

テキストマイニングによる地質・施工情報の要約と対話システムによるタグ付け

鹿島建設(株) 正会員 ○松川剛一 戸邊勇人 升元一彦

1. はじめに

トンネル、橋梁、およびダムなどの土木構造物を合理的かつ安全に施工するには、施工予定地周辺の地質的特徴やそれに起因する施工上の問題について多くの情報を集約し、その集約された情報を迅速に技術者間で共有することが重要である。しかしながら、このような作業を実施する地質技術者の人的資源は慢性的に不足している。そのため筆者らは、地質技術者の人的資源不足を解決するための技術の一つとして、地質や施工に関する情報を文献から自動的に抽出するシステムの開発を進めている¹⁾。今回、このシステムに対話システムを追加することによって、キーワード間のタグ付けと文献の逆引きを容易なものとすることができたので、その成果について報告する。

2. システムの概要

地質技術者は、施工上の問題が発生したときには、まず周辺の地質や施工記録などの文献から関連情報を抽出することから始めるが、これに多くの時間を費やすことが多い。そのため、地質・施工情報に関する重要な単語としてのキーワードの抽出作業をテキストマイニングによって自動化できれば、地質技術者の負担が軽減できる。

2.1 キーワード抽出のための準備

日本語は、英語などのように単語間に空白を配置しないため、前処理として解析の最初に文を単語（形態素）に分解する必要がある。本システムでは、形態素に分解するソフトウェアに多くの実績がある MeCab を使用した。一方、MeCab は専門性の高い形態素を正確に認識できないため、地学辞典²⁾と土木用語辞典³⁾から約 25,000 語の専門用語を抽出して形態素解析の辞書に追加した。

2.2 キーワードの出現頻度算出（文のベクトル化）

次に、文献から重要な単語としてのキーワードの抽出を行う。その後、その文献の特徴を把握するため、文献内のキーワードの出現頻度を表にまとめる。

表-1 には、複数切羽を採用し高速施工したトンネルの文献（論文 1）⁴⁾と、安全に配慮してインバート施工を行ったトンネルの文献（論文 2）⁵⁾に対し、キーワードの出現頻度を解析した例を示す（上位 10 語）。これによると、論文 1、2 ともに「トンネル」が頻出し、論文 1 では「切羽」が、論文 2 では「インバート」が、それぞれ高頻出のキーワードとなっており、各々の文献の概略的な内容を把握することができた。

2.3 共起ネットワーク解析

文献の内容をより詳細に把握するには、キーワードの出現頻度だけでは不十分である。そのため、多くの実績がある KHCoder⁶⁾を使用し、文中の位置、他の単語との関連性、および出現する文や単語の類似性などを基にキーワード間の共起ネットワーク解析を行う。この解析では高い出現頻度をもつキーワードと関連性の高いキーワードがネットワークで連結された塊（クラスタ）として図化される。

図-2 には、表-1 に示した文献の解析結果を示す。これによると、論文 1 では地表面沈下が発生したこと、空洞が出現したこと、本坑で切羽掘削を施工したこと、多量の湧水が発生したことなどをクラスタから読み取ることができる。同様に、論文 2 ではインバート設置時にガードレールや柵を配したこと、コンクリート覆工時に応力を計測したこと、工事時に交通規制を実施したことなどがクラスタから判読できる。

2.4 データベース化と文献検索

共起ネットワーク解析によって抽出されたキーワードを文献のタイトルとともにデータベース化することにより、キーワードを検索キーとして文献の検索（逆引き）ができる。また、逆引きで得られた文献からさらなる逆引きが

キーワード テキストマイニング, 対話システム, 文献検索, 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 TEL042-485-1111

表-1 キーワード抽出例

論文1		論文2	
単語	頻度	単語	頻度
掘削	39	インバート	38
本坑	35	車線	20
トンネル	33	舗装	20
切羽	30	コンクリート	19
採用	21	状況	19
避難	21	写真	14
施工	20	コア	12
沈下	18	トンネル	12
空洞	17	道路	12
短縮	13	隆起	11

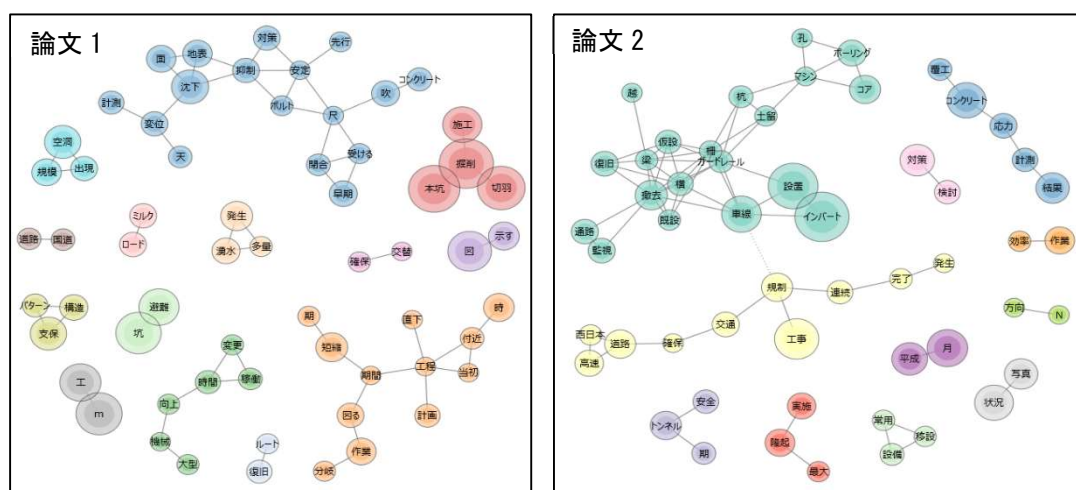


図-2 共起ネットワークの解析

可能になる。これにより、地質・施工情報の集約に必要な時間を大幅に削減することができる。

3. 検索性能向上のための対話システム

3.1 新語登録とタグ付け

文献をデータベース化する場合には、抽出されたキーワードを羅列するだけでなく、新語登録とタグ付け（アノテーション）を行うことにより検索の性能を向上することができる。

新語登録は、形態素解析ソフトウェアに未登録語を追加する仕組みである。たとえば図-2 には「避難」「坑」のクラスタや、「作業」「効率」のクラスタがあるが、これらは本来 1 つのキーワード（「避難坑」「作業効率」）として扱うべきであるため、新語として形態素解析の辞書に追加登録する。

タグ付けは、キーワード間の意味を付与する仕組みである．たとえば図-2 では「空洞」「出現」のキーワードが同一クラスタに存在する．空洞は切羽に出現するものであるため，これらのキーワードの関係性を「切羽に」とタグ付けする（図-3）．これにより，論文1は「切羽に空洞が出現した文献」として，より高度な検索が可能になる．

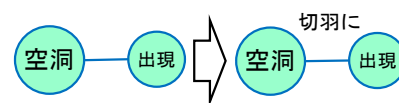


図-3 キーワード間のタグ付け

3.2 対話システムの導入による運用の効率化

前述のタグ付けや新語登録を手動で行うと数分/語 程度の時間を要するため、対話システムによって 10 秒/語 程度まで短縮することができた (図-4)。

4. おわりに

筆者らは、地質技術者の人的資源を補うための技術として、テキストマイニングと対話システムを取り入れた地質・施工情報の自動集約技術の開発を進めており、文献の概略的特徴の把握、共起ネットワーク解析、データベース化、逆引きなどをシステム化することができた。しかしながら、データベースの高精度化のためのタグ付けでは地質技術者との対話によるシステムの学習を必要としているため、完全な人的資源への負担軽減は実現できていない。そのため今後は、タグ付けの完全な自動化について技術開発を進め、より実用性の高い地質・施工情報の要約技術の開発を進めていく予定である。

了解
「避難坑」として辞書登録

図-4 対話システムによるタグ付け
と新語登録

参考文献

- 1) 松川剛一ほか：対話システムによる地質・施工情報の要約技術，2022 年度土木学会年次講演会講演論文集，2022.
- 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) 廣渡学ほか：3.7km の本坑に複数切羽を採用し避難坑とともに 20 か月で掘削，トンネルと地下，5 月号，2020.
- 5) 香川佳隆ほか：安全性と施工性に配慮した連続車線規制によるインバート設置工事，トンネルと地下，6 月号，2020.
- 6) 樋口耕一：社会調査のための計量テキスト分析 第2版，251p，2020.

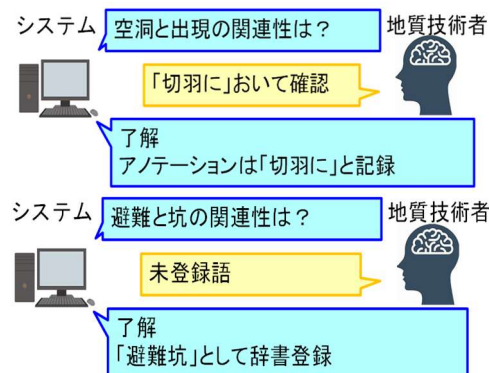


図-4 対話システムによるタグ付け
と新語登録