社団法人 情報処理学会 研究報告
IPSJ SIG Technical Report

2006－ＦＩ－ 84 (5)
2006－ＮＬ－175 (5)
2006／9／12

Web情報活用のためアンカーテキストの分類と利用

吉岡 真治
北海道大学大学院情報科学研究科

概要: Web情報の最も大きな特徴は、ハイパーリンクにより、複数のページが関係付けられ、その関係がアンカーテキストという形で記述されている点にある。本研究では、大規模なWeb文書コレクションであるるNTCIRのnw100gテストコレクションから抽出したアンカーテキストを対象に、アンカーテキストの機能分類を提案すると共に、機能分類の自動化のための方針、機能分類の特徴を生かしたWeb情報活用システムの応用手法について述べる。

Classification of Anchor Text for Web Information Application

Masaharu Yoshioka
Graduate School of Information Science and Technology, Hokkaido University

Abstract: Most significant characteristics of the Web information is description about a relationship among different pages by using hyperlinks and anchor texts. In this research, anchor texts extracts from large web test collection NTCIR nw100g are investigated for a proposal of a functional classification. A strategy for automatic classification of anchor texts in html pages and systems that utilize this classification are also proposed in this paper.

1 緒言

Webに公開されている情報量が飛躍的に増加するのに伴い、これらの情報を活用するための様々なシステムが提案されている。このWebページをデータ源とするデータベースと既存の文書データベースの最も特徴的な違いは、Webページ間の関係を記述するハイーパーリンクの存在である。

Webページにおけるハイーパーリンクの利用法としては、ページの参照関係のグラフ構造に注目したPageRank[1]などの手法だけではなく、html文書におけるアンカー(多くの場合、他のページへのリンク)に対応するテキストであるアンカーテキストを利用する方法が提案されている。

アンカーテキストを利用した先駆的システムとしては、McBryan[2]により提案されたWorld Wide Web Wormというクローラで、アンカーテキストから作成したインデックスをアンカー先のテキストのインデックスとして利用する方法を提案した。現在、この方法は、Googleなどの商用サーチエンジンにおいても広く利用されている。

また、このアンカーテキストは、上記の情報検索への応用のみならず、Webページの分類[3]やWebディレクトリの構築支援[4]など様々な分野で利用されている。

一方、実際にアンカーテキストにどのような情報が記載されている研究は少ない。Eironら[5]は、IBMのイントラネット上に存在する全てのアンカーテキストを分析し、アンカーテキストとqueryの類似性という観点から、アンカーテキストが検索に有用であるという分析を行っている。また、鷲崎ら[6]は、クローリングしたWebページ18万件から、アンカーテキストの表層的な分類(名詞相当・複合名詞など)を行ない、検索に有効なアンカーテキストを生成する

ための補完手法を提案している。

これらの研究では、主に、アンカーテキストに表層的にあらわれる語の種類に注目した研究であると考えることができる。

一方、アンカーテキストが持つ性質を考慮した使い分けを行っている研究がある。例えば、藤井ら [7] は、NTCIR-5 における誘導型検索において、サイト内からのリンクとサイト外からのリンクが持つ性質の違いに注目して、サイト外からのリンクに対応するアンカーテキストのみを用いた情報検索システムを提案し、高い検索性能を実現している。また、鈴木らの研究では、逆にサイト内からのリンクに主に注目することにより、Web ディレクトリ [4] の構築支援を行っている。

しかし、これらの研究では、サイト内からのリンクとサイト外からのリンクといった非常に大まかな分類しかしていない。よって、この分類を詳細に行うことにより、各々のアンカーテキストの性質を考慮した Web 情報活用システムに関する議論が行えるようになると考える。

そこで、本研究では、実際の Web ページにどのようなアンカーテキストが存在することを把握するために、主に日本語の Web ページのコレクションである NTCIR Web テストコレクション (nw100g)[8] を利用し、そこから抽出したアンカーテキストを対象とした分析を行う。また、提案した分類の特徴を生かした利用方法の提案を行う。

2 アンカーテキストの機能分類

2.1 アンカーテキストの抽出

本実験では、nw100g の全ての Web ページ (15,026,516 ページ) から、Perl の TokeParser [1] を用いてアンカーテキストを抽出した。HTML の構文が正しくない場合や漢字コードの問題などにより、多少のエラーを含むと考えられるが、109,133,114 個のアンカーテキストを獲得した。

このアンカーテキストを、URL 中のサイト名が同一、もしくは URL 名が省略されている html ページへのリンクを「サイト内リンク」、その他の html ページへのリンクを「サイト外リンク」、mailto, JavaScript[2] などのリンクを「その他」と分類すると、各々の総数は表 1 の通りであった。

表 1: リンク先のタイプによる分類

リンク先のタイプ	リンク数
サイト内リンク	87,379,245
サイト外リンク	16,438,585
その他	5,315,284

次に、サイト内、サイト外各々のリンクにおける出現頻度の高いアンカーテキストとその回数を整理した (表 2,3)。ここで、[IMG] は、Alt 情報のない画像を表す文字列である。また、表 3 のサイト外リストについては、「ドメイン | 」「登録 | 」とほぼ同数回現れている文字列が他に 5 つ存在したが、基本的に同じテンプレートから作成された BBS サイトからのリンクに対応するアンカーテキストであるため省略した。

表 2: サイト内リンクにおける高頻度アンカーテキスト

[IMG]	10263325
戻る	534805
HOME	343106
ホーム	279649
2	250669
こちら	191530
TOP	187817
3	183851
Home	180190
トップページ	154008
次へ	153011
BACK	150894
4	146803
お問い合わせ	141966
会社概要	139630

どちらのタイプのアンカーテキストにおいても画像だけを示す「[IMG]」や、「こちら」といった指事語が含まれている。

また、サイト内リンクには、「HOME」、「ホー

[1] http://search.cpan.org/~gaas/HTML-Parser-3.55/lib/HTML/TokeParser.pm

[2] JavaScript などにより参照されるページについては、現在扱っていない

表 3: サイト外リンクにおける高頻度アンカーテキスト

[IMG]	1344073
Click here to visit our sponsor	93079
HOOPS!	70013
ドメイン |	69732
登録 |	69727
こちら	51283
ヘルプ	33886
発行者 Web サイト	33736
あ	29869
掲示板	28938
HOME	28151
Yahoo! JAPAN	21453
最新号	21374
バックナンバー	20379
地図	19454

ム」、「TOP」などのトップページを表すテキスト、「戻る」、「次へ」などのナビゲーションを指示するためのテキスト、「2」「3」のようなインデックスを表すようなテキスト、「お問い合わせ」、「会社概要」などのページの機能を表すテキストが存在している。

これに対し、サイト外リンクには、「Click here to visit our sponsor」や「発行者 Web サイト」などのリンク元のページとリンク先のページの関係を表すテキスト、「Yahoo! JAPAN」「HOOPS!」などの、リンク先のページの内容を表すテキストが存在する。また、サイト内リンクと同様の「ヘルプ」、「掲示板」などといったページの機能を表すテキストや「HOME」というトップページを表すテキストが存在する。このようなテキストが、サイト外リンクに存在する理由としては、複数の Web サーバが連携しているために、実質的にサイト内リンクと同じ性質のものが混ざっているためと考えられる。

2.2 アンカーテキストの機能分類

前節でのアンカーテキストに関する議論を受けて、本研究では、アンカーテキストのタイプをその機能に応じて、以下のように分類する。

1. **リンク先の内容を表すテキスト**：「Yahoo! JAPAN」のように、リンク先の名前を示すテキスト
2. **ページの機能を表すテキスト**：サイト全体におけるリンク先のページの機能的役割を示すテキスト
3. **リンク先との関係を表すテキスト**：「発行者 Web サイト」などのリンク先とリンク元のページの関係を示すテキスト
4. **トップページを指示するテキスト**：「HOME」、「ホーム」、「TOP」などのトップページを示すテキスト
5. **ナビゲーションを指示するテキスト**：「戻る」「次へ」、「こちら」などのリンク先のページと関係なく用いられるテキスト
6. **インデックスを表すテキスト**：「1」「2」「3」「あ行」「○」などの、幾つかの関係するページをまとめるためのテキスト
7. **URL**：URL をそのまま利用しているテキスト
8. **その他**：アダルトサイトなどが「18 才以上」を Yahoo にリンクするような、リンク先のページと全く関係ないテキスト

2.3 アンカーテキストの自動分類

前節で提案したアンカーテキストの分類を大規模な Web ページ群に適用し、利用するためには、自動分類を行う方法論の構築が不可欠である。

この自動分類の方法としては、主に具体例を集め、それとの比較に基づいて判別を行う方法である事例収集タイプと、アンカーテキストとリンク元、リンク先のページの情報を比較することによって判別を行うルールタイプの 2 つのタイプを考える。以下では、アンカーテキストの機能分類の性質を考慮した自動分類の方針を述べる。

「4. トップページを指示するテキスト」、「5. ナビゲーションを指示するテキスト」については、バリエーションも少なく、十分な量の事例収集を行うことにより、ほぼ、判別できるようになると考えている。

また、「6. インデックスを表すテキスト」についても、大部分は、事例収集のみで、対応可能である。ただし、「6. インデックスを表すテキスト」については、個別事例に対応した多少のバリエーションが存在するため、一部ルールタイプによる拡張が必要である。具体的には、リンク元のWebページにおいて「6. インデックスを表すテキスト」にマッチする同等のアンカーテキスト(例えば、サイトマップなどにおいて、同じレベルのリスト要素に含まれるアンカーテキスト)において、「6. インデックスを表すテキスト」がある一定の割合以上存在する場合には、そのアンカーテキストを「6. インデックスを表すテキスト」として分類するというヒューリスティックスを利用した判別法が考えられる。

また、「2. ページの機能を表すテキスト」については、かなりのバリエーションが存在することが想定されるため、単純な事例収集では対応できないことが想定される。しかし、様々なWebサイトにおいて標準的に用いられる「2. ページの機能を表すテキスト」が存在する可能性が高いと考えられることから、「6. インデックスを表すテキスト」と同様に、リンク元ページの情報を用いることにより、対応可能であると考えている。

また、「7. URL」に対しては、リンク先のページのURLとアンカーテキストを比較するという単純なルールによって分類可能である。

残りの3種類については、上記の5つの分類で適切に分類できなかったものに対して判別を行う事とする。

「1. リンク先の内容を表すテキスト」と他の2つのテキスト(「3. リンク先との関係を表すテキスト」、「8. その他」)の最も大きな違いは、「1. リンク先の内容を表すテキスト」は必ず、リンク先のテキストの内容と対応付けが可能であるという点にある。よって、アンカーテキストとリンク先のページのタイトル、内容と比較することによって、判別可能であると考える。

最後に、「3. リンク先との関係を表すテキスト」、「8. その他」については、明確な判別規則を作ることが困難である。どちらかの分類について事例を集めて、判別する方法も考えられるが、完全な分類にはならないことが想定される。

2.4 自動分類のための事例収集

次に、実際のアンカーテキストの情報から、「4. トップページを指示するテキスト」、「5. ナビゲーションを指示するテキスト」、「6. インデックスを表すテキスト」、「2. ページの機能を表すテキスト」に関して自動分類のためのアンカーテキストの事例収集を行った。

これらのアンカーテキストの事例については、主にサイト内リンクに多いことが考えられることから、サイト内リンクのアンカーテキストの頻度情報の上位のものから抽出する。ただし、単純なる頻度だと、特定のサイトでテンプレート的に用いられている項目が上位に多く列挙される危険性があること、複数のサイトで共通する事例の方が、広く利用可能な事例であると考えられることから、サイト内リンクにおける各アンカーテキストの異なりサイト出現数を利用する。また、アルファベットの日本語コードと英語コードによる表記、大文字・小文字、文字の間の空白の違いなどは、今回の事例収集にあたり、本質的な違いではないと考えられるので、全て正規化した上で、出現回数を数えた。

このうち、1000以上のサイトで出現している301のテキストについて分類を行ったところ、下記のように分類できた。

1. **リンク先の内容を表すテキスト：** 0
2. **ページの機能を表すテキスト：** 115個(「会社概要」、「リンク」、「リンク集」、「会社案内」、「お問い合わせ」など)
3. **リンク先との関係を表すテキスト：** 9個(「english」、「japanese」、「詳細」など)
4. **トップページを指示するテキスト：** 29個(「home」、「top」、「トップページ」、「ホーム」、「トップページへ」など)
5. **ナビゲーションを指示するテキスト：** 28個(「こちら」、「戻る」、「back」、「next」など)
6. **インデックスを表すテキスト：** 114個(「1」、「2」、「1月」、「2月」、「北海道」、「東京」など)
7. **URL：** 0個

8. その他：6個(イメージのみ、altタグの中身が空白など)

上記の分類をする際に問題になったこととしては、「掲示板に戻る」や「トップページに戻る」といった別のカテゴリーに属するテキストとナビゲーションを指示するテキストの組み合わせというものが存在した。これらのテキストについては、現在、「に戻る」などのナビゲーションを指示するテキストを考慮しない形式で分類を行うこととした。

かなりの数の典型事例を収集できたとも考えられるが、100以上のサイトで出現しているテキストに広げるだけでも、5349個あり、サイト内でのリンクの上下関係などの情報を利用して、事例の収集・整理の支援を行うシステムについて検討をする必要があると考えている。

3 アンカーテキストの機能分類に応じたWeb情報システムの提案

本章では、各アンカーテキストの機能分類に応じた利用法を考える。

3.1 アンカーテキスト中の個物の名前に注目した利用法

「1. リンク先の内容を表すテキスト」は検索において最も役に立つと考えられるテキストである。例えば、藤井ら[7]のNTCIR-5における誘導型検索では、このようなテキストを主に利用して、情報検索システムの検索性能の向上に役立てていると考えることができる。

また、一つのリンク先を一つの特定の個物を表すデータとして考えると、同一リンク先に向けてつけられた異なる「1. リンク先の内容を表すテキスト」は、一つの個物を複数の形式で表現した同実体異表記のデータ群として捉えることができる。

このような異表記を網羅的に頻度つきで集めることにより、藤井ら[7]の研究において、行われている日英の翻字だけではなく、省略や翻訳などのパターンを集める事が可能になると考え、主に省略表記やよくある書き間違いによる異表記の収集を行うシステムを作成した。

本システムは、特定のURLに対し、設定されている頻度つきのアンカーテキストのデータベースを一つの文書と見なし、これに対し、専門用語抽出を行うことにより、特徴的な異表記を取り出す。具体的には、「1. リンク先の内容を表すテキスト」がサイト外からのリンクに多く含まれる事を利用し、サイト外からのリンクのアンカーテキストから、「7. URL」に対応するテキストを除外したものに対し、専門用語自動抽出システム「言選」[3]を用いて、専門用語抽出を行う。

このシステムの動作を確認するために、北海道大学(http://www.hokudai.ac.jp)、国立情報学研究所(http://www.nii.ac.jp/index-j.html)、Yahoo!(http://www.yahoo.co.jp/) の3つのURLに対して、動作させた際に得られた各々の上位10件を表4に示す。

北海道大学の例には、「北大」という略称や英訳名称の一部である「Hokkaido」、「University」などが抽出されている。また、「北海道大学医療技術短期大学部」のように、「北海道大学」の下部組織の名前が見られる。

このようなアンカーテキストは、個物を参照する際に、適切なリンク先を見つけられなかった場合に見られる。次の国立情報学研究所の例には、所属教官の名前を持つアンカーテキストのリンク先として利用されている例があった。「1. リンク先の内容を表すテキスト」を個物に対する異なる表記として利用する必要がある場合には、この辺りの分類を明確に行う必要があると考えられる。

また、国立情報学研究所の例からは、略称の「NII」、旧組織名である「学術情報センター」、旧組織名の略称であるNACSISを含む「国立情報学研究所NACSIS」などが抽出されている。また、間違いやすい表記として「国立情報学術研究所」などが見つけられている。Yahoo!の例では、特定の個物を指すアンカーテキストだけではなく、機能である「検索エンジン」という参照のされ方がなされているため、個物の名前としては適切でない情報を取り出してきている。

このような問題に対しては、アンカーテキストの分類をサイト外・サイト内と分けるだけではなく、今回提案した8つの分類に分けることに

[3] http://gensen.dl.itc.u-tokyo.ac.jp/

より、個物の異表記としてより意味のあるデータを獲得できるようになると考えられる。

3.2 アンカーテキストによる検索結果の絞りこみ

「2. ページの機能を表すテキスト」や「3. リンク先との関係を表すテキスト」はそれ自体のキーワードとしての重みを強くとるとおかしくなるが、サイト全体を表すようなキーワードと組み合わせる場合には、絞り混みに有用なキーワードとなると考えられる。例えば、「北海道大学の入学試験情報を知りたい」という検索要求に対し、「入学試験情報」というアンカーテキストが多くの大学で用いられており、「2. ページの機能を表すテキスト」として認識可能であり、「北海道大学」が「1. リンク先の内容を表すテキスト」として利用されていることが認識可能であったとする。

この場合には、サイト全体を表すキーワードである「北海道大学」によるサイトの絞り込みを行った上で、「入学試験情報」を含むページを探す事により、より適切なページを得られることが期待できる。また、「2. ページの機能を表すテキスト」に関するシソーラスのようなものを作ることにより、キーワードに直接マッチしていない「入試情報」のページを見つけ出すことも可能となる。一般に、このようなシソーラスによる検索語拡張は、精度を下げる危険性があるが、上記のようなサイトを絞りこんだ後の「2. ページの機能を表すテキスト」に関するシソーラスを利用する場合には、有用な情報が存在する確率が高い情報源からのサーチとなるため、精度の低下を起こさずにすむ可能性がある。

このような「2. ページの機能を表すテキスト」に関するシソーラスの作り方については、鈴木ら [4] の Web ディレクトリの構築支援の研究などが参考になると考えられる。

3.3 サイトの単位認定のための利用法

「4. トップページを指示するテキスト」、「5. ナビゲーションを指示するテキスト」については、テキストそれ自体を、有効に利用する方法は見つけられないが、これらのテキストを使うことにより、複数サイトの連携によるひとまとまりのサイトや、一つのサイト内における小さな固まり (geocities や rakuten などの個別のユーザ、店のページ群) を認識する際に利用可能であると考えられる。

「6. インデックスを表すテキスト」についても、基本的にはあまり有用な利用法はないが、「北海道」「東京」といったある種の分類を表しているインデックス群のみに注目し、個別のページを分析すると、特定の概念分類に属するインスタンスの集合のリストが得られる可能性がある。

3.4 その他

最後に、「7. URL」、「8. その他」については、リンク先のページ、リンク元のページに関連する追加情報がアンカーテキスト中に全く存在しないため、有効な利用方法はないと考えられる。しかし、上記のようなアンカーテキストが存在することを前提として、他のアンカーテキストと区別することにより、各々の分類結果をより有意義に利用可能になると考えられる。

4 結言

本研究では、大規模 Web テキストのコレクションである NTCIR Web テストコレクション (nw100g) に含まれるアンカーテキストの具体例を分析することにより、アンカーテキストに関する 8 つの分類を提案した。さらに、その分類を行うための方針を提示すると共に、この分類がどのように Web 情報システムに利用可能であるかについて、試作システムの例などを示しながら説明した。

今後は、アンカーテキストの機能分類の自動化のための枠組を提案するとともに、その機能分類の性質をいかした Web 情報システムの研究を進めたいと考えている。

謝辞

NTCIR コレクションは国立情報学研究所の許諾を得て使用した。

表 4: アンカーテキストからの専門用語抽出の結果

リンク先ページの URL とタイトル	抽出した専門用語
http://www.hokudai.ac.jp 北海道大学ホームページ	北海道大学、ホームページ、北海道大学ホームページ、 Hokkaido、北大、University, 北大ホームページ、 IMG、北海道大学大学院、北海道大学医療技術短期大学部
http://www.nii.ac.jp/index-j.html NII -The National Institute of Informatics-	国立情報学研究所、情報学研究所、国立情報研究所、 国立情報科学研究所、国立情報学術研究所、 学術情報センター、国立情報情報学研究所、NII、 文部省国立情報学研究所、国立情報学研究所 NACSIS
http://www.yahoo.co.jp/ Yahoo! JAPAN	JAPAN、Yahoo、ヤフー、JAPAN トップ、検索、 Japan、YAHOO、ヤフージャパン、yahoo、検索エンジン

参考文献

[1] Sergey Brin and Lawrence Page. The anatomy of a large-scale hypertextual Web search engine. *Computer Networks and ISDN Systems*, Vol. 30, No. 1–7, pp. 107–117, 1998.

[2] Oliver A. McBryan. GENVL and WWWW: Tools for taming the web. In *Proceedings of the First International Conference on the World Wide Web*, 1994.

[3] Dou Shen. A comparison of implicit and explicit links for web page classification. In *Proceedings of the 15th International Conference on the World Wide Web*, 2006.

[4] 鈴木祐介, 松原茂樹, 吉川正俊. アンカーテキストを用いた web ディレクトリの構築. 情報処理学会自然言語処理研究会 ,2005-NL-168, pp. 75–80, 2005.

[5] Nadav Eiron and Kevin S. McCurley. Analysis of anchor text for web search. In *SIGIR '03: Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, pp. 459–460, New York, NY, USA, 2003. ACM Press.

[6] 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. 情報処理学会情報学基礎処理研究会 ,1999-FI-55, pp. 73–80, 1999.

[7] Atsushi Fujii, Katunobu Itou, Tomoyosi Akiba, and Tetsuya Ishikawa. Exploiting anchor text for the navigational web retrieval at NTCIR-5. In *Proceedings of the Fifth NTCIR Workshop Meeting on Evaluation of Information Access Technologies*, pp. 455–462, 2005.

[8] Koji Eguchi, Keizo Oyama, Emi Ishida, Noriko Kando, and Kazuko Kuriyama. An evaluation of the web retrieval task at the third ntcir workshop. *SIGIR Forum*, Vol. 38, No. 1, pp. 39–45, 2004.