文献データベースにおけるキーワード分析の適用

北見工業大学　正員　中岡良司°
北見工業大学　正員　森　弘
北海道大学工学部　正員　五十嵐日出夫
北海道大学工学部　正員　佐藤馨一

1．はじめに

大型コンピュータのネットワーク化やパーソナルコンピュータの普及により、土木技術者や研究者がデータベースを作成したり利用する機会が増えてきている。ところで、統計データベースに代表される数値型データベースにおいては、その利用形態は多種多様であるが、文献データベースに代表される文字型データベースに関しては、多くの場合、蓄積されたデータを検索する利用にとどまっており、その活用方法が期待されるところである。

一方、人文・社会科学の領域においては、言葉の頻度や相互関係からある文章の著者を特定したり、情報の伝達手段に関する深層構造を解明しようとする研究が、内容分析（Content Analysis）という名のもとに古くから行われてきた[1)]。

そこで、本研究は、文献データベースに収められた各論文のキーワードに着目し、その使用頻度や複合構造を分析することによって、データベース全体の研究動向を調べようとするものである。

キーワードとはその論文を最も特徴づける用語である。キーワードは論文著者自身が設定するのが本来であるが、設定レベルのばらつきや記載の無い論文があるなど統一性に欠けている。また、一般に、わが国の学術研究論文においては表題そのものがキーワードの組み合わせとなっている場合が多い。そこで、本研究では、論文の表題からある種の規則に従ってキーワードを自動的に抽出するプログラムを開発した。本研究の方法によれば、論文の表題という最小限のデータ入力によって、データベース全体のキーワードの使用傾向を分析することが可能である。

なお、分析の対象とした文献データベースには、著者らの専門分野である土木計画学に関する論文約1千件を収録した。パーソナルコンピュータを利用したデータベースの構築およびその応用に関しては、既に他の機会で発表してきたであり[2)-6)]。今回のデータベースの構築にあたっては、いかに入力が容易で簡易なデータベースを構築するかという点に主眼をおいた。

ひとつの学問分野の関心の変化を知ることは、研究者の研究課題を全体の中で適切に位置づけるのに役立つばかりでなく、若手研究者においては新たな研究テーマの開拓にも有用であろう。

2．文献データベース

（1）データベースの構築

本研究で分析の対象とした文献データベースは、1979年1月（第1回）から1989年11月（第12回）までに土木学会土木計画学研究発表会で発表された全論文（講演集および論文集）である。論文数は1088件であった。本研究発表会論文を選択した理由は、当該分野が著者らに身近であるというばかりでなく、その領域の広さから、他の分野の研究者にとっては研究対象が分かりにくいものとなっており、本研究はその対象を明確にできるものと考えられたからである。約千件のデータ規模は文献データベースとしては決して多いものではないが、手作業による論文検索と較べれば、コンピュータ利用の効果が現れる十分な規模でもある。

多くの研究者がデータベースの有効性は認めながらも、その入力の負担の大きさによって構築を断念する場合が多い。そこで、本研究ではデータベースの設計にあたって、入力の負担を最小限に抑えた項目の設定とデータ入力法を検討した。一般的に、文献データベースの項目と考えられるのは、出典、表題、著者（ときには所属も含めて）、概要、キーワードなどであり、さらに、概要を目的、方法、結果に分けて入力する場合もある。しかしながら、研究

者が文献調査をする場合、その論文内容をデータベースの段階で理解することはまれであり、当該論文そのものを通読し自己の研究との係わりを判断したり、参考文献から再び文献調査を続けていくのが通常である。その場合、最低限必要なのは論文表題と出典である。そこで、本研究では、データベースの項目を表題と出典関係項目に限定した。データ構造は以下の通りである。

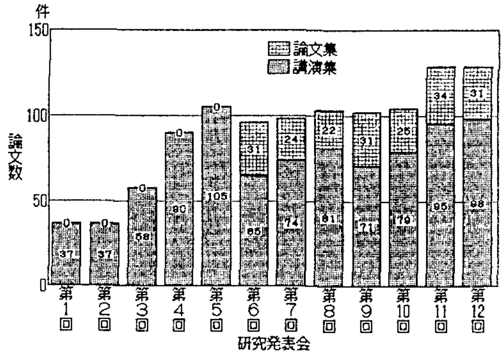
① 発表年 ： 研究発表会回数
② 論文集名 ： 講演集、論文集
③ 論文番号 ： 各号の論文の連番
④ 表　　題 ： 和文（英文は和訳）

データ入力は日本語ワープロソフトで行い、後にデータベースソフトへ変換した。これは、文字の入力に対しては、ワープロソフトが最も操作性に優れているからである。作業者はワープロで文章を作成するのと同様の感覚でデータを入力することができ、共同作業も容易である。以上のデータベース項目および入力方法によって、データベース構築の労力は大幅に低減されたと考えられる。

本研究においては、日本語ワープロソフトには「一太郎」（(株)ジャストシステム）、データベースソフトにはリレーショナルデータベース「桐」（(株)管理工学研究所）、キーワード分析など本研究特有の処理には「ＢＡＳＩＣ／９８」（(株)神津システム研究所）でプログラムを作成した。

（2）データベースの検索

構築したデータベースから１回から第12回までの土木計画学研究論文数の推移を示せば図－１の通りであり、年々増加の一途をたどっている様子がうかがえる。第６回からは審査付論文と自由投稿論文に分けて編集されているが、本研究では特に区分して取り扱っていない。



図－１　『土木計画学』論文数の推移

さて、文献データベースの一般的な利用形態はデータの検索である。大型コンピュータを利用した汎用データベースにおいては、あらかじめ決められたキーワード（検索語）を通じて該当ケースを検索する形態が多いが、キーワードが限定されるなど利用上の制約が大きい。一方、パーソナルコンピュータ用のデータベースソフトの多くはより自由な検索を可能としており実用性は高いが、条件設定数に制約も多い。そこで、本研究では、指示した用語を一部でも含む論文を検索するフリーワード検索プログラムを作成した。本プログラムは本研究の内容に直接係わるものではないが、一例を図－２に示す。

```
Free Word ? = 交通
 Applicable Case =  284         Display or Printer or Continue = C
Free Word ? = 非集計
 Applicable Case =   10         Display or Printer or Continue = P

02,01,025,非集計ロジットモデルによる交通手段選択の分析　―多手段同時選択の場合―
04,01,057,通勤交通手段の予測における集計モデルと非集計モデルの予測精度の比較
05,01,067,選択構造を考慮した非集計交通需要予測モデルについて
05,01,068,非集計モデルを用いた都市内旅客交通需要の推計方法について
05,01,069,非日常的交通行動への非集計モデルの適用　―チョイスベイストサンプルに対
          する推定問題の検討―
06,02,002,意識データに基づく非集計交通手段転換モデルの構築の試み
06,02,005,非集計行動モデルによるＯＤ交通量推計方法
09,01,001,非集計行動モデルを用いた総合交通需要予測システムの開発研究
11,01,067,非集計行動モデルに基づく都市交通需要予測体系の構築の試み
11,02,029,公共輸送機関の経路探索が非集計交通機関選択モデルに及ぼす影響
 ¦  ¦  ¦
 ¦  ¦  └-- 連 番
 ¦  └-- 論文種別
 └-- 発表会回数
```

図－２　フリーワード検索例

図－２は、データベースから"交通"という用語を含む論文を検索した結果である。1088件中 284件が該当することが分かった。そこで、さらに"非集計"という用語を含む論文を検索し10件へ絞り込んだ様子を示している。ここでは、アンダーラインによって、表題の任意の場所に指示した用語が該当している様子がわかる。検索条件としては、指示した用語を含まないという排他的条件の設定も可能である。また、条件数の制約は特に無く、所用時間は約30秒程度である。

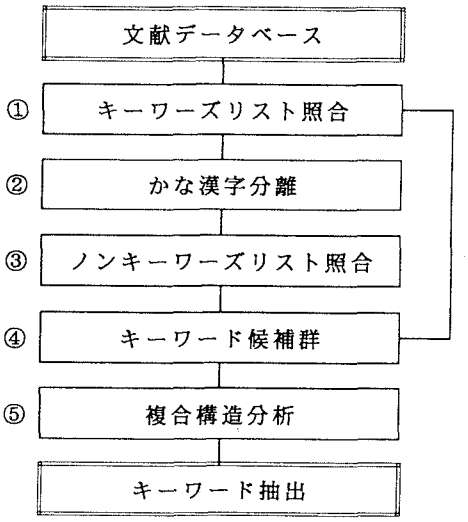
３．キーワード分析

（１）キーワードの自動抽出

キーワード（Key word）とは、本来、文章などの意味を解くための鍵となる言葉を指すが、学術用語としては、その論文を最も特徴づける用語といってよい。論文の表題は、その研究の目的や方法を最も簡潔に記した文章であるから、必然的にキーワードを中心に構成されている。

一般的には、言葉の意味を無視して、任意の文章からキーワードを自動抽出することは不可能であるが、学術研究論文の表題という狭い領域に限るならば、我々は、表題におけるキーワードの切り出しを比較的簡単な手順で行うことができる。当初、キーワードという性格から語の品詞による抽出も検討したが、実用的な基準を設けることはできなかった。

本研究で開発したキーワード自動抽出法のフローを図－３に示す。



図－３　キーワードの自動抽出フロー

表－１　キーワーズ・ノンキーワーズリスト

[Keywords List]

1. 繰り返し	2. ふ頭			

[Non Keywords List]

1. 一	2. 的	3. 法	4. 論	5. 研究
6. 考察	7. 分析	8. 考慮	9. 一考	10. 評価
11. 適用	12. 影響	13. 開発	14. 方法	15. 検討
16. 事例	17. 手法	18. 推定	19. 比較	20. 利用
21. 予測	22. 調査	23. 推計	24. 課題	25. 対象
26. 特性	27. 要因	28. 分類	29. 構成	30. 把握
31. 計測	32. 立地	33. 着目	34. 形成	35. 論評
36. 起因	37. 負担	38. 反映	39. 分割	40. 決定
41. 基礎的	42. 利用者	43. 問題点		

第①段階は、文献データベースに対して、以下の規則で抽出できないキーワードを、事前に設定したキーワーズリストと照合し抽出するものである。例外を排除するため複雑な規則を設けるよりも、例外は事前に設定しておく方が処理は容易である。現時点では、キーワーズリストはわずかに２語を数えるのみである（表－１）。第②段階の「かな漢字分離」は、次の単純な原則に従ったものである。

・ひらがなはキーワードにならない

・漢字１字はキーワードにならない

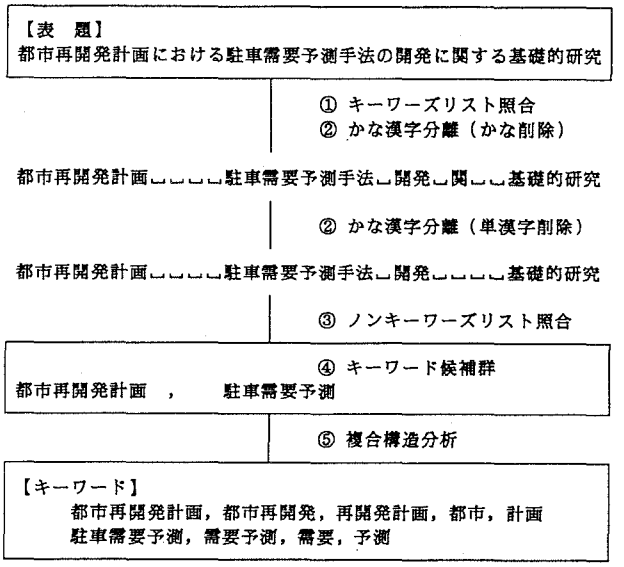
そこで、文字コードを利用して表題を「かな文字列」と「漢字文字列」に分離し、ひらがな部分および漢字１字を削除する。第③段階では、残った文字列に対して、「研究」、「把握」など一般的用語でありキーワードとは認められない文字列をノンキーワードリストと照合し排除し、残った文字列を第④段階のキーワード候補群として抽出する。

以上の過程は、プログラム上で学習機能を持たせ、約１割のデータを処理するなかで実用化していった。その結果得られたキーワーズリスト、ノンキーワーズリストは表－１に示すとおりである。

さて、第④段階で基本的にキーワード群が得られるが、この段階で得られたキーワードによってデータベース全体の中での使用頻度を算出しても有意義な結果は得られない。その理由は、限られた文字数（一般に約40語以内と言われている）の範囲でその論文の特徴を要約している表題においては、キーワードが複合的に使用されているからである。そこで、

第⑤段階として、第④段階で得られたキーワード群すべてを相互に比較することによって、キーワード間の包含関係を分析し有効なキーワードを抽出することにした。

以上の過程を図－４の実例で示そう。与えられた表題に対して、第①段階のキーワーズリストとの照合は該当無しである。第②段階によって、ひらがなの「における」、「の」、「に」、「する」が削除され、漢字１字の「関」も削除される。第③段階によって、ノンキーワーズリストに該当する「開発」、「基礎的」、「研究」が削除されることになる。その結果得られる第④段階のキーワード候補群の「都市再開発計画」および「駐車需要予測」は既にキーワードではあるが、データベース全体のキーワード候補群と照合することによって、その複合構造がより分解された形で最終的なキーワードを得ることができる。ここで、単純に用語の組み合わせであれば、「駐車需要予測」からは「駐車需要」という用語も想定されるが、データベース全体の中でそのような使用例が無かったため出現していない。すなわち、複合構造を分解するといっても実際の使用例に即して分解しているのである。



図－４　キーワード抽出の例

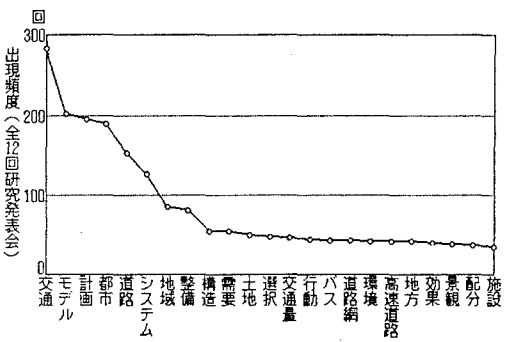
（２）主要キーワード

表－２　キーワードの出現（利用）頻度

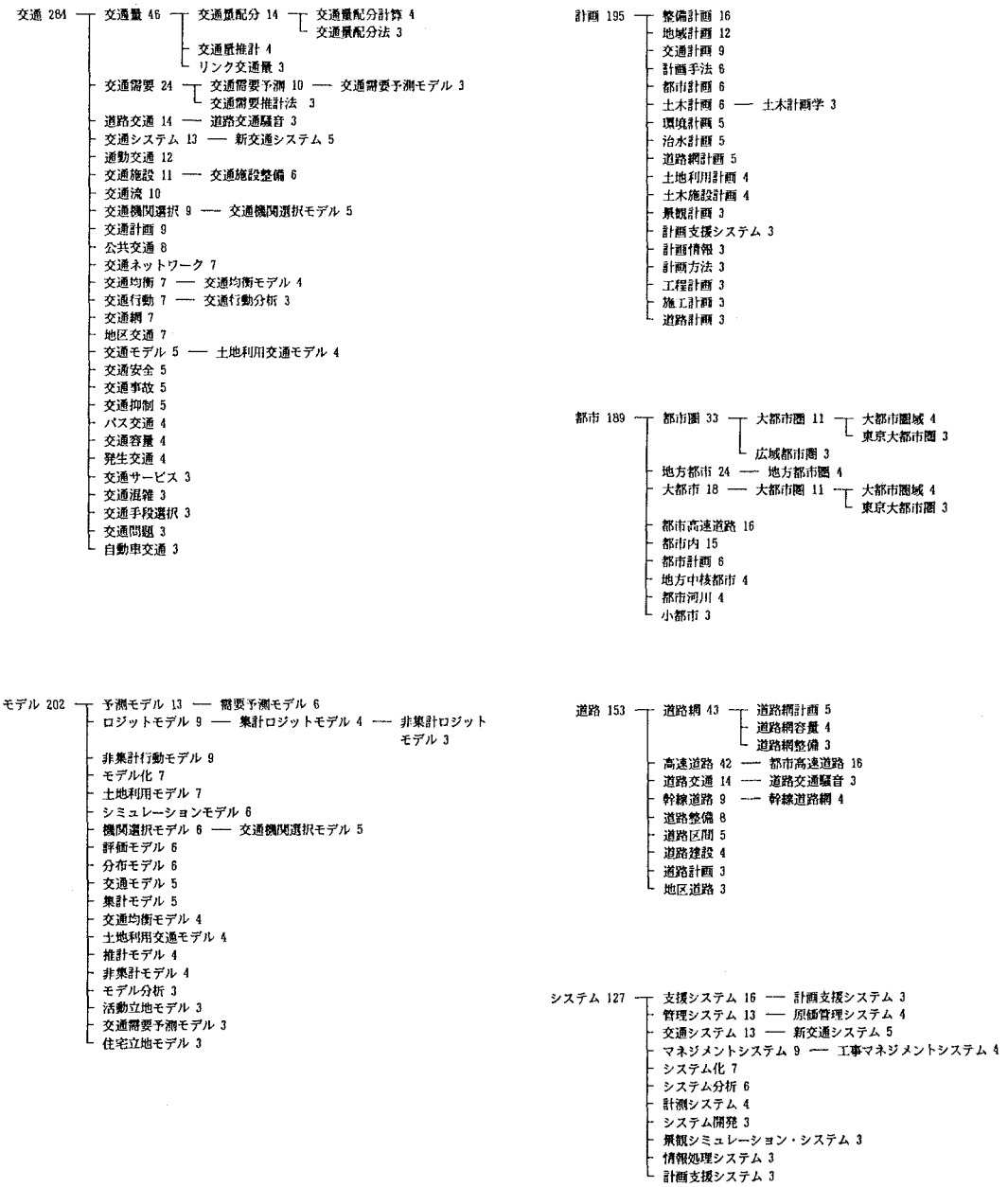
数字は出現頻度（回）

・交通	284	・土地利用	34	・理論	27
・モデル	202	・都市圏	33	・ﾈｯﾄﾜｰｸ	26
・計画	195	・意識	32	・情報	25
・都市	189	・住民	31	・地方都市	24
・道路	153	・高速道路	42	・交通需要	24
・システム	127	・地方	42	・解析	23
・地域	85	・効果	39	・活動	23
・整備	81	・景観	38	・港湾	23
・構造	55	・配分	37	・分布	22
・需要	54	・施設	35	・規模	21
・土地	50	・空間	34	・便益	21
・選択	48	・鉄道	34	・河川	20
・交通量	46	・街路	30	・建設	20
・行動	44	・管理	30	・市街地	20
・バス	43	・住宅	30	・人口	20
・道路網	43	・通勤	30	・ｼﾐｭﾚｰｼｮﾝ	19
・環境	42	・変動	27	（以上、上位50語）	

抽出したキーワード群のうち、上位50語をその出現頻度とともに表－２に示す。図－５は、その一部を図化したものである。図表からも明らかなように、出現頻度 100回以上のキーワード（ほぼ出現頻度は論文数に対応しているが、同一論文中に２回以上使用している場合も有り得るので回と表現している）としては「交通」、「モデル」、「計画」、「道路」、「システム」が特徴的であり、その他のキーワードの出現頻度は相対的に少ない。最も使用率の高い「交通」に関しては1088件中の約３割の論文で使用されており、土木計画学の中心的な課題であることが明確となっている。

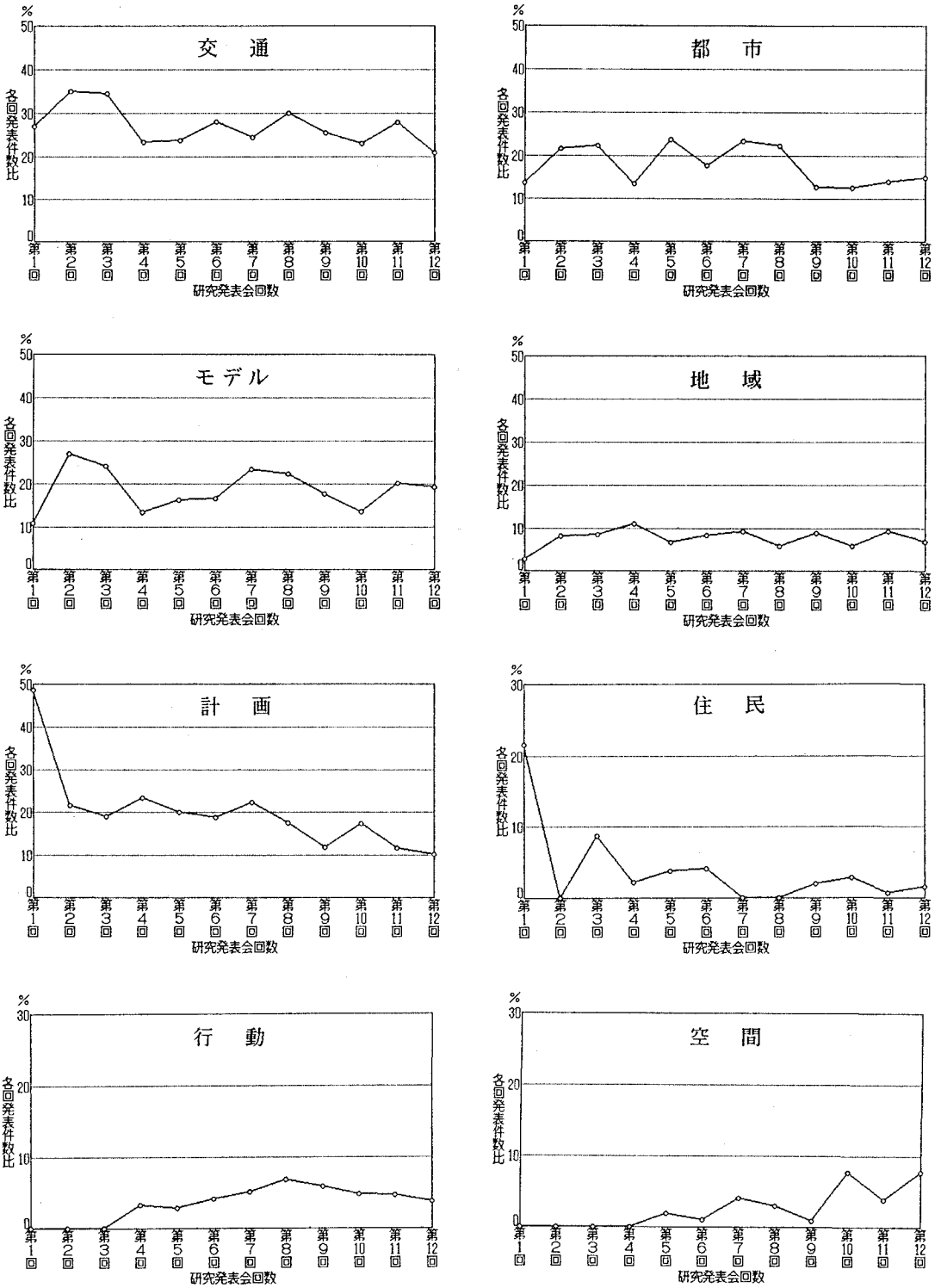


図－５　キーワード出現頻度の分布



※　キーワード末尾の数字は出現頻度

図－6　キーワードの複合構造



図－7　キーワードの利用率の推移

（3）キーワードの複合構造

前項で求めた高頻度のキーワードは、キーワード要素とでも言うべき最小単位でのキーワードであるため、抽象的かつ一般的な用語となりがちである。そこで、より複合化したキーワードとの関係をツリー構造で示したのが図－6である。各キーワード末尾の数字は出現頻度である。ここでは、例えば、「交通」というキーワードが「交通量」、「交通需要」という形で展開し、さらに「交通量」は「交通量配分」、「交通量推計」、「リンク交通量」と次第に具体化していく様子が見られる。仮に、前述の「交通」を扱った論文が最も多いという結果があまりに一般的と考えられる場合は、その中でも「交通量」を対象とした論文が多いという理解をすることが可能であるし、「モデル」には「予測モデル」に関した論文が最も多いということも分かる。

（4）キーワードの利用率

以上の分析は、第1回から第12回までの全論文を一括した結果であるが、ここでは、各回ごとにキーワードの利用率の推移を見てみることにしよう。ここで、利用率とは各回の発表論文数に対する各キーワード使用論文の割合を算出したものである。すなわち、各キーワードを使用した表題論文がその年の論文数の何割あったかを示している。

図－7にその結果を示す。主な特徴を列挙するならば、「交通」は毎年平均して3割程度を占めている。「モデル」、「都市」は増減が激しく変動している。「計画」、「住民」は減少傾向を示している。「行動」、「空間」はともに利用率は低いが増加傾向にある。とりわけ、「空間」という用語は第4回までの論文には使われていなかったという経緯も知ることができる。これらの変化をその背景にある社会状況と対応させることによって、さらに特徴的な経緯が明らかになると考えられるが、現時点の論文数からはその関係を見いだすことはできなかった。

4．まとめ

以上、本研究では構築した文献データベースからキーワードを自動抽出し、その構造と変化を分析してきた。以下に本研究の特徴と成果をまとめる。

① 構築した文献データベースには土木計画学研究発表会の第1回論文から最近までの全論文1088件を収録した。内容は各号を示すデータと論文表題のみであるが、研究者が関係論文を検索するには十分であり、また入力作業は最小限で済む。

② 文献データベースの利用方法としてはフリーワード検索を示した。この検索法は、あらかじめ用意された用語からではなく、自由に連想した用語から論文を検索できる利点がある。

③ キーワード自動抽出法はノンキーワードリストを必要とするなど不完全なものではあるが、比較的簡単な処理によって実用的な方法論になっている。これは、例外的な場合の頻度が少ないため集計段階では表面化しないことによる。

④ 抽出したキーワード群は、「交通」や「モデル」に代表されるように、結果として汎用的な用語が上位を占めた。土木計画学の研究者としては、予想される結果が数的な裏付けとともに得られたことを一応の成果としたい。

⑤ キーワードの複合語構造は本研究によって初めて明確化された。本研究の段階ではキーワードの意味論的処理をいっさい行っていないが、今後、関連するキーワードを体系化してゆくことによって、研究分野の構成を一層明らかにできるものと考えられる。

⑥ 主要キーワードの使用率の推移を実用的に見るには、本研究で作成した論文数では不足するようである。その制約はあっても、土木計画学研究の中にあっては減少する研究分野や新たなキーワードの出現の様子が確認できた。

5．おわりに

以上、本研究は学術研究論文を内容とする文献データベースに対して、また、さらに土木計画学という限られた研究領域に対象を絞って、データベース全体の研究動向を探ってきた。他分野への本研究手法の適用に関しては、キーワーズリストやノンキーワーズリストの整備など注意を要する部分もあるが、基本的な方法論は有用と思われる。

今回使用したデータベースの規模（論文数）は、必ずしも本研究の有効性を実証するには十分ではない。今後は、データベースの一層の蓄積を図る所存である。さいわい、平成3年度から土木学会年次学術講演会講演概要集の内容は学術情報センターデー

タベースに登録する運びとなった。本研究のデータベースとリンクさせていくことによって、より有意義な活用を図っていきたい。

最後に、本研究で使用した主なハードウェアは、パーソナルコンピュータＰＣ－９８０１ＲＡ５１（(株)ＮＥＣ）に増設メモリ８ＭＢを付加したシステムである。

参考文献

1) クラウス･クリッペンドルフ、「メッセージ分析の技法－内容分析への招待－」、勁草書房、1989.8

2) 中岡良司、リレーショナルデータベースによる土木史年表の作成と応用、土木史研究・論文集、No.10、1990.6

3) 五十嵐日出夫・中岡良司、「土木工学ハンドブック（土木史年表）」、技法堂出版、1989.11

4) 中岡良司・森　弘・五十嵐日出夫・佐藤馨一、リレーショナルデータベースによる史的情報管理システムの構築と運用、土木学会第12回電算機利用に関するシンポジウム論文集、1987.10

5) 中岡良司・森　弘・佐藤馨一・五十嵐日出夫、土木史研究データベースの作成と今後の土木史研究について、第７回日本土木史研究発表会論文集、1987.6

6) 中岡良司・森　弘・五十嵐日出夫、リレーショナルデータベースによる非計量データ処理について、土木計画学研究・講演集、第８号、1986.1