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

A study of extracting contents on the Web text based on the position of words

北原 沙緒理 † 波多野 賢治 ‡
Saori Kitahara Kenji Hatano

1. はじめに

World Wide Web 上に存在する Web ページに対して，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 同一 Web ページ内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踏まえた Web 検索が行われてしまうためである．例えば，ミュンヘンにある古城ホテルを探したいときに「古城ホテル」「ミュンヘン」という語をクエリとして使用された単語群 (以下，検索ワードとする) として用いた場合,「ミュンヘンにある古城ホテル」ではなく,「ミュンヘンに滞在し，古城ホテルに泊まった」というドイツ探訪記が検索結果として現れる．これは検索結果として現れた Web ページに「ミュンヘンに滞在した」という内容と「古城ホテルに泊まった」という内容の二つ以上の内容が含まれているためである．

よって，Web 検索を行うユーザを支援するために，Web 上に地の文として存在するテキスト (以下，Web テキストとする) を要約する手法や，Web テキストの重要部分を可視化する手法が提案されてきた [1]．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るため，Web 検索エンジンをそのまま使用するよりも効率よく Web 検索を行うことができるようになった．特に，検索結果の可視化によりユーザが Web テキスト内に存在する要素 (パラグラフ，文など) の重み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．これらのユーザの Web 検索を支援する手法に対して用いられる手法では，Web テキスト中で最終的に比較される要素として文が使用されることが多い．

しかし，Web テキストは誰でも作成できるため，一文中に一つの内容を含まないように作成されている場合もある．例えば，図 1 のように Web テキスト中の文は複数の内容を持つ可能性がある．したがって，Web テキストにおいては一文中においても内容に関する重み付けが変化することを考慮する必要がある．また，ユーザが Web テキストを閲覧する際には，Web テキストを構成する単語を考慮して閲覧していると考えられる．これは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もわかる．この点を踏まえ，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．

本稿では Web テキストにおける単語を考慮する内容抽出手法が，文を考慮する手法よりも Web テキストに対する内容抽出手法として適していることを示すために，これら二種類の手法を比較し、その結果を報告する．

2. 関連研究

Lv らはサイズを固定しないパッセージ検索を実現するための研究を行っている [2]．Lv らがパッセージ検索に用いる手法は，確率的言語モデルを拡張した Positional Language Model である，Positional Language Model を用いることにより，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．このモデルは単語の位置を考慮した重みを用いたモデルであるが，最終的にはテキスト内で算出した重みを足し合わせた指標を一つのテキスト全体を表す指標として用いている．よって，Positional Language Model は最終的に文書を単位とした手法であると考えられる．

また，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 [3]．しかし，このツールにおいても単語の位置を考慮した重みを用いた内容抽出指標を，最終的にはテキスト内で算出した重みを足し合わせて可視化に用いている．よって，西原らの手法はテキスト中の一文を単位とした手法であると考えられる．

一方，田らは検索ワードの出現位置及び出現回数を用い，検索結果のリランキングを行っている [4]．田らは Web テキストを一つの長い文字列として扱い単語単位で見ることにより，Web ページにおける単語二語から成る検索ワード間の単語距離を定義する．その中でも局所的出現密度は検索ワードのうち Web ページ内において最初に出現した単語から最後に出現した単語までの単語距離と，その単語距離までの間に存在する検索ワードの出現回数を用い算出される．しかし局所的出現密度は同じ単語距離及び単語の出現回数であれば単語距離の間で一定の値をとるため，単語単位で内容を考慮することはできても，内容が出現する箇所における内容の重みは一定になる．よって，単語の出現位置によって，単一の Web テキスト内においても内容の重みは変化すると考えられる．その上，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，検索ワードに関する内容が含まれない位置も存在すると考えられる．

3. 比較対象

本節では，Web テキストにおける単語を考慮する内容抽出手法が，文を考慮する手法よりも Web テキストに対する内容抽出手法として適していることを示すために使用する指標について述べる．

† 同志社大学大学院文化情報学研究科，Graduate School of Culture and Information Science, Doshisha University

‡ 同志社大学文化情報学部, Faculty of Culture and Information Science, Doshisha University

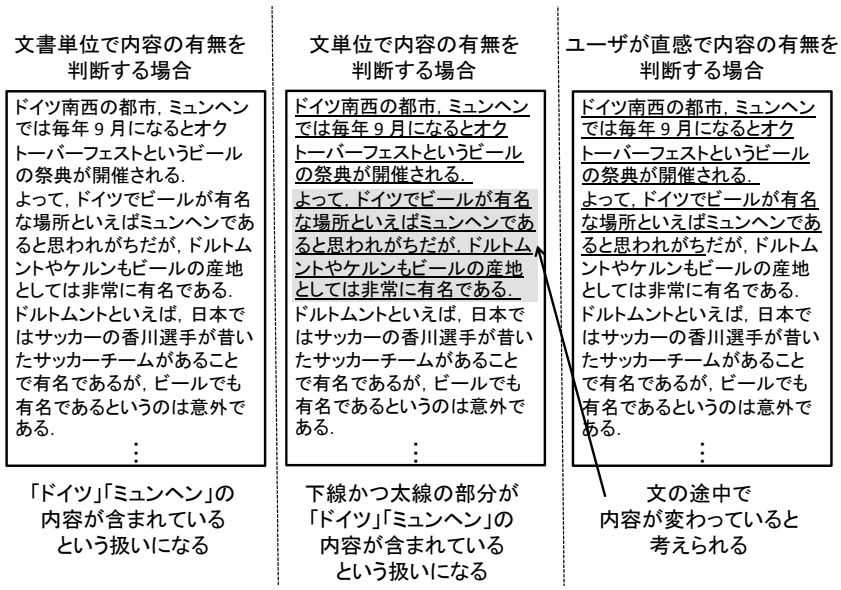

図 1: 異なる単位を用いた内容把握の例

3.1 単語単位で行う Web テキストの内容抽出

本稿では，単語単位で行う Web テキストの内容抽出指標として，以前に我々が提案した内容密度分布を用いる [5]．内容密度分布は，内容の出現範囲及び局所的な内容の影響度変化を表す分布である．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．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 S として付与したものである．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 $s_m(m = 1, 2, \cdots)$，s_m における内容 Q_m に含まれる単語群を t_i^m，t_i^m 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 $t_i^{m.j}$ とすると，$t_i^{m.j}$ が表れる位置 $l[t_i^{m.j}]$ の直前に現れる文の区切りの位置は $a[t_i^{m.j}]$，直後に表れる文の区切りの位置は $b[t_i^{m.j}]$ 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，$t_i^m \subset Q_m(i = 1, 2, \cdots)$ となる．

以上を踏まえると，Web テキスト s_m 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 $hw[t_i^{m.j}]$ 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 $t_i^{m.j}$ の影響度となる．

$$hw[t_i^{m.j}](k) = \begin{cases} \frac{1}{2}(1 + \cos 2\pi \frac{k - l[t_i^{m.j}]}{W}) \\ \qquad (a[t_i^{m.j}] < k < b[t_i^{m.j}]) \\ \frac{1}{2}S(1 + \cos 2\pi \frac{k - l[t_i^{m.j}]}{W}) \\ \qquad (a[t_i^{m.j}] \geq k,\ b[t_i^{m.j}] \leq k) \end{cases} \quad (1)$$

*なお文の区切りには句読点 (。) 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．これは **??** 節において用いられる文の区切りに関しても同様である．

$$(|\ k - l[t_i^{m.j}]\ | \leq \frac{W}{2},\ 0 \leq S \leq 1)$$

ここで単語 $t_i^{m.j}$ の影響は窓の幅 W において存在し，$|\ k - l[t_i^{m.j}]\ | \leq \frac{W}{2}$ の間のみで定義される．重み S がとりうる値の範囲は $0 \leq S \leq 1$ とする．

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，s_m において k 番目の単語が現れる位置における単語 t_i^m における単語密度分布 $hw[t_i^m]$ は式 (2) の通りである．

$$hw[t_i^m](k) = \frac{\sum_j hw[t_i^{m.j}](k)}{\max_k \sum_j hw[t_i^{m.j}](k)} \quad (2)$$

単語一つ一つにおいて単語密度分布 $hw[t_i^m]$ を作成し，s_m における内容 Q_m に関する内容を統合することにより，内容 Q_m における内容密度分布を作成することができる．また，内容密度分布の値に対して閾値を設けることにより，検索ワードに関する内容の抽出を行うことができる．本稿では検索ワードに関する内容密度分布の値が閾値 $E(0 \leq E \leq 1)$ 以上であるテキスト上の位置に検索ワードに関する内容が含まれているとする．

$$hw[Q^m](k) = \begin{cases} \frac{1}{n} \sum_i hw[t_i^m](k) & (hw[t_i^m](k) > E, i = 1, 2, \cdots) \\ 0 & (others) \end{cases} \quad (3)$$

なお，本稿では検索ワードの内容抽出に適した閾値 E を決定するための予備実験も行う

3.2 文単位で行う Web テキストの内容抽出

本稿では，文単位で行う Web テキストの内容抽出としてもっとも単純な手法である，検索ワードが出現する文を抽出する手法を用いる．本稿では，Web テキストの集合中に存在するある Web テキスト s_m において，p 番目に表れる一文 u_p 中にある内容 Q_m に含まれる単語 $t_i^m(m = 1, 2, \cdots)$ が全て含まれるときに，u_p 全体に内容 Q_m に関する内容が含まれているとする．

例えば，図 2 に記されたテキスト s_m に存在する 2 文 u_1, u_p に対して「ドイツ」「ミュンヘン」の内容を抽出すると，u_1 は「ドイツ」及び「ミュンヘン」という単語が含まれているので，2 文には「ドイツ」「ミュンヘン」の内容が含まれると判断される．しかし u_p には「ドイツ」という単語は含まれているが「ミュンヘン」という単語は含まれていないため,「ドイツ」「ミュンヘン」という内容が含まれていないと判断される．

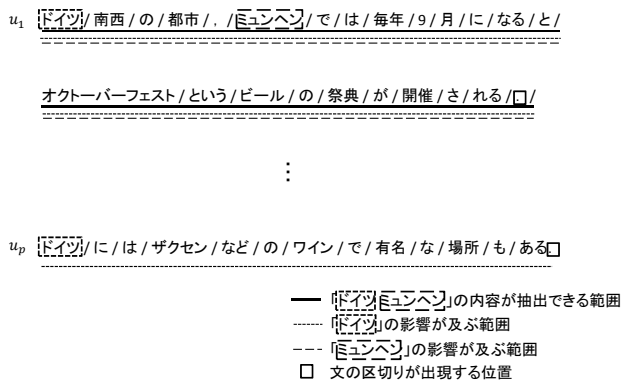

図 2: 文単位で行う Web テキストの内容抽出の例

4. 比較実験

本節では，Web テキストにおける単語を単位とする内容抽出手法が，文を単位とする手法よりも Web テキストに対する内容抽出手法として適していることを示すために，これら二種類の手法を比較する実験を行うための手法について述べる．本稿では単語を単位とする内容抽出手法として 3.1 項で述べた手法を用い，文を単位とする手法として 3.2 項で述べた手法を用いる．また，この比較実験を行うために必要となる内容密度分布の値に対する適切な閾値を求める予備実験に関する手法についても述べる．

なお，比較実験では Web テキストを文を構成する最小要素である形態素にわけ，各形態素を各 Web テキスト上に存在する単語とする．また，比較実験には形態素解析器として MeCab * を用い，内容密度分布の窓の幅 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 倍とする．

4.1 予備実験

まず，既存手法である内容密度分布を文を単位とした手法と比較するために，内容密度分布の値に対して閾値を求める．

*http://mecab.googlecode.com/svn/trunk/mecab/doc/index.html，バージョン 0.98

この閾値は以前に我々が過去に用いた 110 件の Web テキストデータから算出されたデータ [6] を正解データとして算出する．具体的には閾値を 0.1 刻みに設定し，前述のデータを回答とし閾値以上の値を持つ単語と，各正解データとの適合率と再現率の 調和平均を算出し，これらの 調和平均の平均が一番高い閾値を文を単位とした手法と比較するための閾値として設定する．ここで，Web テキスト s_m における，この手法により抽出された箇所の集合を C_m，その集合に含まれる要素の個数を $n(C_m)$ とする．また，正解データとされた箇所の集合を A_m，その集合に含まれる要素の個数を $n(A_m)$，この手法として抽出されてかつ正解箇所とされた箇所の集合に含まれる要素の個数を $n(C_m \cap A_m)$ とすると，Web テキスト T_m における調和平均 F_m は $F_m = \frac{n(C_m \cap A_m)}{\frac{n(C_m)}{2} + \frac{n(A_m)}{2}}$ となる．したがって，今回用いる評価指標である 調和平均の平均 F_{ave} は $F_{ave} = \frac{1}{m} \sum_n F_m$ となる．

以上の結果算出された 調和平均の平均は図 3 のようになる．

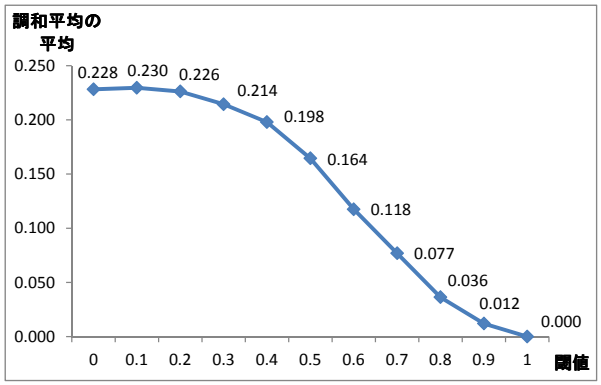

図 3: 各 Web テキストにおける 調和平均の平均

図 3 より，内容密度分布の閾値として適切である値は 0.1 であることがわかる．よって，本稿では内容密度分布の閾値を 0.1 と定める．なお，内容密度分布の閾値が 0.1 と非常に小さいことから，内容密度分布の値がある程度高い箇所のほとんどが，各内容 Q_m を示すために重要であることがわかる．

4.2 本実験

次に，Web テキストにおける単語単位で行う内容抽出手法の方が，文単位で行う手法よりも Web テキストに対する内容抽出手法として適していることを示すために，これら二種類の手法を比較する．

この実験の評価データ作成に使用する検索ワードには NTCIR-5 WEB テストコレクションの検索課題中における「必須」検索課題 400 課題のうち，既知のアイテムに対する複合検索課題 148 課題からランダムに選択した 10 課題を使用した [7]．ここで，既知のアイテムに対する複合検索課題とは，実験協力者が検索したい内容が決まっており，かつ検索ワードが単語 2 文字以上のものを指す．また，既知のアイテムに対する課題であるため，検索課題中に検索課題の対象に関する内容が記述してある．なお，本稿で既知のアイテムに対する検索課題を選択した

理由は複数人いる実験協力者に対して単一の検索課題を出すことができるためである．その中でも複合検索課題を選択した理由は，予備実験で用いたデータも検索ワードを 2 単語以上にしているため，予備実験でのデータの作成条件にできる限り近づけるためである．

また，前述の 10 課題を Google AJAX Search API * を用いて，検索した結果表示される上位 10 件の Web ページのテキストを評価データ作成のための Web テキストとする．

最終的には以上の Web テキスト中のどの部分に検索ワードに関する内容が含まれているかを人手で判断したものを評価データとする．評価データ作成の際には，一つの Web テキストに対して，3 人の実験協力者が評価を行う．具体的には，まず実験協力者に各検索課題の内容が各検索結果の Web テキストに含まれているかを尋ねる．次に，実験協力者が検索課題に関する内容が Web テキストに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マウスにて選択してもらう．ここで選択してもらった単語の位置を評価データとして用いる．そして，3 人中 2 人以上が内容が含まれていると答えた箇所を正解箇所として使用する．

したがって，評価には 75 種類の Web テキストと検索ワードの組を用いる．そして，評価指標に，正解箇所を単語単位に区切った場合の各手法における調和平均の平均により比較を行う．この際に使用する 調和平均の平均を求める導出式は 4.1 項と同様である．

以上の結果算出された 調和平均の平均は図 4 のようになる．

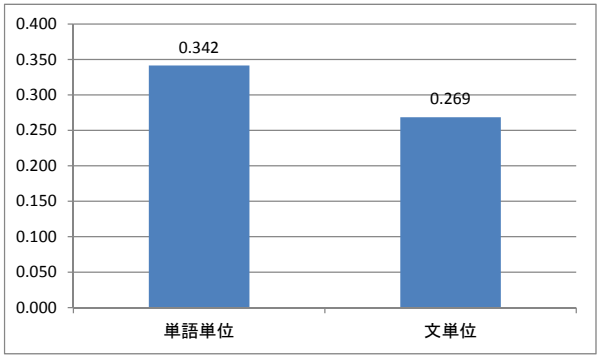

図 4: 単語を単位とした内容抽出手法及び，文を単位とした手法における 調和平均の平均

図 4 より，単語を単位とした内容抽出手法のほうが文を単位とした手法よりも 0.073 調和平均の平均が高いことがわかる．これは，単語を単位とした内容抽出手法では文の範囲を超えた柔軟な内容抽出を行うことができたからであると考えられる．

5. おわりに

本稿では Web テキストに対する内容抽出手法における，Web テキスト上の単語出現位置を考慮し単語を単位として内容抽出を行う手法の優位性を示すため，比較実験とその考察を行った．予備実験から，Web テキスト上における内容密度分布の値がある程度高い箇所のほどんどが，各内容を示すために重要であることがわかった．また，本実験から単語による内容抽出手法が，文による内容抽出手法よりも優れていることがわかった．

今後は内容密度分布を用いた Web 検索支援システムを作成し, ユーザの Web 検索を支援することを考えている．例えば，検索ワードに含まれる単語と各 Web テキストに存在する単語の内容密度分布を作成し Web 検索結果に含まれる様々な内容を俯瞰することを考えている．Web テキストに存在する様々な内容を俯瞰することができれば，ユーザが Web テキスト内に存在する様々な内容の重みを直感的に理解することができるようになる，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ると考えられる．

*https://developers.google.com/web-search/docs/

参考文献

[1] 砂山渡, 高間康史, Danushka Bollegala, 西原陽子, 徳永秀和, 串間宗夫, 松下光範. Total Environment for Text Data Mining. 人工知能学会論文誌, Vol. 26, No. 4, pp. 483–493, 2011.

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

[3] 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.

[4] C. Tian, T. Tezuka, S. Oyama, K. Tajima, and K. Tanaka. Improving web retrieval precision based on semantic relationships and proximity of query keywords. In *Proceedings of the 17th international conference on Database and Expert Systems Applications*, pp. 54–63, 2006.

[5] S. Kitahara, K. Tamura, and K. Hatano. Extraction of Web Texts Using Content-Density Distribution. In *Proceedings of 7th Asia Information Retrieval Societies Conference*, pp. 273–282, 2011.

[6] 北原沙緒理, 田村航弥, 波多野賢治. Web テキストにおける内容密度分布の抽出とその評価. 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.

[7] K. Oyama, M. Takaku, H. Ishikawa, A. Aizawa, and H. Yamana. Overview of the NTCIR-5 WEB Navigational Retrieval Subtask 2 (Navi-2). *NTCIR-5*, 2005.