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

2007－DBS－141 (18)
2007－GN－ 62 (18)
2007／1／26

リンク先の内容を考慮したアンカー文字列のメタデータ生成手法

千葉 將貴† 鈴木 優†† 川越 恭二††

本稿では，アンカー文字列に対応する適切なメタデータを自動生成する手法を提案する．現在，Webページの増加に伴い，Webリンクの信頼性が問題となっている．この問題とは，あるWebリンクを利用者がクリックした場合に，アンカー文字列と全く関連の無いページが提示される問題であり，Webリンクの不整合によって発生するものである．この問題を解決するため，Webページに対するWebリンクが適切かをシステムによって判定することが必要である．そこで，本研究では，適切なWebリンクを選択することが可能となるように，リンク先の内容を反映したメタデータをアンカー文字列に生成する手法を提案する．ここで，本研究では，アンカー文字列と，そのリンク先ページに出現する単語相互の相関は高いと考え，単語の相関の大きさによって，リンク先ページの内容を反映したメタデータをアンカー文字列に生成する手法を提案した．評価実験により，リンク先ページの内容が反映されたメタデータをアンカー文字列に生成することが可能であることを実証した．

An Automatic Generation Method of Metadata for Anchor Text

Masataka Chiba,† Yu Suzuki†† and Kyoji Kawagoe ††

In this paper, we propose an automatic generation of metadata for anchor text on the Web. Recently, the reliability of web page links is one of the main issues, due to increase the number of web pages. This issue is that, when users click web page links, the users cannot browse accurate web page links. To solve this issue, we propose a method for estimating web page links using metadata. Using automatically generated metadata, users can identify whether the link target web pages are relevant to users.

1. はじめに

現在，Webページの増加に伴って，Webリンクの信頼性に関する問題が顕在化している．例えば，あるWebリンクを利用者がクリックした場合に，そのWebリンクに含まれる文字列と全く関連の無いWebページが表示されるといったリンクの不整合がある．リンクの不整合は，SEO (Search EngineOptimization) など，検索エンジンの順位付けアルゴリズムに対する操作を目的としたリンクなどに数多く見られる．リンクの不整合が発生した場合，利用者はWebページから有益な情報を得るために再度リンクの選択をする必要があり，情報検索を効率的に行う上で問題であるといえる．この問題に対応するために，Webリンクが適切かををシステムによって判定することは重要であるといえる．そこで本研究では，適切なWebリンクを選択することが可能となるように，リンク先の内容が反映したメタデータをアンカー文字列に生成する手法を提案する．ここで，適切なWebリンクとは，Webページを閲覧している利用者が，アンカー文字列から推測することができるWebページと実際のリンク先Webページが類似しているようなWebリンクであると定義する．アンカー文字列とは，リンクに含まれる文字列のことである．つまり，アンカー文字列から推測されるWebページの内容を反映したメタデータを自動的に導出することによって，Webリンクが適切かの判断が可能である．

提案手法では，アンカー文字列および，適切であると考えられるWebページに含まれる単語との対応を事前に与えておき，判定対象となるアンカー文字列へ適切であると考えられる単語をメタデータとして導出する．そして，実際のリンク先Webページとメタデータが異なっている場合には，そのWebリンクが適切ではないと判断する．図1では，アンカー文字列に生成されたメタデータから，そのWebリンクが適切であるかを判定している様子を示している．アンカー

† 立命館大学大学院 理工学研究科
Graduate School of Science and Engineering, Ritsumeikan University
†† 立命館大学 情報理工学部
College of Information Science and Engineering, Ritsumeikan University

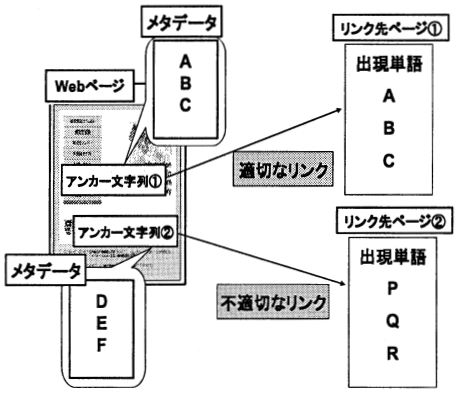

図 1 アンカー文字列にメタデータを生成するイメージ

文字列①には，メタデータ *A*，*B*，*C* が生成されており，メタデータ *A*，*B*，*C* と一致する単語が対応する Web ページに出現しているため，適切なリンクであると判断する．一方，アンカー文字列②には，メタデータ *D*，*E*，*F* が生成されており，対応する *Web* ページに単語 *P*，*Q*，*R* が出現している．ここで，生成されたメタデータが対応するページに出現していないため，適切なリンクではないと判断する．以上の手法を用いることによって，Web リンクの適切性を導出することが可能となる．これによって，利用者はより適切なリンク選択ができ，効率的な検索を実現することが可能となる．なお本稿では，メタデータを生成するまでを行い，メタデータを基にして適切な Web ページかを判定する方法に関しては今後の課題とする．

アンカー文字列は一般的に一つもしくは非常に少ない単語数で構成されていることが多い．既存の研究である情報検索技術を用いることによって，多くの単語から推測される文字列を導出することは可能である．ところが，非常に少ない単語数から推測される文字列を推測することは難しい．提案手法は，アンカー文字列に非常に少ない単語数が含まれている場合であっても利用できる点が利点であると考えられる．

2. 関 連 研 究

2.1 リンクの不整合に関する研究

河合ら[1]は，複数ページ間でのリンク元，リンク先，アンカー文字列の同一性と仲間外れに着目してリンクの不整合を効率よく検出する方法を提案している．これは，複数ページ間で同一のリンク元，リンク先，アンカー文字列の関係が複数回出現したものをグループ化し，異なった関係が検出された場合は仲間はずれとして，仲間はずれとなるリンクを適切なリンクではないと判断するものである．“アンカー文字列と，対応するリンク先の Web ページを用いてリンクの不整合を判断する” という考え方に基づいている点で提案手法と類似しているが，河合ら[1]は，企業の Web ページを対象として，リンク元，リンク先，アンカー文字列の同一性と仲間はずれに着目し，リンクの不整合を判断している．それに対して本稿では，一般的な Web ページを対象としてリンクの不整合を判断している．一般的な Web ページには，様々なアンカー文字列が点在しており，同一のアンカー文字列から同一のリンク先ページにリンクされているとは限らない．そこで，提案手法では判定対象となるアンカー文字列へメタデータを自動生成することによって，リンクの不整合を判断する．これにより，同一のアンカー文字列が同一のリンク先ページにリンクされていない場合においても，リンクの不整合を判断することが可能となる．

2.2 メタデータの生成

メタデータの生成を行うために，様々な研究が行われている．文献[2)3)]では，各特定分野の特徴を反映したメタデータを生成するために，辞書や用語辞書を用いて生成する方式を提案している．これらの方式では，一定の条件を満たす辞書や用語辞典があることを前提としており，辞書や用語辞書がない特定分野については，実現が困難であると考えられる．また，文献[2)3)]ではメタデータの生成を必ず人間による作業が必要であるとしている．提案方式では，辞書や用語辞書を用いず，アンカー文字列と対応する Web ページに出現する単語相互の相関関係を用いることによって，自動的にメタデータを生成することが可能である．そのため，辞書や用語辞書が存在しない少ない単語数で構成されているデータに関してもメタデータを生成することが可能であるため，幅広い分野に応用できることが考えられる．

3. アンカー文字列のメタデータ自動生成方式

本研究では，アンカー文字列と，リンク先ページに出現する単語相互の相関関係を用いることによって，判定対象となるアンカー文字列へ適切な単語をメタデータとして自動生成する方式を提案する．本手法を用いることによって，Web リンクの適切性を導出することが可能となる．

3.1 概　　要

Web 利用者はリンクを辿る際，アンカー文字列を確認することによって，リンク先のページを閲覧するかを判断している．つまり，リンク先ページの内容がアンカー文字列に反映されているため，このような動作につながっていると考えられる．また，アンカー文字列は複数 $(1,2,\cdots s)$ の語から成り立っている．つまり，アンカー文字列を形成しているそれぞれの語が組み合わせって，対応する Web ページの内容を表していると考えることができる．これらの理由から，アンカー文字列を形成している単語それぞれとアンカー文字列と対応する Web ページに出現する単語相互の相関は高いのではないかと考えた．

そこで提案手法ではまず，アンカー文字列を形成している単語を a_i^n とし，a_i^n と対応する Web ページに出現する単語を m_{nj} として，a_i^n と m_j^n に出現する単語相互の相関の大きさを算出する．ここで，n はアンカー文字列と対応する Web ページを特徴付けるためのページ番号を示しており，i はアンカー文字列を形成している単語の番号，j はアンカー文字列と対応する Web ページに出現する単語の番号を示している．相関の大きさには，a_i^n と m_j^n の共起頻度を考慮した重み（関連度），m_j^n のページ内の出現頻度と m_j^n の n ページ間での文書頻度を考慮した重み（特徴量）を用いて算出する．そして，算出された関連度と特徴量の積（特徴関連度）を相関の大きさとし，メタデータを生成するための重みとする．次に，算出した重みを基に a_i^n と m_j^n を対応づけることによって，a_i^n と m_j^n のメタデータ行列を作成する．この時，n ページ間で同一の a_i^n が重複して出現するため，重複した a_i^n それぞれと対応づけられた m_j^n との間でメタデータ行列の統合を行う．メタデータ行列の統合とは，異なるアンカー文字列に含まれる同一の単語から生成された複数のメタデータ行列において，対応づけられたより多くのメタデータ行列に含まれる単語の重みを大きくする処理である．そして，統合されたメタデータ行列それぞれをベクトル化し，判定対象となるアンカー文字列に対応する a_i^n と a_i^n に対応づけられている m_j^n を写像することによって，判定対象となるアンカー文字列のメタデータを生成する．図 2 に提案手法のフローチャートを示す．以降の節では，提案手法のフローチャートに沿って説明を行う．

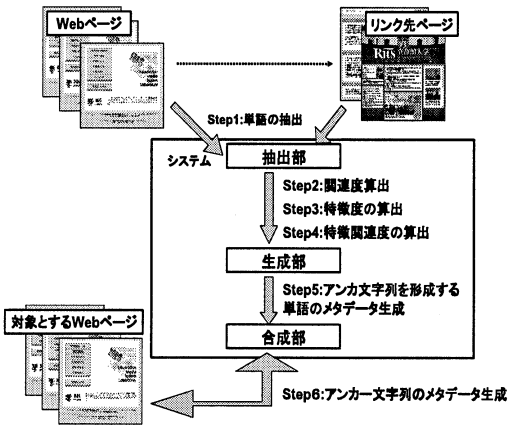

図 2　提案手法のフローチャート

3.2 提案手法の処理手順

Step1 単語の抽出

提案手法ではまず，Web ページからアンカー文字列を形成する単語 a_i^n を抽出する．次に，アンカー文字列と対応する Web ページに出現する単語 m_j^n を抽出する．なおキーワード抽出には，形態素解析を用い，形態素解析器として *sen*[4] を用いる．形態素解析によって抽出するキーワードは，Web ページ間で特徴を表していると考えられる，一般名詞・固有名詞とする．

Step2 関連度の算出

本節では，関連度の算出方法について説明する．関連度は，アンカー文字列を形成する単語 a_i^n と対応する Web ページに出現する単語 m_j^n の共起頻度を考慮した重みであり，a_i^n と m_j^n の関連の大きさを表す．図 3 を用いて，a_i^n と m_j^n の共起頻度を関連の大きさとした理由を示す．図 3 には，“亀田” というアンカー文字列を形成する単語に着目した時に，アンカー文字列と対応する Web ページに出現する単語の様子が示されている．この時，亀田というキーワードと共起して，双方のリンク先ページに “ボクシング” や “ライトフライ級” という単語が出現していることが分かる．つまり，ボクシングやライトフライ級といった単語は亀田という単語の内容を反映していると考えることができる．そこで，a_i^n と m_j^n の複数ページ間での単語の共起頻度から，アンカー文字列とリンク先ページの出現単語の関連の大きさを算出することによって，より適切なメタデータを生成することが可能であると考えた．

提案手法では，関連度の算出に相互情報量[5] を用いることを提案する．相互情報量は，テキス

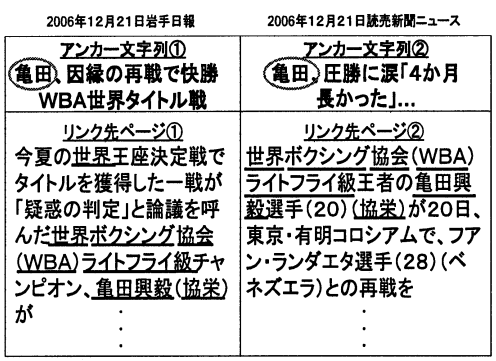

図 3 アンカー文字列に「亀田」を含む 2 つのリンク先ページ

トデータから語の共起情報を抽出する手法として様々な研究[6]で用いられている．相互情報量を用いることにより，a_i^n と m_j^n の関連の大きさを算出することができる．相互情報量は，単語 x と y の出現確率をそれぞれ $P(x)$，$P(y)$ とし，x，y が同時に出現する確率を $P(x,y)$ としたとき，2 語の持つ相互情報量 $I(x,y)$ は以下のように定義される．

$$I(x,y) = log_2 \frac{P(x,y)}{P(x)P(y)} \tag{1}$$

ここで，x の出現確率は x を含む文書数と全文書数 N の比とし，$P(x) = df(x)/N$ とする．相互情報量は x，y が同時に出現する確率 $P(x,y)$ と x と y が完全に独立して出現する確率 $P(x)P(y)$ の比である．

相互情報量の算出式を用いることによって，a_i^n と m_j^n の相互情報量 $I(m_j^n)$ を算出する．相互情報量は a_i^n，m_j^n に対する確率 $P(a_i^n)$，$P(m_j^n)$，$P(a_i^n, m_j^n)$ を算出することで求められる．用語 x の出現する確率は $P(x) = df(x)/N$，2 語 x，y が同時に出現する確率は $P(x,y) = df(x,y)/N$ となるため，文書頻度 df を用いた相互情報量の算出は以下のようになる．

$$I(m_j^n) = log_2 \frac{Rdf(a_i^n, m_j^n)}{df(a_i^n)df(m_j^n)} \tag{2}$$

本稿では，こうして算出された $I(m_j^n)$ を関連度とし，a_i^n と m_j^n との関連の大きさを表す．なお，相互情報量の小さい組み合わせは膨大に存在するため，$I > 1$ を満たさないものに関しては考えないものとする．

Step3 特徴量の算出

本節では，特徴量の算出方法について説明する．特徴量の算出には，$tfidf$ 法[7]を用いる．$tfidf$ 法を用いることによって，アンカ文字列に対応する Web ページに一般的に良く用いられる単語の重みを小さくすることができる．また，ページ内に出現する頻度の多い単語の重みを大きくすることができる．これによって，リンク先ページの内容を反映した単語をアンカー文字列のメタデータとして算出することが可能となる．$tfidf$ の算出は以下のようになる．

$$tfidf(m_j, n) = tf(m_j, n) \cdot idf(m_j) \tag{3}$$

$tf(m_j, n)$ はアンカー文字列に対応する Web ページに出現する単語 m_j のページ n 中での出現頻度，$df(m_j)$ は単語 m_j の出現するページ数である．$idf(m_j)$ は，$df(m_j)$ の逆数であり，$tfidf(m_j, n)$ は $tf(m_j, n)$ と $idf(m_j)$ の積により計算される．$idf(m_j)$ は，以下のように計算する．

$$idf(m_j) = log_2 \frac{K}{df(m_j)} \tag{4}$$

なお，$tfidf(m_j, n)$ としては，以下のような正規化[8]された $tfidf(m_j, n)$ の値を用いる．

$$tfidf(m_j, n) = \frac{tf(m_j, n) \cdot idf(m_j)}{\sqrt{\sum_{j=0}^{k} (tf(m_j, n) \cdot idf(m_j))^2}} \tag{5}$$

本稿では，こうして算出された $tfidf$ 値を特徴量とし，アンカー文字列に適切なメタデータを生成するための m_j^n の重みとする．

Step4 特徴関連度の算出

本節では，特徴関連度の算出方法について説明する．特徴関連度は $Step2$ で算出した関連度と $Step3$ で算出した特徴量の積によって算出される．特徴関連度を算出することによって，アンカー文字列を形成する単語 a_i^n とアンカー文字列と対応する Web ページに出現する単語 m_j^n の相関が大きい単語を a_i^n のメタデータとして m_j^n から選択することが可能となる．

$$w(m_j^n) = I(a_i^n; m_j^n) \cdot tfidf(m_j, n) \tag{6}$$

こうして計算された値を特徴関連度とし，a_i^n のメタデータを生成するための m_j^n の重みとする．

Step5 アンカー文字列を形成する単語のメタデータ生成

本節では，$Step4$ で算出した特徴関連度を用いて，アンカー文字列を形成する単語 a_i^n のメタデー

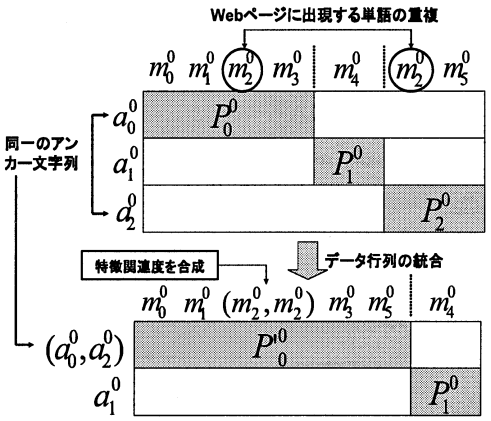

図 4　メタデータ行列の統合

タを生成する方式について説明する．まず，アンカー文字列を形成する単語 a_i^n とリンク先ページに出現する単語 m_j^n を以下のように対応付ける．

$$a_i^n = (m_1^0, m_2^0, \cdots, m_j^n) \tag{7}$$

a_i^n と m_j^n を対応付けることによって，a_i^n のメタデータを生成することが可能となる．次に，a_i^n のメタデータとして適切な m_j^n を選択するために，a_i^n と m_j^n のメタデータ行列を作成する．メタデータ行列は，a_i^n を $(a_1^n, a_2^n, \cdots, a_i^n)^T$ とすることにより，i 行 j 列のメタデータ行列 $P_0, P_1, \cdots, P_i$ を作成できる．同様に，$a_i^1, a_i^2, \cdots, a_i^n$ において，メタデータ行列 $P_i^0, P_i^1, \cdots, P_i^n$ を作成する．ここで，作成した P_i^n において，同一の a_i^n が出現する場合がある．この時，同一の a_i^n と対応づけられている m_j^n からなるメタデータ行列 P_i^n の統合を行う．メタデータ行列の統合とは，同一の a_i^n が出現した際に，a_i^n と対応づけられている m_j^n の重みを大きくする処理である．メタデータ行列の統合を行うことによって，m_j^n の内容を考慮したメタデータを a_i^n に生成することが可能となる．メタデータ行列の統合を行っている様子を図 4 に示す．メタデータ行列を統合する手順は以下のようになる．

[手順 1] アンカー文字列を形成する単語の重複を判断

メタデータ行列 P_i^n において，メタデータ行列を構成している a_i^n の重複の有無を確認する．

図 4 では，データ行列 P_0^0 と P_2^0 において，a_0^0 と a_2^0 の重複を確認したとする．

[手順 2] アンカー文字列を形成する単語に対応づけられている Web ページに出現する単語の重複を判断

手順 1 で重複を確認した a_i^n において，重複を確認した a_i^n それぞれと対応づけられている m_j^n の重複の有無を確認する．次に，a_i^n の重複を取り除き，m_j^n をメタデータとして生成する．

図 4 では，手順 1 で重複を確認した a_0^0 と a_2^0 に対応づけられている $m_0^0, m_1^0, m_3^0, m_5^0$ を a_0^0 と a_2^0 の重複を取り除いた (a_0^0, a_2^0) のメタデータとして生成する．なお，a_0^0 に対応づけられている m_2^0 と a_2^0 に対応づけられている m_2^0 の重複を確認したとする．

[手順 3] 特徴関連度を計算

手順 2 で重複を確認した m_j^n において，重複を確認した m_j^n それぞれが保持している特徴関連度 $w(m_j^n)$ を加算する．なお，重複が確認できた m_j^n が k 個存在した場合には，以下のように計算する．

$$w(m_j^n) = \sum_{k=0}^{k} w(m_j^n) \tag{8}$$

次に，重複を確認した m_j^n の重複を取り除き，加算した $w(m_j^n)$ を重複を取り除いた m_j^n の特徴関連度とし，a_i^n のメタデータとして生成する．なお，計算した特徴関連度の大きい値を保持している m_j^n が a_i^n のメタデータとしてより適切であるとする．

図 4 では，手順 2 で重複を確認した a_0^0 に対応づけられている m_2^0 と a_2^0 に対応づけられている m_2^0 それぞれが保持している特徴関連度を加算し，重複を取り除いた (m_2^0, m_2^0) を (a_0^0, a_2^0) のメタデータとして生成する．

Step6 アンカー文字列へのメタデータ生成

本節では，*Step5* で生成したアンカー文字列を形成する単語のメタデータ行列を用いて，アンカー文字列のメタデータを生成する方式を説明する．判定対象とするアンカー文字列を $G(G \in g)$ とした場合，G は，t 個の語 $g_1, g_2, \cdots, g_t$ から形成されている．この時，*Step5* で生成したメタデータ行列をベクトル化し，判定対象となる g_t に g_t と同一のアンカー文字列を形成する単語 a_i^n と a_i^n に対応づけられている m_j^n を写像する．これにより，判定対象とするアンカー文字列を単語と単語の関係から重みを計算することが可能となるため，アンカー文字列のメタデータを生

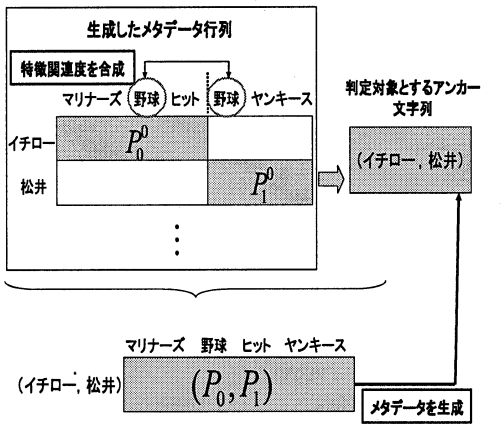

図 5 判定対象となるアンカー文字列のメタデータ生成

成することが可能となる．ここで，アンカー文字列のメタデータを生成する手順を図 5 を用いて具体的に説明する．

図 5 では，判定対象とするアンカー文字列がイチロー，松井から形成されているとする．この時，生成したメタデータ行列において，イチローに対応づけられているマリナーズ，野球，ヒットと松井に対応づけられている野球，ヤンキースを判定対象となるアンカー文字列に写像する．これによって，イチローと松井に対応づけられているマリナーズ，野球，ヒット，ヤンキースをアンカー文字列のメタデータとして生成することが可能となる．ここで，生成したメタデータ行列に野球という単語がイチローと松井双方に対応づけられている．この場合，イチローと松井に対応付けられている野球という単語それぞれが保持している特徴関連度を加算し，重複をなくす．加算の計算は，計算式 (8) と同様である．

こうして生成された判定対象となるアンカー文字列のメタデータの中で，特徴関連度の大きいメタデータをアンカー文字列のメタデータとして採用する．

4. 実　　験

本手法の有用性を検証するために，3 章で示した方式を用いて生成したアンカー文字列のメタデータに対する検証実験を行った．

提案方式は，アンカー文字列とリンク先ページに出現する単語相互の相関の大きさによって，アンカー文字列のメタデータを生成する．そして，アンカー文字列に生成されたメタデータを利用することによって，リンク先のページの内容が適切かどうかを判断することを想定している．そこで本実験では，アンカー文字列に生成されたメタデータが適切なリンク先ページの内容を反映しているかを考察することによって，提案方式の検証を行う．

表 1 ”亀田”に生成されたメタデータ

メタデータ	特徴関連度 ($w(m_j^n)$)
スポーツ	6.62
毅	6.46
浪速	4.43
亀田	4.22
素人	3.87
ジム	3.43
キーワード	3.39
コラム	3.37
再戦	3.15
チャンピオン	3.14

表 2 ”世界”に生成されたメタデータ

メタデータ	特徴関連度 ($w(m_j^n)$)
王者	3.06
世界	3.04
国際	2.53
月	2.44
素人	2.36
ラッシュ	2.14
チャンピオン	2.10
協会	2.08
棒	1.95
サイト	1.93

4.1 実験環境

実験では，Web ページからアンカー文字列と，そのリンク先ページに含まれる単語を抽出し，アンカー文字列のメタデータを生成した．ここで，Web ページから単語の抽出を行うための形態素解析には，3 章の *Step*1 で述べたように *Sen* を用いた．解析時には *Sen* の標準的な辞書を用いたため，実験では解析時に元通りに抽出できずに分解されてしまった用語が存在した．また，Web ページは *googleAPI*[9] を用いて “ボクシング” というキーワードに関するページを 100 ページ取得し，取得したデータを基に実験を行った．

4.2 実験結果及び考察

提案手法を用いて，アンカー文字列を形成する単語にメタデータを自動生成した結果の例を表 1，表 2 に示す．表 1 では “亀田”，表 2 では “世界” というアンカー文字列を形成する単語に対して生成されたメタデータの中で特徴関連度の大きい上位 10 件を示している．表 1，表 2 から亀田という単語の内容を反映したメタデータ，

表 3　アンカ文字列（亀田が世界王者に）のメタデータ

メタデータ	特徴関連度 ($w(m_j^a)$)
スポーツ	9.23
毅	8.21
王者	6.32
素人	6.23
亀田	5.50
月	5.44
チャンピオン	5.24
ジム	5.22
ボクシング	5.06
世界	4.86

世界という単語の内容を反映したメタデータが生成されていることを確認できる．しかし，亀田や世界という単語の内容が反映されていないメタデータとしてコラムやサイトといった単語が生成されている．この原因として，コラムやサイトといった単語はWeb上で一般的に扱われる単語であるため，同一のアンカー文字列と多数共起して出現したために，3章の*Step2*で提案した関連度が大きくなったことが考えられる．

次に表3に，“亀田が世界王者に”というアンカー文字列に生成されたメタデータの中で，特徴関連度の大きい上位10件を示す．表3を確認すると，アンカー文字列の内容が反映された単語が生成されていることを確認できる．また，“亀田が世界王者に”というアンカー文字列からリンクするページには，生成されたメタデータの中で特徴関連度の大きい上位10件の内，7件の単語が含まれていることを確認した．

本実験において，アンカー文字列と，リンク先ページに出現する単語の相関関係を用いることによって，アンカー文字列にリンク先ページの内容を反映したメタデータを生成できることが示された．

5. おわりに

本稿では，アンカー文字列に対応する適切なメタデータを自動生成する方式を提案した．また，提案手法を用いて実際にアンカー文字列へのメタデータ生成を行った．その結果，リンク先ページの内容が反映されたメタデータをアンカー文字列に生成することができた．

今後の予定を述べる．本稿では，同一のアンカー文字列とリンク先ページにアンカー文字列の内容が反映されない一般的な単語（Webページの著作権表示に関する用語や日付）が多数共起して出現したため，上位の結果として算出されてしまう場合があった．そこで，この問題を解決するために，アンカー文字列と，リンク先ページに出現する単語との共起頻度を算出する際に用いた相互情報量の重みを正規化することを検討する．加えて，この問題に関しては，実験データ数を増やすことによって，違った結果が得られる可能性があるため，今後は実験データ数を増やし，この問題を検討していく．また本稿では，アンカー文字列に生成されたメタデータを基にして適切なWebページかを判定する方法に関しては考慮していない．そこで，生成されたメタデータを基に適切なWebページかを判定する方法についても，検討する予定である．

参考文献

1) 河合英紀, 河野泉, 石黒義英, 福島俊一: サイト品質管理のためのリンク不整合検出, 電子情報通信学会第15回データ工学ワークショップ (DEWS2004) 論文集, 5-b-01 (2004).
2) 宮川祥子, 清木康: 特定分野ドキュメントを対象とした意味的連想検索のためのメタデータ空間生成方式, 情報処理学会論文誌：データベース, Vol. 40, No. SIG5(TOD2), pp. 15–27 (1999).
3) 河本穣, 清木康, 吉田尚史, 藤島清太郎, 相磯卓和: 医療分野ドキュメント群を対象とした意味的連想検索のためのメタデータ空間生成方式, 日本データベース学会 Letters, Vol. 1, No. 2, pp. 12–15 (2003).
4) ：形態素解析システム「Sen」. http://ultimania.org/sen/.
5) Church, K. W. and Hanks, P.: Word Association Norms,Mutual Information and Lexicography, *Computational Linguistics*, pp. 76–82 (1989).
6) 松尾豊, 石塚満: 語の共起の統計情報に基づく文書からのキーワード抽出アルゴリズム, 人工知能学会論文誌, Vol. 17, No. 3, pp. 217–223 (2002).
7) Salton, G., Lesk, M.E.(編): Introduction to Modern Information Retrieval, *McGrawhill Book Co* (1983).
8) 北研二, 津田和彦, 獅々堀正幹: 情報検索アルゴリズム, 共立出版株式会社 (2002).
9) ：googleAPI. http://code.google.com/.