情　報　学　基　礎　55－10
デジタル・ドキュメント　19－10
（１９９９．７．１６）

ハイパーリンクの構造を利用した検索結果の選択手法

鷲崎 誠司　村本 達也
NTT サイバースペース研究所
E–mail : {suzaki, muramoto} @isl.ntt.co.jp

概要

WWW 上の情報の増加に伴い，WWW の検索サービスの有効性が相対的に低くなっている．本稿では，利用者が持つ検索主題に対して出力される膨大な検索結果に，それらの情報に関連する WWW のハイパーリンクの構造を利用して，情報の選択要因となる新たな指標としてのアンカー文字列を付加情報として提示することで，利用者の情報選択の認知的負荷を軽減する手法の概略と，その適用例に関して述べる．アンカー文字列は，情報製作者が目的となる情報に対して付与した注釈文章と見做すことが可能で，これを有効に活用すれば，目的となる情報を更に特徴付けることが期待でき，また情報を構造化するための手段となる可能性を秘めた情報である．

A New Decision Factor for IR System extracted from Structure of Hypertexts

Seiji SUSAKI, Tatsuya MURAMOTO
NTT Cyber Space Labs.

Abstract

This paper presents a expansion technique for anchor strings to select useful information from results that WWW search engine finds in database. There are many anchors in WWW information, and anchor strings clarify contents that are referred from other information. We explore anchor strings from contents that the Internet robot extracted from WWW, and apply these anchors to select WWW information from database. Anchor strings are regarded as user's anotation for target information that is connected from anchors.

1　はじめに

WWW (World Wide Web) 上の情報が増加するに従い，WWW 利用者は自分の目的とする情報を発見することが難しくなっている．この問題に対処するために，情報収集プログラム (インターネットロボットとも呼ばれる) が WWW 情報をハイパーリンクを辿りながら自動的に収集し，

それをインデクシングすることで，WWW 情報の部分集合を検索できるようにしたロボット型検索サービスが存在する．しかし，取り扱う情報量が膨大になるにつれて，その有効性が相対的に低くなっているため，様々な手法を用いて検索結果に付加価値を付ける試みが行われている．例えば，概要表示 (AltaVista[1]，Goo[2] 等)，情報を特徴付ける付加情報の表示 (記述言語，発信国などの付加情報の抽出 ： TITAN[3][8]，情報の翻訳 ： AltaVista)，結果の分類 (TITAN[5])，クラスタリング (TITAN)，ハイパーリンクの構造に基づく情報の整理 (InfoSeek[4]，Excite[5])，絞り込み情報の表示 (MONDOU[6]，TITAN)，ディレクトリ型検索サービスとの融合 (InfoSeek) 等がある．

本稿では，上記手法とは異なる手法として，ロボット型情報検索サービスの結果の選別を容易にするために，WWW 情報空間における情報間のハイパーリンク (いわゆるアンカー) が，各々の情報提供者が熟慮した上で付与した利用すべき情報であることを利用する手法に関して説明する．すなわち，アンカーに付随する文字列 (以後，この文字列のことをアンカー文字列と呼ぶ) は目的とする情報に対するアンカー製作者による注釈であると考えることで，各情報間の関係の橋渡しをするだけだったハイパーリンクの構造とアンカー文字列を，より積極的に検索結果の選択に利用する．

2 アンカー文字列の抽出とその拡張

2.1 情報を結ぶハイパーリンク

WWW 上の情報は，それのみが独立して (ハイパーリンクの依存関係が存在しないもの) 存在する場合もあるが，他の情報に対するハイパーリンクが存在する，あるいは他の情報からハイパーリンクで関連付けられている場合が圧倒的に多い．このような WWW 情報空間は，情報をノードとし，ノード同士を関連づけるハイパーリンクから構成される広大な網空間と見做すことができる．WWW 情報の作成者は，独自の情報を発信すると同時に，自分の情報に付加価値を付けるために，他の利用者が作成した情報に対してハイパーリンクを作成し，その際にはそのハイパーリンクを特徴付けるアンカー文字列を付与する．利用者は，WWW 上の検索サービス等から得られる情報を起点にして，ハイパーリンクを辿ることで試行錯誤しながら，自分の探索目的を達成するという，いわゆるブラウジング操作を行う [4](図 1)．

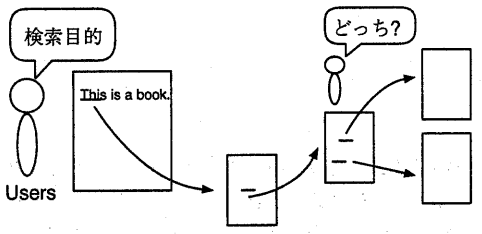

図 1：ハイパーテキスト型の検索過程

アンカー文字列は，情報製作者が利用者のために，リンクされている情報へと誘う窓口に成り得るものであり，各情報製作者はそのような意図を持ってアンカー文字列を付与すると考えることができる (但し，全ての場合でそうとは限らない)．このように製作意図を持つアンカー文字列から関係付けられる情報は，他の情報からも同様にリンクされていることが予想され，複数のリンク元の様々なアンカー文字列により特徴付けられていると見做すことが可能である．

ある任意の情報を考えたとき，全 WWW 情報の中からそれをリンクしている情報が (理想的には全て) 抽出でき，更にそのアンカー文字列を全て抽出することができれば，これらを利用して元の情報を更に特徴付けることが可能になる．様々な利用者により付与されるアンカー文字列は，様々な様式や表現形式が存在することが予想されるが，各々は各情報作成者により主観的に，更には意図的にその情報を特徴付けるのに相応しいと考えたものであるため，情報製作者による注釈文章 (単語，句等) であると見做すことが可能である．

本稿では，検索結果の出力時に，各々の結果に対するアンカー文字列を付加情報として表示する

[1] http://www.altavista.com/
[2] http://www.goo.ne.jp/
[3] http://titan.mcnet.ne.jp/
[4] http://www.infoseek.co.jp/
[5] http://www.excite.co.jp/
[6] http://www.kuamp.kyoto-u.ac.jp/labs/infocom/mondou/

ことを考える．この時，以下のような効用があると考えられる．

1. ハイパーリンク製作者による注釈と考えることができ，信頼できる可能性が高い

2. 該当情報の客観的評価に利用できる
   リンクされている数が多い場合は，情報の信頼性が高いし，有用な情報である確率が高い

3. 情報の構造化に利用できる可能性がある
   アンカー文字列を何らかの基準で整理すれば，情報の目次的な情報 [7] に成る可能性がある

2.2 アンカー文字列の特徴

まず，どのようなアンカー文字列が利用されているかを把握するために，アンカー文字列の特徴に関して調査した結果を述べる．情報収集プログラムにより収集したWWW情報約18万件から抽出したアンカー文字列約99万件 (5.5件 / 情報) の表層的な構造から表1のように分類した．

表1：アンカー文字列の分類

分類	例	出現数
名詞相当	懸賞，Yahoo	128,110
複合名詞	検索エンジン，美術館	251,967
名詞句	検索エンジンの情報	111,472
指示詞	これ，あれ	5,479
文章相当	検索エンジンの情報へ行く	87,214
未定義語	辞書に無い単語やその一部等	24,680
シンボル	●，→ 等	2,524
URL	http://titan.mcnet.ne.jp/	19,980
E-mail	suzaki@isl.ntt.co.jp	2,013
画像	[IMG]	92,873
その他	上記の組み合わせ等	275,389
総計		993,193

表1における名詞相当とは，名詞、英単語、数字、辞書に存在する仮名、片仮名を含む．分類の前処理として，アンカー文字列に対して形態素解析 [6] を行っているため，辞書に登録していない未定義語や、WWW情報中に高頻度で出現する仮名や片仮名文字で構成される語がかなり大量に含まれていることがわかる．更に，「これ」，「それ」等の名詞形態指示詞や，「この」，「その」等の連体詞形態指示詞を含む (あるいはそれが単独でアンカー文字列となるもの) アンカー文字列や，これらを組み合わせたもの (本稿では，その他に分類している) も存在する．また，WWW情報に特有のURL(Uniform Resource Locator)，メールアドレスや，画像をアンカー文字列の代替としたり (我々のHTML解析器では [IMG] として蓄積される)，レイアウト上の制約として，長さを揃えるため単語や文の一部をアンカー文字列にするような場合も見受けられる．特に，単語や文の一部であるかどうかを見分けるための一方法としては，候補と見受けられるアンカー文字列を含む最低限の情報を取得し，それを再度形態素解析した結果が意味を成しているかどうかを調べることが考えられる．

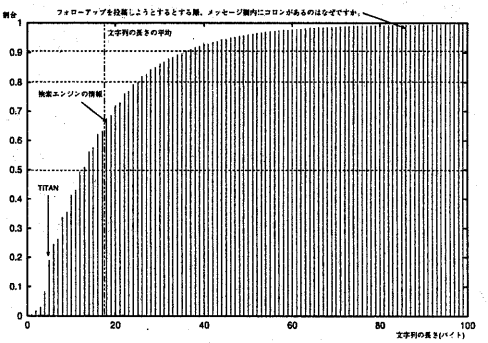

図2：アンカー文字列の長さの累積度数グラフ

図2はアンカー文字列の長さと累積度数の関係を表す．アンカー文字列の平均の長さは17.9バイト，中央値の文字列の長さは13バイト，最頻値の文字列の長さは5バイト (95,078個，英単語が多い) であり，全体の80%は，アンカー文字列の長さが25バイト以下，更に90%は長さが36バイト以下であった．これらの結果から，一般的に簡潔な名詞 / 複合名詞 / 名詞句等がアンカー文字列として利用されていることがわかる．長い文章が利用される場合の典型例としては，FAQ集で質問そのものをアンカー文字列としている場合がある．

2.3 アンカー文字列の拡張

このような特徴を持つアンカー文字列を，情報を特徴付ける文字列と捕らえる際の問題点としては，文字列が指示代名詞 (例：ここに詳細情報があります) のみの場合，メールアドレスやURL等しか含まれない場合，更に画像が用いられる場合

等の取り扱いであろう．これらの情報を，特徴付けるための文字列として利用すると，意味が無いばかりではなく，利用者に不要な情報を提示することになり，利用者の認知的負荷を増大させる結果となる．よって，何らかの手法で，このように情報が不足している文字列を救う必要がある．

また、名詞、名詞句、複合名詞がアンカー文字列になる場合は、リンク先のタイトルをそのまま採用している場合が多く見受けられる。このような場合、対象のアンカー文字列の近傍、あるいはそれを支配しているヘッダ等で情報を補完することで、アンカー文字列の有効性は劇的に変化することがある．

Rowe[3] らは，WWW 上の画像検索のために，文中に含まれるアンカー文字列相当語句の抽出手法を報告している．検索を目的とした画像の特徴付けを行うためには，何らかのキーワードを付与する必要があり，従来は人手によりこれを付与していたが，彼らは使用色数等の画像そのものが持つ特徴と，以下の画像に関する文字情報をパラメータとして，訓練用の WWW 情報から各パラメータの重みを決定し，それらを統合することで精度良いアンカー文字列の抽出を行っている．

- 目標となる画像との距離 (行数)
- 他のアンカー文字列との距離
- 強調文字の存在
- アンカー文字列の長さ
- 特徴的な指示語
- ALT タグの内容
- 意味のある単語の存在 (いわゆるストップワードではない単語)

本稿では，画像がアンカー文字列の代わりになっている場合だけでなく，文字列がアンカー文字列になっているものに対しても拡張することを考えるため，上記パラメータをそのまま利用することはできない．今回は，訓練用 WWW 情報 40 ファイル中から，アンカー文字列を 559 個抽出し，構文解析，意味解析などの処理が重い自然言語処理を用いず，表層的な特徴を用いたアンカー文字列の拡張方法に関して検討した．検討結果とその拡張例を以下に示す．例では，便宜上ハイパーリンクを下線で示している．

1. 名詞相当，複合名詞，名詞句

   step1　アンカー文字列が 10 バイト以上のものはそのまま採用する

   step2　step1 に該当しないものに対して，アンカー文字列を含む前後の文字列を抽出する (句点，改行等区切りと考えられる部分まで)

   step3　step2 までに適切な文字列が抽出できなければ，直近のヘッダ，あるいはタイトルを添付する

   例：「北海道 では蟹が大漁である」の場合，「では蟹が大漁である」も追加する

2. 指示詞
   アンカー文字列を含む前後の文字列の追加

   例：「これ が北海道の情報」の場合，「が北海道の情報」も追加する

3. 文章相当
   そのままアンカー文字列として採用する

4. 未定義語
   アンカー文字列を含む前後の文字列の追加
   例：「北海道の函 館では蟹が大漁である」の場合，「館では蟹が大漁である」も追加する

5. シンボル，URL
   アンカー文字列を含む前後の文字列の追加
   例：「→ 函館の蟹」の場合，「函館の蟹」も追加する

6. E-mail
   アンカー文字列を含む前後の文字列の追加．さもなくば，タイトルを追加する

7. 画像
   Rowe の手法 (パラメータと，それに対する重み) を利用する

例 : 「[IMG] 美瑛の畑の写真」の場合, [IMG] を削除して, 「美瑛の畑の写真」追加する

2.4 アンカー文字列の抽出実験

上記のアンカー文字列の拡張手法の有効性を確認するため, 机上検討を行った. 評価用のデータとして, 訓練用データとは異なる 50 の WWW 情報の中から 793 個のアンカー文字列を抽出した. 表 2 は, アンカー文字列を分類し, 本手法により文字列拡張したものに対して適合率を算出したものである. 適合率は, 本手法により拡張されたアンカー文字列が, リンクしている情報を特徴付けるのに相応しい場合の全体に対する割合である.

拡張文字列が正解である基準としては, 他に最善の拡張方法は存在するが, 拡張されたものが情報を特徴付けるのに有効であるものとした. 文字列を合成する等, 拡張手法を再考すれば, より適切な文字列になる可能性は残っている.

表 2 : アンカー文字列の拡張の適合率

分類	総数	出現割合	正解数	適合率	成功数 / 拡張数
名詞相当	85	10.7	63	74.1	58/71
複合名詞	212	26.7	167	78.8	20/26
名詞句	63	7.9	50	79.4	5/6
指示詞	2	0.2	2	100	2/2
文章相当	149	18.8	133	89.3	0/0
未定義語	12	1.5	6	50.0	6/12
シンボル	0	-	0	-	-
URL	2	0.2	2	100	2/2
E-mail	13	1.6	8	61.5	8/13
画像	94	11.9	51	54.3	51/94
その他	161	20.3	153	95.0	5/5
総計	793	100	636	80.2	147/212

全体としては, 80.2% の適合率が得られたが, 全体に対する出現割合が高い名詞, 複合名詞, 名詞句の拡張正解率が低いことがわかる. 前記拡張ルールを用いて拡張されたものは, 793 個のアンカー文字列の候補中 212 個で, 全体の 26.7% であり, 拡張して成功する割合は 69.3%(画像を除けば 81.4%) であった. アンカー文字列の拡張に失敗した例に関して考察してみると, 以下のような状況が挙げられる.

- キーとなる文字列が, 複数のアンカー文字列を支配している
- 複数の文字列を組み合わせて拡張する必要がある
- HTML の構造を生かした補完が必要である
- リンクの階層構造を考慮した拡張をする必要がある
- 指示代名詞が何を指しているのか不明である
- ゼロ代名詞が存在し, 主題が不明である
- 多くの情報が非文字列で有効に拡張できない

特に, 名詞, 複合名詞, そして名詞句等に対して精度良く文字列を拡張するためには, 各々に対して拡張条件 (文字列の長さの制約等) を十分吟味した上で, リンクのトポロジを考慮した情報の補完, HTML の構造を生かした補完や, 構文解析や照応解析等の自然言語処理などを考慮すべきである. 但し, 文字を補完し過ぎると, 他のアンカー文字列との間の関係が冗長になる可能性や文字数が多くなるにつれて一覧性が低くなる問題があり, 最適な拡張方法に関して議論が必要である.

3 システム構成

3.1 適用例

我々は, 拡張したアンカー文字列を, WWW 検索サービスの検索結果の選択に利用するためのシステムを, 以下のように提案する.

アンカー文字列を収集するために, 全 WWW 情報を収集することは, 収集プログラムの性能や計算機の性能などの内的要因, ネットワーク性能, そしてリンクされていない情報の存在等の外的要因などにより到底無理であるため, 情報収集プログラムを用いて収集した情報を, WWW 情報の部分集合として利用する. 収集された情報はインデクシングし, WWW 上で検索可能とする. 各情報中には, 情報制作者による様々な情報に対するハイパーリンクが存在するため, 各情報を元情報と考え, それに対するリンク元 URL, アンカー文字列等を抽出し, これらを RDB で管理する (図 3). アンカー文字列は, 前章の手法を用いて拡張済みのものである.

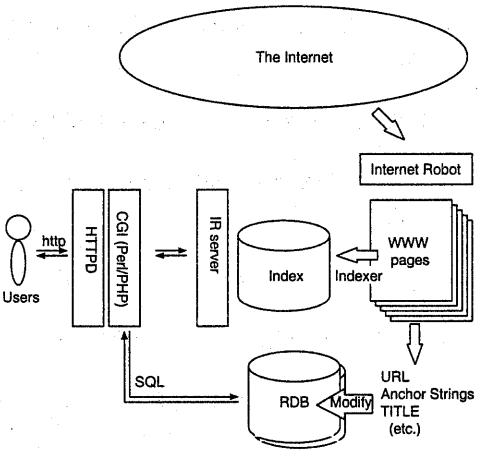

図 3：システム構成

検索サービスにより得られる結果は，使用する検索サービスにより適合度順に整列させられたものであるが，要求があれば各結果に対してそれをリンクしているアンカー文字列を RDB から抽出し利用者に提示する．利用者は，検索サービスが表示する概要などの情報では，選別のための情報として不足していると感じた時には，アンカー文字列を表示することで，各情報製作者が付与した信頼できる注釈情報を参照することができる．

検索結果は，各々の情報が他の情報からより多くリンクされているかどうかを基準に再整列させることも可能である．これは，各々のアンカー文字列が，目的の情報を有用な情報かどうかを測定するための指針と成り得ることを利用している．より多くの情報からリンクされている情報は，情報製作者にハイパーリンクを作成させるだけの高い動機付けを持つものであり，有用である可能性が高いと考えられる．

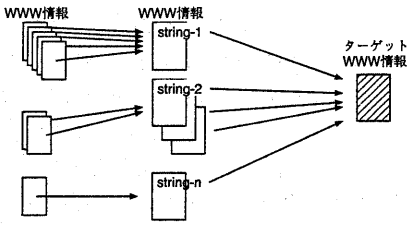

図 4：アンカー文字列により特徴付けられる情報

また，アンカー文字列の表示順序に関しては，図 4 のような多段のグラフ構造が存在する場合，ターゲットとなる WWW 情報をリンクしているアンカー文字列を含む情報が，より多数の情報からリンクされている場合，そのアンカー文字列の重要性が高くなると考え，この重要度順に表示を行うことも考えられる．どの程度の数の情報からリンクされているかに関しては，RDB にアンカー文字列を蓄積する際の前処理として行い，それらを同時に RDB に蓄積しておけば良い．

3.2 基本部分の実装について

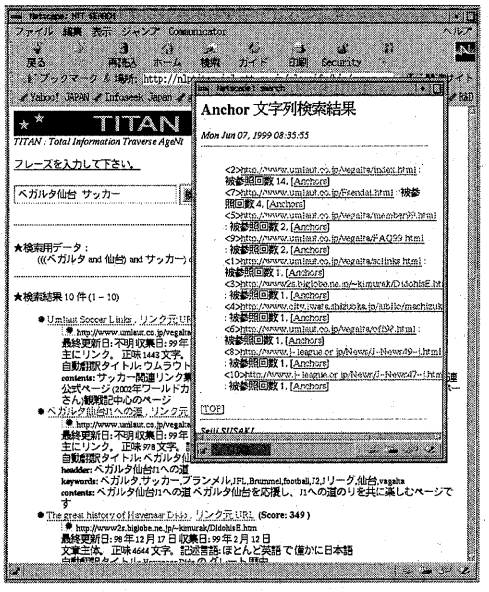

図 5：TITAN 結果出力画面例

我々は，情報ナビゲーションシステム TITAN を 1995 年から試験サービス中であり，この TITAN の情報収集プログラムが収集した情報の中から約 14 万件を抽出して，本手法の適用を試みた．

まず，情報収集プログラムが収集した WWW 情報から，アンカー文字列を抽出した．アンカー文字列は総計約 90 万件存在した．これらのアンカー文字列をリンクしている情報の URL，リンク元の URL 等と共に RDB(PostgreSQL) に蓄積した．この情報を蓄積するのに，1.2GBytes のディスク容量を必要とした．TITAN の検索エンジンは，freeWAIS-sf を利用しており，情報収集プログラムが収集した WWW 情報を TITAN のサービスに見合うように独自に特徴付けて (記述言語や情報の形式など) インデクシングし，WWW を通して検索可能になっている．

利用者は，httpd/CGIを介して検索サーバに対してwaisプロトコルに基づく検索要求を出し，検索サーバはそれに適合する情報を適合度順に整列して返却する．従来のTITANでは，結果をタイトル，概要，記述言語，発信国，情報の型などを各々のURLと共に表示するが，アンカー文字列を利用すると，例えば結果の選別時に以下のような操作が可能になる．

step1　通常検索

step2　被参照数による再整列
リンクされている数が多ければ，その情報は価値がある可能性が高い

step3　アンカー文字列の表示
特定の情報に対するアンカー文字列を表示することで，どのような情報が含まれているかを見当付ける(step1 から直接検索可能)

step4　情報の参照，あるいはstep1，step3へ

step5　必要であれば，リンクを逆に辿り情報をブラウジングする

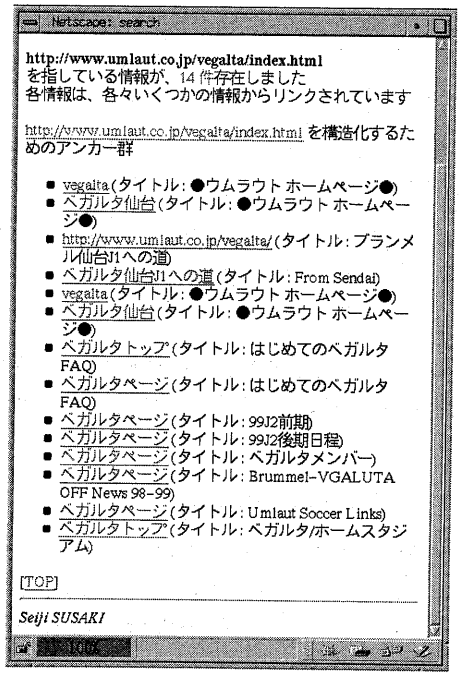

図6：アンカー文字列検索の結果出力例

図5，図6では，検索語句として「ベガルタ仙台，サッカー」を入力した場合の，結果の絞り込み例を示している．図5では，TITANの通常検索結果に対して，アンカー文字列を用いた再整列させた結果を示している．表示されている結果10件のURLを検索キーとしてRDBに問い合わせ，リンクされている数を基準に降順に整列させている．図6では，この結果の中から，最も多くの情報からリンクされている情報に対して，アンカー文字列検索を行い，アンカー文字列14件分を表示している．このようにより多くの情報からリンクされている情報は，有効である可能性が高く，多くの利用者の目に触れていることが予想され，他の情報よりも先に参照すれば良い．

4　考察

アンカー文字列を用いて情報の選別補助をする本手法は，従来まで行われていた概要表示などの補完手法と成り得る．これは，情報作製者が陽に情報を特徴付けたアンカー文字列を利用することで，概要などを機械的に作成するだけでは生まれない主観的な注釈情報の提供になる可能性があるからである．

但し，アンカー文字列の性質上，有用でない形式で提供されるものもあるため，これらをいかに救済するかがポイントであろう．本稿では，いくつかの種類のアンカー文字列を救済するために，単純な手法を用いているが，文字列の表層的な観点から情報の補完をするだけではなく，HTMLの構造や自然言語処理技術を用いて意味的なところにまで踏み込めれば，より有用な情報に成り得る可能性がある．

アンカー文字列の別の側面としては，アンカー文字列を目次情報と見做すことできる可能性があることが考えられる．WWW情報は，情報が断片的に配置されたグラフ構造であると見做すことができるが，実際はこのような情報ばかりでなく，一つの情報の中に多種多様な情報が混在する情報も存在する．例えば，様々な学術情報が含まれる情報があると仮定すると，その中には自然言語処理に関するものもあれば，情報検索関係の情報も存在するであろう．これらをリンクする場合は，アンカー文字列として「自然言語処理の情報」と

いう文字列を使う人もいれば，「情報検索の情報」という文字列を利用する人もいる(個人の興味により異なる)．この場合，目標となる情報は，2つの異なるアンカー文字列により特徴付けられることになり，これらのアンカー文字列を参照すれば，目的の情報は各々の情報が含まれる情報であることが明確になる．

また，断片的なWWW情報を効率的にアクセスするため，そのリンクトポロジを視覚化する手法[2]等も提案されているが，本手法でもハイパーリンクの多段構造を生かした文字列の補完を行うことで，より有用な情報の抽出が期待できるため，今後検討すべきであろう．

5 従来研究

MONDOU，HotBot[7]等のサービスでは，いわゆる逆リンク検索サービスを行っている．これは，URLを検索キーとして，それを参照している情報を表示するものである．参照元情報を表示することに関しては意義があるが，それ以上の付加的なサービスはしていない．

Cakrabarti[1]は，ハイパーリンクの構造を利用して，目的とする情報をリンクしている元情報や，その情報が存在するサーバのリンク構造をapplet上に表示することで，利用者の情報ナビゲーションの利便性を向上させる研究を行った．現在参照している情報に対するリンク元情報は，HotBot等の既存のロボット型検索サービスを動的に検索することで実現している．元情報は，そのタイトルのみを表示しているため，利用者にとっては，どのような文脈からハイパーリンクが作成されているかに関しては不明瞭である．

6 まとめ

本稿では，ハイパーリンクの構造を利用して，情報を指し示すアンカー文字列を，目的となる情報を新たに特徴付ける文字列として利用者に提示することで，検索結果の選別補助と成り得ることを示した．

アンカー文字列は，従来の機械的に生成される概要などとは異なり，情報提供者の主観的な注釈であると見做せば，精度が高くより有用な情報になる可能性がある．アンカー文字列は結果を特徴付ける主情報として利用するのではなく，従来の概要情報を補完するものと考えることで，その価値が出てくるものと期待している．

今後は，現在サービス中のTITANのデータを利用して，本手法の拡張性を確認し，アンカー文字列の信頼性の高い拡張方法を考察する．更に，複数のアンカー文字列を利用した情報の構造化や，リンクトポロジを考慮したアンカー文字列の拡張方法に関して検討する予定である．

参考文献

[1] S. Chakrabarti, et. al., Surfing the Web backwards, The Eighth International WWW Conference, 1999

[2] T. Kopetzky, et. al., Visual preview for link traversal on the World Wide Web, The Eighth International WWW Conference, 1999

[3] N. Roew, et. al., Automatic Caption Localization for Photographs on World Wide Web Pages, Information Processing & Management, Vol. 34, No. 1, pp. 95–107, 1998

[4] R. Horn, ハイパーテキスト情報整理学，日経BP，1991

[5] 巌寺他，検索結果の再構成によるナビゲーション支援，情報処理学会第55回全大Vol.3, pp.424–425, 1997

[6] 渕他，保守性を考慮した日本語形態素解析システム，情報処理学会研究報告，NL117–9, pp.59–66, 1997

[7] 長尾，電子図書館，岩波書店，1994

[8] 鷲崎他，WWW上の情報探索システムにおけるインターラクティブインターフェイス，日本ソフトウエア科学会 WISS96, pp.1–8, 1996

[7] http://www.hotbot.com/