検索拡張生成を用いた砂防技術検索システムの構築への取り組み

岡山理科大学　○佐藤　丈晴，廣田　雅春，小田　哲也

１．目的

近年 AI の発展は、砂防業界にも影響を与えている。砂防学会誌、砂防学会研究発表会概要集にも AI を用いた数多くの研究が実施されているが、画像や地形解析図を用いたリスク評価、被災時の土砂量の算定等の検討が主体であり、技術的な専門用語などの検索技術に関する研究は皆無である。技術職員、民間の砂防エンジニアが、不明な専門用語に遭遇した際、基準書やガイドラインを手に取らず、インターネットで検索した結果をそのまま鵜呑みにしているというケースも報告されている。

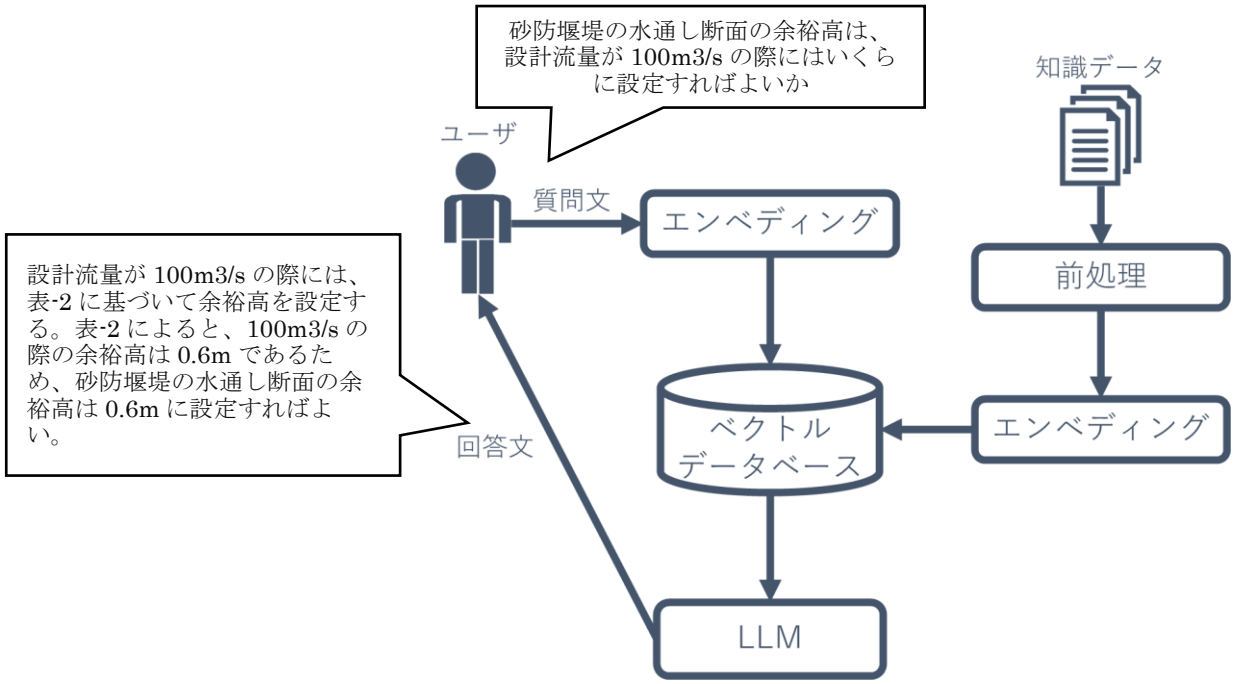
そこで、本研究グループでは、検索拡張生成を用いた基準書・ガイドライン検索システムの開発を開始した。公務員及び民間企業における DX 推進の一助として、職員・エンジニアが砂防の設計基準や解析条件などを基準書やガイドラインに基づいた結果を瞬時に引き出せるシステムを目指している。本報告では、現時点までに得られた結果を提示するものである。

２．開発方針

今回作成したシステムでは、検索拡張生成（RAG: Retrieval Augmented Generation）を用いる。検索拡張生成では、外部知識を用いて、大規模言語モデル（LLM: Large Language Model）の出力を生成する技術である。検索拡張生成による本システムの概要を図－1 に示す。

本システムの事前準備として、外部知識である知識データからベクトルデータベースを作成する。

本研究の知識データは、[土石流・流木対策設計技術指針解説] や、[地すべり等防止法]、[河川砂防技術基準] などの情報を含むファイルである。



図－1　検索拡張生成による本システムの概要

キーワード　検索拡張生成，RAG，砂防技術，基準
連絡先　〒700-0005　岡山市北区理大町 1-1　岡山理科大学　生物地球学部　TEL 086-256-9592

はじめに、これらのファイルをシステムで利用可能な形式に変換するために、それらの文章を段落ごとにチャンクと呼ばれる一定の長さに分割する。それぞれのチャンクに対してエンベディングすることで数値ベクトル表現に変換する。作成されたベクトルの集合をベクトルデータベースとする。

本システムを利用する際には、ユーザは、

[砂防堰堤の水通し断面の余裕高は、設計流量が 100m3/s の際にはいくらに設定すればよいか]

のような、質問文をシステムに入力する。

システムは、質問文を事前準備と同様にエンベディングすることでベクトルに変換する。

次に、そのベクトルと類似するチャンクをベクトルデータベースから取得する。そして、類似するチャンクと質問文 LLM に入力することで、LLM はそれらを組み合わせた回答文を生成し、ユーザに提示する。システムが生成する回答文の例として、上の質問文に対して生成された回答は、

[設計流量が 100m3/s の際には、表-2 に基づいて余裕高を設定する。表-2 によると、100m3/s の際の余裕高は 0.6m であるため、砂防堰堤の水通し断面の余裕高は 0.6m に設定すればよい。]

である。実際に土石流・流木対策設計技術指針解説を確認すると同じ記述があり、正答であった。

また、回答文の生成に用いた情報をユーザに提示するために、そのチャンクが含まれる PDF ファイルのページ番号や、その内容も回答文に合わせてユーザに提示する。これによって、回答周辺の説明文や条件などの確認を可能にした。

３．開発したシステムの概要

知識データは、砂防に関する情報を含む PDF ファイルで、国土交通省の HP で水管理・国土保全の指針・マニュアル・ガイドライン等のページ（https://www.mlit.go.jp/river/shishin_guideline/index.html）に掲載の PDF を主体とした 491 ファイルを学習に用いた。出力は、質問に対する回答と、回答の根拠となった基準書のファイル名、回答するにあたって参考にした部分のテキスト，ページ番号，および類似スコアを出力する。回答は一つの質問に対し、3 通りを出力させた。回答の類似スコアは、0～1 で 1 に近いほど質問に対してそのテキストを回答として用いるのに適切であることを表す。ページ数は、該当するファイルのページ数であり、基準書に記載されているページ数ではない。本システムの実装は、LLM を扱うオープンソースのフレームワークである LangChain を用いた。また、エンベディングには Multilingual-E5-large を、LLM には ELYZA-japanese-Llama-2-7b-instruct を用いた。

４．まとめ

2023 年 12 月の段階では、質問の方法、学習させる PDF ファイルの種類等によって正答率は大きく変化する。また数値を回答する問いに対しては正答率が高いが、文章による要約は十分な検討が必要である。しかし、様々な検討ケースを継続しているうちに徐々にではあるが、正答率が高まってきている。

本システムの目標は、公務員や砂防技術者が実務で容易に検索できるシステムを構築することである。第 3 位までの回答で 100%に近い精度を確保できれば、十分実用に値すると考えている。今後、誤答例などを丁寧に分析し、より良い検索システムの構築を目指したい。

参考文献

砂防に関する情報を含む PDF ファイルで、国土交通省の HP で水管理・国土保全の指針・マニュアル・ガイドライン等のページ（https://www.mlit.go.jp/river/shishin_guideline/index.html）に掲載の PDF を主体とした 491 ファイルを学習に用いた。本概要集掲載原稿に、すべての参考文献を記載できないため、引用した HP 名とアドレスを提示させていただいた。ここに記して感謝申し上げる。