Vol. 29 No. 3 情報処理学会論文誌 Mar. 1988

ショートノート

キーワード密度方式自動抄録法の改良†

——高頻度隣接語による改善——

鈴木康広†† 栃内香次††

文献の抄録を計算機によって自動的に作成する方法としてキーワード密度法が挙げられる．しかし，キーワード密度法にはいくつかの問題点がある．我々は，これを改善するためキーワードの代りに文献の内容に関係の深い語の対（以下，これを高頻度隣接語と呼ぶ）を利用する自動抄録システムの研究を行っており，実際に両方式による自動抄録システムを作成し，情報処理関係の論文など10篇の抄録実験を行った．本論文では，高頻度隣接語を利用した自動抄録方式とそれによる抄録作成実験の結果，およびキーワード密度法との比較実験結果について述べる．

1. はじめに

科学技術文献の急増に伴い，抄録誌などの2次資料の重要性が増すとともに量も増加し，文献抄録の自動作成が課題となっている．文献の自動抄録に関しては，様々な手法が研究されているが，そのほとんどがLuhn の提唱したキーワード密度法をもとにしている[1]．キーワード密度法は文献の内容に関係の深い数語のキーワードを抽出し，これらの語を高頻度で含む文を文献中から抽出して抄録とするものであるが，これには一般に次のような問題点がある．

1. 文献中におけるキーワード分布の偏りが小さいために抄録文の候補が多数になり，その中から適切な抄録文を抽出することが困難である．
2. 図の説明文など主題からはずれた文が抄録文として選ばれることがある[2]．
3. 抽出された抄録文は前後の脈絡が欠けており，全体としてのまとまりを失いやすい[2]．

本論文は，キーワードを拡張した高頻度隣接語の導入による，キーワード密度法における上記問題点の改善について述べるものである．

2. 高頻度隣接語

文章中で隣接または近接している語の組を隣接語と呼ぶことにする．ここで，単語間の助詞や動詞，形容詞の活用語尾などは無視する．これは，助詞や活用語尾などを無視しても隣接語の意味的関係には変化がないと考えられるからである．以下に隣接語の例を示す．

例）～のシステム構成は～ →システム一構成
～のシステムの構成は～ →システム一構成
～のシステムを構成する～→システム一構成

上の例の各文では隣接語はいずれも「システム一構成」となる．

隣接語の中からどのような文献にも一般的に出現する隣接語（「のような一場合」，「必要一である」など）を除いた後，出現頻度の大きいものから数組を取り出したものを高頻度隣接語と定義する．なお，キーワード密度法におけるキーワードは，隣接語ではなく個別の単語について同様な処理を行うことによって抽出される．**表1**は，5編の文献について，各々高頻度隣接語およびキーワードを抽出し，各文献ごとにこれらを含む文の割合を求めたものである．この結果から，高頻度隣接語にくらべてキーワードの方が一文献中に広く分布していることがわかる．このことは，高頻度隣接語を含む文はキーワードを含む文より少数であり，抄録文の抽出が容易であることを意味する．

3. 抄録文の抽出

キーワード密度法は，以下の前提に基づいている[3]．

1. 一つの文献において，その主題と関係の深い語は概して文献中に繰り返し出現する．
2. 一文中にこのような語を多数含む文は，文献の

† Improvement of Automatic Abstraction Using Key-word Density Method —Improvement by "Connective-Word"— by YASUHIRO SUZUKI and KOJI TOCHINAI (Department of Electronic Engineering, Faculty of Engineering, Hokkaido University).

†† 北海道大学工学部電子工学科

表 1 キーワードおよび高頻度隣接語の分布
Table 1 Distribution of key-words and connective-words.

高頻度隣接語またはキーワードの語数		3		4	
文献番号	総文数 (文)	高頻度隣接語含有率 (%)	キーワード含有率 (%)	高頻度隣接語含有率 (%)	キーワード含有率 (%)
1	220	38.2	52.1	43.3	59.9
2	170	19.9	50.0	29.5	53.5
3	144	35.2	59.3	39.4	60.7
4	161	30.6	49.9	31.2	51.0
5	179	12.4	43.0	16.3	50.8

中でも重要な部分で，抄録文として適切な文である．

キーワード密度法による抄録手順を以下に示す[4]．

1. 抄録対象文献の語彙調査を行う．
2. 出現頻度の大きな語の中からどのような文献にも一般的に出現する語（「表す」，「行う」，「である」など）を除いた後，頻度上位のものから数語を抽出し，これをキーワードとする．
3. 文献中の各文について，キーワードの占める比率を求め，その大きさの順に各文に順位をつける．
4. 順位が上位の数文を抄録文として選択する．
5. 選択された抄録文をもとの文献中の出現順に並べて抄録とする．

ここで，キーワード語数および選択抄録文数はあらかじめ定めておく必要がある．本論文で述べる方法は，キーワードに代えて前述の高頻度隣接語を用いるもので，この場合も上記と同様の手順によって抄録文を抽出することができる．以下，これを高頻度隣接語法と呼ぶ．

4. 実験システム

高頻度隣接語法とキーワード密度法各々の実験システムを作成し，自動抄録実験を行った．実験システムは PL/I で書かれており，北海道大学大型計算機センターの HITAC M-680 H 上に作成されている．

前述のように，高頻度隣接語法とキーワード密度法の処理手順は基本的には同一である．以下，**図 1** に示す高頻度隣接語法の処理手順について述べる．

A：文章ファイルの整形

正確な語彙調査を行うため抄録対象文献の文章の整形を行う．これは，下記に示すように漢字の送り仮名の統一や複合語の区切りの統一などを行うものである．

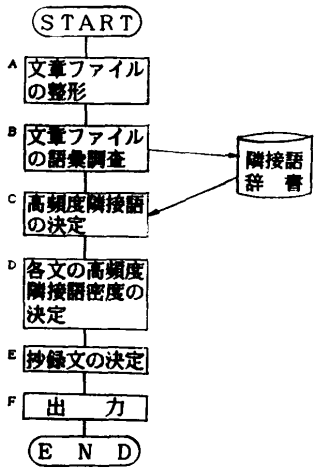

図 1 処理の流れ
Fig. 1 Flow of transaction.

例）表す，表わす →表す
本論文，本／論文→本／論文

B：文章ファイルの語彙調査

文章ファイルの語彙調査を行い隣接語辞書を作成する．隣接語辞書には隣接または近接する語の組と，その単語間の 2 文字以下の文字列が頻度順に記録されており，単語辞書には単語が頻度順に記録されている．隣接語辞書，単語辞書には文献全体を対象として作成した辞書と各章ごとに作成した個別の辞書とがある．

C：高頻度隣接語の決定

Bで作成した隣接語辞書からどのような文章にもよく出現する隣接語を削除した後，頻度上位の一定数を抽出して高頻度隣接語とする．削除すべき隣接語は，抄録対象文献と同一分野の多数の文献について隣接語の頻度を求め，その上位のもの十数語を対象としている．なお，高頻度隣接語の抽出は文献全体を対象とした辞書と各章ごと個別の辞書との 2 種の辞書から独立に行い，各々の語数は実験的に定めている．

D：各文の高頻度隣接語密度の決定

高頻度隣接語を文献全体から M 語，各章から N 語ずつ抽出し各章ごとに $M+N$ 語を用いて各文の高頻度隣接語の密度を算出する．なお，実験は後述のように (M, N) の値が(4, 3)と(3, 3)の場合について行った．

E：抄録文の決定

各章ごとに高頻度隣接語密度の大きいものから一定個の文を抄録文として抽出する．一般に，科学技術文献では一節の最初の部分でその節の主題が述べられ，

表 2 実験資料
Table 2 Experimental data.

文献番号	出典
1	前島 他：情処論誌 Vol. 23 No. 1 p. 16
2	山本 他：情処論誌 Vol. 23 No. 1 p. 58
3	松山 他：情処論誌 Vol. 23 No. 2 p. 142
4	木村 他：情処論誌 Vol. 23 No. 2 p. 162
5	有田 他：情処論誌 Vol. 23 No. 2 p. 260
6	栃内 他：信学技報 EC 82-20
7	栃内 他：情処論誌 Vol. 24 No. 2 p. 209
8	栃内 他：北大工研究報告 119 p. 119
9	栃内 他：情処ソフトウェア工学研報 34-20
10	栃内 他：情処論誌 Vol. 26 No. 4 p. 733

ついでその細部が述べられている場合が多いので，同一密度の文章が複数出現した場合，一節の初めにある文を優先的に抽出する．

F：抄録文の出力

Eで抽出した抄録文をもとの文献に出現した順に並べて出力し抄録とする．

以上，高頻度隣接語法の処理手順について述べたが，キーワード密度法の場合も高頻度隣接語の代わりにキーワード，隣接語辞書の代わりに単語辞書を用いるだけであとは全く同様の処理手順である．

5. 自動抄録実験

情報処理学会論文誌から選んだ論文5編と同一著者によるかな漢字変換に関する論文5編の計10編の文献を用いて以下の実験を行った．表2にこれらの文献の一覧を示す．

5.1 実験手順

● 実験 I

キーワード密度法と高頻度隣接語法の両方式により10編の文献の抄録作成を行い，その結果を比較した．前述のように，両方式ともに抄録対象文献全体から M 語，および各章からそれぞれ N 語のキーワードと高頻度隣接語を抽出し，抄録文の抽出はこれら $M+N$ 語を用いて各章ごとに行った．M, N の値については，$M=3, 4, 5$, $N=2, 3, 4$ の場合について文献1～5を用いて実験を行い，後述の方法により評価値を求めた結果 (M, N) が $(4, 3)$ および $(3, 3)$ の場合が最良であったので，全体の実験はこの2種の場合について行った．抄録文として抽出する文数は，その章全体の文数の15%程度とした．この値は実験に使用した文献に付属している，著者自身が作成した抄録文の文数の実測結果に基づいている．

表 3 両方式で共通に選択された文
Table 3 Commonly selected sentences by both methods.

使用文献	共通文数
文献 1～5	134 文 (64.2%)
文献 6～10	118 文 (52.5%)

● 実験 II

実験 I の結果から両方式で作成された抄録を比較した結果，表3に示すように両方式で選択される文にかなり違いがあることがわかった．そこで両方式を組み合わせた抄録法による実験を行った．実験手順を以下に示す．

1. 高頻度隣接語密度とキーワード密度をそれぞれ文献の各文について求める．
2. 文献の各文にそれぞれの密度の大小により順位を付ける．
3. 両者の平均順位を求め，その順位の高いものから一定文数を抄録として抽出する．

5.2 実験結果

実験システムによって作成された抄録の評価は，抽出された抄録文中に「抄録文としてふさわしい文」がどの程度含まれているか，という尺度で行われるのが妥当である．以下に示す実験結果において，この「抄録文としてふさわしい文」を「評価基準文」と称し，作成された抄録中に含まれる評価基準文の比率を評価値とする．

評価基準文は以下のように定めた．

① 文献1～5については，論文に付属しているアブストラクトの各文と意味的に対応する文を本文から抜き出し，これを評価基準文とした．

② 文献6～10については，論文の本文中から該当する文を抜き出すという方法で著者自身に抄録を作成してもらい，これを評価基準文とした．

実験 I の結果を表4に示す．表4は，作成された抄録中に含まれる評価基準文の比率を示すものである．表4を見ると，両方式でほぼ同一の結果が得られている．しかし，作成される抄録の内容には表3に示すようにかなり相違がある．また，それぞれの方式で作成された抄録の内容を比較すると，抄録文としてふさわしくない文は主観的にはキーワード密度方式によって作成された抄録の方に多く含まれていた．

次に，実験 II の結果を表5に示す．表5を見るとキーワードまたは高頻度隣接語を単体で用いるよりも10%程度良い結果が得られている．このことはキー

表 4 抄録実験の結果
Table 4 Results of abstraction experiments.

(%)

	高頻度隣接語		キーワード	
文献番号	$M=4, N=3$	$M=3, N=3$	$M=4, N=3$	$M=3, N=3$
1	26.3	15.8	26.3	21.1
2	26.7	26.7	33.3	20.0
3	50.0	43.3	25.5	31.3
4	40.0	50.0	30.0	30.0
5	40.0	40.0	26.7	26.7
平 均	36.0	33.3	28.8	25.3
6	20.0	20.0	26.7	26.7
7	32.0	28.8	28.8	28.8
8	31.6	31.6	42.1	47.4
9	17.4	17.4	13.0	13.0
10	26.5	26.5	38.2	32.4
平 均	25.2	24.4	29.8	29.0
全体平均	29.1	27.7	29.1	27.7

表 5 平均順位による抄録結果
Table 5 Results of abstraction experiments by the mean oder method.

(%)

文献番号	$M=4, N=3$	$M=3, N=3$
1	26.3	26.3
2	46.7	40.0
3	43.8	43.8
4	50.0	50.0
5	53.3	53.3
平 均	42.7	41.3
6	33.3	33.3
7	56.0	52.0
8	43.1	47.4
9	21.7	17.4
10	38.2	35.3
平 均	38.2	36.6
全体平均	39.8	38.3

ワード密度方式の自動抄録法に高頻度隣接語を組み合わせて用いることの有効性を示している.

6. おわりに

本論文では, 高頻度隣接語を用いたキーワード密度方式自動抄録法の改善について述べてきた. キーワードの代りに高頻度隣接語を用いることによりキーワード密度法の問題点のいくつかを改善することができた. また, キーワード方式に高頻度隣接語を組み合わせて用いることにより良い抄録結果が得られることが確認でき, 高頻度隣接語の有効性を確かめることができた.

なお, キーワードや高頻度隣接語の密度によって抄録文を選択するアルゴリズムでは作成された抄録は前後の脈絡の欠けたものにならざるをえない. したがって, 今後このシステムに, 抽出された文章の整形等の後処理機能を加え, より自然な形の文献要約作成システムへと発展させることが必要である.

謝辞 本研究を行うにあたり, 終始適切な御示唆をいただいた本学部電子工学科電子機器工学講座各位に感謝します.

参考文献

1) Luhn, P. H.: The Automatic Creation of Literature Abstract, *IBM J.*, Vol. 2 (1958).
2) Oswald, V. A.: Automatic Indexing and Abstracting of the Contents of Documents, RADC-TR 59-208 (1959).
3) 朝倉日本語新講座 6 /運用 II ・人文系研究のための言語データ処理入門, 朝倉書店, 東京 (1983).
4) 水谷静夫: 統計的自動抄録法の問題点, 計量国語学, 27 (1963).

(昭和 62 年 6 月 24 日受付)
(昭和 63 年 1 月 19 日採録)

鈴木 康広 (正会員)

昭和 35 年生. 昭和 57 年北海道工業大学電気工学科卒業. 昭和 60 年北海道大学大学院工学研究科修士課程情報工学専攻修了. 現在同大学大学院博士後期課程在学中. 語の接続関係を利用した日本語情報処理の研究に従事. 電子情報通信学会, IEEE 各会員.

栃内 香次 (正会員)

昭和 14 年生. 昭和 37 年北海道大学工学部電気工学科卒業. 昭和 39 年同大学院工学研究科修士課程修了. 現在同工学部電子工学科教授. 工学博士. 計算機応用, ことに日本語文書処理に興味をもつ. 電子情報通信学会, 日本音響学会各会員.