アンカーテキストとリンク構造解析を用いた Web 情報検索の改善

阿部 匡史† 豊田 正史† 喜連川 優†

†東京大学生産技術研究所 〒153-8505 東京都目黒区駒場 4-6-1

E-mail: †{abe,toyoda,kitsure}@tkl.iis.u-tokyo.ac.jp

概要 Web ページに含まれるアンカーテキストはリンク先の文書を短く簡潔に要約しているものが多い。このため情報検索において有益な情報源となり得る。そこで本研究では、アンカーテキストを用いてコンテンツベースの情報検索の改善を試みた。さらにリンク構造解析と組み合わせることにより、リンクとテキストの両面を考慮した情報検索手法を提案する。評価には第３回 NTCIR ワークショップのデータセットを用いて定量的な評価を行った。またある特定のページに関連したページを抽出する問題に対して、リンク構造解析を用いた手法にテキスト解析を組み合わせ、その改善を試みる。

キーワード Web とインターネット，情報検索，アンカーテキスト，リンク構造解析

Improving Contents-based Web Information Retrieval Using Anchor Texts and Link Analysis

Tadafumi ABE† Masashi TOYODA† and Masaru KITSUREGAWA†

†Institute of Industrial Science, The University of Tokyo 4-6-1 Komaba, Meguro-ku, Tokyo, 153-8505 Japan

E-mail: †{abe,toyoda,kitsure}@tkl.iis.u-tokyo.ac.jp

Abstract Anchor texts that point to a web page tends to simply summarize that page. For this reason, anchor texts may be useful for web information retrieval. In this paper, we propose techniques for improving contents-based web information retrieval using anchor texts and hyperlinks. We first examine the effect of anchor texts, and then integrate link analysis. We evaluate our techniques using the evaluation result of the web retrieval task in NTCIR workshop 3. We also try to improve a related page algorithm that calculates pages related to a given page using link analysis and text analysis.

Keyword Web and Internet，Information Retrieval，Anchor text，Link Analysis

1. はじめに

Web 上には膨大な量の情報が溢れており、この中から必要な情報を探し出すのは非常に困難である。この手助けのために検索システムが存在するが、その精度は十分とは言い難く、さらなる改善が望まれる。一般的な検索方法はユーザからキーワードを受け取り、その語句を含むページを探し出す手法である。しかしテキスト情報のみを用いた手法には限界があり、文書構造やリンク構造解析を効果的に統合することが昨今の課題となっている。有名な検索システム Google[14]においても PageRank[1]を用いた検索を提供している。また異なる検索方法として、ある特定のページに関連したページを検索するという方法がある。これは基本的にリンク構造解析を用いて行われるが、トピックとは異なるページが抽出されてしまうことが多々ある。

そこで本研究ではテキスト解析とリンク構造解析を組み合わせた検索を行う。キーワードによる検索において､既存のコンテンツベース検索システムを基に、その結果の順位の改善を試みる。テキスト解析ではアンカーテキストに注目し、その有効性を示す。さらに HITS アルゴリズム[2]とアンカーテキストの解析を組み合わせた手法を提案する。評価には NTCIR3 で使用されたデータセットを用いて定量的な評価を行う。またあるページの関連ページを抽出する問題に対して本手法を適用し、その効果とアンカーテキストによる問題点を明らかにする。またその改善のため、ヘッディングを用いた関連ページの抽出を試みる。

以下、第 2 章にて関連研究について述べる。第 3 章においてアンカーテキストを用いた Web 情報検索、さらにリンク構造解析との統合についてその実験結果と共に考察する。また関連ページ抽出アルゴリズムとテキスト解析との統合手法について第４章で述べる。第５章にて全体を通した議論を行い、最後に第６章にてまとめる。

2. 関連研究

2.1. アンカーテキスト

アンカーテキストとはHTMLページにおいて、リンクを張る際の<A>タグに挟まれたテキストのことである。これはそのページの作者が人手により作成することが多いが、そのページからリンク先へナビゲートするという機能の性質から、リンク先のページの内容を簡単に表すような文字列を付与する。このためアンカーテキストはそのリンク先のページを簡潔に要約した文であると言え、Webの情報検索において有益な情報源となり得る。このような考え方はMcBryan[7]によって最初に提案され、以後アンカーテキストを用いた情報検索について研究が行われた。この中で第3回NTCIR Workshop[13]のWeb検索タスクにおいてNECより提案された手法[9]についてここで説明する。

この手法はアンカーテキストをそのリンク先のページのコンテンツとみなしてクエリ内の語の検索を行い、クエリ内の語がいくつ含まれているかに応じてランク付けする。またアンカーテキストを重視して、アンカーテキスト内にクエリの語句が含まれていたら、そのページを上位に位置させる。しかし既存のコンテンツベース検索システム（Okapi[10]）に比べ、精度において同程度もしくは劣るという結果が出ている。

だが、この手法ではテキストの長さなどを考慮しておらず、単語が含まれるか否かのみで主なランク付けを行っている。本研究ではアンカーテキストの長さなどを考慮した上でスコアリングを行う。また、アンカーテキストは確かに有益な情報源となり得るが、短い文章であるため、それのみでは情報として不足と考えられる。このためアンカーテキスト内にクエリが含まれれば無条件に上位に置くというこの手法は問題があるように思う。そこで本研究ではコンテンツベースの検索システムを基として、その改善にアンカーテキストを用いるという立場を取る。

2.2. リンク構造解析

Web上のハイパーリンクでつながれたページ群は、そのページをノードとし、リンクをエッジとした有向グラフと考えられる。そしてその構造を解析することで関連したページ群を見つける研究がある[2,3]。ここでは豊田により提案されたCompanion－[11]について紹介する。これはHITSを基に改良を行ったアルゴリズムであり、入力されたURLに対してその関連ページを求めるものである。

まず入力URLに対応するシードページからリンクで結ばれた近隣のページを集め、近傍グラフを作る。近傍グラフにはHITSと異なり、シードページへリンクを持つページとそれらがリンクするページのみを含める。後者のリンクではその出現順序を考慮し、シードページへのリンクの前後R個のリンクのみを含める。また同一の内容を持つミラーページは代表として1つのページを定め、そのページにまとめる。

次にページのサーバ名によってリンクに対する重み hub_wt(p,q), auth_wt(p,q)を計算する。これはそれぞれページpからqへのリンクのハブの重み・オーソリティの重みを表す。同一サーバ内のリンクに対してはこの両方の重みを0にする。あるページpが外部の同一サーバにn個のリンクを持っていれば、そのリンクに当たるhub_wt(p,q)をすべて1/nとする。また同じ外部サーバからn個のリンクがあるページpがあれば、そのauth_wt(q,p)を1/nとする。

ハブ・オーソリティスコアhub(p)・auth(p)を1に初期化し､以下の式を各スコアが収束するまで繰り返す。

$$hub(p) \leftarrow \sum auth(q) \cdot hub\_wt(p,q) \qquad (1)$$

$$auth(p) \leftarrow \sum hub(q) \cdot auth\_wt(q,p) \qquad (2)$$

hub(p)・auth(p)それぞれについて、
その二乗和が1となるように正規化

このようにして得られたauth(p)の高いものからシードページへの関連ページとして出力する。

2.3. テキスト解析とリンク構造解析の統合

リンク構造解析はリンクの多い有用なページを見つけ出すことはできるが、コンテンツ自体を考慮していないために求めるトピックとは異なったページに焦点が移ってしまう可能性がある。そこで、リンク構造解析とテキスト解析を組み合わせることにより、ユーザからのクエリにより関連したページを見つけ出そうとした研究がある[4,5,6]。

Bharatらにより提案された手法[4]はクエリとページのコンテンツからその類似度を計算し、その類似度を用いて近傍グラフの枝刈りやハブ・オーソリティスコアを求める際の重み付けを行った。しかし、この手法ではページ内のテキスト情報すべてを用いるために大量のメモリ・計算時間を要し、またスパムページなどの影響も強く受けると考えられる。そこで本研究ではページ内の語ではなく、そのページへのリンクに対応するアンカーテキストに注目した。これはすでにページ内の解析を行っているコンテンツベース検索システムの改善という目的にも適していると考えられる。

また ARC[5]はアンカーテキストにクエリの語が含まれているかを用いてリンクに重み付けを行った。しかしリンクに対応するアンカーテキストは1つであるため十分な重み付けは行えない。本研究ではページ単位でアンカーテキストを集め、スコアの計算を行う。

3. アンカーテキストとリンク構造解析を用いたWeb情報検索の改善

本研究では既存のコンテンツベース検索システムから検索結果として順位と共にスコアを受け取り、アンカーテキストを用いて検索結果の順位の再ランクを行う。またリンク構造解析とアンカーテキストを統合することにより、さらなる改善を試みる。

3.1. 手法

3.1.1. アンカーテキストを用いた順位付け

まず、あるページpへのリンクに対応するアンカーテキストをWeb上からすべて集める。そのテキスト群を形態素解析して単語に分解し、その出現頻度を要素とした単語ベクトル $\vec{w}(p)$ を定義する。ここで、形態素解析にはJUMAN[15]を使用し、その解析結果において「名詞」又は「未定義語」と判定された単語のみを用いた。またアンカーテキストの数・長さによるばらつきを押さえるため、単語ベクトルはその大きさが1となるように正規化する。

$$\left|\vec{w}(p)\right| = 1 \qquad (3)$$

この単語ベクトル $\vec{w}(p)$ とクエリより作られた単語ベクトル $\vec{q}$ との内積をそのページに対するアンカースコアとする。

$$anchor(p) = \vec{w}(p) \cdot \vec{q} \qquad (4)$$

このスコアを既存のコンテンツベース検索システムから得られたスコア contents(p)とパラメータ α と共に次のように計算し、新たなスコア $S_{anc}(p)$ を求める。

$$S_{anc}(p) = \alpha \frac{anchor(p)}{\max_q(anchor(q))} + (1-\alpha)\frac{contents(p)}{\max_q(contents(q))} \qquad (5)$$

このスコアを基に既存の検索システムの検索結果を再ランク付けし、新たな検索結果とする。

3.1.2. リンク構造解析を用いた順位付け

リンク構造解析には2.2節で説明したCompanion－を用いる。シードページには既存の検索システムの検索結果上位Nページを用いる。またリンクを拡張する際には、シードページへの前後R個のリンクのみではなくすべてのリンクを含める。これは検索結果として得られるページは大概ディレクトリ構造において深部にあり、外部のサーバからのリンクが少なく、結合した近傍グラフの構築ができにくいためである。

ハブ・オーソリティスコアの計算の後、既存の検索システムの検索結果にあるページに対して、そのスコアと得られたオーソリティスコアとをパラメータ α をもって足し合わせて新しいスコア $S_{link}(p)$ を求め、このスコアに従って新たなランク付けを行う。

$$S_{link}(p) = \alpha \frac{auth(p)}{\max_q(auth(q))} + (1-\alpha)\frac{contents(p)}{\max_q(contents(q))} \qquad (6)$$

3.1.3. アンカーテキストとリンク構造解析の統合

リンク構造解析では外部から多くのリンクがあり、より良い情報が含まれるページに高いスコアが与えられるが、コンテンツを考慮していないためにその情報がトピックと一致しない可能性がある。そこでクエリとアンカーテキストから得られたスコアをリンク構造解析と組み合わせることでこの問題の解決を試みる。

リンク構造解析において auth(p)を求める繰り返し計算に 3.1.1.節において説明したアンカースコアanchor(p)を導入する。導入に関して２つの手法について考慮した。以下の式いずれかをオーソリティスコアを求める繰り返し計算(2)式の後に挿入する。

$$\text{(A)}\ auth(p) \leftarrow \frac{auth(p)}{\max_q(auth(q))} + \frac{anchor(p)}{\max_q(anchor(q))} \qquad (7)$$

$$\text{(B)}\ auth(p) \leftarrow auth(p) \cdot anchor(p) \qquad (8)$$

このようにすることによって、トピックに関連したアンカーテキストを持ち、かつ外部からのリンクの多い有用なページが大きなスコアを持つことになる。

以上のようにして得られたオーソリティスコアを既存のコンテンツベース検索システムのスコアとパラメータ α をもって同様に混ぜ合わせ新たなスコア $S_{int}(p)$ を計算し、再ランク付けする。

3.2. 実験環境

実験には第3回NTCIRワークショップ[13]のWeb検索タスクで用いられたデータを使用する。このサブタスクである「検索課題検索」は主にjpドメインから収集したWeb文書約100GBの中から、与えられたトピックについて検索を行う。トピックは一般の大学生へのアンケートから得られたもので、分野は多岐に渡る。実験にはこのタスクで用いられた47個のトピックを使用する。その一例を表1に示す。

正解判定は人手によって行われ、本タスクにおいて各参加者から提出された検索結果の上位100ページに対して判定が成されている。判定には高適合・適合・部分適合・不適合の4段階の評価があり、またその判定に際して「ページ内のコンテンツのみで判断」と「そのページから１ステップのリンク先まで判断基準に含める」の２通りがある。本実験ではリンク先までを判断に含めた高適合・適合ページを正解とした。また評価は平均精度によって行い、既存のコンテンツベース検索システムに対して提案手法による平均精度の増加の割合を精度向上率として定義する。

表１．トピックの例

<TOPIC>
<NUM>0010</NUM>
<TITLE CASE="b">オーロラ, 条件, 観測</TITLE>
<DESC>観測のために、オーロラの発生する条件が知りたい</DESC>
<NARR><BACK>オーロラを観測するために、発生に必要な条件や、発生のメカニズムが知りたい。</BACK><RELE>オーロラ観測記などは、場所と日時が表記されており、発生時の天候・温度等を追跡調査することが可能な物のみ適合とする。</RELE></NARR>
<CONC>オーロラ, 発生, 条件, 観測, メカニズム</CONC>
<RDOC>NW003201843, NW001129327, NW002699585</RDOC>
<USER>大学院修士 1 年, 女性, 検索歴 2.5 年</USER>
</TOPIC>

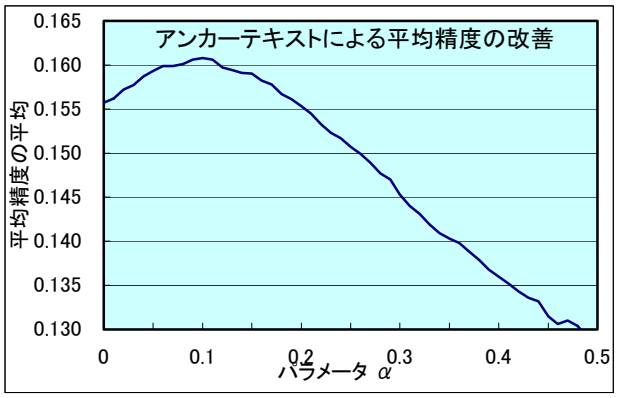

図１．パラメータによる変化

また、実験の基となるコンテンツベースの検索システムには本タスクにおいて最良の結果を得ている、RICOH によって提案された FTS[8]を用いる。これは n グラムモデルに基づいた検索システムであり、初期クエリを用いた検索によって得られたページからクエリの拡張を行った上でもう１度検索を行うことを特徴とする。この FTS を用いて、本タスクにおいて使用された文書集合約 100GB から検索を行った結果の上位 1000 ページを基とし、その改善を試みる。

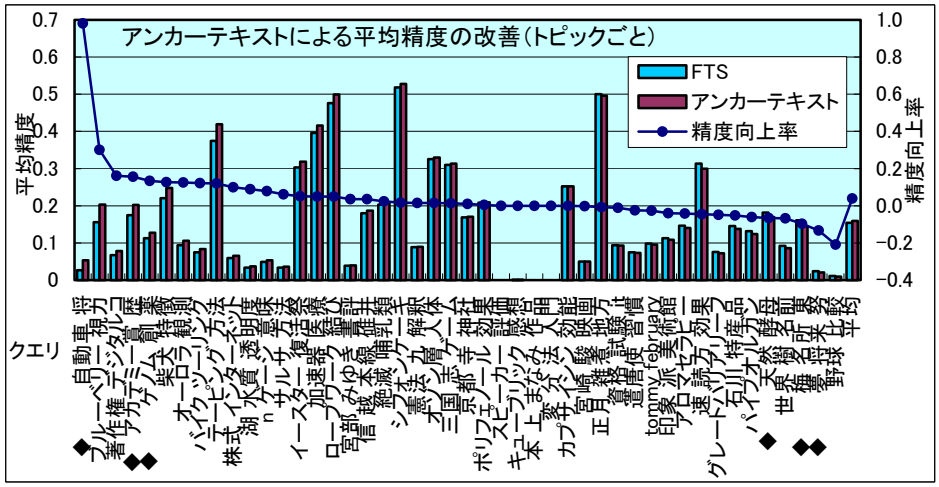

図２．アンカーテキストによる FTS の改善

ここで、アンカーテキストの抽出、リンク構造解析における近傍グラフの作成のためには本タスクにおいて使用された文書集合は十分な大きさではない。そこでこれらにはより大きな、東京大学喜連川研究室において収集した文書集合を用いた。これは主として jp ドメインから 2001 年 10 月初旬に収集しており、約 40M ページ・260GB である。

3.3. 結果・考察

3.3.1. アンカーテキストを用いた改善

クエリとして各トピックの<TITLE>タグ内の語句を形態素解析して用い、出現頻度をすべて１としたクエリベクトルを作る。アンカーテキストによる再ランクを行ったときのパラメータ α による平均精度の平均の変化を図１に示す。但し、トピック「亀」の「寿命」は正解ページが少なく、１つのページの影響が強く出てしまうために一般的な評価には相応しくないと考えられるので、このトピックを除いて評価を行った。

パラメータ α が 0.1 のとき４％程度の精度向上率が得られた。パラメータは異なる評価基準を用いてもほぼ同じ値となり、0.1 が最適であると考えられる。このことからアンカーテキストは情報検索において有効であるが、単独では情報として不十分であると言える。

次にトピック毎の FTS と本手法の平均精度を棒グラフで、また本手法の精度向上率を折れ線グラフで図２に示す。パラメータ α は最大の精度向上率が得られた 0.1 とした。各トピックにおいて、クエリ内の語が正解ページへのアンカーテキストに多く含まれており、コンテンツベース検索システムの改善を果たしている。いくつかのクエリにおいて、正解ページから得られたアンカーテキストの例を表２に示す。

トピック「日本」の「自動車」の「将来」像では、FTS の場合コンテンツ内に「自動車」「将来」という語は含むが、主に排気ガスやエネルギー資源など環境問題について書かれているページが上位にあり、精度を下げている。しかしアンカーテキストにまで「自動車」「将来」が含まれているページは実際に自動車の将来について述べられたページであることが多く、これらのスコアが上がることによって精度の改善が成された。

トピック「アカデミー賞」の「歴代」「受賞」者の場合、FTS では受賞者ではなく映画自体の紹介が多く、これらは不適合となる。しかしこれらのページへのア

表２．アンカーテキストの例

クエリ	アンカーテキスト
自動車 将来像 日本	自動車産業の将来と戦略 日本の自動車産業と自動車部品産業の将来
アカデミー賞 受賞 歴代	★歴代アカデミー賞リスト★ アカデミー賞受賞結果！
ゲノム 創薬 動向	開発の動向 ゲノム創薬 ゲノム創薬キーテクノロジー2001

表３．アンカーテキストとクエリ

評価	アンカーテキスト内にクエリの語がある確率
高適合	32.76%
適合	27.61%
部分適合	24.86%
不適合	11.86%

ンカーテキストは映画の題名になることが多く、アンカースコアは低い値となる。実際にアカデミー賞の受賞者の紹介をしているページには高いアンカースコアが割り当てられ、上位にランクされた。

またページへの正解判定の評価種ごとに、アンカーテキストにクエリ内の語が１語でも含まれている確率を求めると、表３のようになった。トピックへの適合ページほどアンカーテキスト内にクエリの語を多く含んでおり、アンカーテキストの有効性がわかる。

しかしながらトピックによっては平均精度が減少したものも存在する。この原因としては次のようなものが挙げられる。

●ページの一部分に目的の文章がある

トピック「東京」の「梅」の「名所」では、湯島天神や井の頭公園を紹介するページの一部に梅の記述があり、それが正解ページとなっている。しかしこの場合、アンカーテキストには「湯島」や「井の頭」が多く、「梅」などの語はあまり出現しない。一方でアンカーテキストに「梅」を含み東京の梅の名所でないページのランクが上昇し、精度を下げている。この場合、ページ全体の解析が必要となる。

●１つのキーワードに強く影響を受ける

トピック「天然」「酵母」を扱う「店」は店の場所を調べたいクエリだが、酵母の作り方を載せたページのアンカーテキストに「酵母」が多く存在して上位にランクされてしまった。対して「店」はアンカーテキスト中にあまり出現しない。IDF などを用いてクエリ内の語に重み付けすることも可能だが、一般には「店」の方が多く出現するため IDF による改善も見込めない。

●クエリ自体が曖昧

トピック「将来」の「夢」への「努力」では、「将来」「夢」が適合・不適合を問わずアンカーテキスト内に出現したため、適合ページのみランクを上げることができなかった。これはクエリが曖昧であり、目的とは異なるトピックのページにも出現し得る語句である。

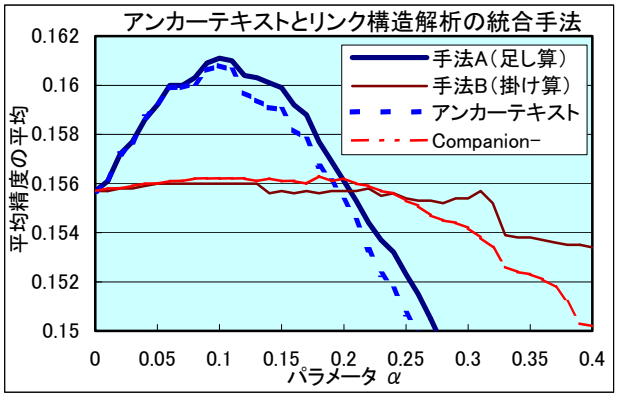

図３．アンカーテキストとリンク構造解析の統合手法

3.3.2. アンカーテキストとリンク構造解析の統合

続いてアンカーテキストとリンク構造解析の統合を試みる。3.1.3 節において述べた手法(A)・(B)、またアンカーテキストのみ、リンク構造解析のみを用いたときのパラメータαによる平均精度の平均の変化を図３に示す。クエリとして<TITLE>タグ内の語を用い、また Companion－のシードには FTS の結果上位 1000 ページをすべて使用した。

リンク構造解析単独ではスコアに変化の有るページは全体の５％以下であった。これは先にも述べたように、検索結果のページがディレクトリ構造の深部に多く外部からのリンクが少ないためである。また HITS アルゴリズムの特徴として数ページの有力なページに高いスコアが割り当てられ、その他のページのスコアは０に近い値となる。このため高スコアを与えられたページによるある程度の改善は得られたが、アンカーテキスト単独ほどの精度向上率は得られなかった。

２つのスコアを掛け合わせた場合（手法 B）はさらに０となるスコアが多くなってしまい、精度の改善は成されなかった。

２つのスコアを足し合わせたとき（手法 A）はアンカーテキスト単独のときの精度を基準として若干の改善が得られた。このとき、トピック毎に平均精度の精度向上率を、アンカーテキストのみ、リンク構造解析のみと併せて図４に折れ線グラフで示す。また FTS の平均精度を棒グラフで示しておく。パラメータはすべて最大の精度向上率が得られた 0.1 を使用した。

トピック「石川」の「特産品」ではアンカーテキストのみの場合、アンカーテキスト内に２語とも含むページは少なく、どちらか片方の語を含む特産品以外の石川情報や石川以外の特産品情報を載せたページが上位に位置し、精度が減少してしまった。しかしそのようなページよりも石川の特産品情報を載せたページ群が密なリンク構造を持っており、高いオーソリティスコアを得ることでアンカーテキストのみのときに比べ精

度の改善が見られた。トピック「バイク」の「ツーリング」「レポート」では、その分野のオーソリティであるが、アンカーテキストからは高いスコアを得られなかったページが上位にランクされた。

一部、組み合わせることによって精度が減少しているトピックがある。トピック「自動車」の場合 Companion－アルゴリズムの結果一つだけが非常に高いオーソリティスコアを持ち、その影響で他のページのスコアが半減してしまったためである。このような特殊な場合を除いた他のクエリにおいては、アンカーテキストとリンク構造解析を統合することによってアンカーテキスト単独の場合に比べて同程度もしくは高い精度が得られた。

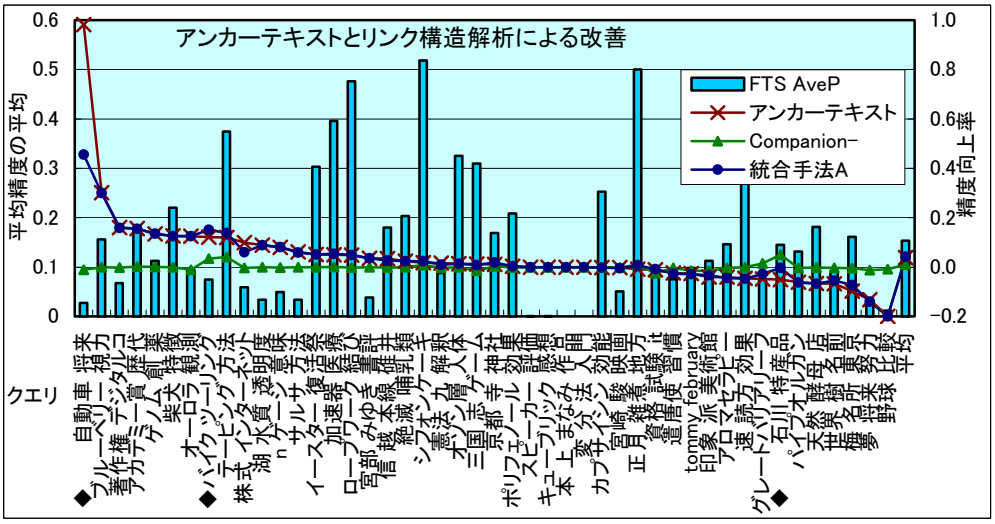

図４．アンカーテキストとリンク構造解析の統合

4. 関連ページ抽出アルゴリズム

一般的な検索はユーザからクエリを受け取り、これに関連したページを検索する。これに対し、ある URL を入力として与え、そのページに関連のあるページを探し出すという検索方法も考えられる。このような検索方法は Google においても関連ページ検索として実装されている。この検索には Companion－のようなリンク構造解析を用いた手法が一般的だが、コンテンツを考慮していないために異なるトピックのページまで抽出してしまうことがある。たとえば "www.jreast.co.jp/"（JR 東日本）をシードページとすると、その関連ページとして JR 西日本など JR 各社の他に ANA などの航空会社を出力してしまう。そこでリンク構造解析とアンカーテキストやヘッディングなどのテキスト情報を組み合わせた手法を用いて、この問題の解決を試みる。

4.1. 手法

4.1.1. アンカーテキストを用いた関連ページ抽出

基礎となるリンク構造解析には Companion－を用いる。入力 URL をシードページとし、2.2.節で述べたように近傍グラフの作成、リンクの重み付けを行う。さらにリンクの出現頻度も考慮し、シードページに対するリンクの前後 10 個以内に出現するリンクのみを拡張の対象とした。ついでハブ・オーソリティスコアを 3.1.3.節で説明したようにアンカーテキスト解析と組み合わせて計算する。しかし関連ページ抽出の場合クエリが存在しないので、シードページに対するアンカーテキストの単語ベクトルをクエリの代わりに用いる。

$$anchor(p) = \vec{w}(p) \cdot \vec{w}_{seed} \qquad (9)$$

このようにすることにより、シードページに関連した内容を持つページのスコアが高くなると考えられる。またアンカーテキストの導入手法には(7),(8)式で示される(A),(B)の二つの手法について実験を行った。

4.1.2. ヘッディングを用いた関連ページ抽出

アンカーテキストはリンク先のページ内容を簡潔に表したものであり、ページのタイトルのような語句が多い。これは関連ページの抽出には適さない可能性がある。たとえば"www.sony.co.jp/"（SONY）をシードとしたとき、その関連ページとして東芝などその他の電気メーカーを抽出したい。しかしアンカーテキストには「ソニー」という語は多いが、上位の概念となる「電気」や「メーカー」などの語句はあまり出現しない。そこでこのようなより上位の概念の語句を含むリンク元のヘッディングを関連ページの抽出に利用する。

他のページへのリンクを持つハブページについて考える。ヘッディングはそのヘッディング以下のいくつかのリンクをまとめた文章である。ここでヘッディングには<H1>,<H2>タグなどで囲まれた文字列を用いる。このタグには重要度によってレベル付けが成されているため、これを利用して一つのヘッディングは、そのヘッディングから、同じまたは高いレベルのヘッディングまでの間にあるリンクに対応するページの関連語と定義する。同じまたは高レベルのヘッディングが無ければページの最後までとした。このようにしてページｐに関連のあるヘッディングをそのページにリンクしているページ群から集めて単語ベクトル $\vec{w}_{heading}(p)$ を作り、これを用いてヘッディングによる

スコア $heading(p)$ を計算する。

$$heading(p) = \vec{w}^{seed}_{heading} \cdot \vec{w}_{heading}(p) \qquad (10)$$

ここで、ヘッディングにはアンカーテキストにはあまり見られない「リンク集」や「ページ」といった語句が特に多い。このような頻出語の影響を抑えるために、人手により選んだ 70 個のストップワードを用いる。また各ベクトルの要素には TF・IDF の値を用いた。

このようにして得られた $heading(p)$ をそれぞれ式(7),(8)の $anchor(p)$ の代わりに用いて、ヘッディング情報を用いた関連ページの抽出を行う。

4.2. 結果･考察

4.2.1. アンカーテキストを用いた関連ページ抽出

シードページとして"www.baystars.co.jp/"（横浜ベイスターズ)を入力したときの各手法の結果上位 10 ページを表 4 に示す。Companion－ではプロ野球の各球団が関連ページとして出力され、良い結果を得ている。

しかし、手法(A)ではディレクトリの深部にあるベイスターズ関連のページが数多く上位に来てしまった。これらのページは外部からのリンクは少ないが、アンカーテキストがシードページと極度に類似しているために高いスコアを得ている。アンカーテキストのみが強く影響し、シードページに特化し過ぎてしまう。

手法(B)の場合リンク構造解析によるスコアとアンカースコアが共に高いものが上位にランクすることになるが、本手法では主として横浜に関連のあるページが出力された。これはシードページへのアンカーテキスト中に「横浜」という語句が大量に出現したためである。また実際に横浜関連をまとめたハブページも存在するためリンク構造解析のスコアもある程度割り当てられ、これらのページが上位に来てしまった。アンカーテキストは主にそのリンク先のタイトルのような情報が多いため、「野球」「球団」といった上位の概念に位置する語句は出現しにくい。

4.2.2. ヘッディングを用いた関連ページ抽出

前節で示したようにアンカーテキストは関連ページの抽出には適していない。そこでより上位の概念の語句が出現すると思われるヘッディングを関連ページの抽出に利用する。シードページ"www.baystars.co.jp/"に対する結果を表 4 の最右に示す。ここでリンク構造解析とテキスト解析の統合手法は(8)式の手法（B）を用いた。ヘッディングを用いることによって「野球」や「球団」といった関連ページの抽出に適切な語句が得られ、他のプロ野球球団が出力された。

4.2.3. 評価

他のシードページに対してもヘッディングを用いた関連ページ抽出手法を適用し、その評価を行う。ある程度知名度のあるサイトのトップページを人手によって 41 個選び、Google の関連ページ検索、Companion－、本手法を適用する。それぞれ得られた上位 20 個のページを関連ページとして相応しいか人手によって判断した。実験に用いたシードページと関連ページの判断基準について、いくつかの例を表 5 に挙げる。次に Järvelin によって提案された DCG[12]を用いて、手法の精度をスコア付けする。これは適合度の重み付き累積であり、上位にランクされたページ程強い影響を及ぼす。順位 i のページの適合度を z_i としたとき DCG は

$$DCG = z_1 + \sum_{i=2}^{20} \frac{z_i}{\log_2 i} \qquad (11)$$

と計算される。ここで適合度には最も適合しているも

表 5．シードページとその関連ページの評価基準

シードページ	評価基準
www.adidas.co.jp/	スポーツ用品を扱う企業
www.nttdocomo.co.jp/	携帯電話会社
www.vector.co.jp/	ダウンロードサイト
tenki.jp/	天気予報、気象情報

表 4．関連ページの抽出結果（シードページ “www.baystars.co.jp/”）

Companion－	＋アンカーテキスト(A)	＋アンカーテキスト(B)	＋ヘッディング(B)
www.baystars.co.jp/	www.baystars.co.jp/	www.baystars.co.jp/	www.baystars.co.jp/
www.hanshin.co.jp/tigers	ifcnet.ne.jp/baystars	www.city.yokohama.jp/	www.dragons.co.jp/
www.dragons.co.jp/	www.lycos.co.jp/cgi-bin/pursuit?	www.marinos.co.jp/	www.fighters.co.jp/
giants.yomiuri.co.jp/	dir.lycos.co.jp/inc2/partner/amazon /dirscrd.html?	yokohamafc.com/	www.buffaloes.co.jp/
www.buffaloes.co.jp/	excite.jp.netscape.com/sports/s_baseball /professional_baseball/central_league/...	home.att.ne.jp/gold/3796-pt	www.hanshin.co.jp/tigers
www.marines.co.jp/	jp.excite.com/sports/s_baseball/profession al_baseball/central_league/yokohama_...	www.so-net.ne.jp/f-marinos /f.htm	www.marines.co.jp/
www.fighters.co.jp/	jp.excite.com/sports/sports_according_to _item/ball_game/baseball/professional_...	messages2.yahoo.co.jp/bbs?	www.carp.co.jp/
www.carp.co.jp/	www.246.ne.jp/~kazoo/yb/baystars.html	www.baystars.net/	www.yakult.co.jp/swallows
www.seibu-group.co.jp/Lions/	ifc.cplaza.ne.jp/baystars/	www4.airnet.ne.jp/nabe /baystars/baystars.htm	www.seibu-group.co.jp/Lions/
www.yakult.co.jp/swallows	dir.lycos.co.jp/hobby_sports/baseball /professional/team/bay_stars/	www.hanshin.co.jp/tigers	www.hawkstown.com/

表６．関連ページ抽出手法の精度

Google	Companion－	提案手法
61.46%	78.58%	82.78%

のから順に 2,1,0 の３段階を用いて評価した。

結果として 41 個のシードページ中、Google の関連ページ抽出に対して 35 個、Companion－に対しては 25 個のページにおいて本手法は優れた評価を得た。またこの３手法から得られたページ群から理想的な出力結果を作成して DCG を計算し、その値に対する各手法の DCG の割合の平均を求めると表６のようになった。Google の関連ページ検索に対して 34.7%、Companion－に対して 5.3%の改善が得られ、ヘッディングを用いた本手法の有効性がわかる。

5. 議論

キーワードによるコンテンツベースの検索では得られたページ内にクエリの語は含むが、目的のトピックではないことがある。しかしアンカーテキストにまでクエリの語が含まれているページは目的のトピックの文書であることが多く、このためアンカーテキストを用いた再ランクによって検索の精度に改善が見られた。またリンク構造解析と組み合わせることでトピックのオーソリティであるがアンカーテキストにクエリの語を含まないページのスコアが上がり、お互いを補完する形でさらに若干の精度の向上が見られた。しかしクエリによる検索の場合、目的とするページはトピックについての詳細な内容を持つページであり、これは大概ディレクトリ構造の深部にある。このため外部からのリンクは少なく、十分なオーソリティスコアを得るページは多くない。よってリンク構造解析単独ではあまり効果は得られなかった。

一方、あるページの関連ページを検索する場合には詳細な内容のページではなく、同じ分野に属するサイトのトップページを求めることが多い。このため外部からのリンクも多く、リンク構造解析は十分に機能する。しかしアンカーテキストはリンク先のページへ特化し過ぎているために関連ページの抽出には適していない。その改善のためには分野を表すような上位の概念の語句を必要とし、リンク元のページのヘッディングを用いることによってページの関連を表す語句が適切に得られ、関連ページ抽出の改善が果たせた。

6. おわりに

本研究ではアンカーテキストを用いて既存のコンテンツベース検索システムの改善を行った。NTCIR3 のデータセットを用いた評価によって４％程度の精度向上率が得られ、情報検索におけるアンカーテキストの有効性を明らかにした。さらにリンク構造解析とアンカーテキストを組み合わせる手法を提案し、５％程度の精度向上率が得られた。またあるページに関連したページを抽出するという検索方法に対して本手法を適用し、その効果とアンカーテキストを利用することによる問題点を示した。さらにその改善のためにヘッディングを用いた関連ページの抽出手法を提案し、既存の検索システムに比べて 35%、リンク構造解析のみを用いた手法に対し５％の改善が得られた。

文　献

[1] Sergey Brin and Lawrence Page, "The anatomy of a large-scale hypertextual web search engine", In Proceedings of the 7th International World Wide Web Conference, 1998

[2] Jon M. Kleinberg, "Authoritative Sources in a Hyperlinked Environment", In Proceedings of the 9th Annual ACM-SIAM Symposium on Discrete Algorithms, 1998

[3] Jeffrey Dean and Monika R. Henzinger, "Finding related pages in the World Wide Web", In Proceedings of the 8th International World Wide Web Conference, 1999

[4] Krishna Bharat and Monika R. Henzinger, "Improved Algorithms for Topic Distillation in a Hyperlinked Environment", In Proceedings of the 21st International ACM SIGIR Conference on Research and Development in Information Retrieval, 1998

[5] S. Chakrabarti, B. Dom, P. Raghavan, S. Rajagopalan, D. Gibson and J. Kleinberg, "Automatic resource compilation by analyzing hyperlink structure and associated text", In Proceedings of the 7th International World Wide Web Conference, 1998

[6] Krishna Bharat, George A. Mihaila, "When Experts Agree: Using Non-Affiliated Experts to Rank Popular Topics", In Proceedings of the 10th International World Wide Web Conference, 2001

[7] O.A. McBryan, "GENVL and WWWW: Tools for taming the Web", In Proceedings of the 1st International Conference on the World Wide Web, 1994

[8] Masashi Toyoda, Masaru Kitsuregawa, Hiroko Mano, Hideo Itoh and Yasushi Ogawa, "University of Tokyo/RICOH at NTCIR-3 Web Retrieval Task", In Proceedings of the 3rd NTCIR Workshop Meeting, pp. 31-38, 2002

[9] Kenji Tateishi, Hideki Kawai, Susumu Akamine, Katsushi Matsuda and Toshikazu Fukushima, "Evaluation of Web Retrieval Method Using Anchor Text", In Proceedings of the 3rd NTCIR Workshop Meeting, pp. 25-29, 2002

[10] S.E. Robertson and S. Walker, "Okapi/Keenbow at TREC-8", In Proceeding of the 8th Text Retrieval Conference, 1999

[11] Masashi Toyoda and Masaru Kitsuregawa, "Creating a Web Community Chart for Navigating Related Community", In Proceedings of the 10th International World Wide Web Conference, 2001

[12] Kalervo Järvelin and Jaana Kekäläinen, "IR evaluation methods for retrieving highly relevant documents", In Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 2000

[13] NTCIR Workshop 3 Meeting, http://research.nii.ac.jp /ntcir/workshop/

[14] Google, http://www.google.com/

[15] 黒橋 禎夫, 長尾 真, 日本語形態素解析システム JUMAN Version 3.61, 京都大学大学院情報学研究科, 1999