修 士 論 文

参照ページからの情報を利用したWeb探索支援

北陸先端科学技術大学院大学
情報科学研究科情報処理学専攻

板橋 英夫

2002年3月

修 士 論 文

参照ページからの情報を利用したWeb探索支援

指導教官　白井清昭 助教授

審査委員主査　白井清昭 助教授
審査委員　島津明 教授
審査委員　鳥澤健太郎 助教授

北陸先端科学技術大学院大学
情報科学研究科情報処理学専攻

910008 板橋 英夫

提出年月: 2002 年 2 月

概 要

本論文では，Web 探索支援を目的とし，ユーザがアンカーをクリックしてアンカーの参照先のページ (対象ページと呼ぶ) を表示する前に，対象ページの内容や第三者による多面的な評価をユーザに提示する手法を提案する．まず，対象ページを指すアンカーを持つページ (参照ページと呼ぶ) を収集する．次に，参照ページから，対象ページに関する内容，意見，感想などを記述した文章の部分 (参照箇所と呼ぶ) を抽出し，これらをユーザに提示する．参照箇所を提示することによって，ユーザが対象ページの有用性を知る手がかりになると考えられる．本論文では，参照箇所を抽出，分類する手法とその結果について報告する．

目 次

図 目 次

表 目 次

第1章　はじめに

1.1　研究の目的と背景

1.1.1　Web 探索支援の必要性

近年，インターネットの普及とともに，オンラインで数多くの電子化されたテキストを入手できるようになった．インターネットの普及とともに情報発信者が増え，WWW 上の情報量は増大し，様々な情報を得ることが可能になった．インターネットで情報を探す一つの方法が，WWW(World Wide Web) 上での情報探索である．しかし，このような大量の情報を一元的に管理する機構，機関が存在しないため，WWW は混沌とした状態になっている．巨大で未整備の WWW では，必要な情報の所在がわからなかったり，閲覧した複数の Web ページ間の関連を把握できなかったり，求める情報を効率的に手にいれることは困難である．

WWW での情報探索において，目的の Web ページを見つけるには 2 つの方法がある．

1. 検索エンジンを用いて目的のページを探す方法
2. Web ページのリンクをたどることにより，目的の情報を得る方法（以下，Web ブラウジングと呼ぶ）

検索エンジンを用いた場合，検索結果にはページの説明文がついていることが多い．例えば Google ではアンカー周囲の何バイトかを提示し，Yahoo ではページの最初から数バイトを表示している．しかし，このような説明文がユーザに必ずしも適切な情報を与えてるとは限らない．また，2 の方法においては，リンク先の Web ページに関して得られる情報はさらに少ない．現在閲覧している Web ページにリンク先の Web ページに関する情報が記述されている場合もあるが，そうでない場合も多い．そのため，ユーザは，とりあえずリンクをたどってその Web ページを見て，自分にとって必要でないと判断したら，また元のページに戻る．この操作を繰り返し行うことは，Web 探索の効率を著しく低下させる大きな要因になっている．そこで，リンクをたどる前に，リンク先に関する適切な情報が得られれば，ユーザは実際にリンクをたどる前にリンク先のページが自分にとって必要かどうかを判断できるため，効率よく情報探索ができる．適切な情報とは，例えばリンク先の Web ページの簡潔な説明や，評価文章などである．これらを獲得し，ユーザに提示することができれば，Web 探索支援において有益であると考えられる．

1.1.2 本研究の目的

前項で述べたように，Web 探索において，アンカーをクリックする前にアンカーの指すページ（以下，対象ページと呼ぶ）に関する情報を与えれば，目的にかなった Web ページを効率的に見つけることができると考えられる．対象ページに関する情報は，例えば対象ページのタイトルや要約を提示するなど，対象ページそのものから獲得することが一般的である．これに対し，本研究では, 対象ページを参照しているページ（以下，参照ページと呼ぶ）に着目し，参照ページから対象ページに関する情報を獲得する．対象ページ，参照ページの関係を図 1.1 に図示する．参照ページには，対象ページそのものから得られる情報とは異なる性質の情報が得られると考えられる．例えば，参照ページの著者が対象ページに関する意見や感想を記述する場合がある．このような記述を複数の参照ページから取得できれば，対象ページに対する第三者の多面的な評価情報を提示することができ，ページの有用性を知る手がかりになると考えられる．

本研究では，複数の参照ページから対象ページに関する情報を自動的に抽出し，ユーザに提示することを提案する．また，どのような種類の情報が参照ページから取得できるかについて，実在する Web ページを対象に分析する．

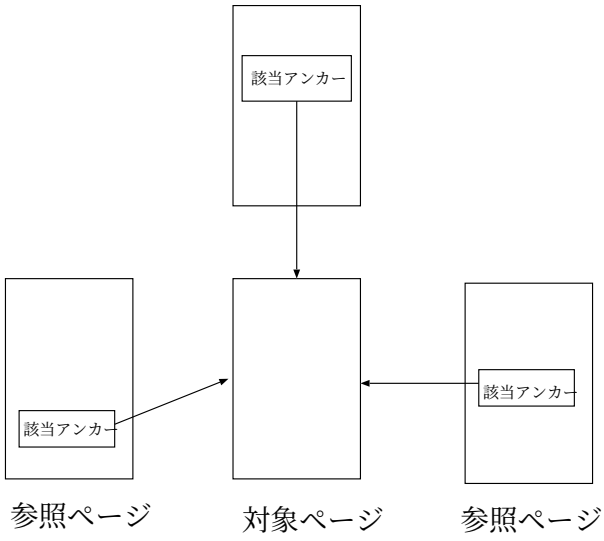

図 1.1: Web ページ間での参照関係

1.2 本論文の構成

本論文では，2 章で Web 探索支援と Web ページからの情報抽出に関する研究を紹介し，本論文との比較を行う．3 章では，対象ページに関する情報を参照ページから抽出する手法を提案する．また，提案手法の評価実験について述べる．4 章では，参照ページからど

のような情報が得られるのかについての分析を行う．また，得られた情報をユーザにわかりやすく提示する手法について考察する．最後に 5 章で本研究の結論と今後の課題について述べる．

第2章　関連研究

2.1　Web探索支援に関する研究

WWW におけるブラウジングは，ページ内に存在するアンカーを選択し，他のページへの移動を繰り返すことによって行われる．その際において問題となるのは次の点である．

- 求める情報が存在しないのに，あると信じて探し続けてしまう
- 求める情報が存在するのに，探し出せない／辿り着けない
- 不必要な情報ばかり収集してしまう

このような問題を解決し，ユーザの WWW のおけるブラウジングを快適にするのが Web 探索支援と呼ばれる技術である．以下，Web 探索支援に関する先行研究を述べる．

1. **協調フィルタリングを用いた研究**
   ユーザへのブラウジング支援の一つとして，協調フィルタリングを用いる方法がある．協調フィルタリングとは，人間の情報収集行動から興味・関心・意図といった問題意識および獲得された情報を収集し，類似の問題意識を持った者に提供することで情報収集活動を支援する手法である [福原 98]. 協調フィルタリングを用いたシステムに，ユーザが見たページに関連するページを推薦してくれるシステムがある [Terveen97]. Terveen らによる PHOAKS[Terveen97] というシステムは，ネットニュースを情報源とし，キーワードと URL を取得する．そしてユーザがキーワードと類似するクエリを入れたとき，その URL を推薦する関連ページとして提示するというシステムである．
2. **共引用を用いた研究**
   Ellen は，ハイパーリンクによる共引用を用いた推薦システム ParaSite[Ellen98] を提案した．共引用とは 2 つの Web ページが同一の Web ページに引用されている状態を指す．図 2.1 は，共引用を図示したものである．この場合において，ページ B とページ C に直接参照関係はない．しかし，ページ B とページ C を共引用するページ A が存在する．Ellen は，ページ A 内において，ページ B とページ C が近い位置で参照されていれば，ページ B とページ C は内容が近いと判断する．よって，ユーザがページ B を閲覧すると，システムは推薦する関連ページとして，ページ C を提示する．
3. **参照ページを用いた研究**
   本研究では，協調フィルタリングや共引用を用いた Web 探索支援ではなく，参照ページを利用した Web 探索支援を目的とする．1 章で述べたように，本研究ではあ

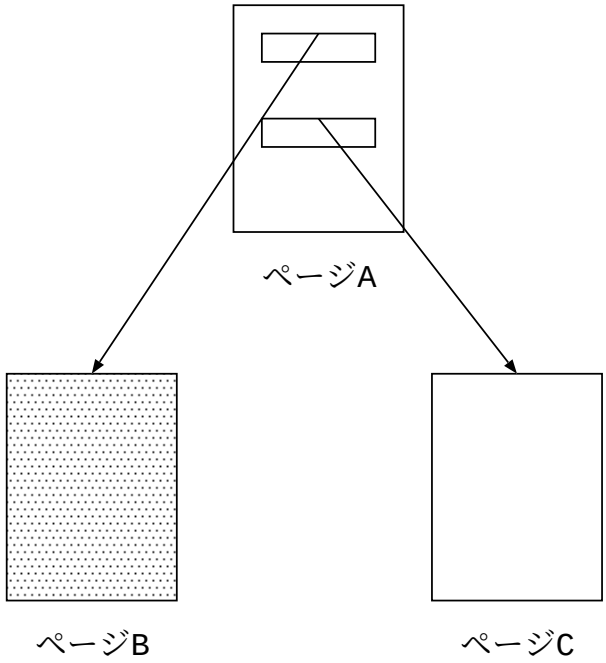

図 2.1: 共引用

るページに対して，その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，参照ページからリンクをはられているページを対象ページと呼ぶ．参照ページを利用して対象ページの評価を行い，Web 探索を支援する研究がいくつか行われている．

検索エンジン，Google[Page98]，CLEVER[Chakrabati99] は，参照リンク数を利用したページのランキングをしている．Google では,「多くの良質のページからリンクされているページはやはり良質のページである」という考え方に基づき，あらかじめ Web ページの被参照数を数える．被参照数が多いページほど重要であると考える．そして，検索クエリが与えられた時，そのキーワードを含むページを探しだし，重要度順にページを並べて検索結果として出力する．つまり，被参照数の多いページから参照されているページほど，評価の高いページであると見なし，上位にランキングする．一方 CLEVER では，ハブとオーソリティーという概念を用いている．ハブとは，リンク集であり，オーソリティーとは良質の情報を含んだサイトのトップページである．よいハブとは，有用サイトをたくさん参照しているページであり，よいオーソリティーとは，より良質の情報を含んだページである．そして，よいハブから参照されているページほどよいオーソリティーであり，よいオーソリティーを参照しているページほどよいハブであるというアルゴリズムを用いて，有用なサイトをランキングする．これらに共通するのは，あるページがあるページを参照するとき，そのページは良質のページとして推薦されていると仮定し，Web ページの評価を行っていることである．しかし，参照ページが被参照ページを常に推薦しているわけではないこ

とは明らかである．また，これらの検索エンジンは参照ページの数だけを利用し，参照ページから直接対象ページに関する情報を取得しているわけではない．

鷲崎らは参照ページに存在する対象ページへのアンカー文字列を，参照ページの筆者の主観的な注釈と見做し，抽出した [鷲崎ら 99]．それらを検索エンジンの検索結果とともに提示し，ユーザが検索結果を選択する為の情報として有効であることを示した．しかし，鷲崎らが述べているように，対象ページに関して記述されている部分は，アンカー文字列以外の部分にも存在する．これに対し，本研究では，対象ページに関する情報としてアンカー文字列だけを取り扱うのではなく，HTML タグを手がかりとして，アンカー文字列の近傍を抽出する．

Amitay は，HTML タグを用いて対象ページについて記述している部分を抽出し，検索エンジンの出力結果の説明文を作成する試みをしている [Amitay00]．Amitay はリンク集などの参照ページに存在する対象ページに関する記述は，対象ページの良質な要約であるという仮定に基づき，これらの記述をタグ情報に基づいて抽出する．抽出した複数の記述を人が読み，最も良いとする記述を選ぶ．その結果を機械学習にかけることによって，複数の参照箇所から最も良い要約結果を含む参照ページを一つ選択し，最終的な要約を出力する．しかし，参照ページにおいて，対象ページに関して記述する文章は，説明だけではない．対象ページを評価し，意見を述べている記述も存在する．本研究では，対象ページを要約しただけでは得られない情報を，参照ページから得ることを試みる．また，Amitay は，一つの参照ページのみから要約を作成しているのに対し，本研究では，複数の参照ページから同時に情報を抽出することにより，対象ページに関する様々な情報をユーザに提示することを目指す．

Web ページと同様に参照 · 被参照関係を持つものとして学術論文がある．難波らは，ある論文を引用している論文から得られる情報に基づいて，論文の要約を作る研究を行っている [難波ら 99]．本研究とは Web ページを対象としていない点が決定的に異なるが，参照論文から被参照論文に関する情報を取得するという考えは共通である．

2.2 Web ページからの情報抽出

本研究は，参照ページから対象ページに関する情報を抽出し，ユーザに提示することを目的とする．これは Web ページからの情報抽出と捕らえることが出来る．情報抽出に関する先行研究をいくつか挙げる．

2.2.1 HTML タグを用いた方法

山田らは，HTML タグや，典型的に現われる表現を手がかりにページをカタログやリンク集といったタイプに分けた後，情報抽出を行っている [山田ら 00]．主にイベント情報，求人情報を対象に情報抽出を行っており，抽出したそれらの情報は，日付，職種等に分類してユーザに提示している．求人情報，イベント情報の抽出精度は 90% に達したと述べ

ている．この研究は，日時，場所などのスロットを埋めるという定型的な情報抽出に関する研究である．しかし，本研究では，特定の情報だけを抽出するのではなく，対象ページに関する情報なら何でも抽出する．

また，吉田らはWebページにおける表からの情報抽出を行っている [吉田ら 00]．さらに，論理的な構造から表を分類し，同様の属性を持つ表同士をクラスタリングを行いまとめている．本研究では，HTMLタグを手がかりとし，対象ページに関する情報を抽出することを考えている．特に，表を表すHTMLタグは，参照ページから情報を取り出す際の有用な手がかりになると考えられる．

2.2.2 言語的な情報を用いた方法

[立石ら 01] では，インターネットに分散して存在する人の意見を一括して検索するシステムを提案している．立石らは，商品等のキーワードをいれ，インターネットからクローラーを用いて関連するページを探し出している．そして，商品名と，印象語を含む一定範囲の文字列を，商品に対する意見として抽出する．次に，印象表現辞書を用いて，得られた評価文字列を分類し，検索語と近接演算することによって，評価文字列の尤もらしさを示すスコアを計算する．スコアの計算は，正規表現でパターンを作り，パターンにマッチする回数を元に計算される．また，商品に対する個々の評価文字列を肯定的·否定的のいずれかに分類する．これは，評価表現にあらかじめ基本属性（肯定，否定）を与え，評価文字列に現われる評価表現の基本属性に従って分類する．さらに，近くに否定語（ない等）があると肯定·否定を反転させる手法を取っている．その結果，スコアを使った手法は，従来のキーワードを用いた手法と比べて，検索結果を15%ほど絞り込めること，検索結果の上位で高い適合率を有することができたと述べている．参照ページ内には，対象ページが紹介している製品などに関する意見の記述が存在する．本研究においても，立石らが着目したような情報を参照ページから抽出し，ユーザに提示することを考える．

2.3 他者の評価を提示する研究

参照ページからは，対象ページに対する第三者の評価が含まれていると予想される．本研究においては，このような他者による評価は，ユーザにとってWebページが重要であるかどうかを判断する有力な材料になると考える．ここでは，他者の評価を提示する手法を取り入れたいくつかの先行研究について述べる．

月出らは，TV番組に対するアンケートにおける5段階評価と自由回答文との相関関係を分析した [月出ら 00]．計算機用日本語基本型形容詞IPALを元に印象語辞書を作成し，TV番組に対する印象を分類している．そして，アンケートで得た5段階評価と照らし合わせ，印象語と5段階評価の相関関係を分析している．その結果，スポーツ番組において，印象語の出現率と5段階評価値に対応がみられたと述べている．

乾らは，文末表現に着目し，自由回答アンケートの分類を行っている [乾ら 98]．分類結

果の再現性と客観性を高めるため，最大エントロピー法を用いている．その結果，人手の分類方法と大きな相違なく自動分類を行えると報告している．

第3章　参照箇所の抽出

あるWebページが他のWebページを参照する場合，参照先のページ（対象ページ）について記述した箇所が存在する．その箇所から得られる情報を収集し整理することで，対象ページに関する様々な情報が得られる．これらはWebページを効率的に探索するための重要な情報になると考えられる．すなわち，参照ページを収集し，対象ページについて書かれている部分を自動的に獲得することができれば，Web探索の際に有効な情報を得ることができる．本節では，参照ページから対象ページに関する情報を自動的に獲得する方法について述べる．

3.1　参照箇所の定義

本論文では，参照ページ内にある対象ページに関する情報が記述された部分を参照箇所と定義する．Webページ間における対象ページ，参照ページ，参照箇所の関係を図3.1に示す．

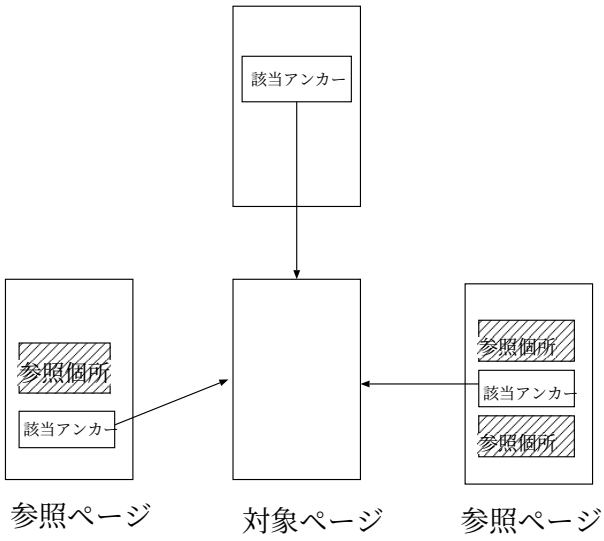

図 3.1: 対象ページ，参照ページ，参照箇所の関係

参照箇所の記述の例を以下に示す．対象ページがチャットサイト「CHAT.CO.JP」のとき，参照ページの１つには以下の様な参照箇所がある．

吹き出しチャットの「CHAT.CO.JP」などはがんばっていると思うがユニークすぎて日常的に使うのは面倒だ.

また，対象ページが商品評価サイト「リブラ」のとき，以下の様な参照箇所がある．

車・バイク，家電製品，パソコン，AV・通信機器のカテゴリーに登録された約一万アイテムから検索し価格や機能等から比較できる．お得な情報満載のコラムも購入の参考になる．

Windows 用のソフトウェアサイト「窓の杜」に対しては以下の様な参照箇所がある．

代表的シェアウェアサイト–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，日本の代表的シェアウェアサイト．10 ケ所以上のミラーサイトがある．ソフトウェアの新着情報等を交換できる ML もあり．

3.2 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 · 被参照関係にある 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. これら 2 つのクエリを選んだ理由は以下の通りである.「チャット」の場合，対象ページはチャットサイトであることが多い．また，その参照ページは個人ページが多く，チャットサイトを直接的に評価する参照箇所が多く存在する傾向が見られた．また，クエリを「窓の杜」とした場合，参照ページにおいて，対象ページを説明する記述が多く現われる傾向が見られた．ここでは，これらのような情報を含む参照箇所を収集できる可能性があると考え，前述の 2 つのクエリを選んだ．収集した対象ページ, 参照ページの数を表 3.1 に示す.

表 3.1: 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

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．本研究で提案する抽出アルゴリズムを図 3.2 に示す．次に，図 3.2 に示した処理の流れについて説明する．

3.2.1 参照箇所抽出アルゴリズム

図 3.2 に示したように，最初に行う処理は，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ことである．ナビゲーション目的での参照とは，主にサイト内リンクを指す．例えば，あるサイトのページにおいて，アンカー文字列に「戻る」「トップへ」などと書いてそのサイトのトップ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を判定し，その場合には参照箇所は存在しないとみなす．本論文で用意した文字列のリストは以下の通りである．

- 「トップへ」で終わる文字列
- 「ホーム（へ)」で終わる文字列
- 「ホームページへ」で終わる文字列
- 「ホームページ」
- 「戻る」で終わる文字列
- 「もどる」で終わる文字列
- 「top（へ)」で終わる文字列
- 「back」
- 「home（へ)」で終わる文字列
- 「home page へ」で終わる文字列
- 「home page」

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．

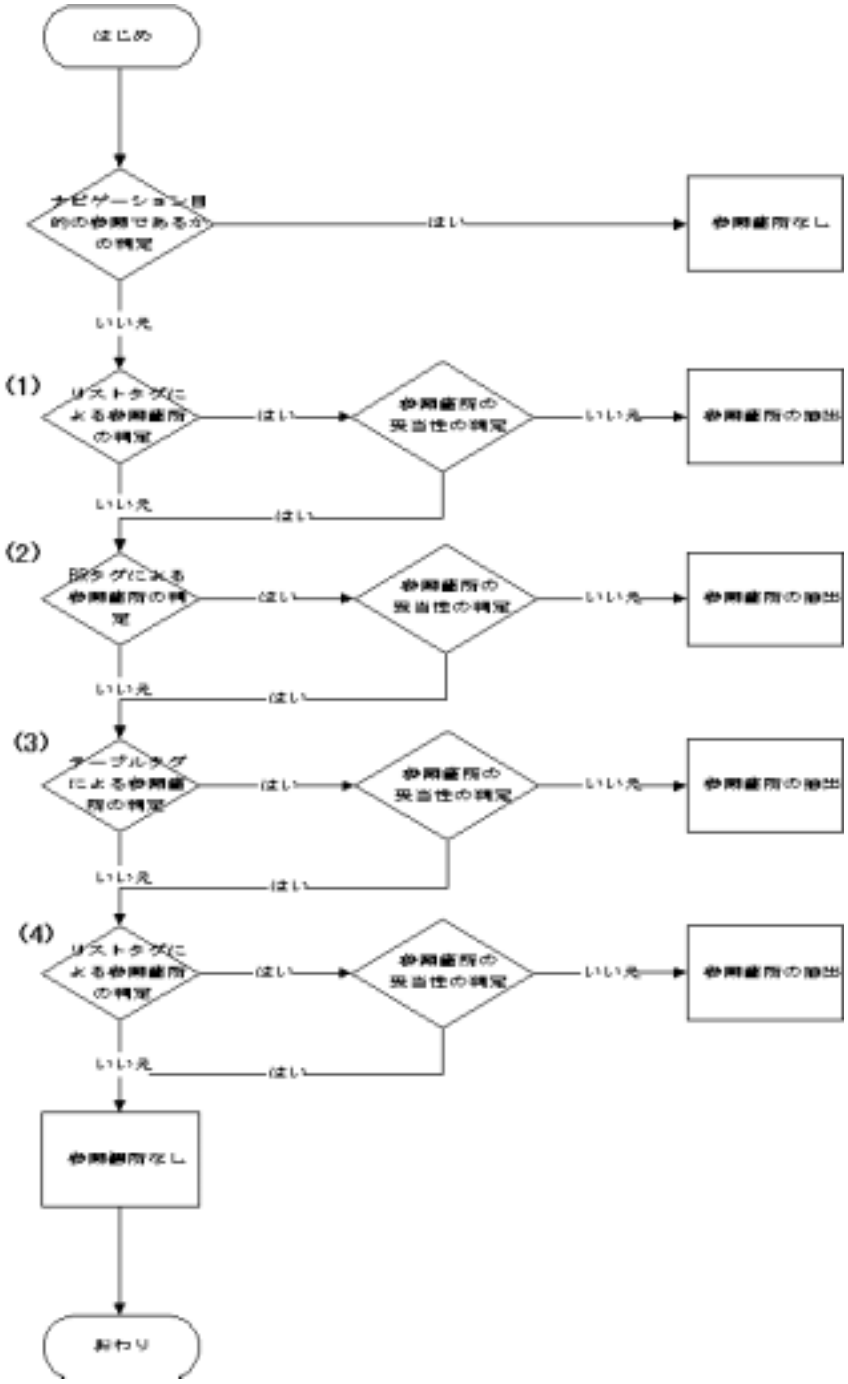

図 3.2: 参照箇所抽出フローチャート

(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．なお，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．

⟨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．⟨dt⟩ タグ ⟨dd⟩ タグの順序は決まっていない．よって，これらのタグのペアは ⟨dl⟩ タグから始まる最初の 2 つの ⟨dt⟩ タグ ⟨dd⟩ タグの並びを判断することにより決定する．

また，⟨dt⟩⟨dd⟩ が入れ替わっても良い．

⟨dt⟩ チャットポータル（http://www.chat.co.jp）
⟨dd⟩ 国内のチャットサイトをジャンル分け

この場合，⟨dd⟩ タグから次の ⟨dt⟩ タグまでを抽出する．

(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 ...

このように，該当アンカーの前後に，該当アンカーを含めて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．実験用データを観察した結果，該当アンカーを含むセルの下または右のセルに参照箇所が存在することが多い傾向があった．ここでは，その観察の結果を用いて参照箇所を抽出する．

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

上の例のように，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．ただし，テーブルタグを用いて一行にアンカーとそのリンク先ページの説明を並べている場合でも，まれにアンカー以外の要素が存在するときもある．そこで，該当アンカーと同じ列にあるセルの全てがアンカーでなくても，そのうち 70%以上がアンカーである場合には，その表はアンカーとそのリンク先ページの説明を列挙しているとみなす．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，数は少なかったが，以下のように，アンカーが横に並んでいるタイプも見られた．そこで，該当アンカーと同じ行にあるセルのうち，70%以上がアンカーであれば，該当アンカーの下のセルに記述された文字列を参照箇所として抽出する．

が存在する横のセルにアンカーが 80%以上存在することで判定し，該当アンカーの下のセルを抽出する．

アンカー A	該当アンカー	アンカー B
アンカー A の参照箇所	該当アンカーの参照箇所	アンカー B の参照箇所

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．

窓の杜 ⟨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．

上記 4 つのパターンで参照箇所を抽出した後，参照箇所の妥当性を判定する．(図 3.2 参照)．すなわち，自動的に抽出された参照箇所が対象ページに関する情報として妥当であるかどうかを判定する．本研究においては，抽出された参照箇所が以下の条件のいずれかを満たす場合，参照箇所として抽出しない．

- 英字，記号（例：○×/-☆）からなる 5 文字以下の文字列である．
- 参照箇所が対象ページの URL の部分文字列である（例：「chat.co.jp」）．

以下の例は，チャットソフトを比較するページを示している．アンカー右側の「A」等の記号は，チャットソフトのサンプルへのリンクである．先ほど述べたように，テーブルタグを手がかりとする場合，参照箇所は該当アンカーの右または下に存在するとみなす．しかし，このように該当アンカーのすぐ右もしくは下には，☆等で表したランキング等の無意味な記述が存在する場合もある．このような場合，参照箇所として妥当でないと判断し，他のパターンによる抽出を試みる．ただし，この例に示したようなパターンにマッチする場合には，記号等のさらに右または下を参照箇所として抽出する．

ななせのレンタル	A	30 日間利用なしで削除されます。
Net4u	-	☆非常に高機能なチャットです。 初回の予定ルーム数には限りが ありますのでお早めにどうぞ。 初回募集以降も順次募集していきます。 (**CGI サービス)
Net-BULL	A	(**Net-BULL) タイトル、タイトル色、背景、全体の文字色、名前色、コメント色、カウンター、発言の初期化、発言表示件数などを自分でデザインして利用できます。

3.3 評価実験

3.2 節で述べた手法に基づき，参照箇所を抽出する実験を行った．

3.3.1 参照箇所抽出実験

本節では，3.2.1 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，3.2 節で述べた 2 つのデータセットについて，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するページ「まゆ工房[1]」と，他者による商品の評価を掲載するページ「リブラ[2]」の 2 つの対象ページとし，それぞれの参照ページから参照箇所を抽出した．これらの参照ページは，3.2.1 項の参照箇所抽出アルゴリズムの検討には用いていない．このオープンテストは，手法の汎用性を調べるために行った．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 3.2，表 3.3，表 3.4 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，3.2.1 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．

表 3.2: 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)
– その他のタグ	0.295	0.787	(0.436)

「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.「その他のタグ」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,「その他のタグ」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．例を図 3.3 に示す．全体がシステムが出力した参照箇所であり，下線部が正解である．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

[1] http://mayukoubou.plaza.gaiax.com/
[2] http://www.libra.ne.jp/

表 3.3: 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)
– その他のタグ	0.406	0.754	(0.561)

表 3.4: 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)
– その他のタグ	0.178	0.578	(0.536)

アクセス解析というのは、このHPにいらっしゃった方の数を日単位で集計してくれたり、そのアベレージを出してくれたり、はたまたご来訪者はどのURLを経由しておいでになっているのか、ということを調べてくれる大変スグレモノなプログラムのことです。これを無料でレンタルしてくれる「CGIboy」さんからお借りして、メインページに設置してあるわけです。ただ、このアクセス解析、私の意図としては、「どんな方がどれくらいうちのHPに来てくださっているんだろう」という純然たる興味から設置したわけですが、見方によってはアクセスの経緯をずっと見ているわけだから、何となくのぞき見っぽいと思われても仕方がない部分はあります。ひょっとしたらそういうことをとても不快に思う方だっていると思うんです。

図 3.3: 非常に長い記述が抽出される例

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．

3.3.2 アルゴリズムの改良に関する考察

3.3.1 項において述べたように，参照箇所を抽出するためには，HTML タグを手がかりとするだけでは不十分であり，言語的な情報を用いる必要がある．その手法について検討した．

語彙的連鎖を用いた手法

参照箇所は，Web ページ上において，段落や行などで定義される形式的なセグメントを形成しているとは限らない．もし形成していれば，HTML タグの情報だけでも十分抽出できるだろう．しかし，実験結果を見ると，形式的なセグメントを切り出すだけでは，参照箇所を正確に抽出することはできないとわかった．そのため，形式的なセグメントではなく，もっと細かい単位で参照箇所を切り分ける必要がある．

参照箇所は対象ページに対する記述であるから，一つのまとまりである．したがって，その部分は文中のまとまった部分であるパッセージを形成していると考えられる．そこでパッセージ抽出を考える．

参照箇所は該当アンカーについて述べているのであるから，アンカー文字列，または対象ページのタイトルのなかに対象ページを表す特徴的なキーワードが含まれていると考えられる．そこで，アンカー文字列，対象ページのタイトルを形態素解析し，そのなかの未知語，名詞をクエリにして，パッセージ抽出を行う．この際，語彙的連鎖を構成する基

準には以下の 3 つがある [望月 99].

1. 同一の語の繰り返しに基づく語彙的連鎖
2. シソーラスに基づく語彙的連鎖
3. 語の共起関係を基づく語彙的連鎖

2. のシソーラスを用いる方法では，インターネット上で日々生成される新語には対応できないと考えられる．また，3. の語の共起関係を用いた方法では，語の共起関係に関する大規模な統計情報が必要になる．したがって，期待できる成果と計算コストを考えると，同一の語の繰り返しに基づく語彙的結束性を利用することが最もうまくいくと考えられる．

3.3.1 項で述べたように，HTML タグの情報のみで参照箇所を抽出したとき，参照箇所を含むが，それ以外の記述も同時に抽出されることが多かった．そこで，語彙的連鎖によって参照箇所をパッセージ単位で抽出することにより，より正確に参照箇所が抽出できると期待できる．

手がかり語を用いた手法

[難波ら 99] が論文を対象に用いた手法である．この手法は，文間の結束性に注目している．手がかり語を用いた手法では，ある一文を基本とし，手がかり語の存在によって前後の文がその基本文と結束性があると判断されれば，基本文と連結して，一つのまとまりとして抽出する．

本研究でも，この手法が利用できると考えられる．まず，該当アンカーを含む文を基本文とする．次に手がかり語によって，基本文と結束性があると判断された文は，基本文とあわせて参照箇所として抽出する．

手がかり語は以下のものを考えている．

1. 照応詞（前方および後方照応している語）
2. 接続詞 （「詳しくは」、「また」等）
3. その他の結束性のある語

具体的な手がかり語については，実データから得られる参照箇所を調べて検討する必要がある．

手がかり語を用いた規則には以下のものを考えている．以下では，現時点での参照箇所のなかで最初の文を 最初の文 とし，最後の文を 最後の文 とする．

1. 最初の文 に照応詞がある場合，前の文も抽出する．
2. 最後の文 の次の文に補足型の接続詞（詳しくは）が含まれる場合，その文を抽出する．

参照箇所をより細かく分析した上で，規則を新たに作ることを考える．そして，実験により有効であると判断された規則だけを適用する．

第4章　参照箇所の提示

4.1　参照箇所から得られる情報の分析

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の一部である.

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。そのほかの操作性はパナソニック端末に慣れていれば問題なし。というか、P503i とはまったく同じだ。アイコンとリストメニューを組み合わせていて、ショートカットを設定できたり、メモリダイヤルのグループ別に受信メールを分類できるのは便利だと思う。クリアボタンや i モードボタンなど、他の 503i シリーズに見られるボタンが省かれていたりするが、それでもしっかりと i モードブラウジング中には受話ボタンでスクロールモードに切り替わるなど、いろいろと工夫されている。デザインは見ての通り個性派。個人的にはSO503i よりもネックストラップに似合うと思う。ちょうつがい側にアンテナとストラップ穴があり、ストラップで吊り下げるときにアンテナが上になるのだ。このデザインが気に入ればもはや迷うことはない。ただ、赤外線ポートのデザインが気に食わない。もっと綺麗に処理した方が良いと思うのだが。ちなみに。編集部で調査したところ、Java の実行速度はP503i （初期ロット）に比べると画像描画を中心に大幅に高速化されている。

この例では，ページ全体に携帯端末 P503is に関する記述があり，記述は長文である．

4.2 参照箇所の提示

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．例えば，3.3.1 項で述べた実験では,「CHAT.CO.JP」を対象ページとした場合，42 個の参照ページから参照箇所が抽出される．これら全ての参照箇所をユーザに提示しても，その全てに目を通すのは時間がかかるため，効果的であるとは言えない．そこで本節では，参照箇所をユーザにとって分かりやすくする方法について検討する．

具体的には，以下のことを行うことを考える．

1. **冗長な参照箇所を削除**
   複数の参照箇所からほとんど同じ情報が得られるときがある．対象ページが「窓の杜」の場合の例を以下にのせる．

●ソフトのダウンロード～ 窓の杜 ／ Common Archives Library

シェアウェアやフリーソフトなどのダウンロードが出来る。

パソコン用の各種プログラムがダウンロードできます。（インプレス社）

   この例の場合，パソコンソフトがダウンロードできるという記述が何度も出てくる．ユーザに提示する際には，このような記述は一度述べるだけで良い．そこで，冗長な参照箇所を削ることを考える．そのためには，参照箇所の文章の共通部分を同定する必要がある．柴田は，新聞記事を対象に，文章融合の際に重複文を同定する手法として以下のことを提案している [柴田 97]．ここでは，文章 1 と文章 2 を融合させる際に，重複文を削除することを考える．この際，まず対象となる文章の全文に形態素解析を行い各形態素の出現頻度を調べる．この時，汎用的に用いられる頻度の高い形態素は除外しておく．次に文章 1 と文章 2 の各文の組み合わせについて，一致する形態素を検出し，それらの形態素の出現頻度に応じて以下の式 (4.1) を用いて，二文間の重複度を計算する．

$$\frac{1}{N} \cdot \sum_{\omega} \frac{100}{(\omega\text{の文章 1 中の出現回数}) \times (\omega\text{の文章 2 中の出現回数})} \tag{4.1}$$

   式（4.1）において，ω は 2 つの文に共通して現われる単語で，N はその数である．文章 1 と文章 2 で，一度だけ用いられてる単語を含む文は重複文である可能性が高い．逆に，文章 1 で N 回，文章 2 で M 回というように多く用いられていれば，N×M の組合わせが考えられるので，その分重複の可能性が低くなる．

   この式によって重複文の検出を行った結果，再現率 96%，精度 96%が得られたと述べている．この手法は形態素の出現頻度をもとに重複文を検出しているため，Web

の文章にもそのまま適用できると考えられる．

2. **参照箇所のタイプの分類**

4.1 節で述べたように，参照箇所から取り出される情報は，大きく分けて 3 種類あることが分かった．そこで参照箇所を 3 つのタイプに分類し，タイプ毎に分けて参照箇所を分類すれば，ユーザも対象ページの内容を把握しやすくなると考えられる．具体的には図 4.1 のような出力を目指す．

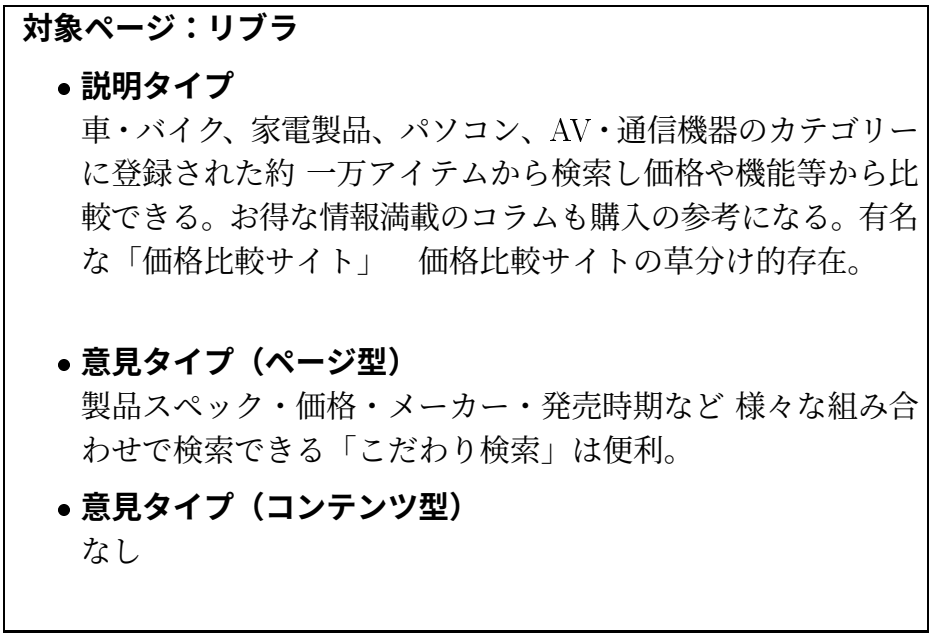

図 4.1: 参照箇所の提示の例

図 4.1 では，説明タイプ，意見タイプ（ページ型），意見タイプ（コンテンツ型）の順に参照箇所を提示している．しかし，どのタイプを先に表示するかは，ユーザにあらかじめ選ばせてもよい．例えば，ユーザが Web ページのデザインやレイアウトのに特に関心があれば，意見タイプ（ページ型）の参照箇所を先に見せる．このようなカスタマイズ機能は，より使いやすい Web 探索支援の要因として重要である．

参照箇所を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．

第5章　結論

本研究では，Web 探索支援を目的とし，参照ページから参照箇所を抽出する手法を提案した．提案手法によって，オープンテスト，部分一致で，参照箇所を再現率 0.69，精度 0.62 で抽出できることを確認した．また，日本語以外の言語で書かれた対象ページの説明や，対象ページそのものや対象ページのコンテンツに関する評価など，対象ページを見ただけではわからない情報を参照ページから取得できることを確認した．しかし，参照箇所が該当アンカーから離れている場合は，参照箇所を抽出することが出来なかった．これは言語情報を用いることで可能になると考えられる．

次に，参照ページのユーザへの提示方法について検討した．参照箇所をその内容に応じて 3 つに分類することにより，ユーザにより見易い提示ができると考えられる．特に，対象ページの説明でなく，対象ページの評価や意見を述べた参照箇所が特定できれば，ある Web ページに対する評判情報をユーザに提示し，Web 探索において有用な情報になると考えられる．

5.1　今後の課題

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．3.3 節で述べたように，HTML タグのみを手がかりに抽出された参照箇所は，ユーザに提示すべきでない情報が含まれていることも多い．そのため，言語情報を用いて，HTML タグで抽出した参照箇所から，提示すべきでない情報を削ることで，より正確に参照箇所を抽出できると考えられる．

また，参照箇所を 4.1 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．

インターネットには，顔文字が頻繁に出現する．顔文字とはつまり，著者の感情表現である．これらを手がかりの一つとして，Web ページや，製品に対する評価を取得できる可能性がある．これらを利用することは，今後の課題とする．

謝辞

本研究を進めるにあたり，終始熱心なご指導を賜りました白井清昭助教授に心から感謝致します．また，東京工業大学精密工学研究所の奥村学助教授には，貴重なコメント，ご指導を頂きました．感謝致します．また，日本学術振興会 特別研究員の難波英嗣氏に，大変お世話になりました．感謝致します．中間審査などのおりに諸先生方から貴重な御意見を頂きました．感謝致します．さらに，貴重な御意見，討論をして頂いた島津明教授，自然言語処理学講座，望月源助手，ならびに研究室の皆様に感謝致します．

最後に，多くの方々の貴重なご援助により，本研究が行うことができましたことを厚く御礼申し上げます．

関連図書

[Page98] Lawrence Page, Sergey Brin, Rajeev Motwani, Terry Winograd, “The PageRank Citation Ranking: Bringing Order to the Web”,1998.
http://citeseer.nj.nec.com/page98pagerank.html
http://google.stanford.edu/~backrub/google.html

[Chakrabati99] Soumen Chakrabarti, Byron E. Dom, S. Ravi Kumar, Prabhakar Raghavan, Sridhar Rajagopalan, Andrew Tomkins, David Gibson, Jon Kleinberg. “ Mining the Web’s Link Structure”. IEEE computer Vol. 32, No. 8, pp60-67. 1999.

[鷲崎ら 99] 鷲崎 誠司, 村本 達也. “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, 1999.

[Amitay00] E.Amitay. “InCommonSense - Rethinking Web Search Results”. ICME 2000.

[難波ら 99] 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.

[福原 98] 福原 知宏. “協調フィルタリングに関する研究動向”

http://db-www.aist-nara.ac.jp/~tomohi-f/
Docs/cf_review/cf_review.html

[Ellen98] Ellen Spertus, Lynn Andrea Stein. “A Hyperlink-Based Recommender System Written in Squeal”. CIKM’98 Workshop on Web Information and Data Management (WIDM’98)November 6, 1998.

http://www.mills.edu/ACAD_INFO/mcs_spertus.html

[船坂ら 96] 船坂 貴浩, 山本 和英, 益山繁. “冗長度削減による関連新聞記事の要約”. 情処研報 NL114-7. 1996.

[月出ら 00] 月出 奈都子, 石坂 俊. “TV 番組に対する自由回答文の印象抽出システム—インターネットアンケート調査による自由回答文の解析—”. 言語処理学会第 6 回年次大会, pp249, 2000.

[山田ら 00] 山田 洋志，福島 俊一，松田 勝志．“Web ページからのタイプ別情報抽出・分類手法”．情報処理学基礎 57-19，自然言語処理 136-19，2000．

[立石ら 01] 立石 健二，石黒 義英，福島 俊一．“インターネットからの評判情報検索”．自然言語処理，144-11，2001．

[吉田ら 00] 吉田 稔，鳥澤 健太郎，辻井 潤一．“表形式からの情報抽出手法”．言語処理学会第 6 回年次大会，pp252，2000．

[乾ら 98] 乾 裕子，内元 清貴，村田 真樹，井佐原 均．“文末表現に着目した自由回答アンケートの分類”．自然言語処理研究会 128-25，1998．

[望月 99] 望月 源．“語彙的連鎖を用いたパッセージ抽出とその応用に関する研究”．北陸先端科学技術大学院大学 博士論文．1999．

[柴田 97] 柴田 昇吾，上田 隆也，池田 裕治．“複数文章の融合”．情処研報 NL120-12，1997．

[Terveen97] Terveen Loren，Hill Will，Amento Brian，McDonald David and Creter Josh “PHOAKS: A System for Sharing Recommendations”．CACM．Vol．40，No．3，pp59-62，1997．

発表論文

(1) 板橋 英夫 望月 源 白井 清昭 奥村 学.“参照ページからの情報を利用した Web 探索支援”. 言語処理学会 第 8 回年次大会 2002（発表予定）