Keywords you should know. 第168回

知っておきたいキーワード

Agentic Search

山本智幸†

† シャープ株式会社 研究開発本部 AI開発センター
"Agentic Search" by Tomoyuki Yamamoto (AI Development Center, Corporate Research & Development Group, SHARP Corporation, Chiba)
キーワード：LLM (Large Language Model)，RAG (Retrieval-Augmented Generation)，AI Agent，Information Retrieval

まえがき

AIが検索・閲覧・比較・検証を自律的に進める「Agentic Search」を解説します．従来の検索エンジンは，人間が検索語を考え，検索結果を読み比べることを前提としていました．RAG (Retrieval-Augmented Generation) はこの流れを一歩進め，検索した文書をLLM (Large Language Model) に渡して根拠付き回答を生成します．一方，Agentic Searchでは，LLMが調査計画を立て，必要に応じて検索語を作り直し，複数の情報源を読み比べ，足りない情報を再検索しながら回答を組み立てます．本稿では，基本的なAgentic Searchの流れについて映像情報メディア領域での活用をモチーフとして説明します．

人の探索行動とAgentic Search

Agentic Searchは人の探索行動のアナロジーで考えると容易に理解できます．膨大な映像コレクションから見たい映像を探す場合にどのように探索するでしょうか？

図1に示すように，探索ツールと探し方を決めて，当たりをつけて探し，候補を確認し，それでよいか，他を探すかを判断して探索を行うのではないでしょうか．このように，目的に応じて探索方針を立て，検索結果を評価しながら検索を繰り返す処理を，AIエージェントによって自律的に実行する考え方がAgentic Searchです．

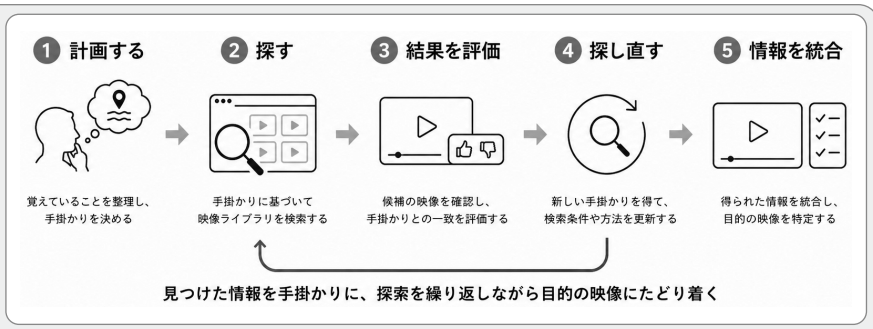


図1　人の探索行動をモチーフとしたAgentic Searchの概念フロー

映像情報メディア学会誌 Vol. 80, No. 6, pp. 732～734 (2026)

RAGとAgentic Search

LLMを用いた対話型の情報検索システムでは，従来RAGが用いられてきました．RAGとAgentic Searchはどのような関係にあるでしょうか？

RAGは，LLMが外部の情報を参照しながら回答を生成するための仕組みです．LLMが学習時に獲得した知識だけでは，最新情報，社内情報，個別の製品情報などを正確に扱えないため，多くの実用的なAIシステムでは外部情報に基づいて回答する仕組み（グラウンディング）が必要になります．

広い意味でのRAGは，データベースや文書，検索エンジンなどから取得した情報をLLMの入力に追加し，その内容に基づいて回答を生成する方式全般を指します．RAGの情報取得部分に，検索計画，検索結果の評価，再検索などのエージェント的な制御を導入することもできます．このように，RAGの処理の一部または全体をAIエージェントが動的に制御する構成は，一般にAgentic RAGと呼ばれます．Agentic Searchは，特に情報探索の計画と反復に焦点を当てた概念であり，Agentic RAGを実現する主要な要素の一つと捉えられます．



図2　単発検索型RAGとAgentic Searchの処理フローの対比

一方で，RAGの代表的な実装では，図2の左側に示すように，ユーザの質問を埋め込み表現（Embedding）によってベクトル化し，意味的に近い文書の断片をベクトル検索で取得します．その検索結果をコンテキストとしてLLMに与え，回答を生成します．

例えば，映像コンテンツを検索する場合は，質問に相当するベクトルと，字幕，文字起こし，動画説明，メタデータなどのベクトルを比較することで関連情報を取得して回答に利用します．このような単発検索型のRAGは，処理がシンプルで高速である一方，最初の検索で適切な情報を取得できなければ，回答品質も低下します．

図2の右側に，Agentic Searchにより同様の質問に回答する処理フローの例を示します．この例では，まず字幕，動画説明など，比較的低コストで検索可能な情報で候補を絞り込みます．その結果だけでは場面を特定できないと判断した場合に，OCR（Optical Character Recognition：光学文字認識）や画像解析を追加で実行し，得られた情報を最終回答に利用します．

Agentic Searchを実現するための技術

Agentic Searchは，AIエージェントを中心に，検索ツール，外部システムとの接続機構，複数エージェントの連携基盤などを組み合わせて実現されます．LLM単体がすべての処理を行うのではなく，目的に応じて適切な機能を選び，実行結果を受け取りながら探索を進めるシステムとして構成されます．

中心となるAIエージェントは，ユーザの依頼や探索途中の状況を解釈し，次に利用するツールや処理を決定します．そのため，利用可能なツールの用途，入力形式，実行上の制約などがあらかじめ与えられます．例えば，字幕やメタデータには全文検索やベクトル検索，映像内の文字にはOCR，人物や物体の確認には画像・映像解析をツールとして与えます．このほか，Web検索，SQLによるデータベース検索，画像類似検索，音声認識など，検索対象に応じた機能をツールに追加します．

AIエージェントと外部ツールを接続する仕組みとして，MCP（Model Context Protocol）などの標準化されたインタフェースも利用されます．MCPを用いると，検索システム，データベース，ファイルストレージなどが提供する情報や機能を，共通のプロトコルを通じてAIエージェントに提供できます．これにより，システムごとに独自の接続方式を設計する負担を軽減し，利用可能な情報源やツールを追加しやすくなります．

Agentic Searchは，探索処理を複数のエージェントに分担させる構成でも実現できます．例えば，検索計画，映像候補探索，候補評価，最終回答生成といった複数のAIエージェントに役割を分ける方法です．複数エージェントの連携には，情報交換の仕組みが必要であり，A2A（Agent2Agent）と呼ばれるプロトコルが開発されています．ただし，役割を細かく分けすぎると処理が複雑になるため，対象業務に応じた設計が重要です．

近年，AIエージェントを安定して動作させるための実行環境や制御機構の設計は，ハーネスエンジニアリングと呼ばれることがあります．☞

☞ 図3の例でいえば，検索・解析ツールや連携エージェントを整備することで実行環境を整えます．また，探索の停止条件，実行回数や時間の上限，エラー時の再試行，処理履歴と根拠情報の記録，安全性担保の仕組みなどを設計します．

Agentic Searchの性能はLLMの能力だけで決まるのではなく，検索ツールの品質と，それらを安全かつ確実に利用させる実行基盤によって大きく左右されます．注意点として，対応すべき検索ツールや情報源の種類，評価すべき観点を増やすほど，AIエージェントの頭脳であるLLMの負担は増加し，実行時間やコストの増加を引き起こします．

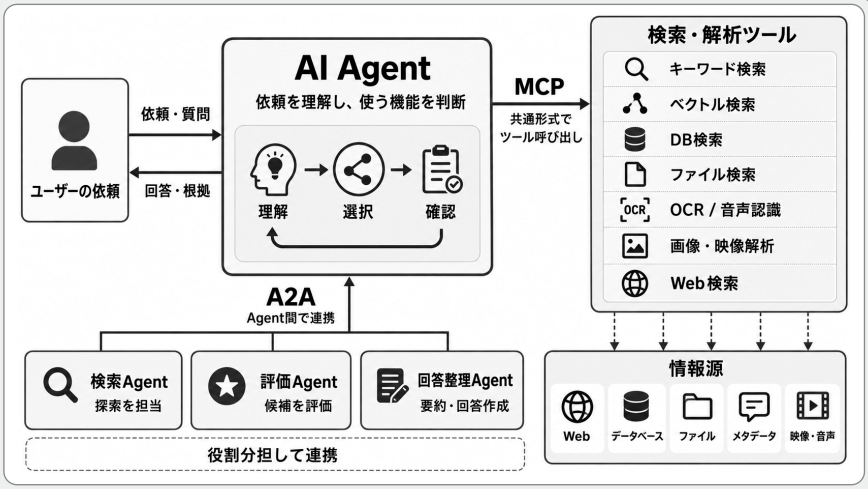


図3　AIエージェントを用いたAgentic Searchの構成

そのため，対象となる情報源や検索ツールを選別することで，探索範囲，停止条件，コストを管理しやすくすることが実用上は重要です．

むすび

Agentic Searchは，人のように多様な情報源を探索し，目的に応じて情報を活用できる可能性を持っています．RAGや従来の検索技術を置き換える技術というよりは，それらをツールとして利用することで自律的かつ柔軟な検索を実現します．

映像情報メディアの領域では，解析対象となるデータ量が大きいため，すべての映像を一律に詳細解析することは現実的ではありません．字幕検索，映像検索，OCR，物体検出などを段階的に使い分けるAgentic Searchにより，検索精度と処理コストを両立することが期待されます．

Agentic Searchをさらに学びたい読者には，次の文献を推奨します．IRCoTは，推論と検索を交互に実行する基本的な枠組みを示した先駆的研究です[1]．Search-o1は，大規模推論モデルの推論過程に動的な検索を組み込んだ代表例です[2]．また，Agentic RAGの全体像については，サーベイ論文[3]が参考になります．

（2026年7月31日受付）

参　考　文　献

1) H. Trivedi, N. Balasubramanian, T. Khot and A. Sabharwal: "Interleaving Retrieval with Chain-of-Thought Reasoning for Knowledge-Intensive Multi-Step Questions", Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics, 10014-10037 (2023)
2) X. Li, G. Dong, J. Jin, Y. Zhang, Y. Zhou, Y. Zhu, P. Zhang and Z. Dou: "Search-o1: Agentic Search-Enhanced Large Reasoning Models", Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, 5420-5438 (2025)
3) A. Singh, A. Ehtesham, S. Kumar, T.T. Khoei and A.V. Vasilakos: "Agentic Retrieval-Augmented Generation: A Survey on Agentic RAG", arXiv:2501.09136 (2025)



山本（やまもと）　智幸（ともゆき）　2004年，東京大学大学院情報理工学系研究科修士課程修了．同年，シャープ（株）入社．同社にて，H.265/HEVCを代表とする映像符号化の標準化活動に従事．近年は，生成AIを用いた対話システムや映像情報システムに関する要素技術開発に取り組む．