テキストマイニングによる地質・施工文献の定量解析とデータベース化

鹿島建設(株)　正会員　○松川剛一　戸邉勇人　升元一彦

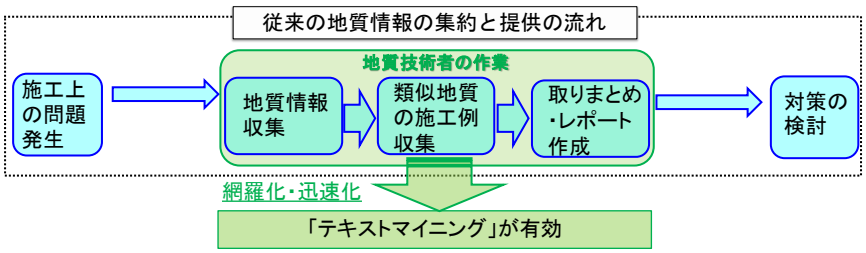
1．はじめに

土木構造物は周辺の地質に基づいて設計・施工されているため，その合理的かつ安全な施工のためには，施工予定地周辺の地質や施工に関する情報を網羅的かつ迅速に集約することが重要である．一方，日本の地質は複雑であるため，わずかな位置の違いにより地質性状に大きな差異が生じやすい．そのため，設計時に想定されていた地質と実際の施工時に確認される地質との間にはしばしば相違が生じ，この相違は施工に大きな影響を及ぼすことがある．したがって，地質とそれに起因する施工上の問題を関連付け，その情報を必要に応じて迅速に提供することが必要であり，このことが地質技術者の重要な役割の一つである．しかしながら，我が国の地質技術者の人的資源は不足しており，これを補うための仕組みが必要である[1)]．我々はその一つの解決方法として，施工に関連する地質情報を自動的に集約するためのシステムを開発している．今回その一部として，施工文献からテキストマイニングによってキーワードを自動抽出し，そのキーワードから地質文献を逆引きすることが可能なデータベースシステムのプロトタイプを開発することができた．本稿では，その成果について報告する．

2．地質文献解析の自動化

本開発では，少ない人的資源で地質・施工文献解析とその情報の集約を行うため，文献解析のフローでボトルネックとなる箇所を自動化した．地質技術者は施工上の問題が発生したとき，まず文献や施工記録をもとに地質情報を収集し，ここから類似した地質における施工例を集め，その結果を取りまとめた上で対策の検討を行っている（**図－1**）．この一連のフローにおいて重要な情報はキーワードである．すなわち地質技術者は「地山の地質」，「地山の強度」，「施工法」といった項目に関連するキーワードを文献から抽出している．したがって，キーワードの抽出作業を自動化することにより，地質技術者の負担が軽減できると考えられる．

文献から自動的にキーワードを抽出する技術としてテキストマイニングがある．この技術には様々な手法が存在するが，本稿では以下に示す（1）～（4）の手法を採用し，これらの手法を順次実行するシステムを開発した．



図－1　施工における地質・施工情報の集約と情報の流れ

（1）形態素解析

形態素解析は，文献中の単語を分解するための解析である．日本語は英語のように分かち書きを行わないため，文を単語に分解するには AI 技術の一つである形態素解析が必要となる．形態素解析を行うソフトウェアには多種のエンジンが存在しているが，本開発では無償かつ高速な MeCab を使用した．また，形態素解析では地質や施工の専門用語を正確にソフトウェアに認識させる必要があるため，地学辞典[2)]と土木用語[3)]をもとに専門用語を約 25,000 語抽出し，辞書ファイルへの登録を行った．

（2）文のベクトル化

文献から分解・抽出された単語と出現頻度は表にまとめられる．この表は，単語の種類を次元，出現頻度を要素とするベクトルと等価とみなすことができる．そのため，この手法は文のベクトル化と呼ばれている．**表－1** には低速度地山の出来形測量に関する施工文献[4)]をベクトル化した結果を示す（名詞の出現頻度上位 10 語を例示した）．

表－1　文のベクトル化

名詞	頻度
トンネル	49
地山	49
切羽	47
出来形	27
早期	27
パターン	24
強度	19
安定	14
工法	14
断面	14

キーワード：テキストマイニング，施工文献，施工の合理化
連絡先　〒182-0036　東京都調布市飛田給 2－19－1　鹿島建設(株)技術研究所　TEL 042-489-6594

（3）共起ネットワーク解析

文献中の単語の重要性は，出現頻度のほかに多くのパラメータ（例えば，文との関連性，文中での出現位置など）を基にして，多変量解析により算出される．この解析は共起解析と呼ばれており，解析結果は共起ネットワークと呼ばれる．なお，本システムではこの解析に無償かつ高速な解析が可能な KHCorder[5]を用いた．

（4）データベース化

文献ごとに共起解析の結果をデータベース化することにより，単語の関連性を検索キーとして文献を検索（逆引き）することができる．また，逆引きの結果を検索キーとしてさらに別の文献を検索することができる．これにより，施工上の問題が発生した場合に必要な多くの文献を即時に引用できるだけでなく，その文献に記述されている内容を即時に判断することができるため，地質・施工情報の集約に必要な時間を削減できる．

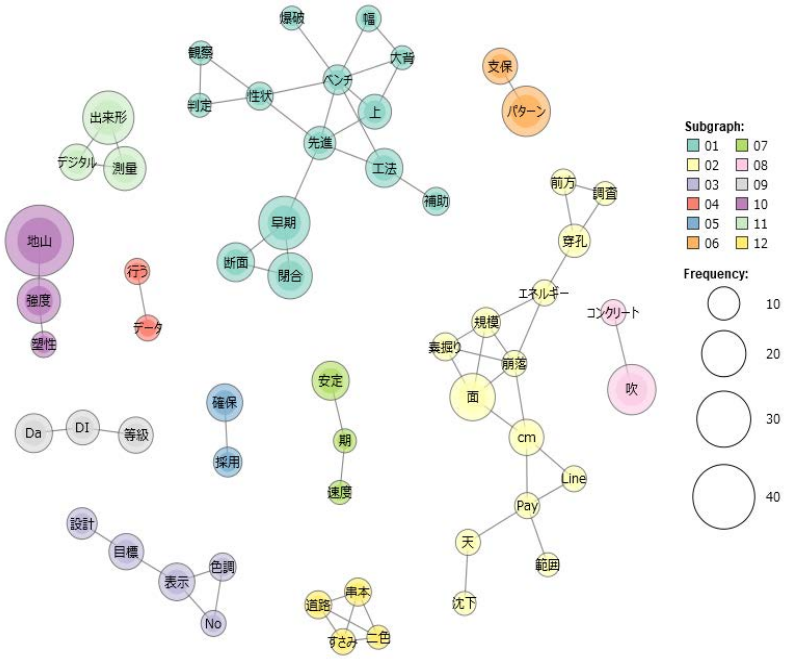
3．抽出キーワードの比較

本システムを検証するため，地質技術者によって抽出されたキーワードと，システムによるキーワードの抽出結果を比較した．文献から地質技術者が抽出したキーワードを**表－2**，本システムで解析した結果を**図－2** に示す．**図－2** 中の色は関連性の強い単語の集合，線は単語間の関連性の強さ，円の大きさは出現頻度を示す．この結果から，地質技術者が抽出したキーワードのうち泥岩を除くキーワードは本システムによって抽出されており，それ以外の多くの単語の解析結果も抽出されることがわかった．

表－2　地質技術者によるキーワード抽出

キーワード抽出
上半先進ベンチカット工法
デジタル出来形測量
塑性地山
早期閉合
泥岩

本システムで地質的なキーワード（泥岩）が抽出されなかった原因は，地質技術者が文献を読解する際には，施工位置における地質的な単語を優先的に抽出する傾向があるためと考えられる．テキストマイニングには，予め決めた基準によって単語の解析をするアプローチ（Dictionary-based）と，出現単語の位置や頻度を多変量解析で抽出するアプローチ（Correlational）があり，それぞれ独自の発展をしている [5]．地質技術者による解読は前者のアプローチに近く，本システムのアプローチは後者のそれである．このことから，本システムに前者の Dictionary-Based アプローチを取り入れることにより，より地質技術者による解読結果に近い自動文献解析が可能になると考えられる．



図－2　施工文献の共起ネットワーク解析結果

4．おわりに

本稿では，地質情報の自動集約技術の一部として共起ネットワーク解析を用いた文献解析システムを開発した．このシステムによる解析結果を地質技術者による抽出結果と比較すると，地質技術者より多くのキーワードを抽出できたが，施工に関連のある地質的情報を取りこぼすことがあることも明らかになった．今後は，この改良のため Dictionary-Based のアプローチをシステムに組み込む予定である．

参考文献

1) 戸邉ら：合理的かつ安全な施工を目的とした地質情報の自動集約システムの開発，2020 年度土木学会年次講演会講演論文集，2020.
2) 地学団体研究会：新版地学事典，平凡社，1443p，1996.
3) 土木用語辞典編集委員会：土木用語辞典，コロナ社，1421p，2000.
4) 笹嶋ら：低強度地山の出来形測量に 3D レーザースキャナーを試行ーすさみ串本道路二色トンネルー，トンネルと地下，5 月号，2020.
5) 樋口：社会調査のための計量テキスト分析　第２版，251p，2020.