北海道大学工学部研究報告
第88号（昭和53年）

Bulletin of the Faculty of Engineering,
Hokkaido University, No. 88 (1978)

学術文献における情報構造 I

——重要語句の自動抽出について——

前田 隆* 桃内佳雄** 沢村 一**

（昭和52年12月28日受理）

Information Structure of Scientific Documents I

—On an Automatic Extraction of Significant Phrases—

Takashi MAEDA, Yoshio MOMOUCHI and Hajime SAWAMURA

(Received December 28, 1977)

Abstract

In this paper, we have presented an automatic method for extracting significant phrases in the Titles and Abstracts of scientific documents. The method is based on the connection with a manner of representing document information from a view point of hierarchical semantic structure analysis and a method of text structural analysis of the Abstract.

Experiments were made on 5 sets of scientific documents in the A.I. research area, using a relatively small dictionary consisting of eliminable words and some selected concepts following the knowledges about the research area and its description language. The results show that significant phrases are effectively extracted in all cases and their numbers for each document and the processing time are fairly satisfactory.

1. まえがき

文献情報検索システムを構成し，これを効率的に維持・運営していく上で基本的に重要な問題の一つは，増加の一途を辿る原情報をいかに簡潔かつ能率的にその内部表現に変換・蓄積するかという問題である。我々はこれまで学術文献（論文）の抄録を主な情報源とする文献の情報化表現を提案し[1]，これに基づく文献情報検索システムを構成してきた[2]。文献情報には種々のレベルを考えることができるが[3]，ここではとくに抄録が持っている次のような特徴を基礎とした情報化表現を採用している。

(1) その研究の主要な内容（目的・方法・結果・結論など）を簡潔に表現している。
(2) 最近ではほとんどの文献が著者による報知的な抄録をつけている。
(3) 抄録と標題は文献内容の表現として相補的な場合がある。

さらに処理対象として技術的には次の点をとりあげることができる。

(4) あまり長くない（ふつう欧文で300語以内）。
(5) 抄録の各文は抄録に特有の文章表現が顕著にみられる。
(6) 各々の表現形態は各文の抄録における構成上の役割（機能）を理解する手がかりとなる。

* 工業数学講座（数物系共通）
** 情報システム工学講座（情報工学専攻）

ここでは (6) での役割として (1) を考慮して① 主題, ② 方法, ③ 結果, ④ 結論の四つをとり[4], この役割とこれに対応する文における重要語句の対によって, その文献を表現するものとする。そこで問題は抄録の各文に対するこの役割の同定およびその文における重要語句の抽出の方法となる。本稿では主として, 抄録文における重要語句の抽出方法およびこの方法による計算機実験について取り扱う。

文献内容に関する手がかりとしての重要語あるいはキー・ワードの割り当ては「索引づけ」(Indexing) とよばれ, 伝統的かつ経験的な図書の索引法, 主として語の生起頻度などに基づく統計的方法, さらに自然言語処理とも関連する言語学的方法などの種々の方法によって研究されてきている。しかしこれらのほとんどは計算時間とメモリー, 実用性や索引の質などの点でいずれも不充分な結果に終っているとされ, これに対して文献抄録を対象とし (主として上記 (2), (4) を根拠として), 語用論的知識を盛り込んだ大規模な辞書を用いた自動索引の研究が精力的に進められている[5,6]。

本稿においては, 文献における重要語句はその学問分野によって異なり得ることを考慮して, 対象分野を定め, そこにおける重要語句を抽出する方法を考察する。まず一定の言語的知識および対象分野に関する知識を用いて, 比較的小規模な辞書を構成する。次にこの辞書を参照しつつ, 重要な概念は名詞句化することを考慮して, 抄録文から名詞句を抽出する。抄録は単なる文の集合ではなく, テキストとしての構造[7]を持っており, この情報を利用して名詞句からさらに重要なものを選定することができる。この際上記 (3) により標題からの情報を援用する。重要語抽出に関して対象分野 (領域) を考慮して統計的手法を適用した研究が, 日本語文献に対してなされているが[9], 本稿で用いるテキスト分析的手法はいわゆる構文論的および意味論的処理のいわばマクロなアプローチとして重要であると考えられる。

2. 名詞句の抽出

2.1 全体の概要

本稿での実験に用いた文献データは図1に示すように, (1) 文献番号と標題, (2) 著者名, (3) 雑誌名, 巻, 発行年, (4) 抄録, (5) 著者によるキー・ワード (つけられている場合) の5種類のカード群からなっており, 抄録はさらに各文ごとに識別できるようになっている。名詞句抽出のため処理される対象は標題と抄録文であり, 著者によるキー・ワードはリストアップされて, 後で比較のために用いられる。これらの処理プロセスの概要を図2に示す。

2.2 単語ストリングの抽出

処理単位は標題は全体, 抄録は各文ごとであり, それぞれカード上の第6カラム, 第7カラム

```
(1)   7323 SYSTEM ORGANIZATIONS FOR SPEECH UNDERSTANDING: IMPLICATIONS OF NETWORK AND
          MULTIPROCESSOR COMPUTER ARCHITECTURES FOR @AI
(2)   L. D. ERMAN, R. D. FENNELL, V. R. LESSER AND D. R. REDDY
(3)   * IJCAI,3,1973
(4)   (001) THIS PAPER CONSIDERS VARIOUS FACTORS AFFECTING SYSTEM ORGANIZATION FOR
          SPEECH UNDERSTANDING RESEARCH.
      (002) THE STRUCTURE OF THE HEARSAY SYSTEM BASED ON A SET OF COOPERATING,
          INDEPENDENT PROCESSES USING THE HYPOTHESIZE-AND-TEST PARADIGM IS
          PRESENTED.
      (003) DESIGN CONSIDERATIONS FOR THE EFFECTIVE USE OF MULTIPROCESSOR AND NETWORK
          ARCHITECTURES IN SPEECH UNDERSTANDING SYSTEMS ARE PRESENTED; CONTROL OF
          PROCESSES, INTERPROCESS COMMUNICATION AND DATA SHARING, RESOURCE
          ALLOCATION, AND DEBUGGING ARE DISCUSSED.
(5)   /*    SPEECH RECOGNITION, SPEECH UNDERSTANDING, SYSTEM ORGANIZATION, NETWORKS,
      /*    MULTIPROCESSORS, PARALLEL PROCESSING, REAL-TIME SYSTEMS, HARDWARE FOR AI,
      /*    SOFTWARE FOR AI,
```

図1 文献データ

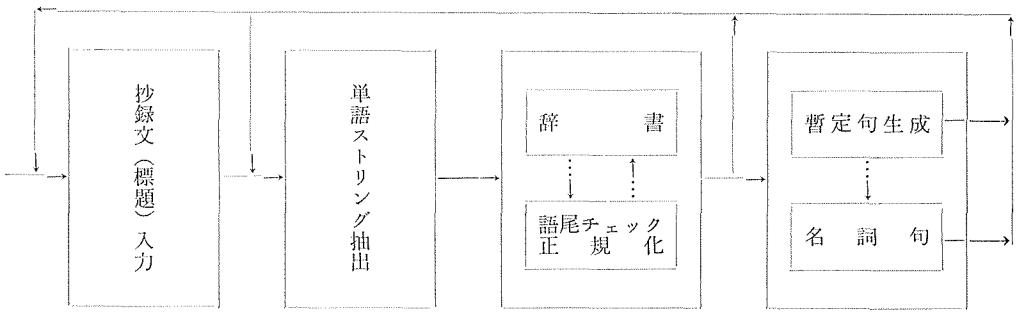

図2 名詞句抽出の概要

以降のストリング列である。単語ストリングのデリミター（区切り記号）としては「空白」の他いわゆる特殊記号をとる。但し，引用符（'），コンマ（,），ハイフン（-），アスタリスク（*）はその直後の一文字を参照して実際の区切りとするかどうかを判定して処理する。数式あるいはこれに準じた表式はデータ作成の段階でこれを識別する記号（¥）をそう入しておき処理上は除去することにし，また数字だけからなるストリングも除去することにした。さらに固有名詞（人名，装置・システム名など）の一部についてもやはり識別子（(@)）を入れ（図1における標題の末尾の '(@AI' の(@)，識別子の次からのストリングの末尾に '-(@' なる記号列を結合し，通常の単語ストリングと区別できるようにした。これは後述の辞書ひきや語尾処理で識別される必要があるからである。こうして抽出されたストリングは単語として認定され，次のステップへと引き渡される。

2.3 辞書の構成

ここで用いる辞書は，まえがきでも述べたように対象分野および言語的知識を前提として，有用な名詞句抽出のためのみの目的に対して構成されたものであって，①規模を大きくしない，②新語への対処，などの理由から「除去語」を主体として次の三種類の語が登録されている。

(1) 無条件除去語……いわゆるストップ・ワード，動詞，副詞，一般的名詞など。

(2) 選択的除去語……形容詞および弱名詞の一部（他の語と結合したとき除去されない）。

(3) 非除去語……後述の語尾チェックなどで変形もしくは除去されてはならない語。

辞書項目の各々は図3に示すように，その項目の取り扱いのタイプ，すなわち除去，文脈的処理，非除去などに応じたフラグがつけられており，処理はこれに従ってなされる。辞書項目にな

No. ＼ 項目	フラグ	登録個数	説明
1	␣␣␣	1,116	除去語(␣は空白，－は省略を示す)
2	␣*＋	2	AND，OR（2.5 末尾参照）
3	␣AJ	326	(2) の形容詞
4	␣AG	66	(2) の ING 形
5	␣NP	683	(2) の弱名詞
6	－－F	4	(2) の弱名詞（7 の役割も果す）
7	－－N	26	文の役割を示す名詞 } 今回使用せず
8	－－V	47	文の役割を示す動詞 }
9	␣NG	19	ING 形で自立的な語
10	␣NN	9	自立的名詞
計		2,298	52.12.22 現在

図3 辞書の構成要素

```
**********************************
WORD SELFCTION DICTIONARY :  77.12.21
**********************************

ALPHARETICAL :                                          BY FUNCTION AND FREQUENCY :
 NP 0000  ABBREVIATION   AJ 0000  ABDUCTIVE                  0599  THE        0509  OF
    0006  ABILITY           0013  ABOUT                     0244  TO         0214  IN
    0003  ABOVE          AJ 0001  ABSOLUTE                  0201  IS         0132  FOR
 NP 0000  ABSTRACT       NP 0003  ABSTRACTION              0099  ARE        0088  THAT
    0000  ACCELERATE        0002  ACCEPT                    0084  THIS       0083  WHICH
 NP 0005  ACCESS            0000  ACCOMMODATE               0081  AS         0075  BY
    0000  ACCOMPANY          0000  ACCOMPLISH                0072  WITH       0066  AN
    0000  ACCORD             0003  ACCOUNT                   0062  BE         0052  IT
 NP 0002  ACCURACY       AJ 0000  ACCURATE                  0049  CAN        0049  ON
 AJ 0000  ACCUSATIVE        0003  ACHIEVE                   0044  THESE      0035  FROM
    0000  ACHIEVEMENT     AJ 0004  ACOUSTIC                  0033  USE        0030  ITS
    0000  ACQUIRE         NP 0000  ACQUISITION              0029  SOME       0028  HAS
    0003  ACT             NP 0007  ACTION                   0026  BEEN       0026  TWO
    0000  ACTIVATE        NP 0000  ACTIVATION               0020  INTO       0020  NEW
 NP 0000  ACTIVITY       NP 0000  ACTUATOR                  0019  ONE        0019  SUCH
 AJ 0000  ACUTE          AJ 0000  ACYCLIC                   0018  HAVE       0016  AT
    0002  ADAPT           NP 0000  ADAPTABILITY             0016  WE         0015  ALSO
 AJ 0001  ADAPTIVE          0000  ADD                       0014  MORE       0014  WILL
    0005  ADDITION           0000  ADDITIVE                  0013  ABOUT      0013  HOW
    0001  ADDRESS         AG 0000  ADDRESSING                0013  MANY       0013  NOT
```

図4 辞書内容の一部（AI-1-6 処理後，数字は参照された頻度を表わす）

い単語は，次に語尾チェックが行なわれる。辞書の内容の一部を図4に示す。

2.4 語尾処理と正規化

語尾チェックを受けていない段階で辞書になかった単語ストリングは，語尾が次の各文字であるとき，対応した処理がなされる。なおこれらの語尾を持つ語で処理されてはならない単語は識別されるか，辞書に登録されていなければならない。

(1) 'S'……名詞の複数形あるいは「三単現」の S（ないしは ES）とみなされ，正規化の処理が行なわれる。正規化された単語ストリングは再び辞書ひきにまわされる。この処理は一度だけ行なわれる。

(2) 'Y'……'-LY' 形（の副詞）であるとき除去語とみなされる。

(3) 'G', 'R'……'-ING' 形（の現在分詞形），'-ER' 形（比較級）であるとき，さらにこのストリングの中にハイフン '-' を含めば弱名詞，含まなければ除去語とみなされる。

(4) 'D', 'N'……'-ED' 形，'-EN' 形（の過去分詞形）であるとき，さらにハイフンを含めば形容詞，含まなければ除去語とみなされる。

(5) 'L', 'C', 'T'……それぞれ '-AL' 形あるいは '-FUL' 形，'-IC' 形，'-EST' 形（の形容詞）であるとき，(4) と同じ処理。

(6) 'E'……① '-BLE' 形，'IVE' 形（の形容詞）であるとき (4) と同じ。
② '-LIKE' 形，'-TYPE' 形のとき，形容詞とみなされる。

上記にあてはまらない単語ストリングは自立的名詞として扱われる。

2.5 名詞句の構成

原データから抽出された単語ストリングは辞書ひきおよび対応する語尾チェック，さらに必要に応じて再度の辞書ひき等の処理を受けた時点で，結局次の四種類のいずれかに分類されることになる。

① 除去語 (D)，② 形容詞系 (A)，③ 弱名詞系 (P)，④ 自立的名詞系 (N)。

ここで構成する「名詞句」というのは，処理の技術的観点からみるならば，「記号ストリング列（1枚のカードからなるのではなく，1つの文全体を構成している記号列）を上の除去語および単独生起の形容詞系ないしは弱名詞系の単語ストリングをデリミターとして区切った部分ストリングである」ということになる。但し形容詞系は後続部分ストリングに対して文脈的取り扱い

を受けるものとする。すなわち，上の記号を用いて形式的には次のように表現することができる。ここで各々のタイプの単語ストリングは複数回生起してよく，P と N は互いに交じり合ってもよいとする。

(1) A+P+N (P, N のいずれか一方は空であってもよい)

(2) P+N (P は空であってもよい)

尚，(1) の形式において，例えば次のような処理が行なわれる。

原データ：'… syntactic, semantic and pragmatic information of …'

抽出名詞句：'syntactic information', 'semantic information', 'pragmatic information'。

3. 重要語句の選定

3.1 抄録のテキスト構造

テキストとしての抄録が持つ構造として，まえがきでも述べたようにここでは各々の抄録文が抄録の構成上果している役割（機能）およびその役割の下でその文が含んでいる概念（重要語句）の組によって表現するものとする。テキスト構造の内容的（意味的）情報を担うものとして，次に述べるように，語の反復的な生起に注目する。

抄録のこのようなテキスト構造にもとづく，文献の情報化表現を得るプロセスは図5のように表わすことができる[1]。

図5において，A–B の流れはいわば意味論的処理であり，C は文字通り構文論的処理となっているが，いずれも語用論的知識の一部を積極的に利用する。C–D の流れは本稿では取り扱わない[4]。

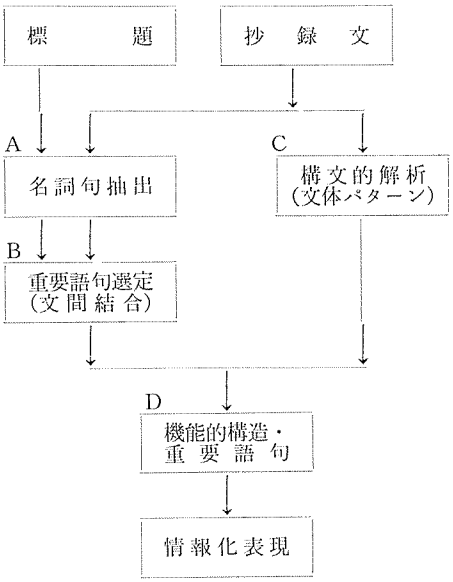

図5 情報化表現のフロー

3.2 文間結合と反復語

テキストにおける各文の間の関係を示す「手がかり」（インジケータなどとよばれる）には種々のタイプが考えられるが，概念を表わす語の反復による結合，すなわち反復型の結合が非常に重要でテキスト構造の実質的な担い手と考えられている[7]。文献7）では日本語について述べているが，この点は英語においても同様であると考えられるので，そのまま援用することにする。

語の反復には次の三つのレベルが考えられている。

レベル 1：同一の語句の反復。

レベル 1.5：レベル1に加えて，語句の構成要素の一部の反復。

レベル 2：意味論的に関連する語句，すなわち同義・類義・内包・外延・対義関係のいわばシソーラス的関係のある語彙要素の反復を含む。

これに対して本稿では，① 反復語を名詞に限らない，② 語句の構成要素の反復には，その要素ストリングを先頭から含む場合のみを取り扱う，という方法をとる。また反復の範囲をその文から抽出した名詞句の列の前半部分とする。すなわち各文はその名詞句の列の前半部分において，先行する文から「既知」（主題，Topic）を受けとり，後半部分で「新知」（解説，Comment）をつけ加え，後続する文へとひきつぐ[7]と仮定する。

3.3 重要語句の選定

抽出された名詞句の集合から重要語句を選定する方法は上で述べた，概念的に文間結合を担っている語句，すなわち反復語をその構成要素として含む名詞句をその抄録における重要語句とする。さらに標題との相関をとり，文間結合の反復語にない語が同時生起しているならばこれを上の反復語と同様に扱い，標題による重要語句の補充に用いることにする。

以上述べてきた方法による実験結果の一部について検討する。対象とした分野は情報科学の中でとくに人工知能研究を中心とする，次の5種類のデータ群である。

① IJCAI 69……Proceedings of International Joint Conference on Artificial Intelligence, 第1回 (1969年)
② IJCAI 71……同上，第2回 (1971年)
③ IJCAI 73……同上，第3回 (1973年)
④ AI-1-6……Artificial Intelligence Vol. 1 (1971年)～Vol. 6 (1976年)
⑤ J. ACM 75…Journal of the Association for Computing Machinery Vol. 22 (1975年)

表1に示すように，文献総数346件に対して，抄録文数は平均5.0，語数は平均108であるが，それぞれ文数1から12，表にはないが語数は13から300と比較的ばらついている。処理対象としての長さは一応適当な範囲に入っているとみなしてよいと考えられる。

文献データに対する処理は文献ごとになされ，図1のデータに対する結果の一部が図6に示されている。(1) は標題と標題から抽出された名詞句が1から5まで出力されている。(2) は抄録文3個に対する処理結果であり，計10個の名詞句，〈WORD PROJECTION〉は反復語を示し，002においては1の SYSTEM が反復していること，003においては1の SPEECH, UNDERSTANDING, SYSTEM が，2の SYSTEM がそれぞれ反復していることを表わしている。(3) は著者がつけておいた重要語句，(4) は (2) の名詞句の異なり語を集め（この場合は全部が異なり語で頻度も等しく1である），(5) は反復語の異なり語を集め（この場合は SYSTEM が3回生起している），それぞれアルファベット順にソートして出力してある。(6) は標題における名詞句に現われる異なり語のうち抄録にも生起する語を出力しており，相関をみている。(7) は本稿における方法によって，(4) の中から選定された重要語句である。標題による補充が

表1 文献データ群の統計

項目＼データ群	IJCAI 69	IJCAI 71	IJCAI 73	AI-1-6	J. ACM 75	計(平均)など
文献数 (著者キーワード付)	63 (12)	65 (16)	85 (33)	83 (1)	50 (50)	346 (112)
平均文数 (最大：最小)	5.7(10：1)	4.8(11：1)	4.5 (9：1)	5.0(10：1)	5.4(12：1)	5.0(12：1)
総語数 標題 (平均)	513 (8.1)	545 (8.4)	666 (7.8)	637 (7.7)	407 (8.1)	2,768 (8.0)
総語数 抄録 (平均)	7,915(125.6)	6,905(106.2)	7,829 (92.1)	8,826(106.3)	5,835(116.7)	37,309(107.8)

表2 抽出語句等に関する統計

語数（番号は図6に対応）＼データ群	IJCAI 69	IJCAI 71	IJCAI 73	AI-1-6	J. ACM 75	計(平均)など
(1) 標題 (平均)	117 (1.9)	133 (2.0)	170 (2.0)	173 (2.1)	94 (1.9)	687 (2.0)
(4) 抄録 (平均)	720(11.4)	635 (9.8)	772 (9.1)	836(10.1)	482 (9.6)	3,445(10.0)
(7) 重要語句 (平均)	388 (6.2)	303 (4.7)	404 (4.8)	454 (5.5)	288 (5.8)	1,837 (5.3)
(3) 著者選定 (平均)	89 (6.2)	94 (5.9)	214 (6.5)	4 (4.0)	268 (5.4)	669 (6.0)
(8) 単純生起名詞句 (平均)	418 (7.4)	333 (5.1)	398 (4.7)	470 (5.7)	294 (5.9)	1,913 (5.5)
(5) 反復語 (平均)	210 (3.3)	149 (2.3)	187 (2.2)	224 (2.7)	140 (2.8)	910 (2.6)

```
(1)  7323   SYSTEM ORGANIZATIONS FOR SPEECH UNDERSTANDING: IMPLICATIONS OF NETWORK AND
            MULTIPROCESSOR COMPUTER ARCHITECTURES FOR @AI
               1 SYSTEM ORGANIZATION                    2 SPEECH UNDERSTANDING
               3 NETWORK                                4 MULTIPROCESSOR COMPUTER ARCHITECTURE
               5 AI-@

             AUTHOR :  L. D. ERMAN, R. D. FENNELL, V. R. LESSER AND D. R. REDDY

(2)     001     1 SYSTEM ORGANIZATION                    2 SPEECH UNDERSTANDING

        002     3 HEARSAY SYSTEM                         4 HYPOTHESIZE-AND-TEST PARADIGM

              < WORD PROJECTION >
                1    1 : SYSTEM

        003     5 MULTIPROCESSOR                         6 NETWORK ARCHITECTURE
                7 SPEECH UNDERSTANDING SYSTEM            8 DATA SHARING
                9 RESOURCE ALLOCATION                   10 DEBUGGING

              < WORD PROJECTION >
                1    1 : SPEECH          1 : UNDERSTANDING          1 : SYSTEM
                2    1 : SYSTEM

(3)       KEY-PHRASE SET BY AUTHOR :
                  1       SPEECH RECOGNITION
                  2       SPEECH UNDERSTANDING
                  3       SYSTEM ORGANIZATION
                  4       NETWORKS
                  5       MULTIPROCESSORS
                  6       PARALLEL PROCESSING
                  7       REAL-TIME SYSTEMS
                  8       HARDWARE FOR AI
                  9       SOFTWARE FOR AI

(4)       ABSTRACT PHRASE SET :
                  1    1  DATA SHARING
                  2    1  DEBUGGING
                  3    1  HEARSAY SYSTEM
                  4    1  HYPOTHESIZE-AND-TEST PARADIGM
                  5    1  MULTIPROCESSOR
                  6    1  NETWORK ARCHITECTURE
                  7    1  RESOURCE ALLOCATION
                  8    1  SPEECH UNDERSTANDING
                  9    1  SPEECH UNDERSTANDING SYSTEM
                 10    1  SYSTEM ORGANIZATION

(5)       PROJECTION WORD SET :
                  1    1  SPEECH
                  2    3  SYSTEM
                  3    1  UNDERSTANDING

(6)       TITLE : ABSTRACT CORRELATION :
                  SYSTEM                    ORGANIZATION              SPEECH
                  NETWORK                   MULTIPROCESSOR            ARCHITECTURE
              * T/A RATIO :  7/ 9 = 77.77 %                         UNDERSTANDING

(7)       PROJECTED PHRASE SET :
                  1  HEARSAY SYSTEM
                  2  SPEECH UNDERSTANDING
                  3  SPEECH UNDERSTANDING SYSTEM
                  4  SYSTEM ORGANIZATION

              * COMPLIMENTAL PHRASES BY TITLE :
                  5  NETWORK ARCHITECTURE
                  6  MULTIPROCESSOR

(8)       PHRASE SELECTION BY SIMPLE ITERATIVE WORDS :
                  1       HEARSAY SYSTEM
                  2       SPEECH UNDERSTANDING
                  3       SPEECH UNDERSTANDING SYSTEM
                  4       SYSTEM ORGANIZATION
```

図6 図1のデータに対する処理結果の一部

2個の場合である。(8) は抄録の名詞句の中で，単純に複数回生起した語を含む名詞句で，この例の場合は，標題による補充を除いて本稿での結果と同じものとなっている。以上の結果のうち，(1)，(3)，(4)，(5)，(7)，(8) および総出現単語，さらに (3) の構成要素から，**2.3** で述べた辞書における除去語以外の語の8種類のデータは，対応するファイルに頻度を加算して書

き加え（変え）られる。

抽出された名詞句，重要語句などに関する統計を表2に示す。表2および処理結果の観察から概ね次のようなことがいえる。

(i) 標題からの抽出数は平均2個であり，標題における名詞句の約80%が1～2語から成っていることを考慮すると，文献内容の表現としてはやや情報不足である。

(ii) 抄録全体からの抽出は平均10個であり，著者によるものの1.5倍以上となっている。

(iii) 本稿での方法による重要語数は5.3個であり，著者によるものとほぼ同数であって平均にはよい結果を示している。内容的にも図6にみるように，重要な概念に関連する語句は選定（補充）されているとみなし得ると思われる。補充された語句は平均1個である。図7にファイルに書き込まれた重要語句を頻度順（FRCY は文献あたりの，SFRC は単純な頻度をそれぞれ表わしている）およびアルファベット順にソートして出力したものの最初の25語句が示されている。(AI-1-6 の場合)

(iv) 抽出された名詞句の構成要素における単純な複数回生起語を含む名詞句も平均して5.5個であるが，標題とは無関係にとっているので，抄録の書かれ方，とくに一つの文で繰り返し生起する語がある場合などによる影響をより受けやすい特性を持っていると考えられる。

(v) 反復語は平均2.6個であり，図6の例においてもみられるように，一応その文献のキー的な概念の構成要素となり得る語が，文間において反復使用されていると考えられる。その意味で本稿における方法の妥当性の一つの根拠をも示しているとみてよいであろう。図8に，アルファベット順，頻度順（F/S は文献あたり，F/T は単純な頻度）にソートしたものの最初の25語を示す。(AI-1-6 の場合)

3.4 重要語句間の関連と語彙調査

本稿における重要語句抽出は一定の言語的および対象分野に関する知識を利用して，対象分野が限られていることもあるが，わずか2,300語ぐらいの辞書によって，比較的よい抽出が行なわれていることを示してきた。処理時間は8種類の直接編成ファイルへのアクセスをも含めて，1文

KEY-PHRASES SET : A.I. 1-6 : BY PROJ. WORD - FRQ. ; - SEQ. ; 77.12.22

NO	FRCY	SFRC	PHRASE
1	14	14	PROGRAM
2	7	7	ALGORITHM
3	6	6	HEURISTICS
4	6	6	SEARCH
5	5	5	KNOWLEDGE
6	5	5	LANGUAGE
7	3	6	ARTIFICIAL INTELLIGENCE
8	3	5	HEURISTIC SEARCH
9	3	4	COMPUTER PROGRAM
10	3	3	CLAUSE
11	3	3	GRAPH
12	3	3	REPRESENTATION
13	3	3	RESOLUTION
14	2	6	AUTOMATIC THEOREM PROVING
15	2	4	LINEAR RESOLUTION
16	2	4	PREDICATE CALCULUS
17	2	4	PROBLEM SOLVING
18	2	4	SEARCH ALGORITHM
19	2	3	LINE DRAWING
20	2	3	NATURAL LANGUAGE
21	2	3	PLANE GEOMETRY
22	2	3	PROBLEM-SOLVING PROGRAM
23	2	3	PROGRAMMING
24	2	3	SEARCH SPACE
25	2	2	LEARNING

(a) 頻度順

NO	PHRASE
1	A-ORDERED RESOLUTION
2	ABSTRACTION SPACE
3	ABSTRACTION SPACE HIERARCHY
4	ACOUSTIC ANALYSIS
5	ACOUSTIC DATA
6	ADAPTIVE-CONTROL PROCEDURE
7	ALGORITHM
8	ALGORITHMIC KNOWLEDGE
9	ALGORITHMIC STRUCTURE
10	ALGORITHMIC WORLD KNOWLEDGE
11	ALPHA-BETA PROCEDURE
12	ALPHA-BETA TECHNIQUE
13	ANALOGY
14	ANALYTIC METHOD
15	ANAPHORIC INFERENCE PROBLEM
16	AND/OR GRAPH
17	AND/OR SEARCH TREE
18	ARRAY
19	ARTIFICIAL INTELLIGENCE
20	ARTIFICIAL INTELLIGENCE PROBLEM
21	ARTIFICIAL PARANOIA
22	ASSEMBLY MANIPULATION
23	ASSOCIATION MEASURE
24	AUGMENTED RECURSIVE TRANSITION NETWORK PARSER
25	AUTOMATED GENERATION

(b) アルファベット順

図7 重要語句の一部

PROJECTIONAL WORDS SET OF THE A.I. 1-6 77.12.22

NO	WORD	F/S	F/T	IDNO
1	ABSTRACTION	1	1	0051
2	ACOUSTIC	1	2	0050
3	ALGORITHM	9	63	0008
4	ALGORITHMIC	1	11	0080
5	ALPHA-BETA	1	1	0073
6	ANALYSIS	2	10	0023
7	ARRAY	1	1	0038
8	ARTIFICIAL	1	1	0070
9	AXIOMATIC	1	1	0004
10	BASE	1	1	0060
11	BREADTH-FIRST	1	3	0037
12	BUILD-@	1	7	0047
13	CAI-@	1	1	0070
14	CALCULUS	2	7	0016
15	CASE	1	1	0074
16	CHARACTERIZATION	1	3	0077
17	CHESS	1	4	0030
18	CHROMOSOME	1	1	0008
19	CHRONOS-@	1	3	0025
20	CLAUSE	2	10	0009
21	COMMONSENSE	1	1	0080
22	COMPLETE	1	4	0032
23	COMPREHENSION	1	6	0080
24	COMPUTER	3	4	0054
25	CONCEPTUAL	1	1	0060

(a) アルファベット順

F/S	F/T	WORD	IDNO
16	32	PROGRAM	0003
9	63	ALGORITHM	0008
8	31	SEARCH	0006
7	10	RESOLUTION	0009
6	28	LANGUAGE	0002
4	18	HEURISTIC	0006
4	11	HEURISTICS	0003
4	21	KNOWLEDGE	0057
4	30	PROBLEM	0044
4	5	SYSTEM	0047
3	4	COMPUTER	0054
3	16	GRAPH	0010
3	3	LINEAR	0021
3	4	NETWORK	0028
3	3	SPACE	0037
3	4	STRIPS-@	0019
3	3	THEORY	0004
2	10	ANALYSIS	0023
2	7	CALCULUS	0016
2	10	CLAUSE	0009
2	4	FUNCTION	0006
2	26	GAME	0035
2	2	INFERENCE	0005
2	4	METHOD	0045
2	3	PREDICATE	0016

(b) 頻度順

図8 反復語の一部

```
254  KNOWLEDGE
       .                                  KNOWLEDGE
       .                                  KNOWLEDGE  STATE
       .                    ALGORITHMIC   KNOWLEDGE
       .              ALGORITHMIC WORLD   KNOWLEDGE
       .                     CONCEPTUAL   KNOWLEDGE
       .                         GLOBAL   KNOWLEDGE
       .                      HEURISTIC   KNOWLEDGE
       .                   WEAK INDUCTIVE KNOWLEDGE
255  LANGUAGE
       .                                  LANGUAGE
       .                                  LANGUAGE   COMPREHENSION
       .          BUILD-@'S IMPLEMENTATION LANGUAGE
       .                       CASE-LESS  LANGUAGE
       .                         ENGLISH  LANGUAGE
       .                           INPUT  LANGUAGE
       .                        INTERNAL  LANGUAGE   PROBLEM
       .                         NATURAL  LANGUAGE
       .                         NATURAL  LANGUAGE   ANALYSIS
       .                         NATURAL  LANGUAGE   COMMUNICATION
       .                         NATURAL  LANGUAGE   PROCESSING
       .                         NATURAL  LANGUAGE   PROCESSING SYSTEM
       .                         NATURAL  LANGUAGE   SENTENCE
       .  NONDETERMINISTIC PROGRAMMING    LANGUAGE
       .             PATTERN DESCRIPTION  LANGUAGE
       .                            @A4-@ LANGUAGE
256  LANGUAGE-FREE
                                          LANGUAGE-FREE
```

図9 KWIP の一部

献あたり1秒弱の短い時間で処理されている。個々の文献についてのこのような処理とその分析と同時に，対象分野おける基本的な語彙に関する調査・分析も重要である。そこで抽出された名詞句についてその構成要素となっている語ごとに，その語が名詞句の中でどのような部分においてどのような意味あいで使われているかをみるために，KWIC にならって KWIP (Key Word In Phrases) ともいうべきものを構成した。図9にその一部を示すが，このような調査は重要語句間の関連性，すなわちシソーラス的関連の分析，およびその自動構成[8]への応用などにとっても重要な課題である。

4. あ と が き

学術文献の情報化表現およびその自動化について，本稿ではとくに抄録における重要語句の自動抽出の問題を取り扱い，対象分野および一定の言語的知識を利用した小規模な辞書にもとづき，さらにテキスト分析的手法を応用した方法を考察してきた。実験結果は数量的に一応満足される程度の重要語句の抽出がみられ，抄録と標題そのものからの自動的抽出方法としては，内容的にも良好な結果を与えていることを示している。一方，抄録について**1.** でみたような共通的な特徴がある反面，必ずしも統一的な形式にのっとって書かれているわけではないので，特殊な形式をもつ文献に対する対策を考えなければならない。この問題についてはまだあまり研究されていないが，対象分野に関する知識を援用する方法，すなわちいわゆるシソーラス的知識の応用の方向がある。本稿での考察・実験等を通じて，学術文献における情報構造，その主要な担い手である反復語，これをキーとする重要語句およびこれらの間の諸関係などを一層深く分析・検討していくための貴重なデータを収集・集積することができた。

最後に，日頃種々の点でご助言をいただく本学部情報数理工学第一講座河口至商教授に深く感謝致します。本稿における実験は北大大型計算機センター FACOM 230-75 により，主として PL/I 言語によって計算されたが，この間プログラム作成やファイル利用などに関して長田博泰氏，相良砌氏，またとくに SORT/MERGE プログラム等に関して貝田辰雄氏らのセンター関係者に多大のご援助をいただいた。ここに記して厚く感謝の意を表します。

参 考 文 献

1) 前田 隆：情報処理学会第17回全国大会講演論文集（1976年），58.
2) 前田 隆，桃内佳雄，沢村 一：北大工学部研究報告，第84号（昭和52年），p. 101.
3) 桃内佳雄，前田 隆，沢村 一：北大工学部研究報告，第84号（昭和52年），p. 113.
4) 前田 隆，桃内佳雄，沢村 一：昭和52年度電子通信学会情報部門全国大会講演論文集（昭和52年），96.
5) Paul. H. Klingbiel: Inform. Stor. Retr., 9 (1973), p. 79-84.
6) Paul. H. Klingbiel: Inform. Stor. Retr., 9 (1973), p. 477-494.
7) 坂本義行，岡本哲也，谷津直和：第10回情報科学技術研究集会発表論文集（1973年）.
8) Stephan Braun: IFIP-INFOPOL-76, J. Madey. ed. (1977年) p. 187-203.
9) 長尾真，水谷幹男，池田浩之：情報処理，Vol. 17, No. 2 (1976年)，p. 110～117.