# LE-9　タ－ム群の出現密度分布を用いた重要文抽出方式

## －タ－ム種別重みの評価実験－

## Sentence Extraction based on Terms Frequency Density Distribution

栗山 義明†*　絹川 博之†
Yoshiaki Kuriyama　Hiroshi Kinukawa

## 1．はじめに

コンピュータとネットワークの普及により，膨大なテキストの電子化が進み，テキスト中の重要な情報を表す文のみを抽出し参照できることが強く求められている．

本研究では，テキスト中から複数のタームを選定し，複数のターム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し，抄録を作成する方式を提案する．

## 2．タ－ム群の出現密度分布を用いた重要文抽出方式

テキスト中の重要な文には，重要なタームが多く含まれると考えられるとの仮定から，従来，出現重要タームで重み付けする重要文抽出方式が知られている．

一方で，ある指定したキーワードについて重要な説明箇所を特定する方法として，出現密度分布を利用する方法が提案されている[1]．これを発展させ，複数の重要なタームの出現密度の高い文は重要な文ということができると考えられる．この点に着目し，以下の重要文抽出方式を提案する．すなわち，

まず与えられたテキストから重要と考えられる複数のタームを選定する（これを以下，タ－ム群とよぶ）．

次に選定したタ－ム群の出現密度分布の高い文を重要文として抽出する．

## 3．タ－ム群選定方式

テキストは，タイトル，サブタイトル，目次，本文から構成される10000語程度の日本語Webテキストとする．ただし，サブタイトル，目次は無いものもある．

中頻度語に重要な語が多い[2]という点を日本語に適用し，以下の手順でタ－ム群を選定する．

(1) ChaSen[3]を利用してテキストを形態素解析し，高頻度の機能語を削除する[4].
(2) 不要語リストを利用して内容語から高頻度の不要語を削除する[4].
(3) (1)，(2)の結果残った語をタ－ム群候補とする．
(4) 語の頻度を利用してタ－ム群候補に重要度を付与する．ある文書 $d$ 中に出現するターム $t$ の重要度を $w_t^d$，頻度を $tf(t,d)$，ターム種別重みを $T_w$（5章参照）とする．

$$w_t^d = tf(t,d) \times T_w \qquad (式1)$$

(5) タ－ム群候補を重要度の降順に整列し，異なり語で上位10分の1をタ－ム群として選定する．

## 4．重要文抽出方式

### 4.1　ハニング窓関数について

重要な文章がある程度まとまって収まるような一定の範囲を設定する．

- 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出現位置重みを軽くする．
- 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．

ハニング窓関数[1]は以上の二つの性質をもつ．出現位置重みを与える窓の幅を $W$，窓の中心位置を $l$ とすると，ハニング窓関数 $h_l(i)$ は次式により与えられる．

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

本方式ではハニング窓関数を用いてタームの出現密度を計算する．なお，窓の幅 $W$ はテキストの文字数の10分の1とした．

### 4.2　ハニング窓関数を用いた密度計算

タームの出現密度計算は以下のアルゴリズムで行う．

(1) テキストを一本の文字列とみなし，位置 $l$ を先頭としてターム $t$ が出現する場合 $a(l) = w_t^d$，そうでない場合 $a(l) = 0$ とする．ここでタームを1語，$W$ = 1500，ターム $t$ が出現する場合 $a(l) = 1$ とすれば[1]の手法になる．
(2) テキストの先頭から順番に各位置をハニング窓の中心位置 $l$，ハニング窓の幅を $W$，テキストの文字数を $L$ とすると，$l$ の前後それぞれ $W/2$ の範囲のタームの出現密度 $d(l)$ は次式により与える．

$$d(l) = \sum_{i=l-\frac{W}{2}}^{l+\frac{W}{2}} h_l(i) \cdot a(i) \qquad (式3)$$

($i<0$ または $i \ge L$ では $a(i) = 0$)

### 4.3　抄録作成方法

タームの出現密度分布を求めると，密度が高い出現位置が重要箇所であろうということがわかる．出現密度分布の面積を求め，面積上位 20%を閾値とし，閾値以上にあるタームを含む文を重要文として抽出し，抄録を作成する（図1）．

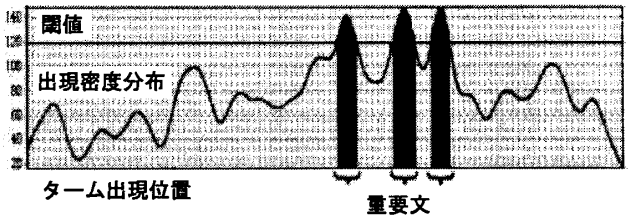


図1：重要文抽出方式

†東京電機大学工学部

*株式会社日立情報システムズ（現在）

## 5. ターム種別重み $T_w$の分析実験

以下の(1)～(3)の語はテキスト中の重要な概念を表しているという仮定をもとに，（式 1）のターム $t$ に対するターム種別重み $T_w$について分析する．

(1) $t$がタイトル中に出現する場合

(a)タイトル中に出現するターム：$T_w = 1.0 \sim 2.0$

タイトル中に出現しないターム：$T_w = 1.0$

(b) タイトル中に出現するターム：$T_w = 1.0 \sim 2.0$

タイトル中に出現しないターム：$T_w = 0.0$

(a)，(b)二通りの場合について $T_w$の値を 0.1 刻みで変化させて重要文を抽出し，結果を比較する．すると，$T_w$の値が 1.0～1.5 のときは(a)，(b)の結果が異なるが，$T_w$の値が 1.6～2.0 のときは(a)，(b)の結果が等しくなる．つまり，$T_w$の値が 1.6～2.0 のときはタイトル中に出現するタームの重要度が強すぎて，それ以外のタームの影響がないということである．また，結果として，抽出された文のほとんどがタイトル中に出現するターム群の説明文であった．

同様に，(2)，(3)について求める．

(2) $t$がサブタイトル及び目次中に出現する場合

$T_w$の値が 1.4～2.0 のとき，サブタイトル及び目次中に出現するタームの重要度が強すぎ，抽出された文のほとんどがサブタイトル及び目次中に出現するターム群の説明文であった．

(3) $t$が固有名詞（人名，企業名，地名）の場合

$T_w$の値が 1.3～2.0 のとき，固有名詞であるタームの重要度が強すぎ，抽出された文のほとんどが固有名詞であるターム群の説明文であった．

## 6. 重要文抽出実験

### 6.1 実験データ

提案した方式の有効性を調べるため，Web 上のオープンソースに関する論文 30 文書を用いて実験した．

### 6.2 ターム種別重み $T_w$の設定

重要文抽出時の（式 1）におけるターム種別重み $T_w$は，以下のように設定した．

表 1：設定したターム種別重み一覧

| タームtの種別 | 重み $T_w$ |
|---|---|
| タイトル中に出現する | 1.5 |
| サブタイトル及び目次に出現する | 1.3 |
| 固有名詞（人名，企業名，地名） | 1.2 |
| 上記以外 | 1.0 |

### 6.3 評価方法

次のような 3 段階の評価基準を考えた．

○：重要である．

△：ある程度重要だが，抄録として絶対に提示しなければならないという部分ではない．

×：重要でない．

この 3 段階の評価を 30 文書の個々の文に対して 2 人の人間が別々に行った．そして，2 人の評価が一致すればそのまま，一致しない場合には○と×は△，○と△は○，△と×は×として人間による評価値を決定し，以下のように定義される精度と再現率を求めた．

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

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

### 6.4 実験結果

実験 1　ターム種別重み $T_w$を表 1 の設定による場合

実験 2　ターム種別重み $T_w$をすべて 1.0 にした場合

実験 1，実験 2 の二通りの場合について比較実験し，精度及び再現率をそれぞれ求めた．

表 2：重要文抽出実験結果

| | 実験 1 | | 実験 2 | |
|---|---|---|---|---|
| | ○を正解 | ○と△を正解 | ○を正解 | ○と△を正解 |
| 精度 | 0.815 | 0.795 | 0.775 | 0.765 |
| 再現率 | 0.705 | 0.755 | 0.700 | 0.745 |

## 7. 考察

5 章，6 章の実験結果から，重要文抽出において重要なターム群の出現密度分布の高い文への着目が有効であることがわかった．

また，本方式におけるターム種別重み $T_w$を表 1 の設定にすることにより，$T_w$をすべて 1.0 にする場合と比べて，精度が 0.03～0.04，再現率が 0.005～0.01 向上することがわかった．

さらに，重要文抽出において，タイトル中に出現するタームの $T_w$を 1.6 以上，サブタイトル及び目次中に出現するタームの $T_w$を 1.4 以上，固有名詞であるタームの $T_w$を 1.3 以上に設定すると，これら以外のタームの効果がなくなることがわかった．

本実験では $T_w$の値を表 1 による設定としたが，今後 $T_w$の最適な値の組合わせを求め，精度，再現率ともに向上させること，テストコレクションを用いて評価することが必要である．

## 8. おわりに

本研究では，黒橋らの方式[1]を拡張し，テキスト中のターム群の出現密度分布を利用して重要文を抽出し，抄録を作成する新しい方式を提案した．本方式の有効性は 30 文書に対する実験によって示した．また，本方式におけるターム種別重み $T_w$の適用可能範囲を特定した．

ターム種別重み $T_w$の最適な値の組合わせを求め，精度，再現率を向上させること，テストコレクションを用いて評価することが今後の課題である．

## 9. 参考文献

[1] 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

[2] H. P. Luhn : The Automatic Creation of Literature Abstracts, 1958

[3] http://chasen.aist-ara.ac.jp/index.html.en

[4] 長尾眞：自然言語処理，岩波書店，pp417-421，1996