

IV-85

キーワード索引による土木用語の類型化に関する研究

北海道大学 学生員 細 川 寛
 北海道大学 正 員 高 野 伸 栄
 北海道大学 正 員 佐 藤 馨 一
 北海道大学 正 員 五十嵐日出夫

1. 本研究の背景・目的

土木用語の定義や関係は非常にわかりにくい。この理由の一つとして、土木技術の一部が経験的なものであるということがあげられる。そのために用語を厳密に定義できない。よって、工学の技術用語といえども、曖昧さがあって仕方がないといえる。しかし、土木用語のわかりにくさは、土木の世界にマイナスのイメージをもたらす。土木関係者がこれからの時代に「開かれた土木」を目指すとき、土木用語の特徴をしっかりと把握することが必要であろう。

そこで本研究は、土木用語のキーワードを抽出することによって、土木における専門分野ごとの用語の使われ方にはどのような特徴があるのかを分析し、その傾向の類型を把握することを目的とする。

2. 既存の構造化手法の問題点

(1) シソーラス

シソーラス(thesaurus)は最近の情報検索の分野においては最も広く用いられている索引言語であり、言葉とその意味の対応関係を明確にし、概念間の関係を体系化しようとする。しかし、階層関係を重視しているため連想関係でつながらない部分が出てきたり、上位概念の拘束により概念を越えた関係をひとまとまりに扱えないことがある。

(2) リレーションチャート

森若による方法で、連想関係を重視し、言葉を理解するためのものである。特徴点としては、用語の関係が「ある／ない」を文章が成立「する／しない」で判断すること、ISM手法を用いて構造決定をしていることが挙げられる。しかし、扱う領域を広げすぎると逆にわかりにくくなる、検索能力がないなどの問題がある。

3. キーワード索引

(1) キーワード索引

対象とする文献のタイトルからつくられる索引を指す。執筆者はタイトルによってできるだけ主題内容を明らかにしようと意図することが多いので、タイトルは検索上の重要な手がかりとなるはずである。このような仮定にもとづいて、タイトル中のキーワード(Keyword)を使って索引をつくることが考えられる。ここでいうキーワードとは、文献の内容を表現している単語であって、階層関係その他の関係をもたない自然語である。このようなキーワードを使って作成した索引のことをタイトルからのキーワード索引(Keyword-From-Title-Index)という。

計画 支援システム
 計画 人口
 道路 計画
 土地利用 計画
 特別都市 計画 法
 都市基本 計画
 都市 計画
 都市 計画 基礎調査
 都市 計画 公園

図-1 キーワード索引の例

(2) 本研究におけるキーワード

本研究における「キーワード」とは、土木用語の「その分野において意味を持ちうる区切り語で、しかも出現頻度が分野ごとに定める基準より多いもの」を言う。よってキーワード索引における「キーワード」とは若干の意味の差がある。

4. キーワードの抽出方法

(1) 用語の分割

本研究では、採用する用語はあらかじめ各分野の専門家によってその分野において意味を持つ最小単位に区切られる。これにより、一つの用語にその分野で意味のある言葉が少なくとも一つは存在し、どの語もキーワードによって拾われる可能性をもつことになる。

(2) コンピューターによる用語の検索

コンピューターに区切った用語のデータをインプットすると、アウトプットとして①「区切り語の数」、②「区切り語の出現頻度」、③ある出現頻度以上の区切り語をキーワードとした場合、それらの語で全体の用語の何%を網羅しているかを表す「網羅率」の各数値が得られる。

(3) キーワードの決定

キーワードを選ぶ際に、数の基準を設けるため、土木の7分野の用語についてコンピューターによる検索を行った。各出現頻度までの累計区切り語の割合は図-2の通りである。

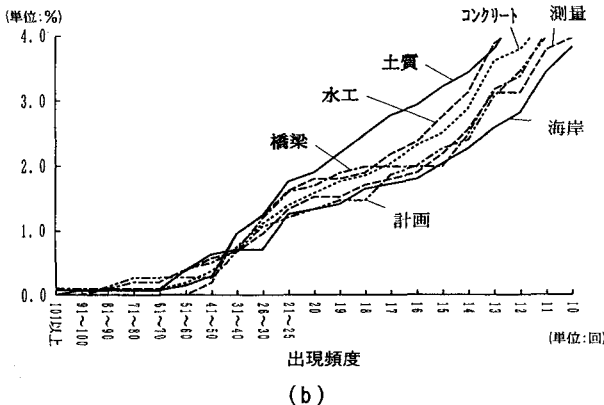
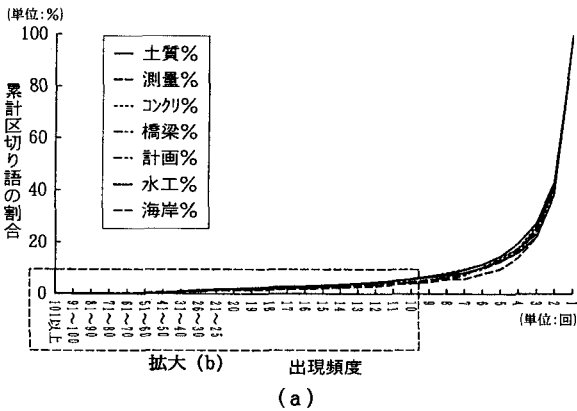


図-2 各出現頻度までの累計区切り語の割合

出現頻度だけを見ると分野ごとに大きな違いはない(図-2(a))。そこで、分野ごとの網羅率を調べてみた(図-3)。ここでは土木の基礎的な学問分野「水工学」「土質力学」「橋梁・鋼構造」「コンクリート工学」の用語が高い網羅率を示した。この結果から、

キーワードの抽出には、網羅率を用い、網羅率60～70%を得る出現頻度の区切り語をキーワードとする。

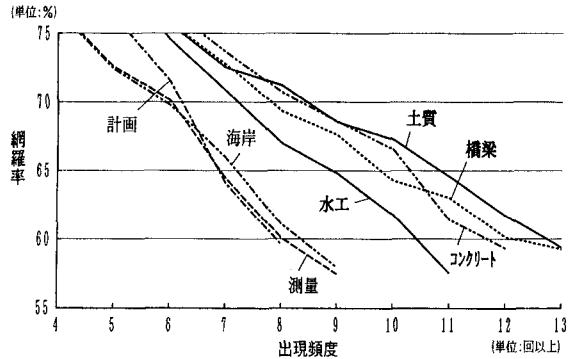


図-3 7分野の各出現頻度における網羅率

5. キーワードによる類型化

出現頻度の高い区切り語が多い分野は「水工学」「土質力学」「橋梁・鋼構造」「コンクリート工学」である。つまり網羅率の高かった分野では、その分野で意味を持ちキーワードとなり得る重要な語が高い出現頻度で存在している。また、網羅率が高い分野の用語は出現頻度の高い区切り語の重複が少なく、ひとつの用語にひとつのキーワードが存在するという洗練された用語が多いと言える。逆に「都市計画」「海岸」のように網羅率が低い分野の用語は、キーワードとなり得るような区切り語が少ないということになる。

6. おわりに

本研究ではキーワード抽出のアルゴリズムを構築し、網羅率という基準によりキーワードを求めて用語の分析を行った。その結果土木用語は各分野ごとに特徴があることが分かった。専門的な言葉が高い頻度で出現して、それにより多くの用語を網羅しているようなキーワード性の高い分野があれば、一般用語を合成したような用語が多く、キーワードとなり得る語が少ない分野もある。用語にその学問の成熟度や歴史が反映されていると考えられる。

参考文献

- 森若峰存 『土木用語の構造化に関する基礎的研究』 北海道大学大学院修士論文 (1993)