出現密度分布を用いた図表付き抄録の自動作成

1U－10

野久 仁志　黄瀬 浩一　松本 啓之亮
大阪府立大学 工学部 情報工学科

1 はじめに

抄録は文書の概要を把握する上で有用であり，従来から自動作成の研究が盛んに行われている．従来法の多くは文書中のテキスト部分だけを対象としている．しかし多くの文書ではテキストの他に図や表が使われる．図表は複雑な構造をわかりやすく表現できるので，抄録に文書の概略を表す図表が加われば，より強力に読者の概要把握を支援できる．

本稿では，概略を表すテキストと図表を文書から抽出し，図表付き抄録を自動作成する手法を提案する．本手法の特徴は，テキストや図表の重要さをテキストにおけるキーワード出現密度を用いて判定する点である．

2 キーワードの出現傾向と重要箇所

本手法では，主題を説明する重要テキストと概略を表す重要図表を文書から選別して，図表付き抄録を作成する．重要テキストと重要図表の選別にはテキストでのキーワードの出現傾向を利用する．一般に“あるキーワードを説明する文では，そのキーワードを頻繁に使う”という傾向がある．この傾向を利用すると，キーワードの密集部分からそのキーワードの説明文を特定できる [1].

そこでテキストや図表を適切なキーワードの集合に置き換え，キーワードの出現傾向から重要箇所を特定する．以下に重要テキストと重要図表の選別法を示す．

重要テキスト　重要テキストは文書の主題を説明する文の集合である．したがって主題に関連する語をテキストキーワードとすると，その出現傾向から重要テキストが特定できる．本手法では文書タイトルと章節タイトル中の自立語をテキストキーワードとする．

重要図表　重要図表は文書の概略を表す図表の集合である．図表自体の解析は困難であるので，本手法では図表説明文を利用する．まず図表本体と図表キャプション中の自立語を図表キーワードとして，その出現傾向から図表説明文を特定する [2]. そして図表説明文中に重要な文を多く含む図表を重要図表とする．

3 図表付き抄録作成

図表付き抄録作成手順を図1に示す．

3.1 キーワード出現密度計算

本手法では，キーワードの出現傾向をハニング窓関数による出現密度で表現する．ハニング窓関数 $h(i)$ は式

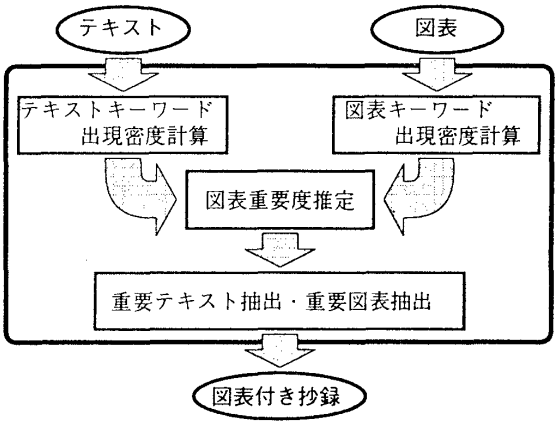

図1　図表付き抄録作成

1で表され，中心が最も高く，窓幅 W 以外では0となる．

$$h(i) = \frac{1}{2}\left(1 + \cos 2\pi \frac{i}{W}\right) \quad \text{ただし} |i| \leq W/2 \quad (1)$$

具体的には以下の手順で出現密度を計算する．

1. 形態素解析… テキスト部分を形態素解析し，単語列 $a(l)(0 \leq l < L)$ に分解する．L は全単語数を示す．
2. キーワード検出… キーワードを単語列 $a(l)$ から捜し，出現を $k(l)$ に記録する．$k(l)$ は位置 l にキーワードが出現する場合に1，出現しない場合に0となる．
3. 出現密度計算… 位置 l における出現密度 $d(l)$ は式2で定義される [1]. ただし $l < 0$ または $l \geq L$ では $k(l) = 0$ とする．

$$d(l) = \sum_{i=-\frac{W}{2}}^{\frac{W}{2}} h(i) \cdot k(l-i) \quad (2)$$

ハニング窓関数の特徴により，キーワード密集部分では出現密度が高くなる．図2，図3にテキストキーワード出現密度分布と図表キーワード出現密度分布を示す．キーワード出現密度が高い部分では，キーワードが説明されている可能性が高い．したがって，テキストキーワード出現密度はテキストの主題を説明する文が存在する可能性を表し，図表キーワード出現密度は図表説明文が存在する可能性を表すと捉えることができる．

3.2 図表重要度推定

各図表が文書の概略を示す度合を図表重要度で表す．本手法では，主題に関連する文を図表説明文中に多く含む図表ほど重要となる．また主題に関連する文と図表説明文が存在する可能性は，テキストキーワード出現密度分布と図表キーワード出現密度分布で表現されている．そこでテキストキーワード出現密度分布と図表キーワード出現密度分布の一致具合を，出現密度分布の積を使って数値化する [3]. 具体的には以下の処理を施す．

Automatic Generation of Summary with Figures and Tables Using Density Distribution of Words
Hitoshi Nohisa, Koichi Kise and Keinosuke Matsumoto
Department of Computer and Systems Sciences, College of Engineering, Osaka Prefecture University

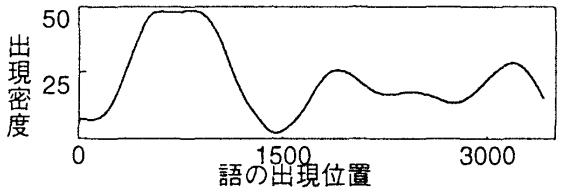

図 2　テキストキーワード出現密度分布

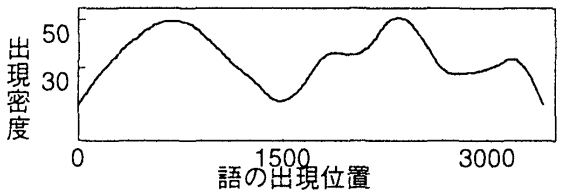

図 3　図表キーワード出現密度分布

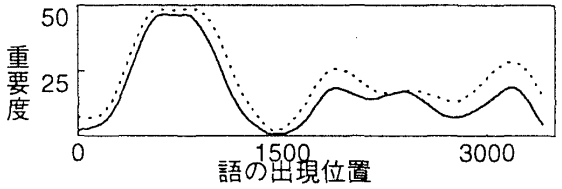

図 4　図表重要度分布

1. 正規化… 図表キーワード出現密度の最大値を求め，最大値で図表キーワード出現密度分布を正規化する．
2. 図表重要度分布の作成… 各位置でテキストキーワード出現密度と正規化した図表キーワード出現密度を乗算し，図 4の実線のような図表重要度分布を作る．
3. 図表重要度計算… テキストキーワード出現密度分布 (図 4: 破線) の面積に対する図表重要度分布 (図 4: 実線) の面積の比を求め，図表重要度とする．

図表キーワード出現密度の最大値で正規化するため，起伏が激しい分布と平坦な分布では，後者の方が図表重要度が大きくなりやすい．

3.3　重要部分抽出

算出された図表重要度に基づいて全図表を順位付けする．上位の図表ほど重要であるので，本手法では最大の図表重要度を持つ図表を重要図表とする．

重要テキストには，主題に関連する文とともに重要図表の説明文を加える．重要図表の説明文を考慮するために，重要図表の図表重要度分布 $d_i(l)$ をしきい値処理する．しきい値を T, $d_i(l)$ の最大値を d_{max} とすると，式 3を満たす語の出現位置 l を全て求める．

$$T \leq d_i(l) / d_{max} \qquad \text{ただし} 0 \leq l < L \qquad (3)$$

位置 l に存在する単語を含む文が重要テキストとなる．

4　実験および考察

文書サンプル 20 個を用いてパラメータ設定実験を行った．パラメータとして，テキストキーワード出現密度計算時の窓幅 W_t と図表キーワード出現密度計算時の窓幅 W_f をそれぞれ 100 語～ 1000 語の範囲で幅 100 語，しきい値 T を 0.05 ～ 0.3 の範囲で幅 0.05 で変化させた．重要テキストの再現率 R と適合率 P, 重要図表の抽出率 E

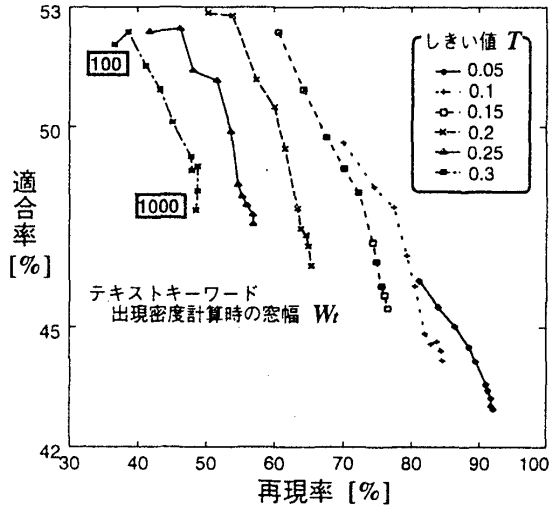

図 5　W_f=100 において W_t と T を変化させた場合の重要テキストの再現率と適合率

を調べ，それらの和 P+R+E を最大にする (W_t, W_f, T) の組を最適パラメータとした．形態素解析器は JUMAN 3.5[4]，正解データは人手で作成したものを用いた．

実験の結果，最適パラメータは $W_t = 1000, W_f = 100$, $T = 0.05$ となった．主要なパラメータを変化させた場合の再現率 R と適合率 P の変化を図 5に示す．図 5が示すように，W_t を大きくすると再現率重視となる．一方 W_f を変化させても，R, P, E はほとんど変化しなかった．また図 5で T が大きいほど再現率が下がるのは，最大値に対するしきい値処理で重要テキストを抽出するためである．人間は文書全体から少しずつ重要文を抽出する傾向がある．そのため重要テキストが最大値周辺に集中する本手法では，良い結果が得られない．これは極大値に対するしきい値処理で改善できると考える．

図表抽出率はパラメータの変化によらずほぼ一定であった．これは，各図表の図表重要度がパラメータに関わらず一定であることが原因である．この結果から図表重要度の設定は妥当であると考えられる．

5　おわりに

本稿では，文書から重要テキストと重要図表を抽出し，図表付き抄録を自動作成する手法を提案した．テキストと図表の重要度推定には，キーワード出現密度を用いた．実験により，本手法の妥当性と有効性が示された．

今後の課題としては，極大値を考慮した重要テキスト抽出と複数の重要図表抽出への対応ならびに対象に応じたパラメータの自動設定があげられる．

参考文献

[1] 黒橋，白木，長尾: “出現密度分布を用いた語の重要説明箇所の特定”，情処学論，Vol.38, No.4, pp.845-854(1997).
[2] 水野，黄瀬，松本: “単語の出現密度分布を用いた図表と説明テキストの対応付け”，第 57 回情処全大 4V-1(1998).
[3] 野久，黄瀬，松本: “図表と説明テキストの対応付けを利用した重要図表抽出”，第 57 回情処全大 4V-2(1998).
[4] 黒橋，長尾: “日本語形態素解析システム JUMAN version 3.5”，京都大学工学部大学院工学研究科 (1998).