E-019

出現密度分布を用いた重要文抽出方式
Sentence Extraction based on Terms Frequency Density Distribution

渡辺修司† 栗山義明‡ 絹川博之†
Shuji WATANABE Yoshiaki KURIYAMA Hiroshi KINUKAWA

1 はじめに

テキストは情報を表現する最も汎用のメディアであり，計算機の急速な普及によりテキストの電子化が進み，膨大な情報が計算機上でアクセス可能になれば，テキスト中の重要な情報を伝えている文のみを参照するという利用法が頻繁で重要になるだろう。本研究では，テキスト中から索引語を選定し，索引語の出現密度分布を調べ，その高密度な出現位置を重要文として抽出する方式に関し NTCIR ２テストコレクションを用いて実験評価する。

2 索引語の重要度付け方法

①tf-idf ② tf / TF ③ SMART の３つの方法で，索引語の重み付けを行う。また、文書 d における索引語 t の重要度を w_t^d と表す。

3 重要文抽出処理方式

3．1 ハニング窓関数について

- 範囲の中心付近の出現を重視し，中心から離れるにしたがって重みを軽くする。
- 範囲の両端付近と，範囲の外側(出現を考慮しない部分)との差を連続的にする。

このような性質の重み付けの関数として，ハニング窓関数がある。窓の幅（重みを与える幅）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。

$$h_l(i) = \frac{1}{2}(1 + \cos 2\pi \frac{i-l}{W}) \quad (|i-l| \le W/2) \qquad （式１）$$

本手法ではハニング窓関数を用いて索引語の出現密度を計算する．窓の幅 W は 5 文字、10 文字、50 文字 100 して実験した。

3．2 ハニング窓関数を用いた密度計算

ハニング窓関数を用いた索引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一本の長い文字列（長さ L 文字）とみなし，テキスト中での索引語の全ての出現位置を調べる。位置 i を先頭として索引語 t が出現する場合 $a(i) = w_t^d$，そうでない場合 $a(i) = 0$ 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 に対する索引語の出現密度 $d(l)$ 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 $W/2$ の範囲の索引語の出現を次式によって足し込む。

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

（$i < 0$ または $i \ge L$ では $a(i) = 0$）

3．3 重要文抽出方法

各文の出現密度の最大値により文を順位づけ、上位から要約率に設定しただけ重要文として抽出するという方法とした。図１に重要文抽出までの流れを示す。また，図２に密度分布の様子を示す。

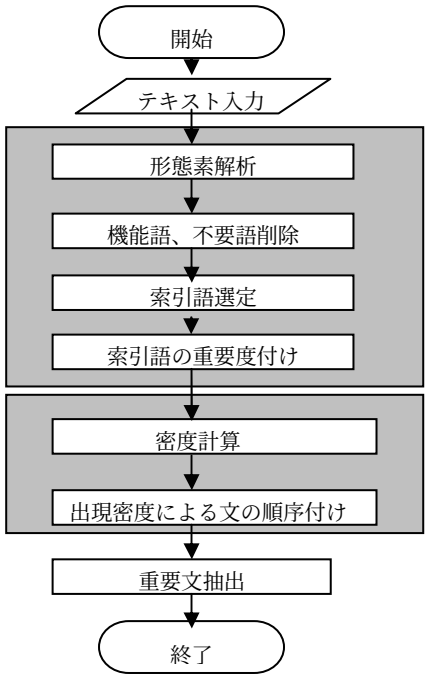
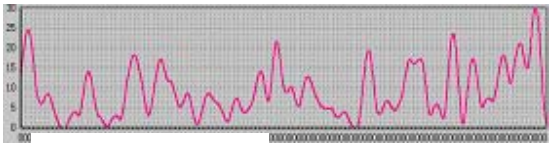


図１．重要文抽出の流れ

† 東京電機大学大学院工学研究科
‡ 株式会社日立情報システムズ



索引語の出現位置

図２．密度分布の様子

４　実験

４．１　実験データ

提案した手法の有効性を調べるために NTCIR2 テストコレクションのうち毎日新聞の社説で、Formalrun 用の１５記事を使用した。

４．２　評価方法

以下のように定義した精度 P と再現率 R を求める。

$$p = \frac{\text{抽出された適合重要文数}}{\text{抽出された重要文候補数}}$$

$$R = \frac{\text{抽出された適合重要文数}}{\text{抽出されるべき重要文数}}$$

ただし、今回の実験では抽出する文数を抽出されるべき文数に合わせた為、P と R は等しくなる。よって、以後これらの値は F 値で表すことにする。

$$F = \frac{2RP}{R+P} = R = P$$

４．３　評価結果

表１　ＳＭＡＲＴ法の時

	要約率１０％	３０％	５０％
５文字	0.116	0.306	0.511
１０文字	0.154	0.309	0.490
５０文字	0.143	0.307	0.475
１００文字	0.163	0.290	0.471

表２　tf-idf ,tf/TF 法の時

	要約率１０％	３０％	５０％
5 文字	0.106	0.334	0.517
10 文字	0.128	0.342	0.524
50 文字	0.111	0.288	0.504
100 文字	0.124	0.269	0.500

5　考察

ハニング窓関数を用い、索引語の出現密度を求め高密度の箇所を重要と考え、その箇所を含む文を重要文として抽出する手法は、NTCIR コンテスト結果と比べ有効ではないという結果になった。

(1)原因として、まず窓幅を広くした場合を考える。窓幅を広くするとまず密度が平均されてしまう。つまり広い範囲で密度の差が出なくなってしまう。

また、文書を一本の長い文字列として考えて、索引語の出現密度を求めるところにも、原因があると考えられる。つまり、各文の前半部分の密度は、１つ前の文の後半部分が影響してしまい、各文の後半部分の密度は、１つ後ろの文の前半部分が影響してしまう。そうすると、１文単位で抽出するにもかかわらず、他の文の影響がでるということなので、それは問題であると考えられる。

(2)次に、窓幅を狭くした場合は索引語の重要度の影響が強すぎると考えられる。つまり、重要度が高い索引語が出現するだけで、そこが重要箇所として選ばれてしまうので、他の索引語がほとんど考慮されることがない。特に要約率が１０％の時などは、索引語の重要度のトップが含まれるだけで、その文が選ばれるようになるので、良い精度を得るのが、難しくなると考えられる。

(3)また、根本的に索引語の出現密度が高ければ重要な箇所という考えがそれほど有効ではないと考えられる。

6　終わりに

当初は、ハニング窓関数を用いて索引語の出現密度を求め、密度の高い文を、重要文として抽出するという手法は有効なのではないかと考えていた。しかし、評価実験を行い精度を求めてみたところ NTCIR の他の方式の精度と比べて、劣っているということが分かった。

7　参考文献

[1]奈良先端科学技術大学 松本研究室 URL：
http://chasen.aist-nara.ac.jp/index.html.ja

[2] NTCIR 情報検索システム評価用テストコレクション構築プロジェクト URL：
http://research.nii.ac.jp/ntcir/index-ja.htm

[3]徳永健伸：“ 情報検索と言語処理. ”
東京大学出版会（1999）

[4]黒橋禎夫，白木伸征，長尾眞：出現密度分布を用いた語の重要説明個所の特定，情報処理学会研究報告 96-NL-115，43—50

[5]栗山義明，絹川博之：ターム群の出現密度分布を用いた重要文抽出方式，FIT2002 情報科学技術フォーラム情報技術レターズ　LE-9