ハイパリンクを考慮した Web ページからの内容抽出とその評価

北原沙緒理† 波多野賢治††

† 同志社大学大学院文化情報学研究科 〒 610–0394 京都府京田辺市多々羅都谷 1–3
†† 同志社大学文化情報学部 〒 610–0394 京都府京田辺市多々羅都谷 1–3
E-mail: †kitahara@ilab.doshisha.ac.jp, ††khatano@mail.doshisha.ac.jp

あらまし 本稿では Web ページにおけるアンカーテキストを利用した内容抽出法を提案する．Web ページはハイパリンクを有しており，ハイパリンクの情報は Web 文書検索等の分野で使用され，Web ページを扱う際に非常に有用な指標であることが示されてきた．しかしこれまで行われてきた Web ページにおける内容抽出法では Web ページ内のテキスト情報のみが使用されているため，Web ブラウジングの際にリンク元の Web ページで閲覧していた内容がリンク先の Web ページに及ぼす内容の影響度を考慮することができない．そこで本稿ではアンカーテキスト上での内容の影響度をリンク先の Web ページに継承することで，リンク元の Web ページの内容を考慮したリンク先の Web ページの内容抽出法を提案する．また，本稿で提案した内容抽出法の優位性を示すため，既存のハイパリンクを考慮しない内容抽出法との比較実験および結果の考察を行う．

キーワード 文書内容抽出，ハイパリンク，内容密度分布，情報抽出，Web マイニング

Extraction of Web Page Contents based on Hyperlinks

Saori KITAHARA† and Kenji HATANO††

† Graduate School of Culture and Information Science, Doshisha University
1–3 Tatara Miyakodani, Kyotanabe, Kyoto 610–0394, Japan
†† Faculty of Culture and Information Science, Doshisha University
1–3 Tatara Miyakodani, Kyotanabe, Kyoto 610–0394, Japan
E-mail: †kitahara@ilab.doshisha.ac.jp, ††khatano@mail.doshisha.ac.jp

1. はじめに

現在，World Wide Web (WWW) 上には膨大な Web ページが存在しており，WWW 上に存在する様々な種類の内容を含む Web ページからユーザが必要としている内容が含まれる Web ページを見つけるために，検索エンジンを用いる．しかし，既存の検索エンジンを用いて Web 検索を行う際，Web 検索結果にユーザが必要とする内容が含まれない場合がある．これは Web ページ内のテキスト (Web テキスト) 中における単一の文中に複数の内容を含むことを考慮していないため，ユーザが意図した内容とは異なる内容を含んだ文を基にした Web 検索結果が出力されてしまう可能性があるためである．

よって，Web 検索を行うユーザを支援するために，Web テキスト内の内容を抽出するための手法 (内容抽出法) が提案されてきた．例えば，Web テキストを要約する手法 [1] や，Web テキストの重要部分を可視化する手法 [2] が内容抽出法として挙げられる．

これらの手法によって，一回の Web 検索時にユーザが手に入れるデータが多くなったため，昨今よりも効率よく Web 検索を行うことができるようになった．特に，これらの手法により，ユーザが Web テキスト内に存在する要素 (パラグラフ，文など) に含まれる内容を直感的に理解することができるようになったため，ユーザが Web 検索を行う際に閲覧するべきデータを取捨選択することが容易になった．

従来の Web ページに対する内容抽出法では，基本的にはテキストのデータである Web テキストのみを用いて内容抽出を行っている [2]．しかし，Web ページにはテキストのデータ以外にも様々なデータが含まれている．特に，ハイパリンクに関するデータは PageRank [3] や HITS [4] などの検索アルゴリズムにて，Web ページを代表するデータの一つとして使用されている．

よって，本稿では従来の Web テキストにおける内容抽出法に，ハイパリンクのデータを踏まえた Web ページからの内容抽出法を提案する．その際，ユーザが Web ページ中に内容が

含まれ手いると判断した箇所を抽出できるか否かを内容抽出の基準として使用する．

ここで，各 Web ページの内容を俯瞰することが出来るような付加情報を各 Web ページに付与することが出来れば，ユーザが必要とする情報に簡単にアクセスできるようにすることが可能になると考えられる．よって，本稿で提案する研究は，Web 検索結果を閲覧するユーザや WWW のブラウジングを行うユーザに対して，各 Web ページの内容を俯瞰することが出来るようにするための基礎となる研究である．最終的には，本稿で提案する手法を Web 検索結果と併用し，検索ワードなどの現在閲覧中の Web ページにおいて注目している内容の可視化を自動的に行うことにより，ユーザの Web 検索および WWW のブラウジングを支援する予定である．

2. 関 連 研 究

本節では，以下の観点を基にした関連研究について述べる．

（1） ハイパリンクを用いた内容抽出法に関する研究

（2） 従来の Web テキストを用いた内容抽出法に関する研究

2.1 ハイパリンクを用いた内容抽出法に関する研究

田村らは確率的言語モデルを用いた情報検索モデルに対してハイパリンクの情報を適用することにより，Web ページに対する情報検索を行っている [5]．しかし，Web ページ中の一つ一つの索引語に対する Web ページ全体の重みを計算しているため，この研究で用いている手法では，Web ページ中の局所的な内容の関連度を求めることができない．

また，阿部らは Web ページ中のアンカーテキストと注目している Web ページに関するリンク構造を考慮した Web 検索についての研究を行っている [6]．この研究では検索クエリ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が多い Web ページを有用なページであると定め，該当の Web ページが大きなスコアを持つようなランキング指標を考案した．その結果，アンカーテキストのみを考慮した場合よりも高い検索精度を得ることができた．しかし，アンカーテキストが必ずしもリンク先の内容を表しているとは限らない．例えば，アンカーテキストが何らかの事象の具体例を表しており，リンク先の Web ページがその事象の概要についての説明であった場合，アンカーテキストがあらわしている「具体例」に関する内容は Web ページ内には存在しない．

一方，荒木らはリンク集ページに掲載されたアンカーテキストの周辺にある文字列から各 Web ページの内容を表すテキストである「アスペクト」を抽出する研究を行っている [7]．この研究では外部の Web ページへのリンク数が既定値以上の Web ページをリンク集ページと呼び，アスペクトを抽出したい Web ページ (対象ページ) へのリンクに書かれたアンカーテキストをキーアンカーと呼んでいる．そして指定した条件に合致する文字列を見出しの役割を果たす内容候補である「見出しコンテンツ」とみなした．これらのコンテンツをキーアンカーと他のアンカーテキストまでの距離を用いて算出したものをクラスタリングし，各クラスタにおける特徴量の大きいものをアスペクトとする．しかし，この手法においては，対象ページとなる Web ページのアスペクトを抽出することが目的であり，Web ページ中の内容の出現位置を見ることは目的ではないため，本稿で提案する手法とは異なる．

2.2 従来の Web テキストを用いた内容抽出法に関する研究

西原らは Web テキスト中に存在する主題と関係がある部分と関係が無い部分を可視化することで，ユーザの Web 検索を支援するツールを作成している [2]．このツールではテキスト内の単語の位置を考慮した重みにより算出された指標を足し合わせ，文単位で主題との関連度を計算する．そして，関連度の大きさに基づいて Web テキストの背景色を，主題と関係がある部分ほど明るく，主題と関係が無い部分ほど暗く，文単位にて変更することで主題と関係がある部分を可視化する．

また，Woodruff らは，Web ページ中に含まれる他の単語よりも大きく表示したサムネイルである enhanced thumbnail を提示することにより，ユーザが Web ページ中における重要視したい内容が存在する位置を把握できるようにする研究を行っている [8]．enhanced thumbnail は重要視したい内容と関係がある単語，およびその関連語を大きく表示した Web ページの画像であるため，ユーザは通常のサムネイルやそのまま Web ページを読む場合よりも enhanced thumbnail を使用した場合の方が素早く重要視したい内容を見つけることができる．

これらに対して，我々は検索ワードの組に関する Web テキストにおける内容の出現範囲および影響度変化を抽出するために，検索ワードの組に関する内容密度分布を抽出する研究を行っている [9]．内容密度分布は，内容の出現範囲および局所的な内容の影響度変化を表す分布であり，内容を形成する単語群に属する各々の単語が影響を及ぼす範囲である単語密度分布により作成される．そして，内容密度分布により表現された内容の出現範囲および影響度変化を用い，Web テキスト中から内容が存在する箇所を抽出する．

しかし，これらの手法による内容抽出では Web ページ中のテキスト情報のみを扱っており，ハイパリンクの情報は取り扱っていない．

3. 提 案 手 法

本稿では 2.2 で述べた従来の内容密度分布抽出法に対して，ハイパリンクの情報を踏まえた Web テキストにおける内容抽出法を提案する．ここで，現在使用されている Web 検索におけるクエリの入力方法のほとんどが単語列であることから，本稿では内容を「Web テキストに存在する単語集合の部分集合」と定義する．これは Web 検索を行う際には単語が重要であると考えられるためであり，ユーザが多くの Web 検索エンジンを用いる際に，自身が閲覧したい内容を表すために単語を用いるためである．内容の定義を単語集合とすることによって，ユーザが閲覧したい内容を自由に定義することができる．

しかし単語一語の場合は多義語などの語における意味の揺れを考慮することが難しく，このような意味の揺れを考慮しない場合，文書を取り扱う研究では意味分類等の側面で精度が低下

するため，これらの意味の揺れを考慮する必要がある [10]．したがって，意味の揺れを考慮するために，内容を形成する単語集合に含まれる単語数は二語以上とする．

ここで，ある Web ページを閲覧する際に注目している内容に関する内容密度分布の値がアンカーテキストの上にも存在しているとき，アンカーテキストが記述されたリンク先の Web ページにも，その内容に関する内容密度分布の値が考慮されると仮定する．

例えば，図 1 の左側の Web テキスト s_x から，図 1 における左下の図のような内容 Q_r^x における内容密度分布が抽出されており，内容密度分布の値が太字で表示されている箇所から抽出した値が左側の Web テキスト s_x 中に存在するアンカーテキストが存在する位置での内容密度分布の値であるとする．Web テキスト s_x に記述されたハイパリンクからは Web テキスト s_y へのリンクが貼られているとすると，Web テキスト s_x 上の内容 Q_r^x における内容密度分布の値が何らかの形で Web テキスト s_y 上の内容 Q_r^y における内容密度分布の値に反映されると考えられる．

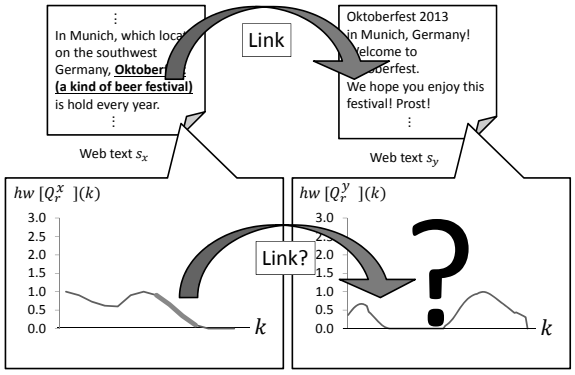

図 1　アンカーテキスト上の内容密度分布の値を用いた内容密度分布の拡張

本節ではまず 3.1 において従来のハイパリンクを考慮しない内容密度分布作成法について説明を行う．次に，アンカーテキスト上における内容密度分布の値を考慮するための要素と，ハイパリンクを考慮するための手法を 3.2 にて提案する．そして，内容密度分布の値を考慮するための要素およびハイパリンクを考慮するための手法を組み合わせることによりハイパリンクを考慮した内容密度分布を提案する．

3.1　従来の内容密度分布作成法

内容密度分布を抽出するためには，内容を形成する単語群に属する各々の単語が影響を及ぼす範囲である単語密度分布を重みつきハニング窓関数によって算出し，これらの範囲を組み合わせる [9].

重みつきハニング窓関数とは通常のハニング窓関数の値に，文の区切りでは前の文に表れる単語の影響度が減少すると考えられることから，単語の影響度の変化を重み D として付与したものである．なお文の区切りには句読点 (。) と全角および半角のピリオド (.)，エクスクラメーションマーク (！)，クエスチョンマーク (？) を用いる．ここで，ある Web テキスト集合に含まれる Web テキストの一つを $s_x(x = 1, 2, \cdots)$，s_x における内容 $Q_r^x(r = 1, 2, \cdots)$ に含まれる単語群を t_i^x，t_i^x のうちテキストの最初から数えて j 番目に出現するものを $t_{i,j}^x$ とすると，$t_{i,j}^x$ が表れる位置 $l[t_{i,j}^x]$ の直前に現れる文の区切りの位置は $a[t_{i,j}^x]$，直後に表れる文の区切りの位置は $b[t_{i,j}^x]$ と表すことができる．また，内容とは単語の集合であるため，$t_i^x \subset Q_r^x(i = 1, 2, \cdots)$ となる．

以上を踏まえると，Web テキスト s_x において k 番目の単語が現れる位置における重みつきハニング窓関数 $hw[t_{i,j}^x]$ は式 (1) のように算出され，この数値が k 番目の単語上における $t_{i,j}^x$ の影響度となる．

$$hw[t_{i,j}^x](k) = \begin{cases} \dfrac{1}{2}(1 + \cos 2\pi \dfrac{k - l[t_{i,j}^x]}{W}) \\ \qquad (\, a[t_{i,j}^x] < k < b[t_{i,j}^x]) \\ \dfrac{1}{2} D(1 + \cos 2\pi \dfrac{k - l[t_{i,j}^x]}{W}) \\ \qquad (\, a[t_{i,j}^x] \geqq k,\, b[t_{i,j}^x] \leqq k) \end{cases} \tag{1}$$

$$(\mid k - l[t_{i,j}^x] \mid \leqq \tfrac{W}{2},\, 0 \leqq D \leqq 1)$$

ここで単語 $t_{i,j}^x$ の影響は窓の幅 W において存在し，$\mid k - l[t_{i,j}^x] \mid \leqq \frac{W}{2}$ の間のみで定義される．重み D がとりうる値の範囲は $0 \leqq D \leqq 1$ とする．本稿では，Web テキスト中の内容抽出に対して最適であるパラメータとして，内容密度分布の重み D を 0.6，窓の幅 W の値を各 Web テキストに現れる文に含まれる平均単語数の 3 倍とする [9].

また，前述の通り単語一つ一つにおける単語密度分布とは内容を形成する単語群に属する各々の単語が影響を及ぼす範囲であり，各単語の出現箇所における重みつきハニング窓関数の値を正規化したものである．したがって，Web テキスト s_x において k 番目の単語が現れる位置における単語 t_i^x における単語密度分布 $hw[t_i^x]$ は，式 (2) の通りである．

$$hw[t_i^x](k) = \frac{\sum_j hw[t_{i,j}^x](k)}{\max_k \sum_j hw[t_{i,j}^x](k)} \tag{2}$$

そして，単語一つ一つにおいて単語密度分布 $hw[t_i^x]$ を作成し，Web テキスト s_x における内容 Q_r^x に関する内容を，各内容密度分布の値が存在する部分において統合することにより，式 (3) のように内容 Q_r^x における内容密度分布を作成することができる．

$$hw[Q_r^x](k) = \frac{1}{n} \sum_i hw[t_i^x](k) \tag{3}$$

$$\left(t_i^x \in Q_r^x,\ (\forall i)\ hw[t_i^x](k) > 0\right)$$

また，式 (4) のように各単語の出現箇所における内容密度分布の値に対して閾値を設けることにより，内容を形成する単語群 Q_r^x 関する内容抽出法として内容密度分布を用いることができる．

$$hasContent[Q_r^x](k) = \begin{cases} 1 & (hw[Q_r^x](k) > E) \\ 0 & (others) \end{cases} \quad (4)$$

本稿では Q_r^x に関する内容密度分布の値が閾値 E $(0 \leqq E \leqq 1)$ 以上であるテキスト上の位置に Q_r^x に関する内容が含まれているとする．なお，本稿では評価実験において，この閾値 E を 0.0 から 1.0 まで変化させることにより，内容密度分布における内容抽出を行う．

3.2 ハイパリンクを用いた内容密度分布の作成

本稿ではリンク元の Web テキスト s_x 上における内容 Q_r^x の内容密度分布の値を，s_y 上の内容 Q_r^y に反映する．その際に，s_x 上においてハイパリンクが存在する箇所を代表する値として用いる指標として，データの分布に関する特徴を表す値である分位数として用いられる [11]，以下の三種類の指標を考える．

- ハイパリンクが存在する範囲における内容密度分布の値の最大値
- ハイパリンクが存在する範囲における内容密度分布の値の中央値
- ハイパリンクが存在する範囲における内容密度分布の値の最小値

またリンク元の Web テキスト s_x での内容 Q_r^x における指標のうちいずれかを，式 (5) のように rep_r^x と表す．なお，Web テキスト s_y へのハイパリンクが存在する範囲の集合を $Link_y$ とする．

$$rep_r^x = \begin{cases} \max\limits_{k \in Link_y} hw[Q_r^x](k) \\ \mathrm{med}\limits_{k \in Link_y} hw[Q_r^x](k) \\ \min\limits_{k \in Link_y} hw[Q_r^x](k) \end{cases} \quad (5)$$

ここで，リンク元の内容密度分布の値 $hw[Q_r^x](k)$ に rep_r^x を適用した，リンク先の Web テキスト s_y 上の内容 Q_r^y における内容密度分布の値を $hw'[Q_r^y](k)$ とする．そして，これらの値を s_y 上の内容 Q_r^y における内容密度分布の値に対して，元の $hw[Q_r^x](k)$ よりもリンク元の内容 Q_r^x に関する内容の影響度を踏まえたことにより $hw'[Q_r^y](k)$ の方が内容の影響度が増加したことを反映する手法を考える．ここで，リンク元の Web テキスト s_x のハイパリンクはリンク先の Web テキスト s_y の全体を指しているため，リンク元の内容 Q_r^x に関する内容の影響度はリンク先の Web テキスト全体に掛かるとする．したがって，リンク元の内容密度分布の値 $hw[Q_r^x](k)$ に rep_r^x を適用した値をリンク先の内容密度分布の値 $hw[Q_r^y](k)$ に適用するには，$hw'[Q_r^y](k)$ の方が $hw[Q_r^y](k)$ よりも内容の影響度が増加する四則演算である乗算もしくは加算を適用する必要がある．本稿ではリンク元の内容を表す指標をリンク先の内容密度分布に対して反映する二種類の手法を提案する．

3.2.1 単純なリンク元内容反映手法

リンク元の内容 Q_r^x に関する内容の影響度を踏まえたことにより $hw'[Q_r^y](k)$ の方が内容の影響度が増加したことを単純に反映するための手法は，以下の二種類が考えられる．

- Q_r^y における内容密度分布の値に (指標 + 1.0) を乗算する手法 (乗算)

$$hw'[Q_r^y](k) = hw[Q_r^y](k) \times (hw[Q_r^x](rep_r^x) + 1.0) \quad (6)$$

- Q_r^y における内容密度分布の値に指標を加算する手法 (加算)

$$hw'[Q_r^y](k) = (hw[Q_r^y](k) + hw[Q_r^x](rep_r^x)) \quad (7)$$

式 (6)，式 (7) はリンク元の内容密度分布の値がリンク先の内容密度分布の値を必ず増加させるという仮定の下，リンク元の内容密度分布の値を生かしてリンク先の内容密度分布の値を増加させることができる式である．本稿ではこれらの式により抽出された内容密度分布による内容抽出に関する評価を行い，ハイパリンクを考慮した内容密度分布の式として最適な要素および指標の組合せはどの組合せであるかを判断する．ここで，従来のハイパリンクを考慮しない内容密度分布と式 (6) の乗算手法を用いてハイパリンクを考慮した内容密度分布を比較した図は図 2 のようになり，式 (7) の加算手法を用いてハイパリンクを考慮した内容密度分布を比較した図は図 3 のようになる．

図 2 において従来の内容密度分布を用いる手法と，乗算を用いてリンク元の内容密度分布の値を考慮する手法を比較することにより，乗算を用いてリンク元の内容密度分布の値を考慮する手法は，リンク先の内容密度分布の値が存在する箇所における，リンク先の内容密度分布の値を増加させる効果があるといえる．したがって，乗算を用いてリンク元の内容密度分布の値を考慮する手法は，リンク先の内容密度分布の値を強調するという形でハイパリンクを考慮して，内容が存在する箇所を際立たせることができると考えられる．また，図 3 において従来の内容密度分布を用いる手法と，加算を用いてリンク元の内容密度分布の値を考慮する手法を比較することにより，加算を用いてリンク元の内容密度分布の値を考慮する手法は，リンク全体における，リンク先の内容密度分布の値を増加させる効果があるといえる．したがって，加算を用いてリンク元の内容密度分布の値を考慮する手法は，リンク先の内容密度分布の値にリンク元の内容密度分布全体に付与するという形でハイパリンクを考慮することができると考えられる．

ただし，リンク先から別のページに遷移すればするほど，元のページとリンク先のページに含まれる内容の類似度が減少するため，本稿ではリンク元のページからの内容密度分布の値のみを考慮する [12]．これらのハイパリンクを考慮した内容密度分布の値に対しても，内容抽出時に用いるパラメータである閾値 E $(0 \leqq E \leqq 1)$ を適用し，Web ページからの内容抽出に用いる．その際，閾値として適切な値は従来の内容密度分布による内容抽出時とは異なる可能性があるため，各要素および手法を考慮した内容密度分布に対して，閾値 E を 0.0 から 1.0 まで変化させることにより内容抽出を行う．また，本稿ではリンク元にはリンク元の，リンク先にはリンク先の内容密度分布による内容の出現分布が存在しており，リンク先においてハイパ

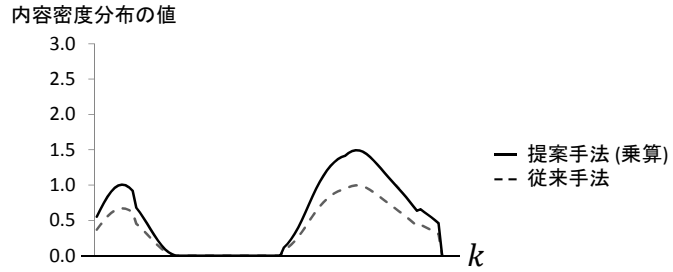

図 2　ハイパリンクを考慮しない内容密度分布 (従来手法) と
乗算によりハイパリンクを考慮した内容密度分布 (提案手法)

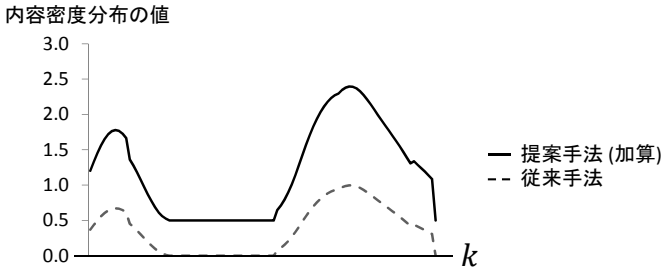

図 3　ハイパリンクを考慮しない内容密度分布 (従来手法) と
加算によりハイパリンクを考慮した内容密度分布 (提案手法)

リンクを考慮する前の内容密度分布と，リンク元の内容密度分布が存在するとする．したがって，リンク元およびリンク先における，内容抽出時に用いる閾値 E を除いた，内容密度分布自体の値を決定するためのパラメータ W および D はリンク元における内容密度分布においても，リンク先における内容密度分布においても，3.1 で使用した Web テキストに対して最適であると考えられる同一のパラメータを用いる．

3.2.2　二重の閾値を使用したリンク元内容反映手法

3.2.1 で提案したリンク元内容反映手法では内容密度分布の値を増加させたことにより，ハイパリンクを用いない内容密度分布を用いた場合では「内容が存在しない」と判断されるような，元々内容密度分布の値が極端に小さい箇所を，内容が存在するとされる箇所として抽出してしまう可能性がある．この問題を回避するために，リンク先の内容密度分布に対して要素となる代表値を考慮する前にも新たな閾値を設け，さらにハイパリンクを考慮した内容密度分布の値に対しても閾値を設けることができれば,「内容密度分布の値が極端に小さい箇所」を考慮することができる可能性がある．したがって，単純なリンク元内容反映手法を用い，リンク元の内容密度分布の値をリンク先に付与することにより，ユーザが内容が存在しないと判断する箇所を内容が出現する箇所として取り扱うことを回避するため，内容密度分布の閾値をハイパリンクを考慮する前にも一度適用する手法を提案する．

ここで，ハイパリンクを考慮する前の閾値を E_1 $(0 \leqq E_1 \leqq 1)$，考慮した後の閾値を E_2 $(0 \leqq E_2 \leqq 1)$ とすると，単純なリンク元内容反映手法と二重の閾値を使用したリンク元内容反映手法は図 4 のようになる．ここで，リンク元からの内容反映手法は，3.2.1 で紹介した乗算などのリンク先の内容密度分布の値を増加させることができる内容反映手法を用いる．

これらの二重の閾値を使用したリンク元内容反映手法を適用した内容密度分布の値に対して，Web ページからの内容抽出に用いる際においても各要素および手法を考慮した内容密度分布に対して，閾値 E_1 及び E_2 を 0.0 から 1.0 まで変化させることにより内容抽出を行う．また，単純なリンク元内容反映手法と同様に，リンク先にはリンク先の内容密度分布による内容の出現分布が存在しており，リンク先においてハイパリンクを考慮する前の内容密度分布と，リンク元の内容密度分布が存在するとする．したがって，リンク元およびリンク先における，内容密度分布自体の値を決定するためのパラメータ W および D はリンク元における内容密度分布においても，リンク先における内容密度分布においても，3.1 で使用した Web テキストに対して最適であると考えられる同一のパラメータを用いる．

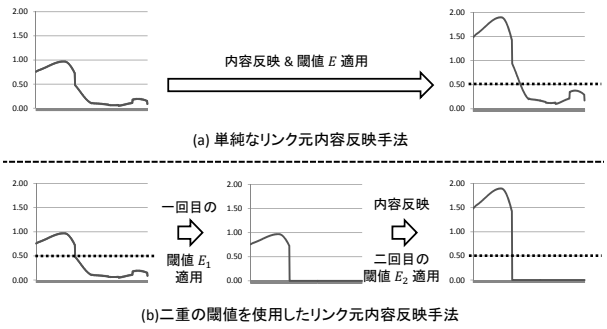

図 4　単純なリンク元内容反映手法と二重の閾値を使用したリンク元内容反映手法の違い

4.　評 価 実 験

本稿では 3. で述べたハイパリンクを考慮した内容密度分布と，テキストのみを考慮した従来の内容密度分布を用いて内容抽出を行った結果を比較することにより，内容抽出にふさわしい内容密度分布を判断する．

そして，本稿で行う実験には我々がかつて行った評価値の平均を用いる [13]．評価値とは，各正解箇所と各手法により抽出された内容が取りうる箇所との精度と再現率の F 値 [14] を基に算出される．この指標の算出方法については 4.1 にて述べる．

また，Web ページにおける内容抽出法による内容抽出結果として抽出された Web ページ中の箇所を比較するためには，ユーザによって内容が存在すると判断された箇所を正解箇所として用いる必要がある．したがって，このようなデータを持たない既存のテストコレクションでは評価を行うことができない．よって，本稿で行う評価実験では人手で作成したテストコレクションを用いる．このテストコレクションの詳細については 4.2 にて述べる．

4.1　評価指標の算出

本稿では，内容抽出にふさわしい内容密度分布はどの要素お

よび手法を考慮したものであるかを，各要素および手法を用いた内容密度分布により，Web ページ中において内容が存在する箇所として抽出された箇所により比較を行う．その際，以下の二観点で同時に量的に評価できるようにするために，情報検索の分野における精度と再現率の F 値 を基にした値を評価指標として使用する．

- ある Web ページ中において注目している内容をどれだけ網羅できているか (網羅性，comprehensiveness)
- Web ページ中から「注目している内容が存在する箇所」として抽出した箇所にどれだけ対象の内容が含まれているか (正確性，exactness)

評価指標を用いる理由は，Web ページの要約や，Web ページ内の重要部分を可視化する手法においては，内容として抽出される箇所が網羅的かつ正確に抽出されなければならないためである．よって正確性単体および網羅性単体での評価は行わない．

ここで，Web テキスト s_x における，ある内容抽出法により内容が含まれる箇所として抽出された箇所の集合を C_x，その集合に含まれる要素の個数を $n(C_x)$ とする．また，正解データとされた箇所の集合を A_x，その集合に含まれる要素の個数を $n(A_x)$，ある内容抽出法により内容が含まれる箇所として抽出された，かつ正解箇所とされた箇所の集合 $C_x \cap A_x$ に含まれる要素の個数を $n(C_x \cap A_x)$ とすると，Web テキスト s_x における評価値 f_x は $f_x = \frac{n(C_x \cap A_x)}{\frac{n(C_x)}{2}+\frac{n(A_x)}{2}}$ となる．したがって，今回用いる評価指標である評価値の平均 $\overline{F}$ は $\overline{F} = \frac{1}{n}\sum_{n} f_x$ となる．

4.2 テストコレクションの構築

本稿で行う評価実験では人手で作成したテストコレクションを用いる．この評価実験に用いるテストコレクションは，以下の三種類のデータから成る．

- Web ページ群
- 内容を形成する単語群
- 各 Web テキストから人手により決定した内容が含まれると判断される箇所 (正解箇所)

本稿ではテストコレクションとして用いる Web テキスト群として「ある一塊の Web ページ群 (Web サイト)」を用いる．これは，Web サイト中に含まれる多くの Web ページ同士はハイパリンクを用い，ページ間を行き来することができるためである．本稿では「同志社大学大学院文化情報学研究科 (注1)」の Web サイトを用いて正解箇所を人手で作成した．

また，内容を形成する単語群として，本稿では同志社大学大学院文化情報学研究科のトップページ中の description タグに含まれる名詞 (注2) のうち，文化情報学研究科の Web サイトで取り扱っている内容を表す単語群としてふさわしいと考えられる「文化 情報 研究 科」を用いた．

さらに，正解箇所は Web ページ群中に含まれる各 Web ページ中のどの部分に内容を形成する単語群中の内容が含まれているかを実験協力者が以下の手順に従って判断した．その際，一つの Web ページに対して，3 人の実験協力者が内容が含まれているか否かの判断を行う．

（1） 実験協力者がトップページから Web ページの閲覧を行い，対象とする内容が各 Web ページに含まれているかを尋ねる．

（2） 実験協力者が検索課題に関する内容が，現在着目している Web ページに含まれていると評価した場合，その内容が Web ページのどの位置に含まれているかを単語単位でマウスにて選択してもらう．

（3） 選択が終了した後に，次の Web ページに移動する．

そして，3 人中 2 人以上の実験協力者により選択された箇所を正解箇所とした．

なお，この正解箇所の算出においては，同じ Web ページに対しての評価であったとしても，対象となる Web ページの前に閲覧していた Web ページが異なる場合は，リンク元が異なるため別の評価として取り扱った．

4.3 実 験 結 果

本稿では，4.2 で述べたテストコレクションを用い，3. で述べたハイパリンクを考慮した内容密度分布と，テキストのみを考慮した従来の内容密度分布を内容抽出法として用いてテストコレクション中の Web サイトに含まれる Web ページに対して内容抽出を行った．

4.3.1 単純なリンク元内容反映手法を用いた実験結果

まずは．単純なリンク元内容反映手法を用いた実験結果について述べる．ここでは，閾値 E を 0.1 ずつ変化させ，内容抽出にふさわしい内容密度分布はハイパリンクに関するどの要素 (最大値，中央値，最小値) および手法 (乗算，加算) の組合せを考慮したものであるかを判断する．

各要素および単純なリンク元内容反映手法を用いた内容抽出結果に対して評価実験を行ったところ，各要素および手法を用いた内容密度分布により算出された評価指標 $\overline{F}$ は図 5 のようになった．

図 5 より，$E \leqq 0.7$ とした場合は従来手法の方が単純なリンク元内容反映手法を用いた提案手法全種よりも $\overline{F}$ が高いことがわかる．これは，$E \leqq 0.7$ とした場合は，従来手法では網羅的に正解箇所を抽出することができているが，$E > 0.7$ とした場合はユーザにより内容が存在すると判断される箇所が，従来手法により内容であると判断される箇所から除外されてしまうためであると考えられる．

一方で，$E > 0.7$ とした場合は従来手法や手法を加算とした場合よりも，乗算とした場合の方が $\overline{F}$ が高いことがわかる．なお，図 5 より手法を乗算とした場合は，手法を加算とした場合よりも常に $\overline{F}$ が高いことがわかる．これは，手法を乗算とした場合は従来よりも内容の出現箇所における内容密度分布の値を増加させる効果があるため，内容密度分布の値を強調するという形でハイパリンクにより内容が存在する箇所を際立たせることができるからであると考えられる．

また，図 5 より，$E = 0.7$ とした場合に各要素および手法を

(注1)：http://www.cis.doshisha.ac.jp/gs/，2013 年 1 月 17 日閲覧

(注2)：「同志社，大学，文化，情報，学，研究，科，ホームページ」 の 8 種類．2013 年 1 月 17 日閲覧

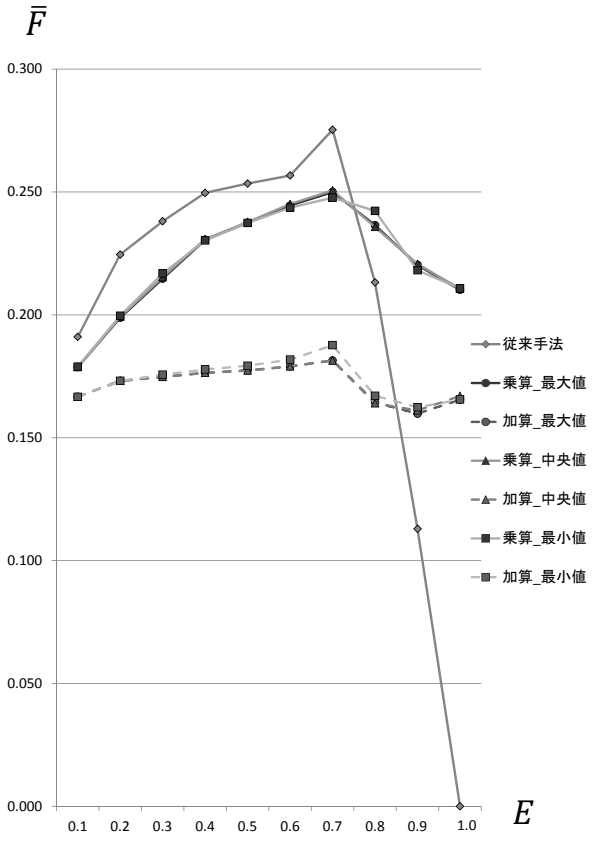

図 5　各要素および手法を用いた内容密度分布より算出された $\overline{F}$

用いた内容密度分布より算出された $\overline{F}$ は最大となる．$E = 0.7$ とした場合における $\overline{F}$ の値は表 1 のようになる．

表 1　各要素および手法を用いた内容密度分布より算出された $\overline{F}$ の最大値

	最大値	中央値	最小値
乗算	0.250	**0.251**	0.248
加算	0.181	0.181	0.188

従来手法	**0.275**

表 1 より，$E = 0.7$ とした場合は従来の内容密度分布から算出された $\overline{F}$ が一番高く，次に要素を中央値，手法を乗算とした内容密度分布から算出された $\overline{F}$ が，他の要素を用いた内容密度分布から算出された $\overline{F}$ よりも高いことがわかる．これは，最大値や最小値は分布における外れ値になる場合がある分位数であるが，中央値は分布の偏りを考慮した分位数であるため，アンカーテキスト上に存在する内容密度分布に外れ値があった場合を考慮できているためであると考えられる [15].

本来は正確性と網羅性をどちらも考慮した内容抽出法が望ましいため，正確性単体および網羅性単体での評価は行わないが，このような実験結果が出た要因を考察するために，正確性および網羅性を以下の式のように定義する．

- 正確性：$\dfrac{n(C_x \cap A_x)}{n(C_x)}$
- 網羅性：$\dfrac{n(C_x \cap A_x)}{n(A_x)}$

実験データより，網羅性はハイパリンクを考慮しない従来の内容密度分布から算出された $\overline{F}$ よりも，ハイパリンクを考慮した内容密度分布から算出された $\overline{F}$ の方が高い．しかし，正確性ではハイパリンクを考慮した内容密度分布から算出された $\overline{F}$ よりも，ハイパリンクを考慮しない従来の内容密度分布から算出された $\overline{F}$ の方が高い．これは表 2 に記した $E = 0.7$ とした場合の正確性および網羅性からもわかる．よって，内容密度分布の値を増加させたことにより，従来手法では内容が存在しないと判断されるような，元々内容密度分布の値が極端に小さい箇所を，内容が存在するとされる箇所として抽出してしまうことがあることがわかる．

表 2　各要素および手法を用いた内容密度分布より算出された正確性，網羅性，$\overline{F}$

要素	手法	正確性	網羅性	$\overline{F}$
乗算	最大値	0.172	0.783	0.250
乗算	中央値	0.173	0.781	**0.251**
乗算	最小値	0.172	0.742	0.248
加算	最大値	0.108	0.883	0.181
加算	中央値	0.108	0.883	0.181
加算	最小値	0.112	0.870	0.188
従来手法		0.220	0.600	**0.275**

4.3.2　二重の閾値を使用したリンク元内容反映手法を用いた実験結果

リンク先の内容密度分布に対して要素となる代表値を考慮する前にも新たな閾値を設け，さらにハイパリンクを考慮した内容密度分布の値に対しても閾値を設けることができれば，「内容密度分布の値が極端に小さい箇所」を考慮することができる可能性がある．しかし，4.3.1 で行った実験より，加算手法は従来手法では内容が存在しないと判断されるような，元々内容密度分布の値が極端に小さい箇所を，内容が存在するとされる箇所として抽出する可能性があるため，適切でないことがわかる．また，各要素においても中央値を用いる場合が適切であることが 4.3.1 の結果からわかる．よって，3.2.2 にて提案した二重の閾値を使用したリンク元内容反映手法に対して，要素として中央値，内容反映手法として乗算を用いた実験を行う．

閾値 E_1 および E_2 に対して評価実験を行ったところ，各要素および各閾値を用いた内容密度分布により算出された評価指標 $\overline{F}$ は表 3 のようになった (注3)．表 3 より，$E_1 = E_2 = 0.7$ のとき $\overline{F}$ は最大となる．また，この値は従来のハイパリンクを考慮しない内容密度分布に対して $E = 0$ を用いた結果，すなわち $E_1 = 0.7, E_2 = 0.0$ とした際の結果と同じ値になる．これは，図 5 において全要素及び手法を用いた場合においても $\overline{F}$ が最大になることから，閾値 0.7 が今回用いたテストコレクション中における Web ページ群における閾値として適切であるためであると考えられる．

(注3)：$E_1 = 0$ の場合は単純なリンク元内容反映手法を用いた実験結果と，$E_2 = 0$ の場合は従来の内容密度分布による実験結果と同様であるため，表 3 からは省略した．

表 3　二重の閾値を使用したリンク元内容反映手法による $\overline{F}$

$E_1 \backslash E_2$	0.1	0.2	0.3	0.4	0.5	0.6	**0.7**	0.8	0.9	1.0
0.1	0.190	0.198	0.214	0.230	0.237	0.244	0.249	0.235	0.220	0.210
0.2	0.224	0.224	0.225	0.230	0.237	0.244	0.249	0.235	0.220	0.210
0.3	0.237	0.237	0.237	0.239	0.241	0.243	0.246	0.232	0.219	0.208
0.4	0.247	0.247	0.248	0.249	0.251	0.252	0.252	0.227	0.205	0.197
0.5	0.250	0.250	0.251	0.252	0.253	0.254	0.254	0.225	0.194	0.185
0.6	0.253	0.253	0.254	0.254	0.255	0.256	0.256	0.226	0.190	0.185
0.7	0.270	0.270	0.271	0.272	0.273	0.274	**0.275**	0.245	0.209	0.192
0.8	0.211	0.211	0.212	0.213	0.214	0.215	0.216	0.213	0.153	0.116
0.9	0.121	0.121	0.122	0.123	0.124	0.125	0.126	0.123	0.113	0.060
1.0	0.014	0.014	0.015	0.016	0.017	0.018	0.019	0.015	0.005	0.000

5. おわりに

本稿では，従来の Web テキストにおける内容抽出法に対し，ハイパリンク上に存在するデータを適用することにより，Web テキストとハイパリンクを考慮した Web ページにおけるアンカーテキストを利用したユーザの判断に近い内容抽出法を提案した．

4. の実験を行うことにより，ハイパリンクに関する要素と手法の組合せを用いた内容密度分布のうち，ユーザの判断に一番近いものはどの組合せを用いた内容密度分布であるか検討を行った．その結果，要素を中央値，手法を乗算とした内容密度分布が Web ページにおけるハイパリンクを考慮した内容抽出法としてふさわしいことがわかった．また，内容抽出のために二重の閾値を設けることにより，従来のハイパリンクを考慮しない内容密度分布を内容抽出法として用いた場合と同等の大きさとなる評価指標の値を得ることができる．したがって，要素を中央値，手法を乗算とした二重の閾値を用いた内容密度分布を内容抽出法として用いることが望ましいことがわかった．

今後は，内容密度分布の適切な閾値を自動的に決定する方法を考えた実験を行う必要がある．また，適切な要素の組合せを用いた Web 検索支援システムを作成し，ユーザの Web 検索を支援を行う．具体的には最適な要素の組合せを用いた内容密度分布を用い，Web ブラウジングの際にユーザが手動で入力した内容に対する内容に関する出現位置および影響度を可視化する方法を考えている．例えば， UserHeat (注4) など，既存の Web ページ中におけるユーザによって読まれた部分を可視化する手法で用いられているような，内容の出現位置および影響度を踏まえて Web ページ中の背景色を変更する手法が考えられる．そして，以上の手法により内容密度分布を Web 検索結果および Web ブラウジングと併用することにより，検索およびブラウジングする人が必要とする情報に簡単にアクセスできるようにすることを最終目標とする．

謝辞　本研究の一部は，独立行政法人日本学術振興会 科学研究費補助金 基盤研究 (A) (課題番号：22240005) 及び，独立行政法人日本学術振興会 科学研究費補助金 若手研究 (B) (課題番号：22700248) によるものである．ここに記して謝意を表す．

(注4)：http://userheat.com/, 2012 年 1 月 3 日閲覧

文　献

[1] C. D. Manning, P. Raghavan, and H. Schuetze. *Introduction to Information Retrieval*. Cambridge University Press, 2008.

[2] 西原陽子, 佐藤圭太, 砂山渡．光と影を用いたテキストのテーマ関連度の可視化．人工知能学会論文誌, Vol. 4, No. 2, pp. 479–487, 2009.

[3] L. Page, S. Brin, R. Motwani, and T. Winograd. The PageRank Citation Ranking: Bringing Order to the Web. Technical Report 1999-66, Stanford InfoLab, 1999.

[4] J. M. Kleinberg. Authoritative sources in a hyperlinked environment. *Journal of the ACM*, Vol. 46, No. 5, pp. 604–632, 1999.

[5] 田村航弥, 波多野賢治, 宿久洋．リンク情報に基づく周辺文書の索引語尤度を考慮した文書検索手法の提案と評価．第 3 回データ工学と情報マネジメントに関するフォーラム (DEIM Forum 2011) 論文集, 2011.

[6] 阿部匡史, 豊田正史, 喜連川優．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．電子情報通信学会第 14 回データ工学ワークショップ (DEWS2003) 論文集, 2003.

[7] 荒木良, 是津耕司, 角谷和俊, 田中克己．リンク参照と文書構造に基づく Web ページのアスペクト抽出．電子情報通信学会第 14 回データ工学ワークショップ (DEWS2003) 論文集, 2003.

[8] A. Woodruff, A. Faulring, R. Rosenholtz, J. Morrsion, and P. Pirolli. Using thumbnails to search the Web. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 198–205. ACM, 2001.

[9] S. Kitahara, K. Tamura, and K. Hatano. Extraction of the contents in the web texts by content-density distribution. *International Journal of Knowledge Engineering and Soft Data Paradigms*, Vol. 3, No. 2, pp. 108–120, 2011.

[10] 上嶋宏, 三浦孝夫, 塩谷勇．同義語, 多義語の考慮による文書分類の精度向上．電子情報通信学会論文誌, Vol. 87, No. 2, pp. 137–144, 2004.

[11] 鄭躍軍, 金明哲, 村上征勝．データサイエンス入門．勉誠出版, 2007.

[12] 杉山一成, 波多野賢治, 吉川正俊, 植村俊亮．ハイパリンクで結ばれた隣接ページの内容に基づく Web ページのための TF-IDF 法の改良．電子情報通信学会論文誌. D-I, Vol. 87, No. 2, pp. 113–125, 2004.

[13] 北原沙緒理, 波多野賢治．単語位置を考慮した単語単位で行う Web テキストの内容抽出に対する一考察．平成 24 年度情報処理学会関西支部支部大会講演論文集, 2012.

[14] R. Baeza-Yates and G. Navarro. *Modern Information Retrieval*. Addison-Wesley, second edition, 2011.

[15] 白旗慎吾．統計解析入門．共立出版, 1992.