCommonSense - Rethinking Web Search Results"ICME 2000

[5] 難波英嗣, 奥村学. 論文間の参照情報を考慮したサーベイ論文作成支援システムの開発" 自然言語処理, Vol.6, No.5