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

鹿島建設(株)　正会員　○戸邉勇人　金子弘幸　升元一彦　松川剛一

１．はじめに

トンネル，ダム，および橋梁などの土木構造物を合理的かつ安全に施工するには，事前に施工予定地周辺の地質工学的な情報を網羅的かつ迅速に集約することが重要である．日本の地質は一般的に複雑であるため，その情報の集約には地質技術者による判断が必要である．その一方，地質情報の集約には多くの時間を要するため，地質技術者の人的資源の不足[1)]が問題となっている．

そのため我々は，地質情報の網羅的な集約を少ない人的資源で実施可能なシステムの開発を進めている．その一部である地質文献のテキストマイニングシステムはシンプルな仕組みながら，地質文献から有効なキーワードを迅速かつ自動的に抽出することができたので，本稿では，このシステムについて報告する．

２．地質文献の解析法のシステム化

2.1　キーワード抽出の重要性

地質技術者の負担を軽減するためには，地質技術者による文献の解析作業を客観的に見直し，ボトルネックとなる箇所を自動化することが重要である．地質技術者は，まず文献からキーワードを選別している．すなわち「砂岩」「泥岩」「丹波帯」「三波川帯」などの岩種・地質帯，「硬岩」「軟岩」などの硬軟，「黄鉄鉱」などの含有鉱物，「節理」「片理」「断層」などの地質構造といったキーワードを，最初に文献から抽出している．次にこれらのキーワードと関連の強い，施工上発生しうる問題点を過去の施工事例の報告書から読み取っている．そして最後に，地質文献から抽出したキーワードと施工事例の報告書から抽出したキーワードとを合わせて地質情報を集約している．

この手順から,地質文献の情報集約作業ではキーワードの抽出が重要であり,その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．

2.2　テキストマイニングの概要

文章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存在する．この技術は以下に示す(1)～(3)の手順により文章の特徴を解析する．

（１）形態素解析

文章を単文さらに単語に分解する．この形態素解析を行うプログラムには多種のエンジンが存在するが，本研究では無償かつ高速な MeCab を使用した。

（２）文のベクトル化

分解された単語の出現頻度を単文ごとに表にまとめる．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，これは文のベクトル化と呼ばれている（**図－1**）．

例文：

裏庭には二羽，
庭には二羽，
鶏がいる。

単語の種類	出現頻度
裏庭	1
庭	1
鶏	1
羽	2
二	2
には	2
が	1
いる	1

図－1　文のベクトル化

（３）数値解析

ベクトル化された文を数値的に解析する．たとえば文献全体の単語の出現状況は，文献中の全ての単文ベクトルを加算し，次元間の長さを比較することにより求められる．このとき，文献中の一部の段落だけを加算し，他の段落の加算結果と比較すれば，段落ごとに単語の偏在性が解析できる．

出現頻度の高い単語は重要性が高く，また出現頻度の高い単語と関連の強い単語も重要性が高いと考えられる．そのため，こうした特徴をもつ単語は文献全体を代表するキーワードとなる可能性が高い．

2.3　開発するプログラムの仕様

本稿では，このテキストマイニングを応用し，文献からのキーワード抽出を自動化した．なお，テキストマイニ

キーワード：テキストマイニング，地質文献，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

ングのシステムとしては，Web やクラウド上で解析を実行するものがすでに多数存在している．しかし施工に伴う情報は，情報管理上 Web やクラウドへのアップロードが不可とされがちである．そのため本システムは，ネットを介さずスタンドアロンで実行可能な仕様とした．

例文：第四紀は完新世と更新世から構成される

単語	品詞1	品詞2
第	接頭詞	数接続
四	名詞	数
紀	名詞	固有名詞
は	助詞	係助詞
完	名詞	一般
新	接頭詞	名詞接続
世	名詞	一般
と	助詞	格助詞
更新	名詞	サ変接続
世	名詞	一般
から	助詞	格助詞
構成	名詞	サ変接続
さ	動詞	自立
れる	動詞	接尾

登録前の解析結果

登録

単語	品詞1	品詞2
第四紀	名詞	一般
は	助詞	係助詞
完新世	名詞	一般
と	助詞	並立助詞
更新世	名詞	一般
から	助詞	格助詞
構成	名詞	サ変接続
さ	動詞	自立
れる	動詞	接尾

登録後の解析結果

図－2　地質用語の辞書登録による形態素解析の適正化

３．試適用事例

3.1 キーワード抽出

地質文献のキーワード抽出は下記の手順で実施した．はじめに，地質文献（PDF 文書）よりテキストデータを抽出した．次に，テキストデータを形態素解析によって単語ごとに分解し助詞を除去した後，全単語の出現頻度を算出した．さらに，出現頻度の高い単語と同一文に出現する単語を抽出した．最後に，下記の式で単語の重要度を算出し，重要度が高い単語を上位から 5 つ抽出した．

重要度＝(出現頻度)＋(出現頻度の高い単語と同一の文に出現する頻度)/2

標準状態の MeCab は地質用語を認識できないため，解析前に地学事典 [2]から引用した地質学用語約 2 万語を辞書登録した．これにより，スタンドアロンでも地質用語の正確な解析が可能になった（**図－2**）．

3.2 抽出例

上記の方法により，複数の地質文献からキーワードの抽出を実施した．解析に用いた文献は，A トンネル，B トンネル，C 橋梁，D ダムの地質調査報告書のうち地形・地質を記した章（A4 版のサイズ×7～10 頁程度）である．これに対し，同一の章を地質技術者が精読して抽出した重要キーワードを比較した（**表－1**）．

表－1 によると，岩種については両者の結果が比較的一致していた．その一方で走向傾斜・崩落・剥落・座屈といった岩盤の構造に関連する情報の抽出については，両者に差が生じる結果となった．

表－1　抽出キーワードの比較

構造物	地質技術者による抽出キーワード				
Aトンネル	丹波帯	頁岩	砂岩	層理	剥落
Bトンネル	根田茂帯	付加体	緑色岩	断層	走向傾斜
C橋梁	三波川帯	結晶片岩	片理	走向傾斜	崩落
Dダム	花崗岩	溶結凝灰岩	風化	節理	座屈

構造物	システムによる自動抽出キーワード				
Aトンネル	調査地	丹波帯	頁岩	砂岩	地層
Bトンネル	調査地	根田茂帯	付加体	緑色岩	断層
C橋梁	調査地	三波川帯	結晶片岩	風化	片理
Dダム	調査地	花崗岩	溶結凝灰岩	風化	節理

地質技術者による結果と一致したキーワード

3.3. 考察

前述の差異が生じた原因は，地質構造については文章よりも文中の地質図に記述されている情報が多いためと考えられる．本システムは，地質図に示された情報（走向傾斜などの方向）を判読できない．そのため，地質技術者による抽出結果との間に差異が生じたものと考えられる．このことから，地質図の自動解釈技術もあわせた総合的な文献解析システムの開発が必要であり，それにより精度の高い自動文献解析が可能になると考えられる．

４．まとめ

本稿では，施工の合理化・安全化を目的とした地質情報の自動集約技術の一部として，テキストマイニングのシステムを製作・試行した．その結果，文章のキーワード抽出については，地質技術者による結果と遜色ないことが明らかになった．今後は，このシステムを発展させ文献の要約文の抽出を試みる予定である．また，本システムにより多数の文献から自動的にキーワードを取得できるため，文献の自動的な逆引きも可能になると考えられる．これは，施工現場のような地質技術者の人的資源が限定されている環境において，地質文献情報を集約する場合にも効果を発揮すると予想される．ただし，その実現には文献のデータベース化など，多くの課題が存在するため，課題の解決に向けてさらなる検討を進める予定である．

参考文献

1) 旺文社教育情報センター：日本の大学数は 774 私立大が 8 割，旺文社教育情報センターの web サイト（http://eic.obunsha.co.jp/），2019.
2) 地学団体研究会：新版地学事典，平凡社，1443p，1996.