調査研究実績からみた行政課題の抽出方法に関する研究*

-行政文書を素材とするテキストマイニングアプローチ-

Research about the Method Extracting the Administrative Subject from the Surveillance Study *

Text-Mining Approach with Administrative Document

松本浩和**

By Hirokazu MATSUMOTO**

１．はじめに

大都市では多くの先駆的な調査研究が行われ、行政機関には多量の文書データが蓄積されている。それら調査研究の目的や内容は、その時期における行政ニーズを表すものと考えられ、それらの情報は将来のニーズへの対応や新施策のための課題探しとして有用であると考えられる。

また、文書データに対して、保管コスト・人件費の削減や情報の共有、劣化防止、データの再利用などを目的とした電子化が各方面で進められており、データベースの作成が行われている。行政機関においても同様の動きが見られ、近い将来、調査研究の報告書だけでなく、既存の紙ベースの資料のほとんどは電子化されることになると予想される。これにより多くの電子化されたテキストデータが存在することになる。

しかしながら、上記のような目的を持ったデータベース作成では、ユーザーの使いやすさや検索方法に関する議論はなされるものの、それらデータの有効活用に関して論じられることは少ない。こうした過去のデータは知的財産として再活用できる可能性があり、データベース化が進められる現在において、しっかりと議論する必要があろう。

一方、コンピュータやインターネットの普及にともない、ウェブページをはじめとしたテキストデータも急増している。こうした中、テキストデータのより高度な活用を実現するために、自然言語処理とデータマイニング技法を結合したテキストマイニングと呼ばれる技術が注目されつつある。

そこで、本研究ではテキストマイニングの手法を導入した行政ニーズの抽出を目的として、大都市における道路担当部局での調査研究事例を収集・整理し、各報告書に記載されている特定の単語の出現頻度やそれらの相関を見ることにより、調査研究が行われた当時の時代背景やニーズを抽出する手法について検討を行った。

*キーワード：財源・制度論，情報処理

**学生員、大阪市立大学大学院工学研究科

（大阪府大阪市住吉区杉本 3-3-138

TEL 06-6605-2731、FAX 06-6605-3077）

２．コーパスの作成

（１）文書データの収集

本研究はケーススタディではあるが、その適用範囲を狭めることがないように注意する必要がある。そこで、取り扱う研究対象は、継続的に行われてきており、また全国的にも広く行われているものでなければならない。

こうした理由から、本研究では道路関連の調査研究事例を取り上げ、昭和46年度から平成16年度までの２０冊の調査研究報告書を対象とした。

（２）テキストデータの電子化

a）原稿の取り込みと認識

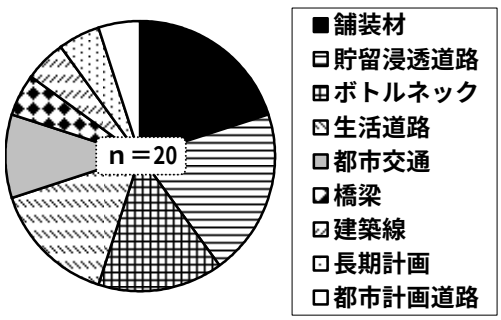
各報告書より、全ページをスキャンしてパソコンに取り込んだ後、ＯＣＲソフトを用いてテキストデータを出力した。なお、図や表、写真等に関してはテキストデータの抽出が困難であり、またその量も少なく影響が少ないと考え、本研究においては対象から除外した。

b）形式の公正

抽出したテキストデータはテキストエディタ上で整理を行った。これは、ＯＣＲソフトで出力された形では改行やスペースが入り組んでおり、適切なキーワードの抽出が困難なためである。また、誤認識に関しては目視によるチェックを行い、適宜修正を行った。

c）コーパス概要

作成したコーパスの総文字数は５８７５５９文字である。図－１に作成したコーパスの基となった報告書の分野の内訳を示す。



図－１　コーパスの内容内訳

３．単語の出現頻度からみた行政課題の抽出

（1）考え方

時代背景やニーズを表すような単語（以下キーワード）の出現頻度は、その時代ごとに異なっており、それらキーワードの出現頻度の変化をみることにより行政の取り組みや課題を抽出できる可能性がある。そこで、本研究ではケーススタディを行うことにより、その方法について検討を行った。

（2）形態素解析

キーワードの出現頻度を調べるには、テキストデータを品詞ごとに分解する必要がある。そこで、形態素解析ソフトを用いてコーパスのテキストデータを品詞分解し、各単語ごとにそれらの出現頻度を調べた。なお、キーワードを抽出する上でその品詞は意味を持たなければならないため、連体詞や助詞や助動詞などの非自立語は対象から除外し、一般名詞と固有名詞のみについて検討を行った。

（3）同系列の調査研究における変化

同系列の調査研究において、各キーワードの出現頻度がどの様に変化しているかを調べた。本研究で取り扱った報告書の中からボトルネックに関する調査研究報告書（平成二年度～四年度）の三冊を対象とし、キーワードの変化からその内容の変化や違いについて推測した。

なお、出現頻度は『（報告書Ａのキーワード Ｘの出現回数÷報告書Ａの全キーワード数）×１００』で求めたものを使用した。これは、報告書ごとに文書の量が異なるため、分量による影響を受けることが懸念されるため、それを取り除く必要があるからである。

代表的なキーワードの年度ごとの値を表―１に示す。

表―１　年度別キーワード出現頻度

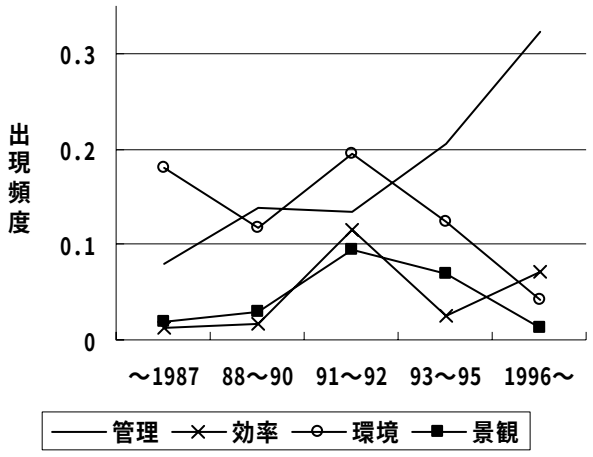
キーワード	二年度	三年度	四年度
施設	1.904	0.058	-
用地	1.114	-	0.074
構造	0.143	0.993	0.405
交差点	0.971	1.110	0.773
交通	0.978	5.257	3.054
ボトルネック	0.135	2.629	1.840
ネットワーク	0.060	0.935	1.582
鉄道	0.226	0.234	0.405
踏切	0.038	0.409	0.589
渋滞	-	2.804	2.281
原因	-	1.110	0.662
分析	-	0.643	0.258

「交差点」や「交通」といったキーワードは全ての年度に満遍なく出現しているのに対し、「鉄道」や「踏切」では増加傾向がみられ、踏切周辺部が課題となりつつあったことがわかる。また、初期においてはあまり着目されていなかったネットワークについて、対策の必要性が認識されたといえる。一方、「施設」や「用地」といったキーワードは初期の調査でしか取り扱われておらず、他の報告書と異なることが明らかである。

（4）各キーワードの経年変化

キーワードは同系列内の調査研究の傾向だけでなく、社会全体の傾向も示すと考えられる。そこで、特定のキーワードを設定し、時系列的にその出現頻度がどの様になっているかをみることにより、行政の主題の流れを見ることを目的とする。

報告書の発行年度を適当な単位で区切り、分類ごとに一冊あたりのキーワード出現頻度をもとめた。図―２はそれを時系列的に並べたものである。



図－２　キーワードの経年変化

これを見ると、「管理」のキーワードが９３年以降増加しており、維持管理が行政の主な課題となっていることが推測できる。また９０年を境に、それ以前はほとんど出現しなかった「効率」が現れていることから、９０年以降行政に効率化が求められているということも推測できる。逆に、「環境」は減少傾向にあり、行政の調査研究としてはあまり行われなくなりつつあることが予想される。

４．まとめ

本研究では、調査研究報告書のテキストデータからキーワードを抽出することにより、調査研究と行政全体におけるおおまかな傾向が抽出できた。

今後は、道路関連以外の事例にも範囲を広げ分析を行い、その妥当性を検討する。また、得られた結果から具体的な施策を導くための方法についても検討する。