解　説

自動索引付け研究の動向†

諸　橋　正　幸††

1. はじめに

自動索引付けは文献検索システムを構築する際の重要課題の一つであり，特に大量の文献を収集し，検索システムを提供する機関にとっては，省力化と最新情報の反映という目的から必要性の高い研究テーマである．自動索引付けは文献中から検索のために有効な索引(index)を抽出することが目的であり，索引は少なくとも次の二つの特性を備えている必要がある．

（1） 文献あるいは文献集合の内容を的確に表現していること（表現力）．

（2） 文献あるいは文献集合間の識別が十分行えること（識別力）．

識別力（ある文献集合を他から区別するのに充分な索引であること．例えば，科学技術文献の検索において「科学」は収容文献全体を指すであろうから識別力はない．）は実用的な検索システムを作る上で重要な尺度で，利用者は適度な量の文献を検索できるようになる．また，この特性は検索システムがカバーする分野に依存する．例えば，「木」という索引語はコンピュータ科学の分野の検索システムでは識別力を発揮するが，動植物学の検索システムにおいてはほとんど識別力はなく，索引語としての価値は低い．

上の例では索引語について述べたが，索引に用いられる情報について，もう少し触れておく．

最も一般的な索引は索引語（index term）で，個々の文献の内容を的確に表現する用語のことである．索引語間の関係は，シソーラスの場合のように必ずしも定義されている必要はない．論文等の場合は，検索システムのためにこれを著者があらかじめ与えることを強制していることもある．また，通常索引語は標題，抄録等に直接現われることが期待されるので，自動索引付けの対象としてよく取りあげられる．

シソーラスは，あらかじめ分類された分野を前提に各分野の内容を代表する用語(descriptor)を集めたもので，分野を限定し，それに属する文献を集めるのに役立つ．descriptor は階層化されているので，選んだ文献が検索者の意図する範囲とくい違いを生じた場合は階層を上下に辿ることで検索意図に近づけることができる．また，うまく階層化できない概念については related term が用意されているので，検索を行う前にシソーラスで検索したい分野を限定するというのが典型的な利用法である．シソーラス自身を自動的に作成することについては筆者の知る限り，いまだ研究はないが索引語を付与する際にシソーラスを利用しようという試みはいくつかある[1),2)]．

用語以外の索引情報として，引用索引(citation index)が知られている[3)]．これは，その名が示す通り論文の引用文献を調べて論文相互の関係を有向グラフにしたものである．このグラフから相互参照の多い部分(cluster)は同一分野で，読んでおく必要のある重要論文群であると見なせる．また，矢印を逆に辿ることで研究の契機となった論文がわかり，矢印を順に辿ることで研究の最新の動向を摑むことができる．引用索引は，着眼が面白く利用法も上記のようにいろいろあるが，自動索引付けという観点からは単純な手法で実現できるので，本論の主旨には沿わないであろう．

以上紹介した索引のうち自動索引付けのアプローチが多岐にわたり，研究の数が多いのは索引語の抽出なので，ここでは自動索引付けを索引語の自動抽出という範囲で摑えて，以下にその主な手法と特徴を紹介する．

2. 自動索引付けの手法

自動索引付けとして提唱されている研究の多くのものは，文献の内容を表現する単語列（抄録，標題等）を索引語抽出のための出発点とし，これを入力した時に的確な索引語の集合を出力として得られるような関数（あるいはモデル）を定義する形をとる．したがっ

† Automatic Indexing Survey by Masayuki MOROHASHI (Science Institute, IBM Japan, Ltd.).

†† 日本アイ・ビー・エム(株)サイエンス・インスティチュート

て，抄録，標題等の入力となる単語列には直接現われない用語は，索引語にはなり得ないことになる．自動索引付けの研究者は，それゆえ，次のような前提の上に研究を始める．

前提：文献で記述されている内容あるいはその一部を的確に表現する語は必ず文献中に現われる．

各文献のどの部分を調べれば効率的に索引語を拾えるかが次に問題となるが，限られた長さ（単語数）で内容全体を効率よく記述している部分として抄録や標題が多く使われる．抄録や標題には本文中で述べている研究そのものの記述と同時に，研究の占める学問上の位置付けや応用面の価値なども記述されている．検索の立場からは，それら周辺の記述も充分役立つ索引語に成りうるものであり，これらの部分が抽出の対象として使われる理由となっている．

索引語抽出のためのモデルは，大きく二つに分けられる．一つは索引語の構造や表現形式等の特性，あるいは索引語が使われる文の特性を利用して索引語を抽出するモデルであり，もう一つは語の出現頻度特性を利用したモデルである．以下の2章でこの二つのモデルの一般的特徴と代表的な研究を紹介する．

3. 表現形式，構文構造を利用した索引語抽出

前述の通り，語の構造（主に複合語内の語構成と前後の語との関係）や，文の構造（文を支配する動詞と格の関係）から索引語を抽出する．この手法は局所的情報によって抽出を行うため，「はじめに」で述べた索引が持つべき二つの特性のうち識別力に関して十分な結果を得られないという共通した欠点をもつが，次に述べる出現頻度によるモデルに比較して，表現力は逆に優れている．

3.1 原始的アプローチ

言語学的情報を積極的に利用する手法について紹介する前に，ごく原始的ではあるが，実際的で有効な手法を述べておく．これは，IBM の STAIRS (STorage and Information Retrieval System)[4] で最初に実現されたもので，小中規模の商用情報検索システムで有効であり，コンピュータ・メーカの多くが同様のシステムを提供している．手法は，各文献の対象とする部分（抄録，標題など）に現われる全単語から冠詞や代名詞など表現力，識別力のまったくない不要語（これらを stop word と呼ぶ）を除いたものすべてを索引語とするものである．不要語は検索システム構築の際に定義することができるから，例えば，電気分野の検索システムで electric, electronic などを不要語として索引語の対象からはずすことも可能である．不要語の指定はシステム管理者に任されているが，不要語指定にあたって文脈に依存した指定ができないため complex number（複素数），proper noun（固有名詞）などの用語が使われる可能性のある場合は，形容詞 complex, proper を不要語とするわけにはいかなくなる．また，抽出にあたって，各語の形態素解析はしないから，用言を不要語にする場合には，すべての活用形を登録する必要がある（例えば，be 動詞について，be, am, are, is, was, were, been, being）．このように，抽出方法が原始的であるため，検索コマンドが高度になっており，それによってシステム全体の精度を上げるよう工夫されている．その一例に単語の部分一致が挙げられる．前にも述べた通り索引語は形態素解析が行われないので，用言の語幹だけの一致を可能にすることは必須である．また化学分野における索引などでは基を知るために語の最後尾の一致を必要とする．そのため，部分一致は前方，後方一致ができるようになっている．また索引語の組合せも検索精度を上げるために重要で，複合語の指定は複数個の単語が順に並んでいるという形の検索条件で行う．索引語の組合せは隣接以外にも同一文内，同一段落内など，より緩い条件も指示できる機能もある．

STAIRS では索引付けの機能に関しては，その原始的手法のゆえに自動という言い方をせず，単に蓄積 (storage) と称しているが，他の自動索引付けの手法の多くが研究段階にある中で実用的な手法として評価できる．ただし，索引語が多いことと，その中に多くのノイズ（索引として役立たない語）があるという欠点に起因する，

（1） 索引に要する記憶装置（磁気ディスク）を多く必要とする

（2） 検索条件をうまく設定しないと不要な文献を多く含む答えしか得られない

などの問題があるので，運用を工夫する必要がある．

日本語においては，単語の分かち書きという習慣がないために STAIRS のようなアプローチをとることが困難であるが，漢字一文字を単語とみなして同様の処理を行う検索システムが存在する[5]．このシステムにおいては，すべてのひらがなとカタカナを索引の対象からはずし，漢字一文字を索引語（字？）としている．本来の意の索引語は従って漢字の隣接関係を使って指示する方法をとる．用語をひらがな書きされたり

(当用漢字，常用漢字による漢字使用の制限で専門用語のかな書きやまぜ書きが多くなった)，外来語のカタカナ表記に対処できないという欠点のため，印欧語を扱う STAIRS に比べ検索精度が下がるが，同一の発想にたって日本語を処理した例として注目すべきシステムである．なお，このシステムは法令検索を目的としているため，かな書きや外来語の問題は一般の文献に比べると影響が少ない．

3.2 DDC の機械補助索引

索引語の語構成に対する一般的特徴を利用して索引付けを行う例の代表として，米国防省のドキュメンテーション・センター (DDC—Defence Documentation Center) の機械補助索引システム[6)~8)]を紹介する．なお，このシステムは JICST の中井氏により，参考文献 8) で他の機械補助索引システムと共に詳しく紹介されているので，興味のある方は一読されたい．

このシステムもまた，機械補助索引 (MAI—Machine-Aided Indexing) という名付け方をしており，厳密な意味での自動索引付けとは区別すべきではあるが，人手による処理は特徴抽出とそれを特殊な辞書に反映させる作業であり，索引付け自体は機械的に行われるので，ここでは自動索引付けの一つとして紹介する．

この手法においては，まず国語辞典の見出しのような単語辞書 (Recognition Dictionary) を用意する．辞書中の各語には処理ルーチン番号が記載されている．処理ルーチンは，その語に品詞を付け，複合語の単位をみつける仕事をする．品詞が直接辞書に記載されていないのは，品詞が語に一対一対応でつくのではなく文脈に依存しているためで，処理ルーチンは前の処理の状態を次に引き継ぐことで文脈に依存した品詞付けが行える．付与される品詞は，文法論で伝統的に扱われる区分とは異なり，索引語抽出を目的とした独自のものである (表-1)．この他に，不要語にあたるものがあるが，これは処理ルーチンによって捨てられる．

こうして得られた複合語の品詞列を，もう一つの辞書であるフォーマット辞書(Format Dictionary, 表-2) と比較し，該当するパターンが存在するならば索引語とする．例えば，Protective clothing は

Protective ⟶ A

表-1 DDC の MAI における品詞

品 詞	説 明
A	adjective
N	stand alone noun
+	"and"
Z	weak noun
P	"of"
X	"or"
Y	"other"
B	special adjective class
R	special noun class

(文献 8) より)

表-2 フォーマット辞書

Rank	Frequency	Format	Rank	Frequency	Format
1	26,085	ZZ	40	167	NZZZ
2	20,539	AZ	40	167	ANN
3	18,633	N	41	160	ZR
4	6,433	AZZ	42	159	ZZPZ
5	6,155	ZZZ	42	159	ZZZZZ
6	4,957	NZ	43	156	ZNZZ
7	3,635	AN	44	151	AZAZZ
8	3,185	ZN	45	128	NZN
9	2,274	ZPZ	46	122	AZZN
10	2,043	AAZ	47	108	A+AN
11	1,439	ZAZ	48	105	ZZPN
12	1,433	AZZZ	49	104	AR
13	1,104	NZZ	50	85	AA+AZ
14	1,089	AZN	51	84	ZAZZZ
15	1,085	ZZZZ	52	83	ZNN
16	979	NN	53	68	NAN
17	972	ZPZZ	53	68	ZZZN
18	962	ZPN	54	64	NNN
19	953	AZN	55	42	AXAZ
20	782	ZNZ	56	38	ZZR
21	730	AAZZ	57	32	NNZZ
22	717	A+AZ	58	22	AZR
23	610	ZZN	59	19	AA+AZZ
24	432	ZAZZ	60	17	ANZZZ
25	367	AZAZ	60	17	BA
26	360	AAN	61	14	A+AZZZ
27	286	AZZZZ	62	12	A+YZ
28	284	APAN	62	12	ZNZN
29	273	NAZ	63	10	AA+AN
30	265	ZAN	63	10	AXAN
31	255	ANZZ	63	10	NAAZZ
32	246	BN	64	9	AXAZZ
33	229	AZNZ	65	8	ARZ
34	228	ZPNZ	66	6	A+XAZ
35	227	NPZ	67	3	A+YZZ
36	215	A+ZZ	68	2	A+XAN
37	181	ZZAZ	69	1	A+YAN
38	179	NNZ	70	0	A+YN
39	177	A+AZZ			

Frequency は約 200 万語の処理でマッチした回数 (文献 8) より)

clothing ⟶ N

という対応から品詞列 AN におきかえられ，フォーマット辞書の 7 番目のパターンにより，索引語とみなされる．このシステムで一番重要なことは，抄録，標題にあらわれるすべての語にここで体系づけられた品詞を与えることであり，これは試行錯誤から生まれた経験則である．したがって同じ名詞であっても単数形と複数形で品詞が異なるとか，summary は不要語であるが summarized には形容詞(A)の品詞が与えられるなど実際のデータに則した品詞付与が行われている．

単語辞書とフォーマット辞書の更新は，一致しない語やフォーマットがでるごとに行われており，この件数がどの程度で落ち着くかが実用化の鍵となる．

3.3 JICST の JAKAS

STAIRS の手法を，単語⟶漢字という対応で日本語に応用した例が存在するように，前節の MAI の手法を同様の対応で処理しようとするならば，単語辞書 (Recognition Dictionary) のかわりに漢和辞書を使い，索引語の構造を漢字ごとの品詞列とするフォーマット辞書を用意することで日本語文献の機械補助索引が実現できそうである．

この手法に非常に近い手法として JICST の高野等による JAKAS[9),10)] がある．このシステムは，元来，漢字カナ変換システム (K-KACS)[11)] で開発された手法をもとにしており，索引語にカナをふると同時に索引語抽出をする．JAKAS ではフォーマット辞書にあたるものは存在しない．単語辞書にあたる辞書は 4 種類存在し，これらの辞書を適用して得られた結果は自動的に索引語となる．これらの辞書に共通する作成方針は，例外を並べあげるということで，標準の処理でできないものに対処する．4 種の辞書と処理工程の対応を図-1 に示す．

分かち書き辞書は，入力文中のひらがな文字列につき分かち書き情報を与える．例えば，「が」はその前後の文字が漢字かカタカナの場合だけ格助詞とみなし，前後に空白を入れる．

切断辞書は文字(列)ごとに以下の 4 種の処理のいずれを行うべきかが記述される．

Cut 処理：当該文字(列)の前後で切断

Pass 処理：何もしない

Alter 処理：指定された文字(列)におきかえ

Quench 処理：文脈（字種による）に依存した切断

このうち，Alter 処理は，表記のゆれの統一や切断しすぎた部分の補修に使われる一種の例外処理である．

不要語辞書は STAIRS の不要語と同様，索引としての価値のない語を除去するために使われる．

漢カナ変換辞書は，漢字ごとのよみを持ち，二通り以上のよみを持つ漢字には，どのよみが妥当かを決定する処理名あるいは例外の語を記載してある．例えば，「歯」は漢カナ変換辞書で図-2 のように記述されている．1 行目のシ，ハは音よみ，訓よみを示し，標準処理 B は前後の文字がひらがなならば訓，その他は音よみを与える処理である．B の規則に合わない語が 2 行目以降に示される．

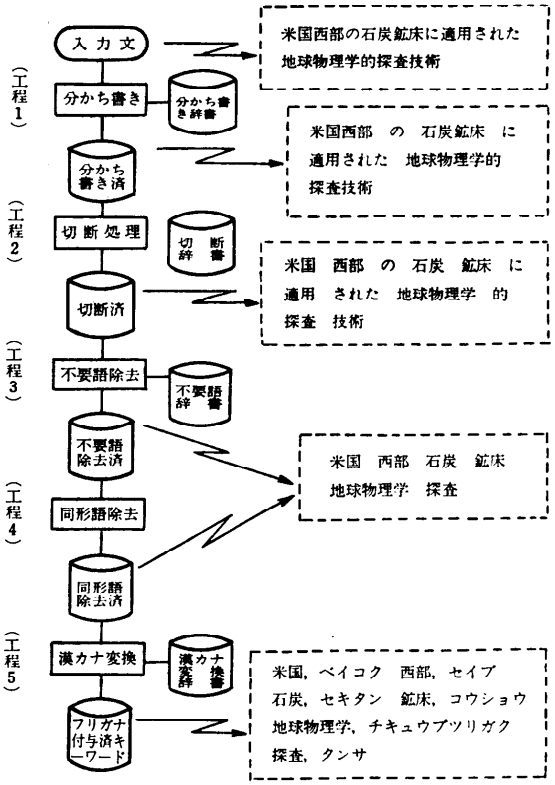

図-1　JAKAS の索引語抽出処理（文献 10)より）

```
歯，シ，ハ，B……標準の処理
  歯そう，シソウ
  歯茎，ハグキ
  歯並び，ハナラビ
  むし歯，バ          } 例外
  ムシ歯，バ
  歯車，ハグルマ
  ……
```

図-2 「歯」に対する漢カナ変換辞書（文献 11) より）

これらの辞書のうち，語数の大きくなる切断，漢カナ変換辞書は漢字を基本的単位とする辞書となっているので単語単位の辞書に比べて，サイズが小さくてすむ利点があり，また，例外のみを辞書に追加するという方針をとるので，DDC のシステムに比べ辞書の更新作業は楽である．また，現在のところ JICST の理工学文献の標題をもとにした特別の辞書を作成，更新しているので，実用に耐えるシステムとなっている．

3.4 絹川の文構造解析による自動索引付け

3.2, 3.3 では，文献抄録あるいは標題という範囲内での言語現象を的確に摑えるような特徴を自ら発見し，組み上げていくという形でのシステムを紹介した．そこで使われる特徴は主に形態素レベルの情報が多いため，索引語抽出対象が変化した場合に同様に適用できる保証はなく，システムの精度を保つには常時辞書の更新を続けていく必要がある．ここで紹介する日立の絹川等による手法[12]は，より一般的な言語モデルに基づいた構文情報を用いて，索引付けを行おうとする試みである．

このシステムでは，茨城大の石綿の動詞による語彙分類[13]が使われ，動詞によって決定されるロール（名詞句の意味的役割）により，索引語抽出と同時にそのロールをも付与する．

ロールには，①主体，②客体，③時，④場所，⑤活動，⑥その他の主題がある．

処理は，まず文中の文節を認定し，文節とそれを支配する動詞との関係を摑む．次に文節末の付属語や，文節内の名詞の意味分類を用いてロールを決定する（**図-3**）．

実験は，外電記事文 281 件について行われ，記事中から，索引語抽出をすると同時に，それらの間の 5W 1H の関係を付与している．こうしたロール付けは，自動抄録のための基礎技術として重要で，単に索引としての用語抽出でなく，用語間の関係から文中に記述されている内容の把握へと進む可能性を持っている．

3.5 形態素，構文情報を用いたその他の試み

一般的なパーザを用いて索引付けを行う試みとして，ロッキードの Earl[14),15)] の報告がある．ここで用いたパーザ PHRASE は，4つのレベルから成り，その最初のレベルで文中の名詞句，動詞句，不定詞句をみつける．英語では多品詞語が多く，特に名詞と動詞の同形が多いため，句の認定には，この品詞認定を正しく行うことが重要である．このために形態素の情報が使われる．

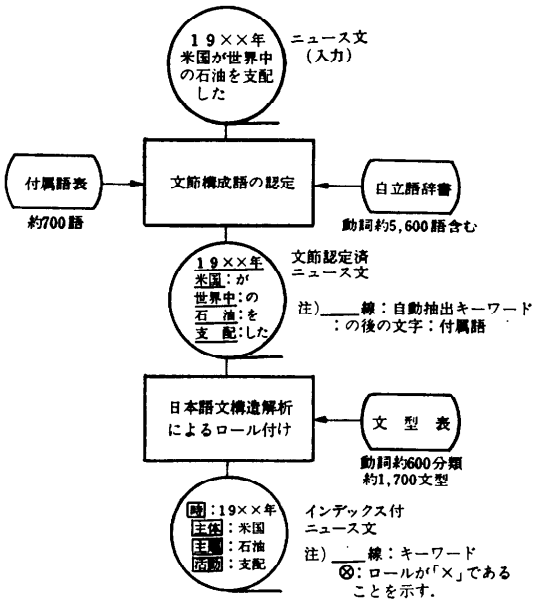

図-3 ロールによる自動索引付け
(情報処理 Vol. 20, No. 10, p. 909)

索引語抽出では，このレベルでの結果が使われ，抽出された名詞句が索引語の候補となる．最終的な索引語は，名詞句中の名詞の出現頻度によって決定される．

構文解析を行うにあたって辞書中にない語が入力文に現われると解析が不能となるが，この事態を避けるために入力文中に現われる全単語を辞書に収めようとすると辞書のサイズは急速に増大する．特に索引語は日常使われない語が多いため汎用の辞書では役立たない．**3.4** および **3.5** で述べた手法はなるべく既存の文法と既存の辞書を使って文献集合の特性に依存しないことを狙っているから，辞書が特殊な用語でふくらむことは避けたい．その点に着目したのが，細野等の試みで[16]，名詞という品詞しか持たない語は辞書に入れない，逆に言うと，辞書にない語は名詞と同等に扱うという方法で，句の抽出を行っている．

その他，索引語の特性を摑むための基礎調査が各所で行われている．特に，日本語においては表記のゆれや漢字使用制限によるまぜ書きの問題が絡むため，実際のデータに基づく用語の調査が重要になっている．姫路短大の田中のカタカナ列の解析[17]や，複合語の語基同志の関係[18]など示唆に富む発表がある．

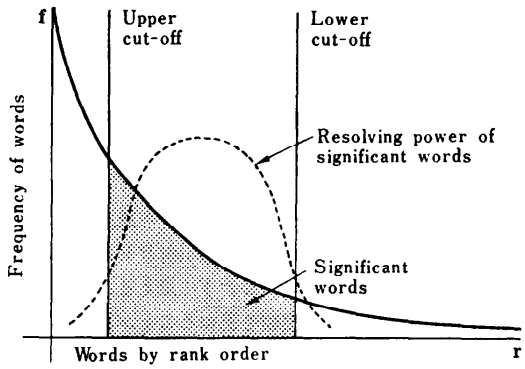

図-4　Luhn の自動索引付けモデル（文献 20) より）

4. 語の出現頻度による索引付け

語の出現頻度による手法は，索引語の出現パターンにはある固有の傾向があるという前提にたち，そのモデル化をはかるものである．その最初の着想は，Luhn によって示された[19),20)]．図-4 は一文献中に使用される語を頻度の高い順に並べた時の出現頻度曲線（実線）で，索引語としての価値は点線のようになると考えた．したがって影のついた部分をとり出せば索引語が決定できるとしている．ただし，その後の研究の主流は，文献集合全体での語の出現頻度パターンの違いに着目するようになってきている．そこでの前提は，索引語は文献集合全体では一様に出現することはないが，それによって代表される適切な部分集合の中では一様に出現する．また，不要語の場合には集合全体で一様に出現するというものである．

言語情報の解析的手法においては，ある語（複合語）が索引語であるという判定がなされた場合，その語が出現した（正確にはその語を抽出した）文献は無条件でその索引語によって得られる文献であるとした．

出現頻度を用いた統計的手法においては，文献集合全体の中での索引語の出現パターンを調べるため，索引語集合を決定することと，索引語を各文献へ付与することが別の問題となる．索引語集合決定のためのモデルは，以下で紹介するように多くの提案と実験があり，抽出された索引語集合の識別力に関して客観的指標を与える．この点では解析的手法より優れている．しかしながら，各文献への索引語付与については具体的方法がない．ここで，解析的手法で行っている語の出現＝その文献の内容を代表する索引語という発想がとれないのは，統計的手法により得られる索引語が計算量の制約から単語になってしまうという事実があり，上記の発想に対し慎重にならざるを得ないという事情による．

4.1　2 ポアソン・モデル

Bookstein と Swanson[21)], Harter[22),23)] によって提案されたモデルである．前述のようにある語が不要語ならば，一定の長さの文献ならばどこでもランダムに出現する．つまり，出現はポアソン分布に従うとみなせる．いま知りたいのは索引語の分布であるが，これも文献集合を二つ，その語を索引として使える文献と，その他に分割すれば，それぞれの部分集合での出現はポアソン分布に従うと思われる．そこで各索引語についてその出現確率は

$$f(k)=\pi\frac{e^{-\lambda_1}\lambda_1^k}{k!}+(1-\pi)\frac{e^{-\lambda_2}\lambda_2^k}{k!}$$

という 2 ポアソン分布であらわせる．$f(k)$ はある語が k 回出現する確率，λ_i は各部分集合での平均出現回数，π は索引として適切な部分文献集合の全体に対する割合である．

このモデルで，$\lambda_1 \gg \lambda_2$ となる語が索引語として望ましいが，望ましさの指標として

$$z=(\lambda_1-\lambda_2)/\sqrt{\lambda_1+\lambda_2}$$

を用いる．

なお，索引語の文献への付与について，検索洩れのコストがノイズのコストよりも大ならば付与するという一般的な示唆が述べられているが，具体的な自動付与の方法は与えられていない．

4.2　Dennis-Salton のモデル

出現頻度の分散を索引語判定の基準とするモデルである[24),25)]．ただし，各語の総出現頻度の違いによる影響を排去した式が提案されている．

$$F^k V^k/(\bar{f}^k)^2 \quad \cdots\cdots\cdots\cdots ①$$

ただし，f_i^k＝文献 i における語 k の出現頻度，n＝文献総数として，

$$F^k=\sum_{i}^{n} f_i^k$$

$$\bar{f}^k=(1/n)F^k$$

$$V^k=(\sum_{i}^{n}(f_i^k-\bar{f}^k))^2/(n-1)$$

①の値が大きい語は文献間の識別力が高いので索引語とみなせる．また，ここでは各文献の長さを考慮した式も提案されている．

4.3　ベクトル空間モデル

各文献 D_i を，索引語を軸とするベクトルで表現するモデルである[26)]．

$D_i=(w_{i1}, w_{i2}, \cdots, w_{it})$

w_{ij}は,各索引語の重みである.このモデルで,内容が類似する文献が互いに近接するように座標軸を決める.この空間での二つのベクトルの距離が,したがって類似度Sという尺度となる.そこで望ましい文献空間は

$$Q=\sum_{i}^{n} S(C_1 D_i) \qquad C: \text{重心のベクトル}$$

の密度Qが最小の時である.ここから語kを除いて求めたQをQ_kとすると,

$DV_k=Q_k-Q>0$

ならば,kは索引語として適切である.

4.4 χ^2による索引語検定

長尾等によって与えられた手法[27]で,文献集合に対しすでに分野のような区分が与えられているという状況のもとで,分野間で各語の出現頻度にかたよりがないという仮定でχ^2値を求め,χ^2値が十分大きければ,索引語として価値があるとみなす(ある分野に集中して現われるから).

ただし,本来の検定の立場から有意水準を越える語だけを選ぶと索引語となる語が非常に少なくなるので,χ^2値を単にかたよりを示す指標として値の大きい語を適当に選んでいる.

4章で述べた他の手法に比べ,これは分野との関連がある程度わかる(かたよりの大きい分野がある)ので,分野を代表する語としてシソーラスの作成にもよい情報を与える.

4.5 その他の試み

統計手法に共通する弱点として,複合語(あるいは句)の抽出ができないという問題がある.Doyle[28]は何らかの方法で得られた索引語(単語)が同一文献あるいは同一センテンス内で共出現すれば,それらの語は互いに関連があり,複合語となる可能性もあることを示した.

また,3章で示したロッキードのEarlは,両方式の折衷の形で,この問題に対処しようとしている.

解析的手法では漢字を単語として扱う試みがいくつかあったが,統計的手法においては複合語の問題のためにこうした試みはあまり役立たない.ただし,漢字出現特性を分野と結びつけた場合には自動分類の可能性があるので,筆者らはその目的で漢字の統計処理を行っている[29),30)].

5. おわりに

自動索引付けの二つの主なアプローチを述べた.実用化という点からは,解析的手法,それも伝統的な言語処理の手法ではなく,実際に収集した文献集合固有の性質を利用した手法が最も効果的である.この手法は初期投資が大きく,継続的な辞書管理も必要なため大規模システムでないと適用しづらい.今後,汎用化が望まれる手法である.他の理論的アプローチは実用段階には達していないが,大量データを扱うための工夫や改良が進めば,汎用な自動索引付けシステムが達成できる可能性を持っているように思える.今後は新たなモデルの提唱より実験を重ねることが重要となろう.

本稿を書くにあたり,細野公男教授(慶大・文)との自動索引付けに関する共同研究から多くの示唆を得た(特に文献31)から多くの資料を得た).共同研究を通じ,ご教示を頂いた細野先生,同研究室後藤智範君に謝意を表する.

参考文献

1) 細野,大河内,諸橋他:文献情報処理におけるキーワード/ディスクリプタ自動変換,IBM TSCレポート,N:G 318-1577 (1982).
2) Field, B. J.: Towards Automatic Indexing, J. Documentation, Vol. 31, No. 4, pp. 246-265 (1975).
3) Garfield, E.: Citation Indexing, John Willy & Sons, New York (1979).
4) STAIRS/VS General Information, IBM Manual, SH 12-5114.
5) 法令検索システムの概要,行政管理庁行政管理局.
6) Klingbiel, P. H.: Machine-aided Indexing of Technical Literature, Inf. Stor. Retr., Vol. 9, pp. 477-494 (1973).
7) Hunt, B. L., Synderman, M. and Payne, W.: Machine-assisted Indexing of Scientific Research Summaries, J. Am. Soc. Inf. Sci., Vol. 26, pp. 230-236 (1975).
8) 中井:機械補助索引(MAI)について[I],情報管理,Vol. 19, No. 4, pp. 247-259 (1976).
なお,No. 5に[II]が掲載されており,他のMAIシステムの紹介とDDCのシステムとの比較評価がある.
9) 高野,荒木,金子,日夏:日本語論文タイトルからのキーワード自動抽出システム(JAKAS),情報処理学会自然言語処理研究会資料 26-3 (1981).
10) 荒木,金子,高野,日夏:日本語キーワード自動抽出システム(JAKAS),第18回情報科学技術研究集会発表論文集,pp. 35-44.
11) 荒木,板山:JICSTの実用的全自動漢字一カナ変換システム,K-KACSについて,情報処理,

Vol. 20, No. 10, pp. 917-923 (1979).
12) 絹川, 橋本, 木村：日本語文構造解析による自動インデクシング方式, 情報処理学会日本語情報処理シンポジウム報告集, pp. 169-175 (1978).
13) 石綿：動詞を中心とした語彙の分類, 国立国語研究所報告51 (1974).
14) Earl, L. L.: Experiments in Automatic Extracting and Indexing, Inf. Stor. Retr., Vol. 6, pp. 273-288 (1970).
15) Earl, L. L.: The Resolution of Syntactic Ambiguity in Automatic Language Processing, Inf. Stor. Retr., Vol. 8, pp. 277-308 (1972).
16) 細野, 後藤, 諸橋他：パターン・マッチングによる重要語の自動抽出, 情報処理学会自然言語処理研究会資料 39-1 (1983).
17) 田中, 長田, 土屋：科学技術文献抄録における片仮名列の解析, 計量国語学, Vol. 14, No. 1, pp. 15-20 (1983).
18) 田中, 水谷, 吉田：語と語の関係について, 情報処理自然言語処理研究会資料 41-4 (1984).
19) Luhn, H. P.: The Automatic Creation of Literature Abstracts, IBM J. Res. Dev., Vol. 2, pp. 159-165 (1958).
20) van Rijsbergen, C. J.: Information Retrieval, Second Edition, Butter Worths, London (1979).
21) Bookstein, A. and Swanson, D. R.: Probabilistic Models for Automatic Indexing, J. Am. Soc. Inf. Sci., Vol. 25, pp. 312-318 (1974).
22) Harter, S. P.: A Probabilistic Approach to Automatic Keyword Indexing Part I, J. Am. Soc. Inf. Sci., Vol. 26, pp. 197-206 (1975).
23) Harter, S. P.: A Probabilistic Approach to Automatic Keyword Indexing Part II, J. Am. Soc. Inf. Sci., Vol. 26, pp. 280-289 (1975).
24) Dennis, S. F.: The Design and Testings of A Fully Automatic Indexing-searching System for Documents Consisting of Expository Test, Information Retrieval—A Critical Review, Thomson book, Washington D. C., pp. 67-94 (1967).
25) Salton, G.: Dynamic Information and Library Processing, Prentice-Hall, p. 82 (1975).
26) Salton, G. et al.: A Theory of Term Importance in Automatic Text Analysis, J. Am. Soc. Inf. Retr., Vol. 26, pp. 230-236 (1975).
27) 長尾, 水谷, 池田：日本語文献における重要語の自動抽出, 情報処理, Vol. 17, No. 2 (1976).
28) Doyle, L. B.: Indexing and Abstracting by Association, Am. Doc., Vol. 18, pp. 378-390 (1962).
29) 細野他：電気工学分野における重要漢字の調査, 三田図書館・情報学会研究大会, pp. 19-22 (1983).
30) 梅田, 細野, 諸橋他：漢字出現頻度に基づいた日本語文献の定量的分析, 情報処理学会第28回全国大会論文集, pp. 1209-1210 (1984).
31) 細野, 諸橋, 大河内他：文献情報を対象とした自動索引の実験, IBM TSC レポート, N：G, pp. 318-1541 (1981).

(昭和59年6月11日受付)