生成AI基礎

7 検索拡張生成(RAG)について理解する

Retrieval Augmented Generation（RAG）とは?

- 外部の情報を活用してユーザーのクエリに回答する技術
- 大規模言語モデル（LLM）の限界を補完
- より正確で関連性の高い回答を提供

仕組み

情報検索（Retrieval）

- ユーザーの質問や命令をクエリに変換
- 関連情報の取得（外部データベース、API、ニュースなど）

生成（Generation）

- データ統合（外部情報とLLMの知識）
- 応答生成（引用を含む場合もあり）

利点

- **最新情報の反映**
  - 常に最新の情報を提供可能
- **精度と信頼性の向上**
  - 情報源の引用で信頼性アップ
- **柔軟性と適用範囲の広さ**
  - チャットボット、質問応答システム、教育アプリなどで活用

課題

- **データ品質と更新頻度への依存**
  - 古い情報が提供されるリスク
- **処理時間の問題**
  - リアルタイム性の遅延
- **誤情報のリスク**
  - 外部データベースの品質管理が必要

Retrieval Augmented Generationの実例

- 検索エンジン
  - 全文検索や画像検索、音声検索に利用
  - キーワードの正確な一致に頼らず関連性の高い結果を提供
- 推薦システム
  - ユーザーの過去の行動や好みをベクトル化
  - パーソナライズされた推薦を行う
- 質問応答システム
  - 文書をテキスト埋め込みに変換し、適切な回答を提供
  - 長い文書から必要な情報を迅速に引き出す

分かち書きと全文検索システムの基本

- 分かち書きとは
  - 文章の中で単語や文節ごとに空白を挿入する方法
  - そもそも英語などのラテン文字を使う言語では、単語ごとに空白を挟むのが一般的
- 形態素解析
  - 日本語の文章を単語や文節に分割する処理
  - MeCabなどの形態素解析器を利用
- n-gram解析
  - 文章をn個の文字や単語に分割して解析する手法
  - 2-gram（バイグラム）や3-gram（トライグラム）などがある

データの前処理

- テキストデータの収集
- MeCabによる形態素解析

検索クエリの作成

- 検索クエリの生成
- クエリを用いた情報検索

検索結果とLLMへの質問の組み合わせ

- 検索結果の抽出と整理
- LLMのプロンプトとの結合

結果の出力

- 回答の生成
- 結果の評価と改善

ベクトルサーチ

ベクトルサーチの仕組み

- ベクトル埋め込み
  - テキストを数値ベクトルに変換
- セマンティック検索
  - キーワードの完全一致に依存しない検索
  - 意味やコンテキストを理解する
- 近似最近傍探索（ANN）
  - 高次元ベクトル空間での効率的な検索
  - 大規模データセットでも高速に動作

ベクトルサーチの利点

- 高精度な検索
  - 意味とコンテキストを考慮するため、より関連性の高い結果が得られる
- 柔軟性
  - 多次元空間でデータを扱い、複雑な関係性を表現できる

ベクトルサーチの課題

- 計算コスト
  - 高次元ベクトルの計算には大きな計算リソースが必要
  - 特にリアルタイム検索ではパフォーマンスが課題

ベクトルサーチでの全文検索とLLMの組み合わせ方

データの前処理

- データを収集し、クレンジング（ノイズ除去、データフォーマットの統一）を行う
- テキストや画像データをベクトル埋め込みモデルを使ってベクトル化する
- ベクトル化されたデータをデータベースに格納

検索クエリの作成

- ユーザーのクエリ（質問や検索ワード）をベクトルに変換
- クエリベクトルとデータベース内のベクトルとの類似度を計算
- 最も関連性の高いデータ（ベクトル）を検索結果として取得

検索結果とLLMへの質問の組み合わせ

- 検索結果から関連情報を抽出し、LLMに渡す
- LLMは検索結果を基に、ユーザーの質問に対する回答を生成

結果の出力

- LLMが生成した回答をユーザーに表示
- 必要に応じて、検索結果の出典情報も提供
- ユーザーからのフィードバックを収集し、システムの精度向上に活用

ハイブリッド検索

ハイブリッド検索とは

- 従来のキーワードベースの検索（全文検索）とベクトルベースの検索（セマンティック検索）の強みを組み合わせた方法
- 検索結果の関連性と精度を向上させる

ハイブリッド検索の主要要素

- **全文検索**：反転インデックスを使用し、キーワードを迅速に検索。構造化されたクエリや正確な用語検索に優れる
- **ベクトル検索**：機械学習モデルで生成されたベクトル埋め込みを使用し、セマンティックな意味とコンテキストを捉える。概念的に類似したドキュメントを検索
- **組み合わせ**：レシプロカルランクフュージョン（RRF）などのアルゴリズムで両方の検索結果を統合しランク付け

Text-to-SQLとは?

- 自然言語の質問をSQLクエリに変換する技術
- ユーザーが平易な言葉でデータベースと対話可能に
- LLM（大規模言語モデル）の進展により性能と汎用性が向上

プロンプトエンジニアリング

- 効果的なプロンプトエンジニアリング戦略
  - Chain-of-Thought
  - Few Shotプロンプティング
- 関連するスキーマ情報や例をプロンプトに含める

検索クエリの実行

- クエリを用いた情報検索

検索結果とLLMへの質問の組み合わせ

- 検索結果の抽出と整理
- LLMのプロンプトとの結合

結果の出力

- 回答の生成
- 結果の評価と改善

まとめ

- 検索拡張生成（RAG）は、外部情報を活用してユーザーのクエリに回答する技術
- ベクトルサーチは、高次元ベクトル空間での効率的な検索を可能にする
- ハイブリッド検索は、全文検索とベクトル検索の強みを組み合わせた手法
- Text-to-SQLは、自然言語の質問をSQLクエリに変換する技術