Vol. 38 No. 4 情報処理学会論文誌 Apr. 1997

出現密度分布を用いた語の重要説明箇所の特定

黒橋禎夫† 白木伸征† 長尾眞†

本論文では，テキスト中に，ある語が複数箇所出現する場合，その中からその語の重要な説明箇所を自動的に特定する方法を提案する．語に対して，重要な，あるいは関連性の高い説明をしようとすれば，必然的にその語を繰り返し用いる必要がある．そこで，テキスト中での語の出現密度分布を調べ，その高密度な出現位置を取り出すことによって，その語の重要説明箇所を特定することができると考えた．密度計算には，ハニング窓関数を用いて，ある範囲の語の出現を重み付きで加算するという方法を用いる．新書10冊，100キーワードに対する評価実験によって本手法の有効性を具体的に示す．

A Method for Detecting Important Descriptions of a Word Based on Its Density Distribution in Text

SADAO KUROHASHI,† NOBUYUKI SHIRAKI† and MAKOTO NAGAO†

In this paper we propose a method for detecting important descriptions of a word in a text. Our method is based on the assumption that an important description of a word can be found in the text segment where the word occurs in high density. The word's density is calculated by counting its occurrences in a certain range of the text with weights assigned by Hanning window function. We report experimental results to illustrate the effectiveness of our method.

1. はじめに

本には少なくとも2種類の利用法が考えられる．1つは本のタイトル，目次などを見て本全体に興味を持ち，そこに書かれている内容全体が知りたいという場合で，このときには本を前から順に通読する必要がある．もう1つは，読者の側にまず知りたい内容があって，それがどこかに書かれていないかと期待して本を調べるという場合である．この場合，読者にとって必要なのは本の中のある一部分だけであり，その部分だけを拾い読みするということが行われる．

これから電子図書館の時代になり，数百万冊単位の本の全文テキストに高速にアクセスできるようになれば，後者の，いわば「しらべもの」的な本の利用法がより頻繁で重要になると考えられる．本をしらべもの的に利用する場合にはそこに与えられている索引が重要な役割をはたす．本の索引には次の2つの情報が含まれている．

(1) 索引語そのもの，すなわち，その本にとって重要な語のリストがあげられているということ．

(2) 各索引語に対する重要な説明が本の中のどの部分にあるかという位置情報．冊子体の本の索引では各索引語に与えられた頁番号に相当する情報．

これらの2つの情報はともになくてはならないものである．ある事柄について説明した本を調べたい場合，第一の重要語リストとしての情報は適当な本を選択するために必要であり，第二の位置情報はその本の中で本当に重要な部分にアクセスするために必要となる．

情報検索の分野では索引情報の自動抽出に関する研究が古くからさかんに行われてきた．しかし，そこでは索引の第一情報，すなわち本を代表する重要語の抽出が中心であり[3),4),7),10),11),13)]，第二の位置情報に関する問題はほとんど扱われてこなかった．これは，これまで計算機に蓄えられてきたテキストが論文のアブストラクトに代表されるような比較的小さなサイズのテキストであったため，索引語による検索で適当なテキストが取り出されれば，そのテキストをすべて読むことも十分に可能であったからである．

ところが，本1冊の全文テキストのように大きなサイズのテキストを扱う検索が一般的になれば，索引の第二の情報，すなわちテキストの中での索引語の重要説明箇所に関する情報がなくてはならないものとな

† 京都大学大学院工学研究科電子通信工学専攻
Department of Electronics and Communication, Kyoto University

る．そこで，本論文ではこの重要説明箇所に関する情報を語の出現分布を用いて自動的に抽出する方法を提案する．

語の重要説明箇所の特定が自動化できれば，索引情報付与のように前もって行われる静的な処理だけでなく，これを動的な処理として利用することもできる．たとえば，ある電子化されたテキストに対して索引情報のない語について調べたい場合，単に全文検索をしてそのすべての出現位置を見ていくというのではなく，自動的に特定された重要説明箇所をまず調べてみるということができるようになる．

また，重要説明箇所の特定はハイパーテキストリンクの自動付与にも利用することができる．電子テキストの有効性を最大限に活用する方法としてテキストをハイパーテキスト化して扱うことが考えられるが，そこではテキストの関連性を示すリンクをどのように自動付与するかということが課題となっている．リンクには様々なタイプのものが考えられるが，その最も基本的なものはある語についてテキスト中のそれほど重要でない出現位置（そこを読んだだけではその語についての意味などは分からない位置）から，重要な説明位置へのリンクである．語の重要説明箇所を自動特定すれば，このようなリンクを自動付与することが可能となる．

2. 語の出現密度の利用

ある語に対する重要な説明というものがどのようなものであるかを厳密に定義することは難しい．ここでは語の定義的説明が与えられている部分や，語に深く関連する内容が述べられている部分を重要説明箇所と呼ぶことにする．また，以下では重要説明箇所を特定すべき語（処理対象の語）を**キーワード**と呼ぶことにする．

キーワードの重要説明箇所を取り出す方法の1つとして，重要と考えられる説明文の表層的特徴を利用することが考えられる[8]．たとえば，

「○○とは … である」

という文パターンは○○の定義を与えると考えられるから，このようなパターンにマッチする部分を重要説明箇所として取り出すという方法である．

このような方法には，重要説明箇所のパターンを網羅的に用意することができるかという問題がある．文パターンによる方法が比較的有効であると考えられるのは専門用語の定義的説明などの場合であるが，たとえそのように対象を限定したとしても種々の専門分野のテキストに一般に通用するパターン・セットを用意することは容易ではない．まして，一般的な語の場合には重要な説明というものをパターン化すること自体が困難である．たとえば『都市の生態学』という本では「川」に関する重要な説明部分として，江戸時代の関東平野の河川改修工事の説明，現代の都市化にともなう河川の水質汚濁についての説明などがあるが，そのような部分を文パターンによって取り出すことはほとんど不可能である．

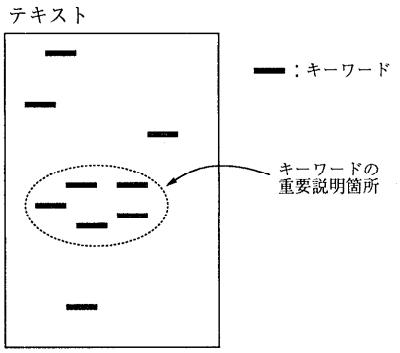


図1 語の出現密度分布と重要説明箇所の関係
Fig. 1 Important descriptions of a word and its density distribution.

そこで我々が考えたのは，キーワードの出現密度の分布を利用するという方法である．あるキーワードに対して重要な，あるいは関連性の高い説明をしようとすれば，必然的にそのキーワードを繰り返し用いる必要がある．そこで，あるキーワードがテキストで複数箇所に現れる場合，重要な説明箇所はそのキーワードがまとまって多数使われている場所，すなわち，キーワードの密度が高い場所であると仮定することができる（**図1**）．この仮定が妥当であるとすれば，キーワードのテキスト中での密度分布を調べることによって，その重要説明箇所を特定することができる．この仮定の妥当性は4章で実験的に示すとして，次章では語の密度分布の具体的計算方法を説明する．

3. 語の出現密度分布の計算と重要説明箇所の特定

3.1 密度計算のための単位の設定

キーワードの出現密度を求めるためには適当な範囲の単位を設定して，その中でのキーワードの出現をカウントする必要がある．単純な方法としては，章，節，段落などのようにテキストに明示的に与えられているまとまりを単位とする方法が考えられる（さらに，出現頻度をその単位の大きさ（文字数）で正規化するということもできる）．しかし，段落のような小さな単位を考えると，語が複数の段落に連続的に出現しても

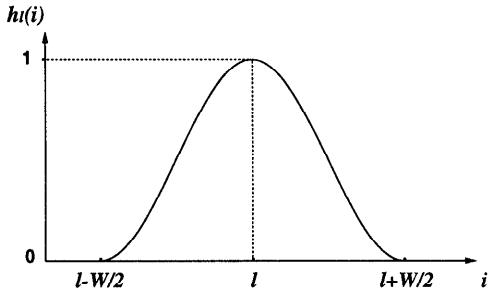


図2 ハニング窓関数
Fig. 2 Hanning window function.

そのことが評価に加味されないという問題がある．逆に，章，節などはかなり大きな範囲となり，またその大きさにもばらつきがあるため，ある部分に非常に密に現れる語であってもその評価が単位の大きさやばらつきでぼかされてしまう可能性がある☆．

そこで，このような問題が起こらない密度計算の方法として，次のような方法を考えた．

(1) 段落よりもう少し大きな，ある一定の範囲（文字数）を設定する．

(2) その範囲内のキーワードの出現を，範囲の中心付近の出現を重視し，中心から離れるに従って重みを軽くするという重み付けで足し込む．

上記(2)のような性質を持った重み付けの関数として，音声認識などで広く用いられているハニング窓関数と呼ばれる関数がある[9]．窓の幅（重みを与える範囲）を W，窓の中心位置を l とすると，ハニング窓関数 $h_l(i)$ は次式によって与えられる（**図2** 参照）．

$$h_l(i) = \frac{1}{2}\left(1 + \cos 2\pi \frac{i-l}{W}\right) \quad (|i-l| \leq W/2)$$

本手法ではこのハニング窓関数をもちいて語の出現密度を計算することにする☆☆．

3.2 ハニング窓関数による密度計算

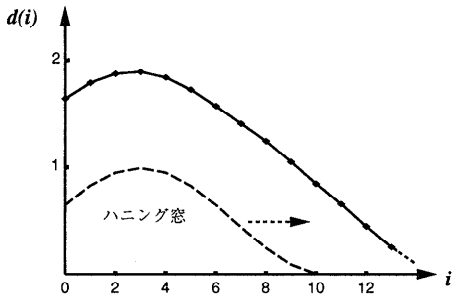
ハニング窓関数を用いた語の出現密度の計算は以下のアルゴリズムで行う．

(1) 与えられたテキストを1本の長い文字列（長さ L 文字）と見なし，テキスト中でのキーワードのすべての出現位置を調べる．位置 l を先頭としてキーワードが出現する場合 $a(l) = 1$，そうでない場合 $a(l) = 0$ と表すことにする．

(2) 位置0（テキストの先頭）からスタートして，順に各位置をハニング窓の中心位置とし，その中心位置 l に対するキーワードの出現密度 $d(l)$ を計算する．ハニング窓の幅を W とすると，中心位置の前後それぞれ $W/2$ の範囲のキーワードの出現を次式によって足し込む．

$$d(l) = \sum_{i=l-\frac{W}{2}}^{l+\frac{W}{2}} h_l(i) \cdot a(i)$$

（ただし，$i < 0$ または $i \geq L$ では $a(i) = 0$ とする）



テキスト：人民の人民による人民のための…

$a(i)$ ： 1 0 0 1 0 0 0 0 1 0 0 0 0 0

図3 キーワードの出現密度の計算例
Fig. 3 An example of calculating the density of a keyword.

図3 に密度計算の簡単な例を示す．W を15としてキーワード「人民」の出現密度を計算する場合，たとえば位置3での密度は

$$\begin{aligned} d(3) &= h_3(0) \cdot a(0) + h_3(3) \cdot a(3) + h_3(8) \cdot a(8) \\ &= 0.65 + 1.00 + 0.25 \\ &= 1.90 \end{aligned}$$

となる．

3.3 説明区間の切り出しと重要度付与

前節の方法でキーワードの出現密度分布が求まると，密度の最大位置あるいは極大位置の付近にキーワードの重要説明箇所が存在するだろうということが分かる．そこで，次に考えなければならないのは，説明箇所をどのような単位で取り出すかという問題である．

キーワードが密集して現れていることは，その部分がキーワードに関連するひとまとまりの説明であることを示唆していると考えられる．しかし，どの程度の密集であればひとまとまりの説明と見なしてよいかと

☆ たとえば，章の中の出現回数を章の文字数で正規化するという方法を考えると，1,000文字の章の中で離れて2回出現する場合と，2,000文字の章の中でまとまって4回出現する場合が同じ密度と計算されていしまい，適当ではない．

☆☆ 「密度」を直感にあうかたちで計算するためには，ハニング窓のように中心からの距離に応じた重み付けを行う窓関数を用いる必要がある．たとえば，窓の範囲内でキーワードが2カ所に現れるとする．単にある範囲の出現を数えるだけの方形窓のような窓関数を用いると，2カ所の出現が窓の範囲内のどこであっても出現密度分布の最大値は一定となってしまう．これに対して，ハニング窓関数を用いる場合には，2カ所の出現が近ければ近いほど出現密度分布の最大値が大きくなる．

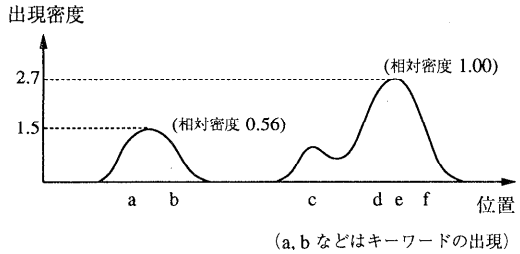


図4 説明区間の切り出しと重要度付与
Fig. 4 Detection and ranking of description segments.

いうことは難しい問題である．たとえば図4に示した簡単な例の場合，c，d，e，fの部分をひとまとまりと見なすか，cとd，e，fを別のまとまりと見なすかという問題である．

これは本来的にはキーワードの密度分布だけでは決定できない問題であるが，ここでは密度分布だけを用いる方法の有効性（と限界）を示すために非常に簡単な基準を導入することにする．すなわち，キーワードの出現密度が0よりも大きい値で連続的に分布する範囲をひとまとまりの説明と見なすことにする．これはハニング窓の幅 W 以内の間隔でキーワードが連続的に出現する範囲に相当する．以下では，この区間を**説明区間**と呼ぶ．各説明区間には，区間内の密度分布の最大値を密度分布全体の最大値で正規化した値（これを**相対密度**と呼ぶことにする）を，重要度の尺度として与えることにする．

図4の例の場合，これが，あるキーワードに対するテキスト全体の密度分布であるとすれば，取り出される説明区間はa，bを含む区間と，c，d，e，fを含む区間の2つとなり，それぞれ0.56，1.00という相対密度（重要度）が与えられることになる．

4. 評価実験と考察

4.1 評価実験

提案した手法の有効性を調べるために，岩波新書10冊，各冊ごとに10キーワード（付録参照）を対象として評価実験を行った．キーワードの選択は著者らが行い，各本のタイトル，目次などを参考にして，各本に関連して調べたいと思うキーワードを10個ずつ考えた．このとき，一般的な語（「貿易」,「色」など）とある程度専門的な語（「ペレストロイカ」,「同音衝突」など）がバランスよく混在するように注意をはらった．キーワードの平均出現回数は40.2回，平均出現段落数は27.9個であった．

実験では，各キーワードに対して本手法による重要説明箇所の特定（窓の幅は1500；4.2節参照）と，人手による重要説明箇所の特定を別々に行い，それらの結果を種々の基準で比較した．

4.1.1 人手による重要説明箇所の特定

人手による重要説明箇所の特定は3人の被験者によって行った．被験者はいずれも文系の大学院生で，比較対象が高密度部分を重要説明箇所と判断する手法であることは知らされていない．

被験者には次のような作業を行ってもらった．

(1) 重要であるかどうかの判断は，キーワードに関連するひとまとまりの説明を単位として行う．そのためにまず，キーワードを含む段落（これを**キー段落**と呼ぶ）を最小単位とし，どの範囲のキー段落（キー段落群）がひとまとまりの説明と見なせるかを判断する．

(2) 説明のまとまりと判断した各キー段落群に対して，重要（○），ある程度重要（△），重要でない（×）という3値の評価を行う．そして，その評価値をキー段落群中の全キー段落に与える．

上記(1)の判断は，非常に高頻度のキーワードで，キー段落が広範囲に連続的に分布する場合には，かなり難しい判断となる．しかし，キー段落数が10～20程度の場合には，いくつかのキー段落がほぼ連続してまとまって現れ，そのようなまとまりがテキスト中に離散的に存在するということが典型的である．そのような場合には単に「キー段落のほぼ連続するまとまり」が説明のまとまりと判断されることが多い．なお，上記(2)のように3値の評価値を設定したのは，説明の重要性に対して重要かそうでないかの2値の判断を行うことが難しすぎるという場合が少なくないからである．

上記の基準で各被験者に作業を行ってもらった後，各キー段落に対して3人の総合評価値を計算した．これは，○を2点，△を1点，×を0点として，3人の評価値の総和が3点以上（{○,△,×}の場合など）であれば総合評価○，3点未満であれば総合評価×とした．**図5**に図4の例に対応する人手評価の例を示す．

4.1.2 段落単位の評価結果

まず，人手による重要性の評価結果と本手法による評価値を，100キーワードの全キー段落について比較した．本手法によるキー段落の評価値は，それを含む説明区間の相対密度とした☆．相対密度は0から1まで

☆ たとえば，図5の例では，aのキー段落とbのキー段落の評価値を0.56，cのキー段落，d，eのキー段落，fのキー段落の評価値を1.00とする．このように，3.3節でひとまとまりの説明（説明区間）であると認識されたキー段落群には同じ評価値を与える．

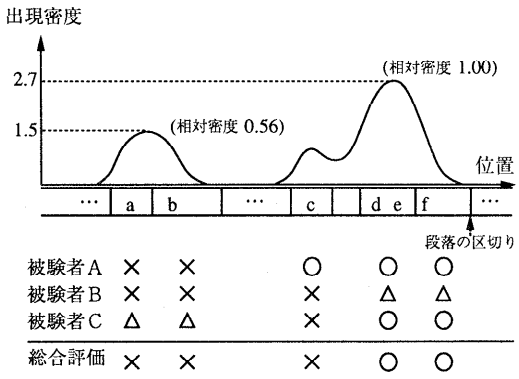


図5 人手による説明区間の評価
Fig. 5 Manual evaluation for description segments.

表1 段落単位の評価結果
Table 1 Results of experiments per paragraph.

人手による評価	○	×
相対密度 0.6 以上	931	591
相対密度 0.6 未満	196	1071

の連続値であるので，重要と見なすかどうかを決定するためには閾値が必要となる．相対密度 0.6 を閾値とした場合の比較結果を**表1**に示す．この結果について再現率・適合率[13]を計算すると以下のようになる．

$$再現率=\frac{相対密度\ 0.6\ 以上で人手の評価○の段落数}{人手の評価○の段落数}=\frac{931}{931+196}=82.6\%$$

$$適合率=\frac{相対密度\ 0.6\ 以上で人手の評価○の段落数}{相対密度\ 0.6\ 以上の段落数}=\frac{931}{931+591}=61.2\%$$

相対密度に対する閾値を 0.05 きざみで変化させた場合の再現率，適合率を**図6**に示す．

人手による総合評価が○の段落は 1 キーワードあたり平均 11.3 個であったが，このうちすべての被験者の評価値が○または△のものは平均 4.4 個であった．このようにだれが見ても，ある程度重要であると判断されるような段落については，平均 3.9 個（88.5%）が相対密度 0.6 以上であった．これらの結果から，語の出現密度が語の説明の重要性の尺度になりうるという仮定がかなりの程度妥当であることが分かる．

提案手法に対する比較対象として本の中でのキーワードの最初の出現段落を重要説明箇所とする方法を考えてみる．出現の順序だけを考慮するとすれば，最初の出現段落に重要な説明が与えられている可能性が高いと考えられるからである．これに対して，本手法

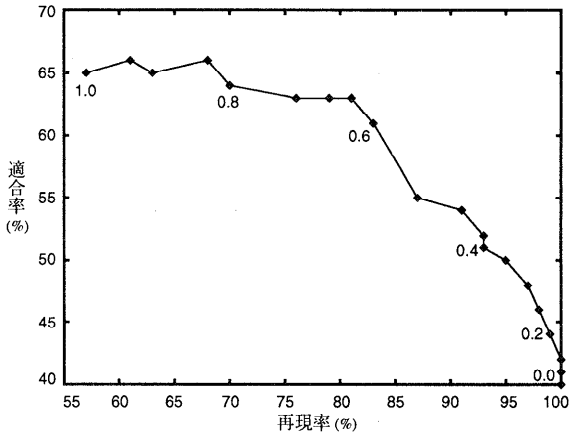


図6 段落単位の再現率と適合率
Fig. 6 Recall and precision per paragraph.

表2 先頭のキー段落と密度最大値に対するキー段落の評価結果
Table 2 Comparison between the first paragraph of a word and its highest density paragraph.

人手による評価	○	×	適合率
先頭のキー段落	40	60	40%
密度最大値に対するキー段落	86	14	86%

によって最も重要な段落を 1 つ選択するとすれば出現密度分布の最大位置を含むキー段落となる．そこで，各キーワードに対してこれら 2 種類の段落に対する人手の評価値を調べたものが**表2**である．この表から，キーワードの最初の出現位置を重要説明箇所とする方法はまったく不十分であり，それに比べて出現密度分布を用いる方法がはるかにすぐれていることが分かる．

4.1.3 説明区間単位の評価結果

本手法によって取り出される説明箇所は 3.3 節で述べた説明区間を単位としたものとなる．そこで，説明区間を単位としてその中のキー段落に対する人手の評価を調べた．自動抽出される説明区間と人手による説明のまとまりの認識は必ずしも一致しないので，1 つの説明区間の中に評価値○の段落と評価値×の段落が混在することもありうる．そこで，説明区間単位の評価では人手の評価との一致を次の 4 つの場合に分類した．

(1) 説明区間内のすべてのキー段落の評価が○
(2) 半数以上のキー段落の評価が○（(1)の場合を除く）
(3) 少なくとも 1 つのキー段落の評価が○（(1), (2)の場合を除く）
(4) すべてのキー段落の評価が×

相対密度の閾値を 0.6 として上記の分類を行った結

表3 説明区間単位の評価結果
Table 3 Results of experiments per description segment.

人手による評価	1：すべての評価○	2：半数以上の評価○	3：少なくとも1つの評価○	(1, 2, 3 の和)	4：すべての評価×
相対密度 0.6 以上	88	52	14	(154)	40
相対密度 0.6 未満	51	15	18	(84)	503

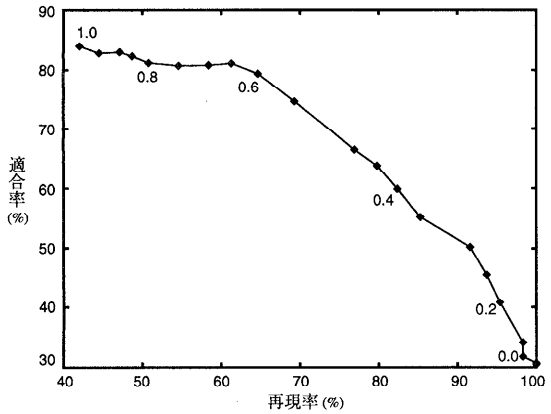


図7 説明区間単位の再現率と適合率
Fig. 7 Recall and precision per description segment.

果が**表3**である．1キーワードに対して相対密度0.6以上の説明区間は平均1.9カ所である．これに対して，評価値○の段落を少なくとも1つ含むかどうかという基準で再現率，適合率を計算するとそれぞれ64.7%，79.4%となる．相対密度に対する閾値を0.05きざみで変化させた場合の再現率，適合率を**図7**に示す．

本手法によって自動的に索引位置を決定するとすれば，相対密度がある閾値以上の説明区間に対して，その中の先頭のキー段落を索引位置とすることが考えられる☆．このような方法による索引付けの妥当性を調べるために，評価値○のキー段落を少なくとも1つ含む説明区間について，先頭のキー段落から数えて（先頭のキー段落を0段落として）何段落目に評価値○のキー段落が現れるかを調べた．その結果は，キー段落だけを数えると中央値で0段落目（平均値1.0段落目），キーワードを含まない段落を含めて数えた場合にも中央値0段落目（平均値4.0段落目）であった☆☆．これは，評価値○の段落を含む説明区間では，半数以上の場合，先頭のキー段落の評価値が○であるということを意味する．これらの結果から，本手法を索引位置の自動特定に実際に利用することも十分可能であると考えられる．

図8に本手法による重要説明箇所の特定の具体例を示す．これは『中国とソ連』という本で「貿易」というキーワードの説明区間を求めた結果である．本手法で最も重要と判断された相対密度1.0の説明区間は1980年代以降の中ソ貿易に関する説明で，この部分に含まれるキー段落に対する人手の評価値は○○○○×××であった．次に2番目に重要であると判断される相対密度0.76の部分は1950年代の中ソ貿易に関する説明で，この部分のキー段落に対する人手の評価値は××○○○○○○○○○であった．相対密度が3番目以下のものについては人手による評価値はすべて×であった．

4.2 ハニング窓の幅について

実験ではハニング窓の幅を1,500文字とした．新書1頁の文字数は約600文字であるから，この値はだいたい見開き2頁（約1,200文字）内のキーワードの出現をある程度の重みで加算するという設定である（中心から600文字離れた位置の重みは0.1であるから，それより外側の影響は非常に小さい）．この幅の値は，実験で用いたテキストとは別の新書1冊について予備実験を行い，500，100，1,500，2,000などの値の中から最も妥当な値として選択した．

窓の幅の最適な値は検索意図やテキストの種類によって異なると考えられる．たとえばあるキーワードについて簡潔な説明がほしいという場合には狭めの幅が良いだろうし，逆にキーワードが大きなテーマとして取りあげられている部分を探したいならば広めの幅が良いであろう．また，一般向けの新書などのように丁寧に（若干冗長に）書かれたテキストの場合と，科学技術論文のようにできるだけ簡潔な表現を用いようとしている場合でも最適値は変わるであろう．

このような問題から，本論文では適当な性質の窓関数を適当な幅で用いた場合の手法の有効性を実験的に示すことを主眼とし，窓の幅の最適値を追求することは対象外とした．

☆ ここでは，最も基本的な場合として索引位置を1カ所指定する場合を考えるが，電子読書環境においてユーザインタフェースを工夫すれば，説明区間全体やその密度分布を表示することも可能である．

☆☆ 中央値とは，変量（各説明区間に対して何段落目か）を大きさの順に並べた場合の中央の値である．頻度が200を超えるようなキーワードの場合には，非常に大きな説明区間が取り出され，その中で評価値○の段落が後ろの方にしか存在しないということがある．上記の中央値と平均値のずれはそのような場合の影響である．実際的には，非常に高頻度のキーワードについては，密度分布の連続とは別の基準で説明区間を分割する必要がある．

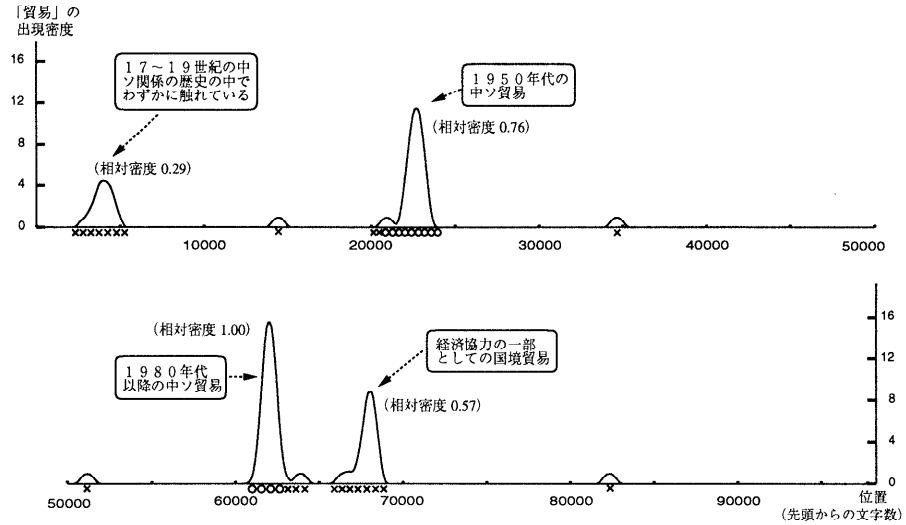


図8 岩波新書『中国とソ連』に対してキーワード「貿易」の重要説明箇所を特定した例
Fig. 8 Results of detecting important descriptions of the word "Boeki (trade)" in the book "China and the Soviet Union."

4.3 誤りについての考察

本実験での解析誤りについて原因を考察する．ここで解析誤りとは，人間は×と判断したのに相対密度が0.6以上となったもの，逆に人間は○と判断したのに相対密度が0.6未満となったものをさすことにする．解析誤り例を調査したところ，その原因は以下のようなものであった．

相対密度0.6以上 ↔ 人間の評価×：

- 本の中で一様に，かつ頻繁に用いられる語の場合，相対密度1.0の説明区間はたしかに重要であるが，2番目，3番目の相対密度の部分はそれほど重要とは考えられない場合が多い．
- 例外的な部分ではあるが，ある語が個条書き部分や引用部分で頻繁に現れるとその部分が非常に高密度となり，他の一般の説明部分よりも上位にランクされてしまう．このような部分が必ずしも不必要であるというわけではないが，上位にランクされ過ぎるという傾向がある．文章構造の情報（SGMLタグなど）を補助的に利用することで対処できる可能性がある．

相対密度0.6未満 ↔ 人間の評価○：

- 検索したい語がテキスト中で同義語あるいは上位語などで代用されている場合には本手法ではうまく処理できない．具体例としては，『やきもの文化史』という本で「茶碗」の重要説明箇所を調べたところ，日本の茶碗と輸入茶碗との対比が述べられていて重要と考えられる部分があったが，そこでは「茶碗」だけでなく「カップアンドソーサー」，「ディナーセット」などの語が用いられていたために「茶碗」自身については相対密度0.6以上とならなかった．

  これに対して，語の省略，代名詞化などが誤りの原因となる例はほとんどなかった．これは，省略・代名詞化は狭い範囲（複文や，隣りあう文間など）では頻繁におこるが，広い範囲でおこり続ける現象ではないからである．つまり，1つの説明区間で見れば，部分的には省略・代名詞化が行われていても，全体としては適当な間隔でその語自身が使われているのである．
- 語の最初の出現で簡単な定義が与えられるが，そこではその語は1，2度使われるだけで，それよりかなり後ろに詳しい説明が現れるという場合がある．簡単な定義とは，たとえば語の直後に括弧に囲まれて定義が与えられるというような場合である．このような場合，後の詳しい説明部分は高密度位置として高い相対密度となるが，最初の簡単な定義部分は低い相対密度にしかならない．
- 相対密度はあくまでも相対値であるので，密度最大値が極端に大きいと，2番目，3番目の極大値は（たとえ重要な説明箇所であっても）相対密度0.6以上にならない．実際には，相対密度0.6以上，あるいは絶対密度5.0以上で上位3番目以内，というように相対密度，絶対密度，順番などを総合的に加味した閾値を考える必要がある．

4.4 関連研究

1章でも述べたように，テキストから重要語（索引語）

を取り出す問題については多くの研究[3),4),7),10),11),13)]があるが，語の重要な説明箇所を取り出すという問題はこれまでほとんど扱われてこなかった．語の出現分布を利用するという点ではHearst[5)]の研究が最も関連するものであるが，Hearstの研究は語の出現分布をユーザインタフェース（検索結果の表示）として利用するものであり，重要説明箇所の特定という問題にまでは踏み込んでいない．語の出現分布を扱ったもう1つの研究としてBookstein[3)]らの研究があるが，これはテキスト中での分布（密集度）を尺度として重要語（索引語）を取り出す研究であり，重要説明箇所に関する研究ではない．

ハイパーテキストリンクの自動付与という観点では，テキスト間の類似度を計算し，類似するテキスト間などにリンクをはる研究[2)]，従来の索引語抽出手法によってテキストから索引語を抽出し，テキストと索引語の間にリンクをはる研究[1)]，語がどのテキストに出現しているかという分布を語の特徴量とし，類似する語の間にリンクをはる研究[12)]などがあるが，いずれも本手法で扱うような語の重要説明箇所に関するリンクを扱ったものではない．黒橋ら[8)]の研究は語の定義部分を含む種々の説明箇所を文パターンで取り出して，それらをもとにリンクを付与するものであるが，文パターンによる方法の問題点についてはすでに2章で議論した．

検索対象のテキストサイズが大きくなるにつれて，テキスト全体ではなく，その一部分を検索の単位とする研究がいろいろと行われている．このうち，Saltonら[14)]やHearst[6)]はテキストを自動的に小さなまとまりに分割して検索単位とする手法を提案している．もし，この分割されたまとまりがちょうどある語の重要な説明箇所の範囲となり，その語をキーとして検索を行ったときにそのまとまりがちょうど取り出されるとすれば，結果的には本手法の重要説明箇所の特定と非常に近いことが行われていることになる．この異なる2つのアプローチの結果を比較検討してみることは重要なことで，今後の課題としたい．

5. おわりに

本論文では，テキスト中の語の出現密度分布を利用して語の重要説明箇所を特定する方法を提案した．密度計算には，ハニング窓関数を用いることによってある範囲の語の出現を重み付きで加算するという方法を用いた．手法の有効性は新書10冊，100キーワードに対する実験によって示した．

今後はこの手法を電子図書館プロトタイプシステムなどに組み込み，実際の運用実験を行う予定である．また，本手法をハイパーテキストのリンク自動付与に利用する実験も行う予定である．

謝辞 評価実験を手伝っていただいた坂口昌子氏，齊藤由衣子氏，小山哲春氏に感謝いたします．また，貴重なコメントをいただいた査読者の方々に感謝いたします．

参考文献

1) Agosti, M., Crestani, F. and Melucci, M.: Design and Implementation of a Tool for the Automatic Construction of Hypertexts for Information Retrieval, *Information Processing & Management*, Vol.32, No.4, pp.459–479 (1996).
2) Allan, J.: Automatic Hypertext Link Typing, *Proceedings of Hypertext '96*, pp.42–52 (1996).
3) Bookstein, A., Klein, S.T. and Raita, T.: Detecting Content-bearing Words by Serial Clustering – Extended Abstract, *Proc. SIGIR '95*, pp.319–327 (1995).
4) 原田隆史，細野公男，野美山浩，諸橋正幸：抄録からのキーワードの自動抽出，情報処理学会研究報告，Vol.94, No.37 (94-IF-33), pp.35–40 (1994).
5) Hearst, M.A.: Tilebars: Visualization of Term Distribution Information in Full Text Information Access, *Proc. ACM SIGCHI Conference on Human Factors in Computing Systems*, pp.59–66 (1995).
6) Hearst, M.A. and Plaunt, C.: Subtopic Structuring for Full-length Document Access, *Proc. SIGIR '93*, pp.59–68 (1993).
7) 木本晴夫：日本語新聞記事からのキーワード自動抽出と重要度評価，電子情報通信学会論文誌 (D-I), Vol.J74-D-I, No.8, pp.556–566 (1991).
8) 黒橋禎夫，長尾 眞，佐藤理史，村上雅彦：専門用語辞典の自動的ハイパーテキスト化の方法，人工知能学会誌，Vol.3, No.7, pp.336–345 (1992).
9) 長尾 眞：パターン情報処理，コロナ社 (1983).
10) 中渡瀬秀一，木本晴夫：統計的手法によるテキストからの重要語抽出メカニズム，情報処理学会研究報告，Vol.95, No.87 (95-IF-39), pp.41–48 (1995).
11) Ogawa, Y., Bessho, A. and Hirose, M.: Simple Word Strings as Compound Keywords: An Indexing and Ranking Method for Japanese Texts, *Proc. SIGIR '93*, pp.227–236 (1993).
12) Rada, R.: Converting a Textbook to Hypertext, *ACM Trans. Information Systems*, Vol.10, No.3, pp.294–315 (1992).
13) Salton, G. and McGill, M.J.: *Introduction to Modern Information Retrieval*, McGraw-Hill, New York (1983).

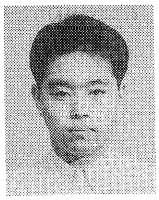
14) Salton, G., Singhal, A., Buckley, C. and Mitra, M.: Automatic Text Decomposition using Text Segments and Text Themes, *Proc. Hypertext '96*, pp.53–65 (1996).

(平成8年8月29日受付)
(平成9年2月 5 日採録)

付録 実験対象の書名・キーワード一覧

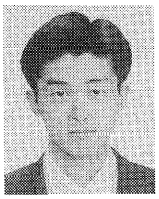
書名	キーワード
白球礼讃 (ベースボールよ永遠に)	愛 (45, 36) フットボール (12, 9) 草野球 (38, 34) グリップ (6, 6) ベーブ・ルース (21, 16) 野球人生 (2, 2) 職人 (23, 18) プロ野球 (35, 26) ルール (20, 18) ソフトボール (6, 5)
ハイテク社会と労働	情報化社会 (11, 10) ME化 (1, 1) ロボット (206, 104) FA (61, 46) インテリジェントビル (39, 30) IDカード (14, 8) OS (20, 16) ソフトウェア (168, 108) メカトロ (61, 41) 職人 (28, 18)
報道写真家	天女 (6, 4) 学生 (37, 24) ジャーナリスト (45, 32) テーマ (12, 9) 戦争 (91, 65) 賞 (21, 6) 政治 (18, 18) 反日 (6, 6) 兵士 (50, 31) デモ (40, 29)
象徴天皇制への道	戦争犯罪人 (6, 6) 穏健派 (71, 52) 自由主義 (17, 15) ポツダム宣言 (27, 24) 軍国主義 (46, 39) 降伏 (43, 33) 憲法草案 (14, 13) 軍隊 (12, 11) 民主主義 (20, 17) 日本国憲法 (15, 13)
リゾート列島	温泉 (8, 7) ゴルフ場 (65, 49) サンゴ礁 (3, 3) サウナ (11, 9) 国立公園 (13, 8) ログハウス (18, 13) バカンス (13, 11) セントラル・パーク (1, 1) 田園リゾート (21, 20) リゾート・オフィス (7, 4)
都市と水	ウォータフロント (1, 1) カスリン台風 (5, 5) 水害 (127, 77) 都市治水 (2, 2) 地下分水路 (12, 10) 節水 (19, 14) 下水道 (53, 30) BOD値 (1, 1) 発電所 (6, 3) 地盤沈下 (14, 11)
中国とソ連	原爆 (8, 8) マルクス・レーニン主義 (10, 9) イデオロギー (31, 26) 文化大革命 (22, 18) 中ソ同盟条約 (6, 6) INF (2, 2) 経済 (161, 99) ペレストロイカ (11, 10) 貿易 (67, 36) 社会主義 (133, 92)
やきもの文化史	青磁 (175, 100) タイル (9, 8) 茶碗 (50, 30) 梅瓶 (10, 9) 顔料 (12, 11) 商人 (24, 22) コレクター (11, 10) 回教 (46, 35) 東インド会社 (12, 11) カオリン (24, 19)
日本語と外国語	外国語運用能力 (1, 1) ドイツ語 (55, 45) 同音衝突 (4, 4) 漢字 (188, 115) イギリス人 (28, 23) 色 (518, 280) 国際化 (9, 9) 表記 (36, 35) 文化 (148, 118) 相互交流 (3, 3)
仏教入門	哲学 (43, 36) キリスト教 (12, 12) サンスクリット (68, 60) 修行 (28, 26) 寺 (89, 53) 中国 (98, 83) 学派 (8, 8) 宗派 (11, 8) 葬儀 (1, 1) 出家 (34, 30)

(括弧内の数字は各キーワードの出現回数と出現段落数)



黒橋　禎夫（正会員）

1966年生．1989年京都大学工学部電気工学第二学科卒業．1994年同大学院博士課程修了．博士（工学）．同年，京都大学工学部助手，現在に至る．自然言語処理，知識情報処理の研究に従事．1994年4月より1年間Pennsylvania大学客員研究員．



白木　伸征

1973年生．1997年京都大学工学部電気系学科卒業．現在，同大学院工学研究科修士課程在学中．自然言語処理，知識情報処理の研究に従事．

長尾　　眞（正会員）

1936年生．1959年京都大学工学部電子工学科卒．工学博士．京都大学工学部助手，助教授を経て，1973年より京都大学工学部教授．国立民族学博物館併任教授（1976～94年），京都大学大型計算機センター長（1986～90年），日本認知科学会会長（1988～90年），パターン認識国際委員会副会長（1982～84年），機械翻訳国際連盟初代会長（1991～93年），電子情報通信学会副会長（1993～95年），情報処理学会副会長（1994～96年），言語処理学会会長（1994～96年），京都大学附属図書館長（1995年～）．パターン認識，画像処理，機械翻訳，自然言語処理などの分野を並行して研究．