Web テキストにおける内容密度分布の抽出とその評価

北原沙緒理† 田村 航弥†† 波多野賢治†

† 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
†† 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
E-mail: †{kitahara,tamura}@ilab.doshisha.ac.jp, ††khatano@mail.doshisha.ac.jp

あらまし 現在 Web 検索結果と同時に表示されるスニペットは， Web テキストにおけるクエリの出現位置周辺に存在するテキストしか表示されない．よって，スニペットだけでは検索結果のリスト内においてクエリに関する内容が取り扱われているかを判断することが困難である．そこで，本稿では検索結果として表示された Web テキストに対して，クエリに該当する内容の出現範囲及び局所的な内容の影響度変化を抽出することを提案する．そのために，クエリに関する内容の出現範囲及び局所的な内容の影響度変化を表す内容密度分布を抽出し，最終的にスニペットが抱える上記の問題を解決する．また作成した内容密度分布と既存のスニペットに対してクエリに該当する内容出現範囲との解答の合致率を比較することにより内容密度分布により抽出される内容の妥当性に関する評価を行った．

キーワード Web テキスト，テキストの内容把握

Extraction of the Meaning–Density Distribution from a Web Document

Saori KITAHARA†, Koya TAMURA††, and Kenji HATANO†

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

1. はじめに

現在，World Wide Web (WWW) 上にはさまざまな種類の内容を含む膨大な Web ページが存在している．そして誰もが容易に Web ページを作成及び公開することが出来るため，WWW 上に存在する Web ページの数はこれからも増加することが予想される．よって膨大な Web ページの中からユーザが必要とする情報が含まれた Web ページを探すための技術を開発することが重要となる．その技術の一つとして，ユーザが必要とする情報に近いクエリを入力し，そのクエリに近い Web ページを算出するという Web 検索がある．しかし Google(注1) など既存の検索エンジンを用い Web 検索を行った場合，検索結果として提示される Web ページにユーザが必要とする情報が含まれていない場合がある．その要因として図 1 のように Web 検索に用いたクエリが，同一 Web ページ内に存在するテキストデータ (Web テキスト) 内にある別々の内容に含まれていることや，Web ページ内にユーザが必要とする情報が含まれると考えられる内容がほとんど含まれないことが挙げられる．また，一つの Web ページ内に存在する内容は一つだけであるとは限らない上に，同一の内容が一つの Web ページ内の複数箇所に存在する場合もある．

よって従来の Web 検索だけでは Web テキスト内に含まれる内容が Web テキスト中のどの範囲に存在しているかを把握することが出来ない．したがって従来の Web 検索結果に加えて，Web テキストに存在する内容の範囲だけでなく，Web テキスト上のある内容における，ある位置での内容の強さ (影響度) を求めることが出来れば，ユーザが Web テキスト上の内容を把握することを助けることが出来ると考えられる．そこで本稿ではクエリとして用いた単語 (検索ワード) の組が一つの内容を形成していると見なし，それらが形成する内容の出現範囲及び影響度を考慮した Web テキストの内容密度分布を作成することを提案する．

以下 2. において Web テキストの内容抽出に関する関連研究について述べ，その後 3. において関連研究の問題点を踏まえ

(注1)：http://www.google.com/

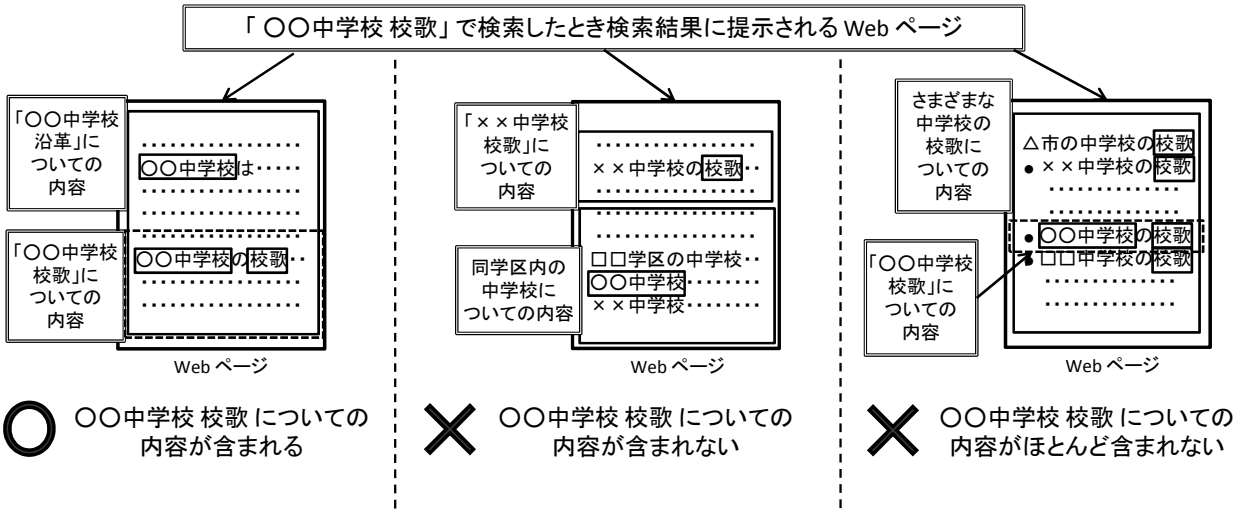

図 1　クエリを 「〇〇中学校」「校歌」として検索を行った例

た Web テキストの内容密度分布作成について提案を行う．その後 4. において Google スニペットと作成した内容密度分布に対して比較を行うことにより，内容密度分布に対して評価を行う．最後に 5. において本稿のまとめと，内容密度分布に関する今後の展望について述べる．

2. 関 連 研 究

本節ではまず， Web テキストの内容把握に関する関連研究として， Web テキストの内容が変化する位置で文章を分割するテキストセグメンテーションに関する研究について述べる．

阿部らは Web テキストを文単位で捉え，各文中の単語及びその関連語に着目したテキストセグメンテーションの研究を行っている [1]．従来のテキストセグメンテーションでは文書中の各文章に含まれる単語の種類及び数の推移を考慮し文書を分割していた．しかし従来の文章に含まれる単語のみを用いたテキストセグメンテーションでは関連語を考慮していなかったため，前の文と関連する単語が次の文に含まれていても，同一内容の文と見なされないことがあった．しかしこの手法では同一の内容が離れた位置にあることを考慮する場合や，各内容間において局所的に他の内容と関連度が変動する場合を考慮することが出来ない．具体例を挙げて説明すると，図 2 の場合は内容 2 の中に内容 1 と関連度が高い部分が存在する．テキストセグメンテーションではこの部分は内容 2 の内容とされるため，内容 1 の内容を含むことを考慮されない．本稿で用いる手法では同一の内容が離れた位置にあることも考慮することが出来る．

田らは局所的出現密度という値を算出することにより， Web テキストの内容出現範囲及び影響度を捉え，Web 検索結果の並べ替えることで検索精度を向上する研究を行っている [2]．局所的出現密度は， Web テキスト内に出現した検索ワードのうち最初に出現した単語から最後に出現した単語までの間に検索ワードの影響度が存在するという考えにより算出され，その値は一定となる．しかし実際の検索ワードにおける内容の影響度は， Web テキスト上に現れる単語の密集度により局所的に変化すると考えられる．局所的出現密度は検索ワードのうち最初

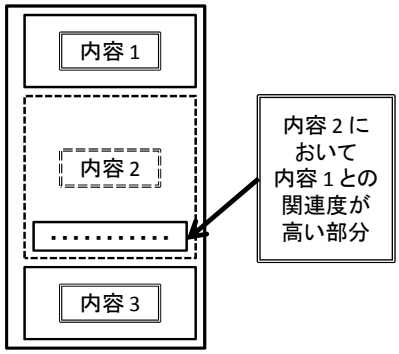

図 2　テキストセグメンテーションの問題点

に出現した単語から最後に出現した単語までの間の総単語数と，その間の検索ワードの出現回数が同じであれば一定の値をとるため，単語距離内における単語の密集度を考慮することが出来ない．本稿で用いる手法では，単語の出現位置により内容の出現影響度が異なるため単語の密集度を考慮した内容抽出を行うことが出来る．

ここで，Web テキスト全体における内容出現範囲及び影響度を捉える研究について述べておく．Boudin らは医療関係の文書を十等分し，各部分文書ごとに PECO と呼ばれる要素が現れる割合を可視化する研究を行っている [3]．PECO の要素は患者の年齢や性別など (P : Patient-problem)，主要な処置方法 (E : Exposure-intervention)，比較対処の処置方法 (C : Comparison)，主要な処置方法による処置結果 (O : Outcome) から成る．しかし，この手法では各部分文書内における PECO の要素に関する内容の範囲及び影響度を把握することができない．本稿で提案する手法では Web テキスト内の各単語の出現位置において，内容の範囲及び影響度を把握することができる．

佐野らは単語の出現密度分布及び出現密度分布から算出される単語間の関連度を用いることで，自身らが提唱した Web 文書検索法である極小マッチ部分グラフによる検索法を改善する研究を行っている [4]．佐野らの手法は単語の出現密度分布は

単語に注目しているため，Boudin らの手法より更に細かい粒度で Web テキスト全体の内容出現範囲及び影響度を捉えることができる．なお極小マッチ部分グラフの特殊な例がクエリに複数の単語を用いた検索である．単語の出現密度分布とは単語が持つ内容の出現範囲及び影響度を捉えるための分布であり，単語の出現位置とハニング窓関数により算出される．単語の出現密度分布の値が高い箇所はその単語の影響力が強い場所となる．しかしハニング窓関数は前後の内容を考慮せず単語の出現密度分布を算出してしまうため，その単語の前後において内容が変化した場合においても，単語の影響があると判断してしまう．よってハニング窓関数を用いて Web テキストの内容分布を表わす際には不都合が生じる．またこの研究では内容の影響度を算出する基本単位として，一単語ではなく一文字を扱っている．したがって，一単語内で取りうる影響度が単語中の文字数によって変化するため，一単語を形成する文字数により内容の影響度が変化するという点が問題となる．本稿で用いる手法では内容が変化する可能性がある箇所では影響度の値が低下する関数を，内容の影響度を算出する基本単位として一単語をそれぞれ用いる．

なお，検索ワードが出現する一部分を検索する研究としてパッセージ検索が存在する [5]．パッセージ検索の目的は，Web テキストの中でも検索ワードに関連している部分のみを抽出し検索結果として提示することである．しかし本稿で述べる手法は Web テキスト内に含まれる内容のうち，検索ワードに関する内容が出現する部分がどのような位置にどのような影響度で存在するかを求めることが目的であるため，一般的なパッセージ検索とは目的が異なる．

Lv らは Positional Language Model を作成することで，パッセージの大きさに決まりの無いパッセージ検索を実現するための研究を行っている [6]．Positional Language Model は確率的言語モデルを用いた情報検索モデルを拡張したものであり，Web テキストにおける検索ワードの出現位置及び検索ワード同士の近さを抽出することができる．しかし Lv らの研究の目的はパッセージの大きさに決まりの無い,「柔軟な」パッセージ検索を行うことである．実際に Lv らは自身らの論文中でモデルの生成及び，モデルの検索スコアへの応用のみを取り上げている．またこのモデルで算出される数値は，ある位置における単語のカウント数の推定値である．つまりこのモデルは，ユーザが Web テキスト内の検索ワードの出現範囲及び影響度を把握するためには使用されていない．対して本稿で提案する手法では，Web テキスト内の単語が形成する内容を把握することが目的であるため，関連研究の目的とは異なる．また，単語のカウント数の推定値を内容の影響度だと見なすと，Lv らが提案した手法では単語のカウント数が 0 となる箇所が存在しない．したがって内容の影響度が 0 となる部分，すなわち内容が存在しない部分を抽出することができないため，本稿で提案する手法と比較することができない．なおこの手法を適合性フィードバックに用いた Positional Relevance Model も存在するが，これは適合性フィードバックの精度向上のための研究であるため，本稿で提案する手法とは目的が異なる [7]．

3. 提 案 手 法

本稿では関連研究の問題点を踏まえ， Web テキストの内容密度分布を作成することを提案する． Web テキストの内容密度分布とは Web テキストにおける内容の出現箇所及び影響度の強さを反映した分布である．Web テキストの内容密度分布を作成する手順は以下の通りである．

（1） Web ページ内の単語の抽出
（2） 検索ワードにおける単語密度分布の作成
（3） Web テキストにおける内容密度分布の作成

ここで，単語密度分布とは各単語の影響が及ぶ範囲及び強さを反映した分布である．なお本稿では Web テキストを構成する名詞，動詞などの単位である一形態素を一単語として扱い，Web テキスト内における形態素数を Web テキストの単語数とする．そして今回取り扱う「内容」を検索ワードの集合とする．

3.1 Web テキスト内の単語の抽出

まず前述の条件を満たしたクエリを用いて Web 検索を行う．その結果表示された Web テキストに対して形態素解析を行う．その際活用がある単語は終止形に変換する．しかし既存の辞書には日々生み出される新語や造語が登録されておらず，これらの品詞を特定することが出来ない．よって，今回は形態素解析及び終止形への変換には辞書やコーパスに依存しないかつ，辞書に登録されていない未知語に対する品詞の類推を行うことが出来るという特徴を持つ形態素解析器 MeCab を用いた(注2)．

3.2 検索ワードにおける単語密度分布の作成

次に検索ワードにおける単語密度分布の値を算出する．ここでは Web テキストを一本の文字列と見なし Web テキストにおける個々の単語を単位とする．

まず Web テキスト一つ一つについて各検索ワードが出現する位置前後における単語密度分布を作成する．そしてその単語密度分布の値を検索ワード一種類ずつにおいて合成することにより，検索ワードごとの単語密度分布が作成される．ここである Web テキストに存在する検索ワード t_i の一つを t_i^j とすると，まずある地点 k における単語密度分布の値 $hw_{t_i^j}(k)$ を算出することになる．その際，各単語の影響が強い部分はその単語の出現箇所であると考えられるため，先頭から t_i^j の出現箇所を $l_{t_i^j}$ とすると，$hw_{t_i^j}(k)$ は $l_{t_i^j}$ において最大値をとる．また，出現箇所 $l_{t_i^j}$ から離れれば離れるほど，単語 t_i^j が及ぼす影響度も減少すると考えられるため $hw_{t_i^j}$ の値は $l_{t_i^j}$ から離れれば離れるほど減少する．

さらに，文中において内容が変化する可能性がある箇所は句読点 (。) などの文の区切り候補であると考えられるため，その部分において文の区切りにおいて影響度の重みが変化するようにする．ここで Web テキストにおいて $l_{t_i^j}$ の直前に存在する内容の切れ目となりうる記号が出現する位置を $a_{t_i^j}$, $l_{t_i^j}$ の直後に存在する内容の切れ目となりうる記号が出現する位置を $b_{t_i^j}$ とすると，$a_{t_i^j}$ から $b_{t_i^j}$ までの間に存在する単語は一つの文章を形成し， $l_{t_i^j}$ は その中の一単語として扱われる．よって，$a_{t_i^j}$

(注2)：http://mecab.sourceforge.net/

及び $b_{t_i^j}$ よりも $l_{t_i^j}$ から離れた位置にある単語密度分布の値はより減少すると考えられる．今回は文の区切り候補として，句読点 (。) と全角及び半角のピリオド (.)，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．以上の点を考慮し，今回は単語密度分布の算出に以下の関数を使用する．

$$hw_{t_i^j}(k) = \begin{cases} \frac{1}{2}(1+\cos 2\pi \frac{k-l_{t_i^j}}{W}) & (a_{t_i^j} < k < b_{t_i^j}) \\ \frac{1}{2}(1+\cos 2\pi \frac{k-l_{t_i^j}}{W}) \times S & (a_{t_i^j} \geqq k,\ b_{t_i^j} \leqq k) \end{cases} \quad (1)$$

(ただし $| k - l_{t_i^j} | \leqq \frac{W}{2}$, $0 \leqq S \leqq 1$)

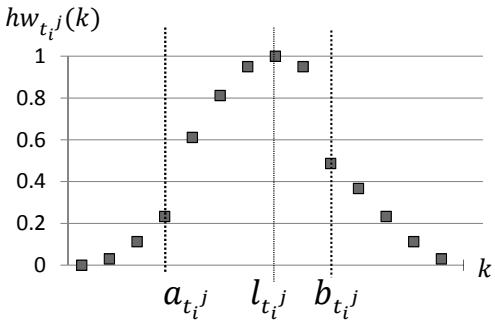

図 3　重みつきハニング窓関数

式 (1) は佐野らの手法で用いられていたハニング窓関数に重みを付与した関数 (重みつきハニング窓関数) であり，単語の出現位置において図 3 のような値をとる．ここで W は窓の幅と呼ばれ，検索ワードの影響が及ぶ範囲を表す．ただしこの関数の影響が及ぶ範囲は $| k - l_{t_i^j} | \leqq \frac{W}{2}$ の間のみである．また重み S がとりうる値の範囲は $0 \leqq S \leqq 1$ とする．なお今回ハニング窓関数を用いた理由は，単語の出現位置から値が緩やかに減少する関数であり，関連研究として挙げた佐野らの研究において使用されていたためである．

次に $hw_{t_i^j}$ を統合し，検索ワード一種類ずつの単語密度分布 hw_{t_i} を算出する．まず各検索ワードごとにそれらの出現箇所における単語密度分布の値である $hw_{t_i^j}(k)$ を足し合わせる．その後足し合わせた値の最大値で，各検索ワード出現箇所における $hw_{t_i^j}(k)$ を足し合わせた値を除算することにより算出した値を単語密度分布とする．つまり単語密度分布は以下の式 (2) で表すことが出来る．単語密度分布を最大値で割り，単語密度分布の最小値を 0，最大値を 1 に正規化することにより検索ワードの出現個数によらず単語密度分布の値を比較することが出来る．検索ワード一つ一つにおける単語密度分布の算出方法ををを図 4 に記す．図 4 では単語 t_i の出現箇所が 2 箇所だった場合の単語密度分布の算出手順を示している．

$$hw_{t_i}(k) = \frac{\sum_j hw_{t_i^j}(k)}{\max_k \sum_j hw_{t_i^j}(k)} \quad (2)$$

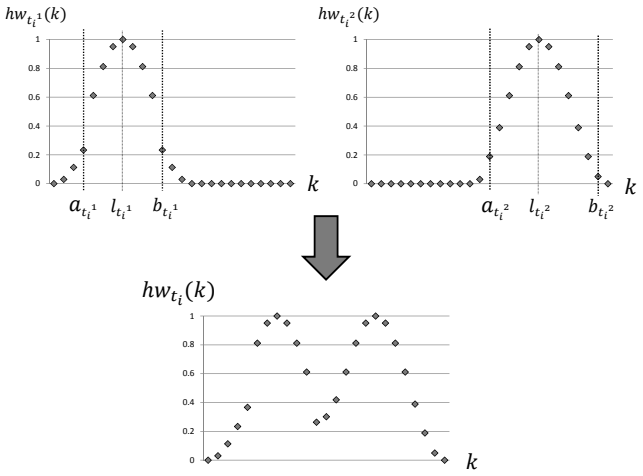

図 4　単語密度分布の算出

3.3　Web テキストにおける内容密度分布の作成

検索ワード一つ一つにおいて単語密度分布 hw_{t_i} を作成し，これらを統合することにより，検索ワードの組 q ($q \ni t_i$) における内容密度分布を作成する．検索ワードが一単語の場合は，単語密度分布を内容密度分布として用いる．しかし検索ワードが二単語以上のときは Web テキスト内において検索ワードの組が内容を形成しない場合もある．よってこのとき内容を形成するか否かは各単語が単語密度分布の影響力をお互いに及ぼしあう範囲 (内容の範囲) が存在するか否かを用いて判断する．そして内容を形成すると判断された単語の組に対して，内容密度分布を作成する．

例えば「同志社」「大学」というクエリを用いて検索を行った際に提示されたある Web テキストに対して，図 5 のような単語密度分布が算出されたとする．このとき「同志社」の単語密度分布と「大学」の単語密度分布は内容の範囲を持つので，この Web テキストには「同志社 大学」という内容が含まれる．

内容密度分布を作成する際，内容の範囲に含まれる各単語出現箇所における単語の強さを足し合わせるだけでは， Web テキスト内において検索ワードが出現する回数が大きく異なる場合，単語ごとの単語密度分布を比較することが出来ない．よって各単語における内容の範囲に含まれる影響度の強さを足し合

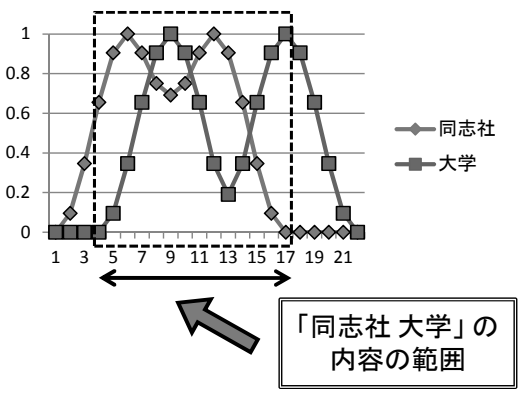

図 5　「同志社」と「大学」の単語密度分布

わせ，各グループに属する単語種類の数で除算し平均を取ることにより内容密度分布を作成する．したがって検索ワードの組 q $(q \ni t_i, i = 1, 2, \cdots, n)$ における内容密度分布は式 (3) で表すことが出来る．なお内容密度分布の算出方法をを図 6 に記す．図 6 ではクエリが検索ワード t_1 及び t_2 であった場合の内容密度分布の算出手順を示している．

$$hw_q(k) = \begin{cases} \frac{\sum_i hw_{t_i}(k)}{n} & (hw_{t_i}(k) > 0, i = 1, 2, \cdots, n) \\ 0 & (others) \end{cases} \tag{3}$$

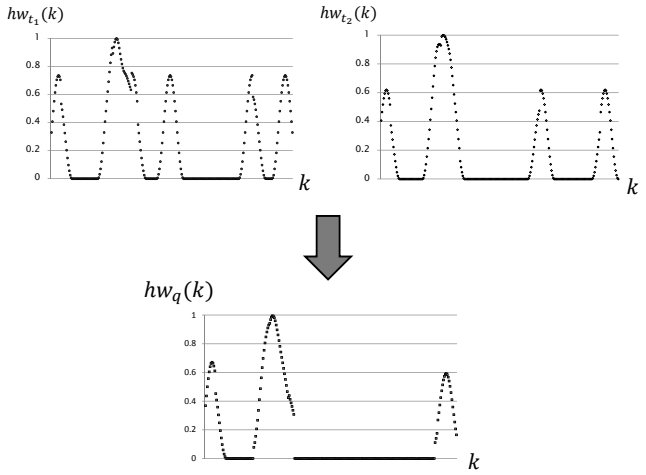

図 6　内容密度分布の算出

4. 評 価 実 験

評価実験では内容密度分布の範囲及び影響度を別々に評価し，内容密度分布の妥当性について考察を行う．本節ではまず評価実験に用いるデータの収集方法について述べる．続いて収集したデータを用いて内容密度分布の範囲が妥当であるかを既存の手法と比較することにより述べる．今回は既存の手法として Google のスニペットを用いる．既存の手法に Google のスニペットを用いた理由は，Google のスニペットが Web ページにおける検索ワードに関する内容を含む部分として，ユーザにより日常的に使用されているためである．Google の検索結果及びスニペットの抽出には Google AJAX Search API (注3) を用いた．

最後に内容密度分布における影響度が妥当であるかについて考察を行う．なお，今回は予備実験を行った結果最も内容と内容密度分布の値が当てはまりがよかったため，窓の幅 W を 0.6，重み S 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 倍とした．

(注3)：http://code.google.com/more/

4.1 評価データの収集

本節では評価実験に用いるデータの収集方法について述べる．評価実験に用いるデータは Web 検索をした結果表示された Web テキストを元に人手で作成する．

しかし形成出来る内容の組合せが多くなりすぎると，各々の内容についての評価が必要となり，人手で評価を行う際に評価を行う実験協力者の負担が増加する．また一形態素の検索ワードで形成されたクエリを用いた場合，その単語が多義語であった際に，実験協力者が単語の多義性を踏まえた評価を行う場合がある．つまり，テキスト内の他の単語を踏まえて内容を補完することにより，純粋にその単語があらわしている内容に対する評価にならない場合がある．よって評価データを収集する際に実験協力者にかかる負担を軽減するため，取り扱う「内容」を二形態素の検索ワードで形成されたクエリとする．そして実験協力者に条件を満たすクエリを考えてもらい Google を用いて検索してもらった．その結果として表示される Web ページの上位 8 件から Web テキストを抽出して作成した評価用の Web ページを表示し，その各々に対して後述の内容の範囲及び影響度に関する評価を行ってもらう．実験に検索結果上位 8 件の Web ページを用いた理由は，Google AJAX Search API で取得できる Web ページが検索結果上位 8 件までであるためである．抽出した Web ページ内にテキストが存在しない場合，Web テキストが抽出できないため，このような Web ページは評価対象から除外する．なお，今回評価に使用した Web テキスト数は 110 である．

4.2 内容密度分布の範囲に対する評価

内容密度分布の範囲に対する評価では，まず実験協力者が選んだクエリが内容として各検索結果の Web テキストに含まれているかを尋ねる．実験協力者がクエリが内容として含まれていると評価した場合，その内容が Web テキストのどの位置に含まれているかを単語単位で選択してもらう．ここで選択してもらった単語の範囲を，範囲に対する評価の正解データとする．

今回はこの正解データを基にして内容密度分布の範囲と検索エンジン Google のスニペットとの比較を行った．その際，評価指標として解答合致率を用いた．解答合致率とは実験協力者が内容が含まれるとした単語数と同一箇所の Web テキスト中の内容密度分布に含まれる単語数との一致度であり，式 (4) を用い算出される．

$$\frac{n(A_{d_m} \cap E_{d_m})}{n(A_{d_m})} \tag{4}$$

ここで A_{d_m} を Web テキスト d_m における評価データに対する解答部分の集合とし，その解答が出現する部分の個数を $n(A_{d_m})$ とする．また E_{d_m} を各評価における影響度が存在する部分の集合とし，$n(A_{d_m} \cap E_{d_m})$ は各評価の影響度及び評価データの解答が出現する部分の個数とする．例えば内容密度分布に対する解答合致率は，解答が出現していてかつ内容密度分布の値を取る単語数を，解答が出現している位置の単語数で除算したものである．それぞれの評価手法においてこの指標を用いることにより，各 Web テキスト内で内容が存在している位置を抽出することができた箇所の割合を測定することができる．

例えばある評価手法における，ある Web テキストの解答合致率が 1 である場合，その評価手法では内容が存在している位置を全て抽出することができたと考えられる．

しかし，スニペットに含まれる単語数は内容密度分布と評価データに含まれる単語数よりも少ない場合がある．また，内容密度分布はスニペットに含まれる単語数よりもはるかに多い単語数で形成されている場合，適切に比較出来ない可能性がある．例えば評価データに含まれる単語数が極端に少なく，内容密度分布に含まれる単語数が多かった場合，$A_{d_m} \cap E_{d_m}$ となる部分がある可能性は単語数が多い分だけスニペットよりも内容密度分布の方が高い．このために条件付提案手法という新たな評価指標を加える．よって比較を行うために，各 Web テキストの正解データに対して内容密度分布 (提案手法)，スニペットの単語数と内容密度分布の範囲に含まれる単語数をそろえた内容密度分布 (条件付提案手法)，スニペット (従来手法) における解答合致率を算出した．そして，各手法における解答合致率の平均値 (平均解答合致率) により，各手法の範囲に対する比較を行った．

表 1　各手法における平均解答合致率の比較

	提案手法	条件付提案手法	従来手法
平均解答合致率	0.4192	0.1705	0.1514

表 1 より内容密度分布の範囲に対する評価においては提案手法及び条件付提案手法のほうが従来手法よりもユーザが挙げた解答に合致していることが分かった．したがってこの評価により，提案手法や条件付提案手法の方が従来手法よりもユーザが想定する内容が存在する範囲を抽出できると考えられる．

提案手法が従来手法よりも解答に合致している理由は，従来手法では単語の出現位置前後を表示する KWIC が用いられており，単語の出現位置の前後のみを考慮に入れているが，提案手法では検索ワードが出現する位置の近さも考慮に入れているためであると考えられる [8], [9]．しかし，従来手法は Web 検索結果と併記するために，ある程度使用出来る単語数に制限があるが，提案手法や解答では従来手法ほど使用出来る単語数に制限はない．この理由により従来手法より提案手法の平均解答合致率が高くなった可能性がある．

また条件付提案手法が従来手法よりも解答に合致している理由は，条件付提案手法では提案手法が従来手法よりも解答に合致している理由に加え，Web テキスト中の内容が変化する可能性がある位置で内容の影響度を減少させているため，Web テキスト中における内容の影響度が高い箇所を的確に抽出出来たためであると考えられる．

4.3　内容密度分布の影響度に対する考察

内容密度分布の影響度に対する考察をするために，後述の正解位置における内容密度分布の値と，内容密度分布における最大値との比較を行った．ここで正解位置とは実験協力者が内容密度分布の範囲に対する評価の際に選択した部分の中で「その内容の中心となっている場所」を単語単位で一ヶ所選択した部分である．式 (5) は各テキストにおける内容密度分布の最大値を基準として，正解位置における内容密度分布の値がどれだけ最大値に近いかを把握するための指標である．つまり本節ではこの評価指標を用いて実験協力者が内容の影響度が最大だと考えた場所における内容密度分布の値と実際の内容密度分布の最大値とのずれを比較することにより内容密度分布の影響度に対する考察を行う．また，内容密度分布により内容の影響度が最大であると判断された部分において内容密度分布は最大値を取る．よってこの指標により，それぞれの評価手法が各 Web テキスト内において内容の中心となっている位置での内容密度分布の値が内容密度分布の最大値とどれだけ離れているかを測定することができる．

$$\frac{hw_{d_m \cdot q}(c_{ans})}{hw_{d_m \cdot q}(c_{max})} \tag{5}$$

ここで $hw_{d_m \cdot q}$ を，Web テキスト d_m における，クエリ q に関する内容密度分布，c_{ans} を正解位置，c_{max} を内容密度分布の最大値が出現する位置とする．この値は 0 から 1 までの間で推移し，正解位置における内容密度分布の値が最大値に近ければ近いほど高い値をとる．例えば解答合致率が 1 である場合，Web テキスト内において検索ワードに関する内容の中心となっている位置における内容密度分布の値と，内容密度分布により内容の影響度が最大であると判断された部分における値が等しいことが分かる．

なお正解位置を選択する際，実験協力者はクエリに関する内容が含まれていないという選択することも可能である．この評価には前述の内容密度分布の範囲に対する評価で用いたデータのうち，影響度に対する評価が行われたデータを用いる．

表 2　内容密度分布における影響度の考察

	0	0〜0.2	0.2〜0.4	0.4〜0.6	0.6〜0.8	0.8〜1.0
度数	32	4	2	8	15	28

表 2 によると影響度の抽出が出来たと判断出来る式 (5) の値が 0.8〜1.0 をとるものが多い．実験協力者が内容の中心と選択した位置の多くは，検索ワードの出現位置もしくは検索ワードの密集位置である．一方，内容密度分布は元々単語密度分布を組合せて作成したものである．そして単語密度分布はクエリに含まれる各検索ワードが出現する位置において値が高くなる．よって内容密度分布の影響度が高い部分は，実験協力者が内容の中心と考える部分を捉えられたと考えられる．なお，式 (5) の値が 0 をとるものも多く見られるがこれらのほとんどは内容密度分布における範囲の評価において解答合致率が 0 か 0 に近い値であった．つまり，内容密度分布の範囲と実験協力者が選択した内容の範囲の合致部分が少なかったため，影響度の値をとる部分が少なかったことが，この結果につながったと考えられる．

したがってこの考察により，内容密度分布が内容が存在する範囲を抽出できた場合，内容密度分布の影響度が高い部分も抽出できることが分かる．つまり，この場合ユーザが内容の中心だと思う場所においては内容密度分布の値も高いということが考えられる．

5. おわりに

本稿では，Web テキスト内に存在する内容の出現範囲及び局所的な内容の影響度変化を抽出するために，Web テキストの内容密度分布を作成することを提案した．そして検索ワードの内容密度分布とスニペットに対し，平均解答合致率による内容密度分布の範囲に対する比較を行った．比較を行った結果，提案手法のほうが従来手法よりもユーザが挙げた解答に合致していることが分かった．また従来手法の条件に限りなく近づけた条件付提案手法においても条件付提案手法のほうが良い結果が得られた．さらに内容密度分布の影響度に関する考察により，内容密度分布の範囲がユーザの上げた解答にある程度合致する場合には，Web テキストにおいてユーザが検索ワードに関する内容の中心だと考える位置は内容密度分布においても影響度が高いことが分かった．

今後の課題の一つは内容密度分布の影響度に対する比較評価を行うことである．そのために内容密度分布の影響度を評価するために妥当な指標を考慮し，この点に関しても早急に比較実験を行えるようにする．また検索ワードだけではなくその関連語も用いた内容密度分布を作成することを目標とする．もし関連語も用いた内容密度分布を作成することが出来れば一つの Web テキスト内における様々な内容を抽出することが出来るようになる．その際，抽出される内容が複数あることが予想されるため，優先的に表示する内容を考慮した内容抽出を行う必要がある．

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

文　献

[1] 阿部直人, 内山俊郎, 内山匡, 奥雅博. “ウェブ検索を利用したしきい値選択型テキストセグメンテーション”. 情報処理学会論文誌 (ジャーナル), Vol. 49, No. 12, pp. 4025–4038, 2008.

[2] 田馳, 手塚太郎, 小山聡, 田島敬史, 田中克己. “質問キーワードの意味的関連と近接性に着目したウェブ検索精度”. 電子情報通信学会 第 17 回 データ工学ワークショップ ・ 第 4 回日本データベース学会年次大会 (DEWS2006), 2006.

[3] F. Boudin, J.W. Nie and M. Dawes. “Positional Language Models for Clinical Information Retrieval”. In *Proc. of the 2010 Conference on Empirical Methods in Natural Language Processing*, pp. 108–115, October, 2010.

[4] 佐野綾一, 松倉健志, 波多野賢治, 田中克己. “部分グラフを基本単位とした web 文書検索：単語の出現密度分布の適用”. 情報処理学会研究報告, Vol. 99, No. 61, 1999-DBS-119, pp. 79–84, 1999.

[5] G. Salton, J. Allan, and C. Buckley. “Approaches to Passage Retrieval in Full Text Information System”. In *Proc. of the Sixteenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 49–58, June/July, 1993.

[6] Y. Lv and C.X. Zhai. “Positional Language Models for Information Retrieval”. In *Proc. of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 299–306, July, 2009.

[7] Y. Lv and C.X. Zhai. “Positional Relevance Model for Pseudo-Relevance Feedback”. In *Proceeding of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 579–586, July, 2010.

[8] C.D. Manning, P. Raghavan, and H. Schutze. “Introduction to Information Retrieval”, pp. 157–159, Cambridge University Press, July 2008.

[9] H. P. Luhn. “Key word-in-context index for technical literature (kwic index)”, American Documentation, Volume 11, Issue 4, pp. 288–295, Wiley Subscription Services, Inc., A Wiley Company, 1960.