論　文

日本語文献における重要語の自動抽出*

長　尾　　真** 水　谷　幹　男** 池　田　浩　之**

Abstract

An automatic method is developed which extracts important words (typical words or possibly key words) from Japanese scientific documents.

The idea is to extract the words which appear frequently in some particular documents, but which appear seldom in other documents. To measure this particularity of word usage we utilized the idea of χ^2 test.

The experiment is done on a text book of chemistry of the middle school, and "Current Bibliography on Science and Technology (Electrical Engineering)" edited by JICST. More than 120,000 Japanese words are handled and the results are fairly satisfactory.

1. はじめに

一口に言語処理分野といっても多くの内容に分けられようが，ごく大まかに分類すると，文字どおり機械に言語を理解させることを最終目的とした意味的，構文解析的処理と，言語を数量的に取り扱った統計的処理との2つが現在の主流といえる．とりわけ後者の統計的処理に対する社会的要請は，情報洪水時代といわれ，とても人間の手にはおえないほどの大量の文献があふれる現代において，ますます急を要するものとなりつつある．

毎月発表される学術的論文にだけついてみても，その数は膨大なものであり，その分野の専門家ですら，すべての文献に目を通すことは不可能という今日の状況を解決するためには，どうしても電子計算機を利用したしっかりした文献検索システムの開発が必要なのである．

さてこの問題を解決するために，これまでにも各種の文献検索システムが考案され，研究されてきたが，その場合，肝要な働きをするそれぞれの文献の重要語（またはキー・ワード）を抽出する段階では（少くとも日本語文献の検索システムに関する限り）ほとんど人間が内容を読んで，判断して選別するといった形で行われてきたのが普通であった．ここでいう「重要語」とは文献内容をよく表わし，検索する際に「見出し語」として使用できるような特徴ある単語のことであり，その他の語は「一般語」ということにしよう．

本論文は，この重要語抽出を電子計算機で自動的に行う統計処理方法を開発し，大量文献に対して試みた結果を示したものである．

現在，各地で研究開発が進められているいろいろな文献検索システムはそのかなりのものが英語文献を処理するものであるが，本研究で取り扱っているのは日本語文献であり，そこには当然英語文献には見られない日本語特有の諸問題がある．そのうちでも特に英語と比較した場合，顕著なものとしては

1. 文が単語ごとに区切られていない．
2. 使用される文字が漢字，ひらがな，カタカナなど数千種にのぼる．
3. 漢字が表意文字で，また熟語，複合語を多く作る．
4. 外来語がカタカナで表わされ，比較的多く現れる．

などがある．

1. に関しては本研究のテーマが重要語抽出であり，

---

* An Automatic Method of the Extraction of Important Words from Japanese Scientific Documents by Makoto NAGAO, Mikio MIZUTANI and Hiroyuki IKEDA (Faculty of Engineering, Kyoto University)

** 京都大学工学部

その点からのみいえば，重要語とはまず名詞とみなしてよいであろうから文中から名詞だけが区別できればそれでよいわけであるが，今後さらに厳密さが要求されたり，また他の方面への幅広い応用などを考えれば，一般にはむろんあらゆる語がはっきりとそれぞれの品詞に区分できることが望ましい．

2. について，本研究の場合は文字は漢字，ひらがなどを問わず，入出力，および途中の処理もすべて漢字テレタイプ・コードで処理したから，文字の種類が多いということで直接的な困難はなかった．むしろ字種の豊富なことは，単語の識別が容易になる，文字の種類から大まかな重要度の判定ができる（例えば，ひらがなの部分は重要度が低いといえる）など種々の利点があって，本研究遂行の上で大いに便利であった．

3. の漢字が日本語文において果たす重要な役割については，一般に文中の漢字だけを拾い読みすれば，その文章の内容の大意がわかるといわれるほど，そのもつ意味には大きなものがある．従って日本語の処理を考えるとき，漢字のこの利点を最大限有効に生かすことが得策といえよう．

4. のカタカナの働きについても，3. と同様のことがいえて，外来語が文字の種類だけで簡単に見分けられるというのも他国語には見られない大きな利点である．とりわけ学術論文などにおいては外来語の占める比重は大きく，重要語抽出に際してもかなりの役割を果たすものである．

2. 重要語自動抽出の考え方

さまざまな文献検索システムが研究開発されているが，重要語抽出を自動的に行っているものはほとんどない．従って現状はまだ暗中模索の状態にあるが，コーネル大学で SMART システムを作った G. SALTON は次のような方法を提案している．

まず前もって人間の判断によって指定されたいくつかの基本的な重要語に適当な重みづけをほどこした文献ベクトルを考える．また，これによって構成される文献空間内で文献ベクトルごとの距離を適当にきめ，文献空間の密度を計算する．そして，ここに新しい単語を入れて，その語を含めた場合の文献空間の密度を計算し，その値が前より大きくなければ一般語，小さくなれば重要語とするというわけである．

しかしこの方式とても，まずある程度の重要語群を最初に用意しておく必要があり，また重要語が増加していったときの計算量はかなり膨大となり困難をともなう．そこで我々は次のような考え方を用いた．すなわち，全文献をいくつかの分野に分け，ある分野にはよく現れるが他のほとんどの分野にはあまり現れない単語をみつける．このような単語はその特定の分野を特色づける単語であると考えることができる．このような単語のみつけ方としては，分野ごとに現れるすべての単語の出現頻度を求め，特定の分野にのみよく現れる単語を調べればよい．これには χ^2 分布による検定法の考え方を用いることができる．

すなわちひとつひとつの単語について，各分野ごとの頻度を標本値とし，帰無仮説として「その単語の出現する確率は全分野をつうじて等しい」と仮定して，χ^2 値を求め，もしそれが十分大きな値となるなら，「分野によって頻度にかたよりがある」，いいかえれば「ある分野にのみ集中して現れている」ということになり，重要語であるといえる．

具体的には単語 i の χ^2 値は次式で算出される．

$$\chi_i^2=\frac{\sum_{j=1}^{n}(x_{ij}-m_{ij})^2}{m_{ij}} \qquad (1)$$

ただし

$$m_{ij}=\frac{\sum_{j=1}^{n}x_{ij}}{\sum_{i=1}^{m}\sum_{j=1}^{n}x_{ij}}\times\sum_{i=1}^{m}x_{ij}$$

m：異なり単語数

n：分野数

x_{ij}：単語 i の j 分野における頻度

m_{ij}：単語 i の j 分野における理論度数

なお本研究の場合，分野数を 10 あるいは 19 ととったため自由度 9 あるいは 18 になるが，そうなると検定に用いる有意水準はきわめて大きな値となって，結果として各分野の上位のほんの数語のみが有意の重要語で，他はみんな有意であるとは必ずしもいえないということになってしまう．いいかえればそれら χ^2 値のあまり大きくない語は，「偶然でもその程度のかたよりは起こりうる可能性がある」というわけである．しかし，かといってこれら有意水準以下の語を無視してしまうと，残る単語はごく少数で実用に耐えないものとなってしまう．

そこで本研究では正確性を犠牲にして，χ^2 値を検定法としてではなく，全く単純に，かたよりの程度を示す指標，度合として用い，有意水準云々の考えは無視することにした．従って正確には，この方法は「χ^2 法」とはいえないが，実用上は良い結果を与えること

がわかった.

このような考え方に基づき日本語文献中の名詞単語を χ^2 値の大きい順に並べることにより，重要語をどの程度抽出できるか実際の文献を用いて行ってみた．第1には中学校理科の教科書を用いたが，サンプルが大きいほど有効であることから，大量文献としては，日本科学技術情報センター（以下 JICST と略す）発行の「科学技術文献速報，電気工学編」の内容を記録してある磁気テープのうちから，各文献の「抄録」の部分を資料として使用した．

3. 中学校理科の教科書における重要語抽出

入力データとしては，中学校一年の理科の教科書「中学理科 1-B」（啓林館）の8章と9章（1ページ～81 ページ）を使用した．その目次は **Table 1** に示す

Table 1 Contents of Chapter 8 and 9 of the textbook.

目　次	実験のため設定した文献区分	目　次	実験のため設定した文献区分
第8章　物質と原子		**第9章　電流のはたらき**	
1. 化合物と元素		1. 電流と電圧	
1. 物質の変化	1	1. 豆電球を流れる電流	7
2. 物質の分解		2. 電池のはたらき	
3. 化合と化合物		3. 金属線を流れる電流と電圧との関係	
4. 物質の変化と分子	2		
5. 元素と元素記号		4. 金属線の電気抵抗	
6. 炎色反応		5. 水溶液を流れる電流	8
2. 化学変化と物質の量		6. 回路の性質	
1. 水の電気分解	3	7. 直流と交流	
2. 水の合成		2. 電気エネルギー	
3. 物質の変化と質量		1. 電流による発熱	9
4. 質量保存の法則		2. 電力	
5. 金属の酸化		3. 電子と電流	
6. 金属酸化物の還元と定比例の法則	4	1. 気体の中を流れる電流	
7. 金属と酸性の物質		2. 陰極線と電子	10
8. 化学変化とエネルギー		3. 二極管と電子	
3. 化学変化と分子・原子		4. 金属内を流れる電流と電子	
1. 分子より小さい粒子	5		
2. 化合物の分子と原子			
3. 原子や分子の大きさ			
4. 化学変化と原子の結合			
5. 物質を表わす式	6		
6. 化学変化を表わす式			

通りである．これを内容によって表に示すように8章を6つ，9章を4つのグループに分けて，合計 10 個の文献とみなした．これらの総語数は 16,363 語，異なり語数は 1,371 であった．

（1） 8, 9章，全部で 10 個の文献区分に対して，前章に述べた方法で行った結果の χ^2 順上位 20 位の語を **Table 2** に示す．目次と χ^2 順の上位の語との関係は，だいたい満足できるものであるが，このままの順位では，総語数の少い文献に出現する単語が強調される傾向がある．一方この χ^2 順位からその単語がピークを示した文献区分番号より各文献区分別に上位4位を取り出すと，**Table 3** のようになり，χ^2 順位よりは納得しうる形となる．

これを見ると，アブストラクトとまではいかないにしても，教科書のそれぞれのセクションの内容を伝え

Table 2 The 20 words selected according to χ^2-value.

χ^2 順位	順頻位度	単　語	頻度	χ^2	χ^2 がピークを示した文献区分
1	47	原　子	59	11.93	5,6
2	15	電　流	145	11.86	7,8,9,10
3	69	元　素	34	11.85	2
4	25	分　子	95	11.27	2,5,6
5	75	塩化ナトリウム	31	10.30	1
6	98	電　子	23	10.14	10
7	112	分子式	18	10.09	6
8	89	豆電球	26	9.75	7
9	84	塩素酸ナトリウム	28	9.69	1
10	82	粒　子	29	9.12	1
11	149	電熱線	13	9.05	9
12	166	電　力	12	8.35	9
13	45	表わす	61	8.31	6
14	108	亜　鉛	19	8.03	4
15	31	電　圧	81	7.99	7,8,9,10
16	19	物　質	120	7.18	1,2,3,4,5,6
17	78	数	30	7.13	6
18	114	乾電池	17	7.11	7
19	124	熱	16	6.80	9
20	162	化学反応式	12	6.73	6

Table 3 The first 4 words selected from each division.

文献区分	上　位　4　単　語
1	塩化ナトリウム，塩素酸ナトリウム，粒子，物質
2	元素，分子，物質，炎
3	物質，水素，上ざら天びん，化学変化
4	亜鉛，物質，水素，発生する
5	原子，分子，物質，結びつく
6	原子，分子，分子式，表わす
7	電流，豆電球，電圧，乾電池
8	電流，電圧，回路，金属線
9	電流，電熱線，電力，電圧
10	電流，電子，電圧，流れる

るものとしてかなり良好な結果であると言える．χ^2 順位で 13 位の「表わす」は，一般語であるようにも考えられるが，文献区分6の目次に2回も使われているのを見ると，局所的には重要語であると考えられる．

逆に，文献区分1の「塩化ナトリウム」「塩素酸ナトリウム」，文献区分3の「上ざら天びん」などはあまりに具体的な名称であり過ぎるきらいがある．しかし文献区分7の「豆電球」などは目次にも使われているわけで，結果の評価は少しやっかいである．

また目次を見ると，本来，文献区分4は「酸化」「酸化物」「酸性」などの言葉が重要語として選ばれるべきであるが，頻度としては「酸」という概念がこれらの言葉に分散して現れた結果，抽出されなかった．こういう場合にこそシソーラスによる単語のグルーピングを行う必要がある．

(2)　8章および9章をそれぞれ独立に調べた結果の χ^2 順上位 20 位の語を **Table 4** に示す．

χ^2 順上位の語を(1)の結果と比較すると，8章では「物質」が脱落し，「分子」の順位が下がったこと，9章では「電流」「電圧」「流れる」が脱落したことが特徴的で，これらは，章を特徴づける単語ではあっても

Table 4 The 20 words selected only from Chapter 8, and those only from Chapter 9.

	第8章のみ　6文献区分					第9章のみ　4文献区分				
χ^2 順位	頻度順位	単語	頻度	χ^2	文献区分ピークを示し	頻度順位	単語	頻度	χ^2	文献区分ピークを示し
1	45	表わす	37	7.18	6	68	発生する	14	3.48	3
2	31	原子	53	6.69	5,6	49	電子	23	3.38	4
3	47	元素	34	6.02	2	40	豆電球	26	3.06	1
4	80	分子式	18	5.61	6	78	電熱線	13	3.01	3
5	63	粒子	25	5.59	1	86	熱	12	2.78	3
6	53	塩化ナトリウム	31	5.31	1	87	電力	12	2.78	3
7	55	塩素酸ナトリウム	28	5.03	1	63	乾電池	16	2.57	1
8	79	亜鉛	19	4.41	4	98	熱量	11	2.55	3
9	17	分子	94	4.33	2,5,6	127	水熱量計	8	1.85	3
10	57	数	27	4.10	6	88	二極管	12	1.76	4
11	105	化学反応式	12	3.73	6	53	◆	20	1.74	3
12	49	熱する	33	3.30	1	77	時間	13	1.71	3
13	104	エネルギー	12	3.20	4	71	電気	14	1.67	4
14	95	上ざら天びん	12	3.16	3	105	電気エネルギー	9	1.62	3
15	77	1個	20	3.09	6	41	金属線	25	1.56	2
16	135	炎	8	2.88	4	76	金属	13	1.53	4
17	127	大きさ	9	2.79	5	67	電気抵抗	15	1.48	2
18	165	炎色反応	6	2.75	2	103	電極	10	1.47	4
19	111	結びつく	11	2.60	5	104	放電	10	1.47	4
20	102	含む	12	2.49	2	106	端子	9	1.45	1

Table 5a The first 4 words in each of the 6 divisions of Chapter 8.

文献区分	上位4位
1	粒子，塩化ナトリウム，塩素酸ナトリウム，熱する
2	元素，分子，炎，炎色反応
3	上ざら天びん，質量，入れる，出入り
4	亜鉛，エネルギー，燃える，出入り
5	原子，分子，大きさ，結びつく
6	表わす，原子，分子式，分子

Table 5b The first 4 words in each of the 4 divisions of Chapter 9.

文献区分	上位4位
7	豆電球，乾電池，端子，回路
8	金属線，電気抵抗，回路，G
9	発生する，電熱線，熱，電力
10	電子，二極管，電気，金属

節を特徴づける重要語ではないという事を示している．また各文献区分での順位は **Table 5** となった．この結果から見ると，8章では「物質」「水素」，9章では「電流」「電圧」が脱落し，それぞれ「炎色反応」「電気抵抗」「二極管」などの一層細い重要語が現れてくる．ここでは，新たに「入れる」とか「大きさ」などの一般語であるべきものまで現れてくるが，この場合は，「入れる」とか「大きさ」を修飾したり，その目的語である，「何を入れる」とか「何の大きさ」の，その「何」が問題であるという事になる．従ってこのような要素まで考えようとすると，ある程度の構文分析が必要となる．

文献区分5と文献区分6には，「原子」「分子」が共通しているが，これはもともと同じ節の文章を無理に2つに分けたもので，元来重要語は共通してもおかしくはなく，この点文献の類似性という事もこれにより判断しうると考えられる．もう1つ問題になるのは，「質量保存の法則」などは重要語としての資格は十分であるが，データ入力の際「質量」「保存」「の」「法則」に分割され，原形の意味が失われてしまって現れてこないという問題がある．これは単語をどこで区切るか，＜複合語の分割＞という問題である．

4. 電気工学関係における重要語抽出

(1) 処理の方式

JICST の磁気テープ1巻には巻によって多少の差はあるが，およそ 1,200～1,300 の文献が含まれており，ひとつの抄録には40～50の単語が含まれている．(ここでいう「単語」とは，後述するように，漢字，

Table 6 The division of 'Current Bibliography on Science and Technology'

A. 電気工学一般	
1. 電気工学一般	2. 電気材料部品
B. 計測・制御	
3. 計測	4. 制御
C. 電力工学	
5. 電力	6. 電力機器 7. 電力応用
D. 電子工学	
8. 電子工学一般	9. 電子部品 10. 電子回路
11. 量子エレクトロニクス	12. 電子技術の応用
E. 通信工学	
13. 通信一般	14. 電波伝搬，アンテナ
15. 伝送方式，機器	16. 通信応用
F. 情報処理工学	
17. 情報処理工学	18. 電子計算機
19. 情報処理応用	

カタカナ，または英文字で構成された名詞形の単語である)．文献内容は JICST により **Table 6** のように予め分類されている．A～F の大見出しの分類ではあまりにも分類が大まかになりすぎて，重要語を出してもあまり意味がないと思われたので，結局，1～19 の小さな区切りを1分野として処理した．1分野あたりの文献数は同じ分野でも巻によってかなり大きな変動があって，それぞれ50～200文献ずつくらいである．

また，重要語となりうる候補として選び出す語としては「抄録」文中の漢字またはカタカナまたは英文字で書かれた単語だけを抽出することにした．これはひらがなで書かれた単語は，一般に助詞や助動詞，または自立語の活用語尾あるいは形式名詞など重要語となるとは考えられないような語がほとんどであって，今これらを無視したところで，結果にはほとんど影響が出ないであろうと判断したのと，こうすることによって単語の切り出しが，きわめて機械的に簡単にできるからである．

(i) 名詞の切り出し

当初は，漢字，カタカナ，英文字で構成されたひと続きの語であるならどのようなものでも抽出し，記録していたが，その結果次のような問題が明らかになった．

① 漢字，カタカナ，英文字のひと続きの語を全く機械的に切り出しているので例えば「パターン」「パターン認識問題」などをそれぞれ全く別々の単語とみなして処理してしまう．この場合，常識的にみて「パターン」が重要語であるような文献でもそれが結果として出ない可能性がある．

② 例えば，「対する」，「対して」などの「対(たい)」と，「対(つい)」とを全く同じ「対」として処理するので，「対(つい)」が重要語となるような文献があっても出ない．

③ 頻度テーブルを磁気ディスク内に作っているので，切り出す単語が多いとディスク入出力に時間がかかりすぎて実用に耐えなくなる．

そこで，これらの問題を解決するため以下のような手段を用いた．

①については，サブルーチンを変えて，漢字とカタカナと英文字とで，それぞれ他種の文字が現れたときはそこで単語を二分する．すなわち前例の場合なら「パターン認識」を「パターン」と「認識」とに分離するようにした．

②，③については，文献中より名詞単語らしいもののみを抜き出すという方法で解決を試みた．このようにすれば，当然，ディスクに収められる単語数は大幅に減少することが予想され，文献読みこみ時間が短縮される．また前例の「対する」と「対(つい)」の問題にしても，動詞である「対する」は読みとばされ，名詞である「対(つい)」は残るから②の問題もかたづくわけである．

文中から名詞らしいもののみを抽出する方法としては，次のような方法を用いた．

すなわち，文中の漢字，カタカナ，英文字で構成された単語の次の文字を見て，それが名詞に接続する助詞(「が」，「の」，「に」，「を」，「へ」，「と」，「は」，「も」，「で」，「や」，「より」，「から」，「こそ」，「まで」，「など」，「さえ」，「でも」，「しか」，「だけ」，「ほど」，「きり」，「ずつ」，「ばかり」，「くらい」，「および」，「ならびに」)であるか，または名詞の後にくるような特殊文字(コンマ，ピリオド，かっこなど)であった場合にのみ，それを記録する単語としてディスクに入れるという方式である．

この方式を採用しても，まだ「最(もっと)も」など不適当な語も抽出されてしまうという問題が残っているが，統計的処理の場合，そこまで厳密性を要求してもあまり意味がないと思われるし，またこのような一般語は χ^2 値も小さいと予想されるから，あとでふるいにかけて落とされ，全体的にたいした影響は出ないと考えられる．

(ii) ハッシングについて

抄録文中から(1)で述べた過程を経て，抽出された名詞単語は，ディスク内に，それ以前に収められた単語と照合され，もしその中に同一単語があればその単語の頻度が1つ加算され，またもし初めて現れた単語であれば，新しく記録され，頻度1とするわけであるが，この照合の際ハッシングを用いて出現単語との比

較の回数を減少させた．ハッシングの方法としては切り出された単語を構成する文字の漢字テレタイプ・コードの総和をとり，それを 9,696（ディスクセクター数）で除したときの余りに1を加えた数をその語の収まるべきディスクセクター番号とするというものである（ここで1を加えるのは単語によってはハッシング・コードが0になる可能性があるからである）．また，ハッシングによって同一番地におちる，いわゆるコリジョンが起きたときはリニアーサーチとした．

この方式による読みこみによって，JICST の磁気テープ1巻あたりミニコンピュータで約1時間半かかった．ハッシングによるコリジョンの回数は読みこみの終わりの部分になっても約2,3回程度であった．

（ⅲ） 全体の処理の流れ

全体の処理の流れは次の通りである．すなわち，JICST 磁気テープから1巻分ずつ読みこんでディスク内に単語の頻度テーブルを作り，それを磁気テープに移す．これをくりかえし，1本の磁気テープに数巻分の頻度テーブルを収めた上で，今度はそれを合成して全巻分の頻度テーブルを作って，χ^2 計算を行い出力するというものである．

プログラム言語はすべて FORTRAN を用い，計算機はミニコンピュータ TOSBAC-40 C を用いた．

1単語の領域としては64バイト（整数型変数32個分）をあて，**Fig. 1** のように，各領域の用途をわりあてた．このようにすれば，当然 10 字より長い単語は記録できないから，それ以上に長い単語は 10 字までカットした．実際の結果でも 11 字以上の単語はのべ数にして全体の約 0.1%（1巻あたりおよそ 240 語）程度にすぎず，また一般にそんなに長い単語はきわめて特殊な単語とみなせるから，全体的な影響はほとんどない．作ったプログラムの主な特徴は以下の通りである．

- 分野数……19
- 分野ごとの χ^2 順出力……可変（一応200位まで）

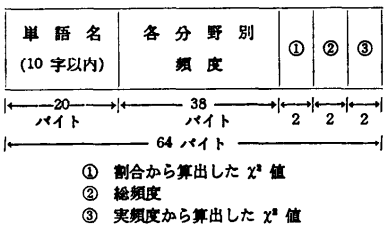

Fig. 1 Entry of occurence of a word.

- JICST 1巻中の異なり単語数……38,784 種まで（実際には1巻あたり 13,000 種程度しかない）
- 同一ハッシング・コード語……100 語まで（実際には最高でも 10 語程度，ふつう 2～3 語）

なお，出力としてはソニーテクトロニクス 4010 ブラウン管にドット方式で漢字を出力できるようにしてあるので，その上に出し，ハードコピーユニットによりコピーをとった．漢字のドット出力は 32×32 ドットで並列出力のためのインターフェイスを作っているので十分な速度で出力できた．

（2） **結果と考察**

実験に用いた資料は次の通りである．

資料：JICST「科学技術文献速報」（抄録部）Vol. 17, No. 8～No. 11,（4巻）

文献数 4,890 文献

No. 8	1,197 文献
No. 9	1,314 文献
No. 10	1,140 文献
No. 11	1,239 文献

文献番号 46009730～46014619

異なり単語数 33,764 個

のべ総単語数 120,050 個

実頻度をもとにした χ^2 値を(1)式にもとづき各分野ごとに算出し，χ^2 順にソーティングして出力したもののうち，上位10位ずつを **Table 7**（次頁参照）に示す．また頻度はある程度大きいが χ^2 値の小さい語一つまり一般語を上位30位まで **Table 8**（次頁参照）に示す．

概観したところ，各分野ともだいたいうまく重要語が上位に出てきている．しかし頻度がきわめて大きな単語の場合，たとえそれが常識的にみて「一般語」であったとしても，どうしても出現頻度にかたよりが生じ，そのため(1)式の性質から，χ^2 値としてきわめて大きな数値をとることがありうる．

その例としては，第1分野「電気工学一般」5 位の「方法」などがあり，この語の場合は，全体で 943 回も出現し，しかも，かなりのかたよりがあるため，χ^2 値が171.18と，大分大きな値になってしまっている．

これと同様の理由で，一般語であるにもかかわらず，χ^2 値が大きくなって上位に出ているものとしては，以下にあげるような単語がある．（χ^2 値が100以上のもの，カッコ内は左から分野番号，その分野での順位，総頻度，χ^2 値）

問題 (1, 7, 551, 170.60)，本文 (1, 19, 379,

Table 7 The first 10 words of each division.

	1. 電気工学一般	2. 電気材料部品	3. 計測	4. 制御	5. 電力
1.	グラフ	エポキシ	測定	制御	MW
2.	メートル	ケーブル	計測	プロセス	送電線
3.	枝	トリー	校正	プラント	系統
4.	定理	マグネットワイヤ	誤差	X	エネルギー
5.	方法	樹脂	測定誤差	タイマ	kV
6.	有向閉路	トリーイング	計器	ゲーム	ガス
7.	問題	内部応力	オシロスコープ	最適制御	電力系統
8.	伝達関数	フェライト	電圧計	計装	発電所
9.	解	耐熱性	量計	調節計	年
10.	Flow	シース	変換器	シーケンス	変電所
	6. 電力機器	7. 電力応用	8. 電子工学一般	9. 電子部品	10. 電子回路
1.	回転子	光源	プリント	ダイオード	トランジスタ
2.	発電機	照明	エッチング	エピタキシャル	回路
3.	MVA	照度	セル	層	フィルタ
4.	トルク	1ε	基板	GaAs	増幅器
5.	断器	デザイン	チップ	発光	周波数
6.	巻線	ランプ	ピット	半導体	発振器
7.	変圧器	器具	マスク	Si	導波管
8.	サイリスタ	照明器具	基板上	シリコン	入力信号
9.	誘導発動機		ボンディング	注入	デルタ
10.	発動機	納入	モジュール	接合	線路
	11. 量子エレクトロニクス	12. 電子技術の応用	13. 通信一般	14. 電波伝搬	1ç. 伝送方式
1.	光	ベッド	信号	ビーム	英国郵政省
2.	ホログラム	テープ	符号	ダイポール	チャネル
3.	Cm	磁気	通信路	スロット	ダイバーシチ
4.	レーザ	生体	情報源	アンテナ	導波路
5.	モード	記録	FM	放射	方式
6.	パルス	音	雑音	アレイ	トラヒック
7.	発振	患者	通信	電離層	中継器
8.	He	血液	系列	指向性	伝送
9.	ポンピング	弾性表面波	エントロピー	プラズマ	dB
10.	メーザ	超音波	率	ダイポールアンテナ	Km
	16. 通信応用	17. 情報処理基礎	18. 電子計算機	19. 情報処理応用	
1.	テレビ	オートマトン	メモリ	計算機	
2.	放送	認識	コンパイラ	プログラム	
3.	カラー	アルゴリズム	プリンタ	CAI	
4.	テレビジョン	集合	コード	データ	
5.	レーダ	パターン	ディスプレイ	システム	
6.	カメラ	言語	ユーザー	ホスト	
7.	画質	チューリング	命令	機能	
8.	VTR	漢字	レジスタ	モデル	
9.	交換機	ストローク	マイクロプロセッサ	サブジェクト	
10.	サービス	コンピュータ	マクロ	自動設計	

Table 8 30 words which are classified as common words.

1. 両 者	16. 詳 細
2. 要 因	17. 配 置
3. 容 易	18. 発 展
4. 共	19. 種
5. 十 分	20. 結 合
6. 応用例	21. 簡 単
7. 内	22. 内 容
8. 必 要	23. 多 数
9. 際	24. 一 般
10. 最 初	25. 可能性
11. 傾 向	26. 役 割
12. 利 点	27. 主
13. 検 出	28. 役 者
14. 研 究	29. 分 析
15. 同 様	30. 使 用

108.87), 温度 (2, 16, 102, 157, 170.79), 特性 (2, 29, 454, 131.76), 装置 (3, 12, 392, 139.45), 原理 (3, 16, 117, 114.42), ディジタル (3, 20, 239, 109.33), % (8, 24, 239, 125.81), パラメータ (13, 17, 311, 129.77), 場合 (14, 23, 963, 150.04), 方式 (15, 5, 379, 204.97)

また, このままの方式で小数点以下2位まで χ^2 値を出そうとしたとき, **Fig. 1** に示すように χ^2 の値は2バイト分しかとらなかったので χ^2 の上限が 327.68 (300 をこえるものは強制的に 300 とした) となって

それ以上の数が表わせず，例えば「11. 量子エレクトロニクス」では，上位8位まで χ^2 値が 300.00（プログラム上の最大値）となって重要度の区別がつかない．

これらの問題点，特に前者の問題を解決するため，次のような手段を考えた．

（1） 各単語の頻度ヒストグラムをなんらかの形で正規化し，ピークの鋭さから，単語の集中度をきめ，χ^2 順と，この順とをかねあわせてランクづけする．

（2） 実頻度からではなく，単語の頻度の分野における割合から，χ^2 値を算出する．この場合，頻度が多くても割合はさほど大きな数字にはならず，従って，頻度が大きいゆえに χ^2 値が大きくなるといった問題はかなり修正されると思われるからである．

（1）の方式は，かなり有効に働き，ピークのなだらかな一般語は除外されることが予想されるが，ここではアルゴリズムの簡単な（2）の方式を試み，結果としてかなりよいものを得た．

一般語は，前の方式より，軒なみにランクが落ち，また上位の重要語の「χ^2 値」もはっきりと出て，その値からだいたいの重要度を直感的に判断することができる．

前方式との比較例として第1分野「電気工学一般」のみの結果を取り出して示すと次のようになっている．（上位 10 語まで，左が前方式，右が改良した方式，数字はそれぞれの χ^2 値）

1. グラフ	300.00	グラフ	16.383
2. メートル	293.40	メートル	6.155
3. 枝	222.21	枝	4.652
4. 定理	177.71	有向閉路	3.590
5. 方法	171.81	定理	3.334
6. 有向閉路	171.15	伝達関数	3.121
7. 問題	170.60	Flow	3.077
8. 伝達関数	158.58	作用素	2.960
9. 解	155.92	問題	2.862
10. Flow	146.70	解	2.812

この結果1位の「グラフ」がとびぬけて重要であることがわかるし，また前方式では5位だった「方法」が10位以下（実際は11位）に落ち，代わって 10 位の「Flow」が7位に，11位の「作用素」が8位に上がって，これらがかなりの重要語であることを示している．

その他一般語として，順位の落ちているものを調べてみると，この第1分野では，主なものだけでも下の通りである．

本文（19 位→24 位），数（21 位→28 位），節点（34 位→51 位），安定性（32 位→55 位），手法（35 位→65 位），解析（58 位→70 位），理論（55 位→66 位），表現（56 位→79 位）

このように一般語のランクが下がったぶんだけ，当然重要語に属するもののランクが上がっているわけである．

5. おわりに

以上の結果を見てわかるように，χ^2 値を指標として用いる本研究の重要語自動抽出法はかなり有効なものであるということができる．少なくとも，常識的にみた場合重要語とみなせるような単語の χ^2 値は，まず例外なく上位に位置している．

また一般語が上位に残っていたり，不適当な単語が抽出されてしまったりする（例えば4.の「15. 伝送方式・機器」1位の「英国郵政省」など）といった問題点は，今後，文献数を大幅に増加していくことによって，その大半は自然に解決されることが十分期待できるし，統計的処理法を用いる場合のひとつの宿命として，どうしても完璧なものは望めない以上，χ^2 法を大量文献の重要語自動抽出の手段として利用するという本研究の試みは，いちおう成功したと結論づけてよいものと思う．本研究の一部は文部省科学研究費補助金によった。

参考文献

1) G. Salton: "Recent Studies in Automatic Text Analysis and Document Retrieval", JACM Vol. 20, No. 2, (1973, April).

2) 長尾真，落合和博，水谷幹男："日本語文献検索におけるカイ2乗を使った重要語自動抽出"，昭和49年度電子通信学会全国大会 No. 1451，昭和49年7月．

（昭和50年7月16日受付）
（昭和50年8月25日再受付）