F-025

社内全文検索システムと生成 AI を組み合わせた検索拡張生成(RAG)システム
RAG on combined with Enterprise knowledge System and Generative AI

齋藤 靖二† 岩元 淳† 加藤 総士† 鈴木 健太郎†
Seiji Saito Jun Iwamoto Sohshi Kato Kentaro Suzuki

1. はじめに

近年、企業内に蓄積された技術情報やノウハウを有効活用し、経験豊富な技術者の継承や不具合の過去事例を活用した品質管理体制の強化を実現する手段として、情報検索技術の重要性が高まっている。検索対象である社内文書は、事業活動の幅広い分野を対象としており、特定の業務分野の専門用語が頻出するという特徴を持つ。文書の形式や構造も統一されておらず、大量の文書が存在する。そのため、多様な文書構造に対応可能で、大量の保有文書から短時間で必要な情報を抽出する検索手法が求められる。一般的な検索手法として、全文検索システム[1]が挙げられるが、検索結果として提示された複数の文書から所望の情報を引き出すには、文書を１つ１つ開いて内容を確認する必要がある。そこで、RAG（Retrieval-Augmented Generation）技術を適用し、検索結果として得られた社内の複数文書を生成 AI に渡して回答させることで、複数文書の中から求める情報を効率的に把握する手法を検討した。社内の全文検索システムを活用しながら生成 AI を組み合わせ、自然言語による質問応答を実現する検索拡張生成（RAG）システムを構築した。

本稿では、提案した社内 RAG システムの設計方針、実装アーキテクチャ、導入における技術的工夫、および実運用に基づく評価・考察について詳述する。

2. システムの概要

2. 1 RAG を活用した検索支援の方法

本研究では、社内の全文検索システムと生成 AI を組み合わせて検索拡張生成（RAG）システムを構築した。利用者の自然言語による質問に対して、関連文書の検索と回答生成を一貫して行うことができる。図 1 に利用者が設計部門の場合を想定したシステムの全体構成を示す。本システムでは、分散して存在する複数の部門サーバから、技術文書や報告書などの情報を定期的にクロールし、全文検索システムの設計部門層データベースにテキストデータとして格納している。これにより、横断的な情報検索が可能となり、複数のサーバに分散した情報がサイロ化することを防いでいる。この全文検索システムをそのまま活用し、以下の処理フローにより、自然言語による質問応答を実現している。

・処理１．対話型検索アプリケーションへの質問入力

利用者は自然言語で質問を入力する。UI はチャット形式で、対話的な操作が可能である。

・処理２．生成 AI による検索式の生成

入力された質問に対して、生成 AI を使って、全文検索システムを検索するための適切な検索式（キーワードと AND/OR を組み合わせた式）を生成する。なお、製品番号や型番など特定の業務分野固有の専門用語は、そのままキーワードとして抽出されるようプロンプトで調整している。

・処理３．全文検索システムによる文書取得

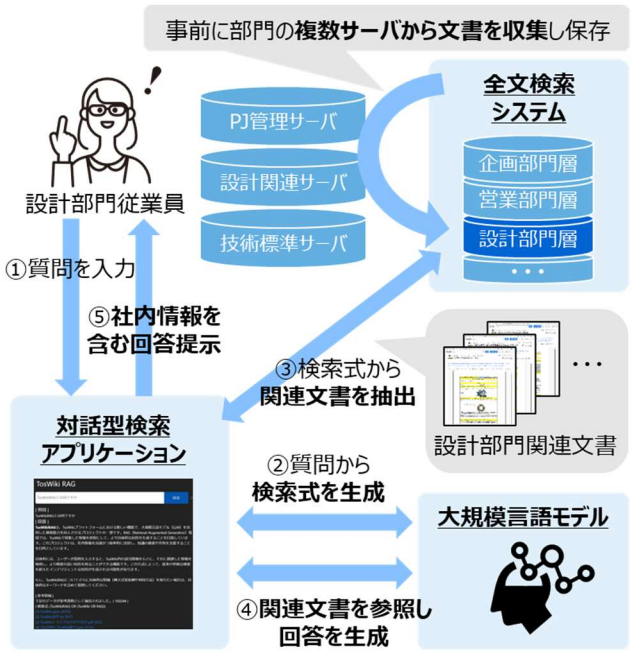

図 1 システムの全体構成

生成された検索式を用いて、社内の全文検索システムから関連文書を検索式との一致スコアが高い順に取得する。すでに構築されているセキュリティ制御により、ユーザ個人のアクセス権のある設計部門層の文書のみ参照できる。

・処理４．関連文書を参照し、回答を生成

元の質問と、全文検索から得られた文書をコンテキストとして生成 AI に入力し、自然言語による回答を生成する。検索結果として取得した各記事は、全文をコンテキストとして渡している。

・処理５．回答提示

取得した文書をもとに生成した回答と、回答生成に利用した文書のリンクをユーザに提示する。文書のリンクを提示することで、情報源をトレースすることができ、効率的にファクトチェックを行うことが可能となる。

この構成により、従来のキーワード検索だけでは困難であった複数文書の概要把握が可能となり、利用者の検索負担を大幅に軽減することができた。

2. 2 全文検索システムとの統合

既存システムの全文検索エンジンを活用することで、セキュリティ対策など、社内で利用するために必要な機能の実装工数を削減した。その実現のため、全文検索システムと大規模言語モデルを連携する方式を採用した。一般的な RAG では、質問とのベクトル類似度が近いチャンクをベクトル DB から抽出し、回答を生成している。そのため、文

章をチャンクという単位に分割し、埋め込みモデルを使ってベクトル化する必要がある。図 2 に示すような、文中に図が挿入されている文書形式は社内文書でもよく見られる。このような文書は、テキストデータ読込み時に文書が分割されてしまい対象情報を含んだチャンクを正しく抽出できないという問題が生じる。チャンク長やレイアウトに応じた調整を行うことで改善できる場合もあるが、社内文書は構造が標準化されていないため、個別の調整は困難である。そこで、チャンク分割を行わず、全文検索で該当するキーワードを含んだ文書の全文をコンテキストとして LLM に渡す手法を採用した。入力文書数が限定されるが、昨今の LLM では入力トークン数が格段に増加してきており、チャンク分割による情報欠落を回避することができる。特定の業務分野の用語についても、類似ベクトルと正しくマッチしないという問題がある。埋め込みベクトルを生成するモデルを再学習することで対応できるが、その都度再学習が必要となる。運用の低コスト化を実現するため、対象文書の抽出は全文検索のみで実施することとした。

2.3 情報開示範囲の制御と参照範囲の絞り込み

社内の情報を取り扱うには情報アクセスに配慮したセキュリティ設計が必要である。本システムでは、全文検索システム側の既存のアクセス制御機構を活用することで、ユーザごとに参照可能な文書を制限している。これにより、生成 AI に入力される文書は、利用者のアクセス権に基づいて適切に制御され、関係者以外への情報漏洩リスクを抑制している。

また、全文検索システムは多種多様な専門の文書を対象としているため、回答範囲の分野を切り分けずに検索を行うと、似たような用語を含んだ全く異なる分野の文書が紛れ込んでしまうリスクがある。その結果、生成 AI に渡す情報に無関係の情報が含まれることになり、回答精度が低下する可能性がある。そこで、検索条件として文書が保存されていた URI やフォルダパスの指定を行って検索することで、生成 AI に与える文書を絞り込むことが可能となり、回答の精度を向上させている。

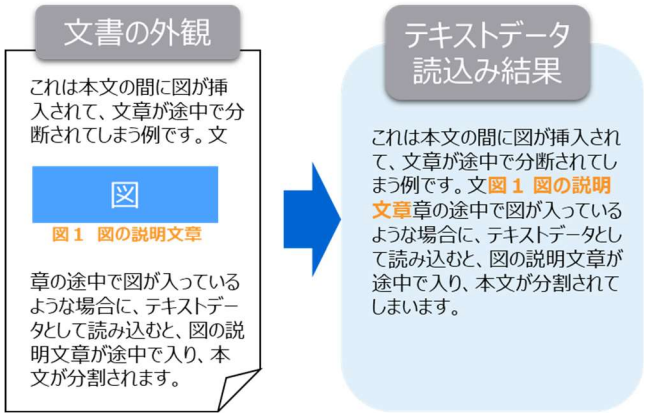

図 2 図の説明文章（黄色文字）で本文文章が分割される例

3. 評価・考察

本システムの有効性について、社内の複数部門で運用し、以下の観点から評価と考察を行った。

3.1 RAG 検索の実運用における有用性

顧客からの要件を元に設計を行う際、過去の設計事例や不具合事例を探し、参考資料として利用することが多い。対象の資料を探す際、結果として表示されたファイルを 1 つ 1 つ開き、内容を確認し、検索ワードを変えて絞込みを行っていた。そのため、必要な情報を含んだ文書に到達するまで非常に時間がかかっていた。本システムの導入により、複数ファイルにまたがる情報を統合的に確認し、短時間で必要な情報へ到達できることを確認した。

実際の社内ユーザによる運用においても、検索効率の向上が業務の効率化に寄与し、社内情報の有効活用が促進されたとの評価が得られた。

3.2　システム構築と運用の容易さ

社内の全文検索エンジンを活用し、生成 AI との連携を行うことで、短期間での構築を実現した。大規模なデータを改めてベクトル DB として埋め込みデータとして構築する時間を短縮できたため、約 2 週間という短期間で検証環境を構築し、社内への迅速な導入を実現した。また、全文検索で対象文書を抽出するため、専門用語などを含んだ新規文書の追加時においても、特定領域に特化した用語もモデルの再学習をせずに検索できる。これにより、運用負荷を抑えつつ、社内用語や特定の業務分野分野における専門用語への対応が可能となっている。

3.3 アクセス制御と参照絞込みによる回答精度向上

既存の全文検索システム側でアクセス制御を行うことにより、セキュリティを考慮した検索システムが実現できた。また、多岐にわたる作業に応じて、利用者自身がサイト名やフォルダを選ぶことにより、生成 AI に参照させる文書を適切に絞り込むことが可能なため、それぞれの作業に沿った有用な回答が得られやすくなり、所望の情報を把握する効率が向上した。

4. おわりに

本研究では、社内の全文検索システムと生成 AI を組み合わせた検索拡張生成（RAG）システムを構築し、自然言語による質問応答を可能とする仕組みを提案した。社内の全文検索システムを活用することで、短期間かつセキュアに構築可能であり、再学習を必要としない柔軟な運用が実現できた。また、フォルダ名の指定による情報の絞り込みにより、生成 AI の回答精度を向上させる工夫も行った。実運用においては、従来の検索では難しかった複数文書の概要把握が可能となり、社内の技術情報活用を促進する効果が確認できた。

参考文献

[1]　佐野恵一，“階層的データ管理と複数データ領域の高速横断検索を実現する社内ナレッジシステム”，FIT2022 0-021
[2]　東芝レビュー，”情報共有基盤 TosWiki™と大規模言語モデルの連携活用で社内情報検索の効率向上を図る検索システム”，https://www.global.toshiba/content/dam/toshiba/jp/technology/corporate/review/2025/02/1-3.pdf