AIエージェント時代の特許調査支援―RAGの設計・実装・評価

―大規模コーパスからハイブリッド検索・分類絞り込みまで―

Patent Search Support in the Age of AI Agents: Design, Implementation, and Evaluation of RAG

アジア特許情報研究会　知財情報解析　グループリーダー／ソフィア・リサーチラボ代表

安藤　俊幸

1985 年現花王株式会社入社、研究開発に従事
1999 年研究所の特許調査担当（新規プロジェクト）、2009 年知的財産部
2011 年よりアジア特許情報研究会所属　知財情報解析グループ
2020 年　特許情報普及活動功労者表彰 日本特許情報機構理事長賞「技術研究功労者」受賞
2024 年　花王を定年退職、第 49 回情報科学技術協会賞「研究発表賞」受賞
2025 年　ソフィア・リサーチラボ　代表
情報科学技術協会、人工知能学会、データサイエンティスト協会　各会員

はじめに

生成系の人工知能（AI: Artificial Intelligence）の急速な発展により、特許調査の在り方が問い直されている。大規模言語モデルが文章を生成する能力に注目が集まる一方で、特許の先行技術調査においては、生成の前段にある検索の精度こそが調査全体の質を左右する。誤った、あるいは網羅性を欠いた検索結果の上に生成を重ねても、調査の信頼性は確保できない。本稿は、生成よりもまず検索の精度を地道に積み上げるという立場をとる。

従来の特許調査は、検索によって得られた多数の文献を調査者が読み込み、関連性を判断する作業に多くの労力を費やしてきた。本稿はこの構図を「読む調査」と呼び、AI エージェントが調査範囲の判断を担い、人間がより高次の判断に集中する「判断する調査」への移行を展望する。

筆者は、日本語特許文書を対象とした先行技術調査支援のための検索拡張生成（RAG: Retrieval-Augmented Generation）[4] システムを研究開発してきた。本稿は、約 400 万件規模の特許コーパスの構築から、疎・密検索の比較、親子チャンク・分類メタデータ・自前特徴語によるハイブリッド検索の統合、生成 AI による重み付けの検証、そして分類予測による調査範囲決定までを、実装と定量評価の観点から総説的にまとめるものである。成功した手法だけでなく、効果を確認できなかった手法も併せて報告し、AI をどの工程に用いるべきかという設計指針を示す。

1 特許調査支援の課題とRAG の位置づけ

特許調査には、出願前の先行技術調査、事業の自由度を確認する侵害防止調査（FTO: Freedom To Operate）、権利を争う無効資料調査など複数の目的がある。これらは求められる精度の特性が異なる。先行技術調査や無効資料調査では、関連文献を取りこぼさない網羅性、すなわち再現率が重視される。一方で侵害防止調査では、抵触の可能性がある権利を広く拾ったうえで精査する必要があり、やはり見落としの回避が要となる。いずれの目的も、調査対象の技術を分類やキーワードで定義し、母集団を抽出し、関連文献を絞り込むという共通の流れをもつ。

図 1　特許調査の目的と RAG の対応

RAG は、検索（Retrieval）と生成（Generation）を組み合わせる枠組みである。特許調査に適用する場合、まず検索によって関連候補を集め、必要に応じて生成

AIが要約や対比を補助し、最終的な判断は人間が担う。本稿が主に扱うのは、この枠組みのうち検索と、検索結果の融合の部分である。生成・検証の工程は実務システムの要件として接続する位置づけとする。

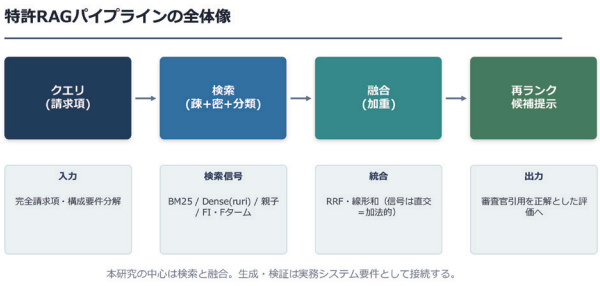

図２　特許 RAG パイプライン全体像

検索の精度が調査全体の質を規定する。請求項という長文かつ専門的な文書から、関連する先行文献を漏れなく拾うには、語彙の揺れや表記の差を吸収しつつ、専門用語の一致を的確に捉える必要がある。本稿は、この検索精度を定量的に積み上げ、どの信号がどれだけ寄与するかを切り分けることを主題とする。

2 大規模特許コーパスの構築

研究の基盤として、約 425 万件の特許公報からなるコーパスを構築した。書誌・請求項・明細書段落・分類・審査官引用を、公報番号を主キーとして単一のデータベースに格納した。

コーパスのデータ規模

項目	値	内容
全件コーパス 公報数	4,347,009	patents_meta（約425万件）
全件DB 容量	315.7 GB	単一DuckDBファイル
全件 請求項数	54,379,966	patents_claims
全件 構成要件数	135,216,761	patents_claim_elements
全件 段落数	227,673,186	patents_paragraphs
G06F 公報数	183,413	主分類G06Fサブセット
G06F 評価コーパス	2,648	評価対象サブコーパス
評価クエリ数	1,267	審査官引用を正解化
正解ペア数	1,434	Ax 1,336 / Ay 98

出典：*inventory / duckdb_tables / 各実験ナレッジ*

図３　コーパスのデータ規模

図４　データベースのスキーマ構成

データの取り込みは、特許庁が提供する公報の拡張可能マーク付け言語（XML: Extensible Markup Language）を解析し、タグの正規化を経て各要素を抽出する流れで行う。新形式と旧形式の双方に対応し、抽出後に正規化を施してデータベースへ格納する。

図５　公報 XML の取り込みフロー

構築の過程で、データ品質が検索精度を大きく左右することが判明した。当初、請求項の多くがタグの見落としにより途中で切断されていた。文末の句点の出現率を完全性の代理指標とすると、修正前は 3.2 パーセントにすぎなかったが、再抽出により 99.1 パーセントへ改善した。この完全請求項化により、検索精度は全件で 46 から 64 パーセント向上した。請求項の完全性は、いかなる検索手法を用いる場合でも前提となる土台である。どれほど高度な検索手法も、入力となるデータの品質を超える精度は得られない。コーパス構築の段階での品質確保が、後続のあらゆる工程の上限を定める。

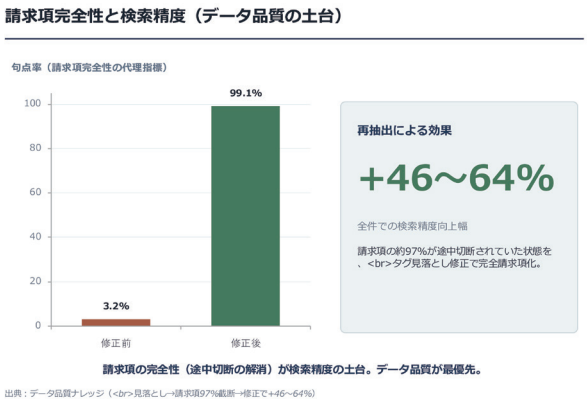

図６　請求項完全性と検索精度

3 評価フレームワークの設計

検索手法を公平に比較するため、審査官が引用した先行文献を正解とする評価枠組みを構築した。指標には、最初の正解の順位の逆数の平均である平均逆順位（MRR: Mean Reciprocal Rank）、上位 100 件の網羅性を測る再現率、上位 10 件の順位品質を測る正規化減損累積利得（NDCG: Normalized Discounted Cumulative Gain）を用いた [9]。

図７　評価指標の定義

正解は、単独で新規性や進歩性を否定しうる X 引用と、複数文献の組み合わせで効く組み合わせ引用に層別した。前者は強い関連、後者は分野横断的で捕捉が難しい関連であり、手法ごとの効き方の違いを観察できる。

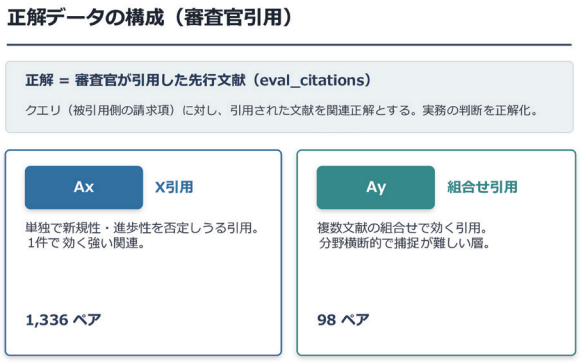

図８　正解データの構成と Ax/Ay

評価母集団は、全クエリから順に条件を満たすものへ絞り込む形で定義した。スコアはコーパスの規模や分野に強く依存するため、絶対値ではなく同一枠組み内の相対比較として解釈することを設計原則とした。

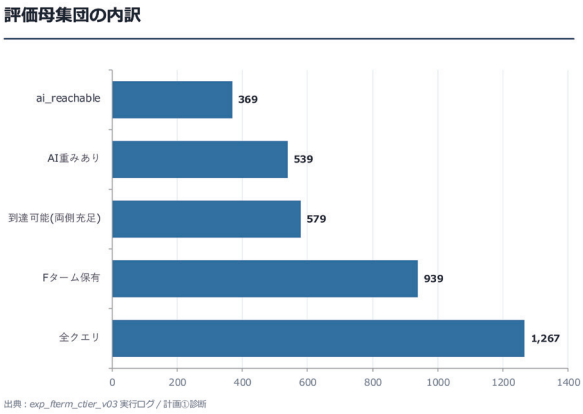

図９　評価母集団の内訳

4 疎・密検索の基礎比較

検索の基礎として、語の表層一致に基づく疎な検索である BM25（Best Matching 25）[1][2] と、意味を埋め込みベクトルで捉える密な検索 [5][6] を比較した。情報処理分野のサブコーパスにおいて、BM25 の MRR は 0.2120 であり、密検索の 0.1041 の約 2.04 倍であった。技術文書では専門用語や型番の完全一致が強い信号となるため、疎検索が優位となる。

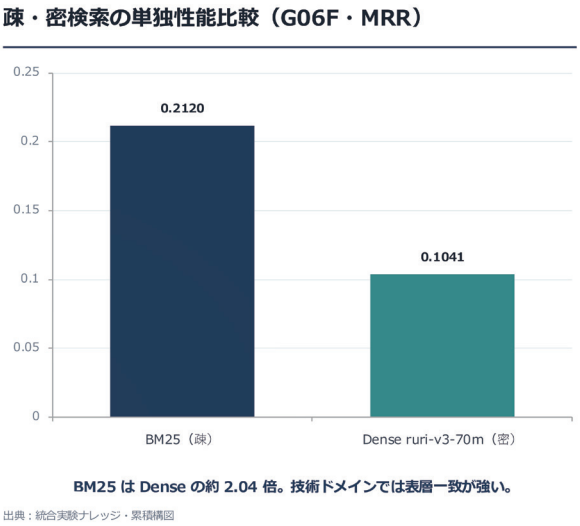

図 10　疎・密検索の単独性能比較

密検索のモデルには、日本語汎用テキスト埋め込みモデル Ruri[3] を用いた。Ruri は近年の ModernBERT[7] を基盤とし、外部の分かち書きを要さない設計をもつ。軽量版と大規模版を比較すると、統合後の MRR は軽量版が 0.2896、大規模版が 0.2862 とほぼ同等である一方、索引構築時間は約 7.3 時間と約 29 時間で約 4 倍の差があった。品質とコストの両面から軽量版を採用した。

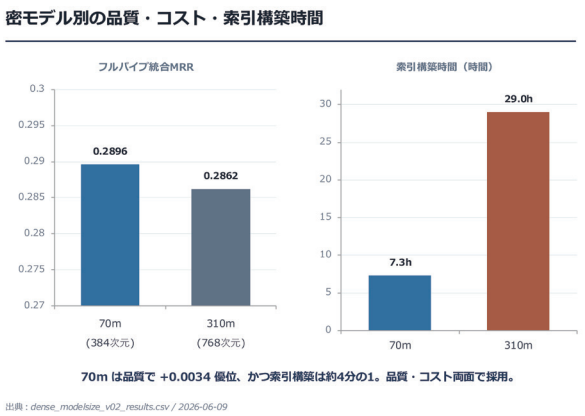

図 11　密モデル別の品質とコスト

両者は排他的ではない。疎検索は完全一致に、密検索は言い換えや同義に強く、誤りの傾向も異なる。両者は直交する信号として、後述の融合により互いを補完する。

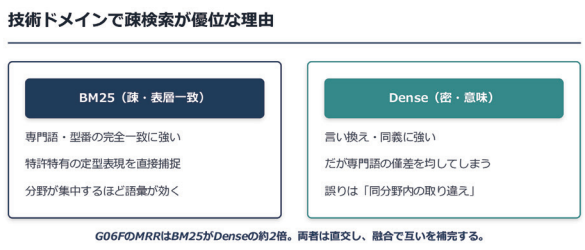

図 12　疎検索が優位な理由

5 親子チャンクによる構造的検索

請求項は複数の構成要件から成る。そこで、構成要件を子チャンク、請求項全文を親チャンクとし、子で検索して親へスコアを集約する構造的検索を導入した。細かい単位で一致を捉えつつ、請求項単位で評価する考え方である。

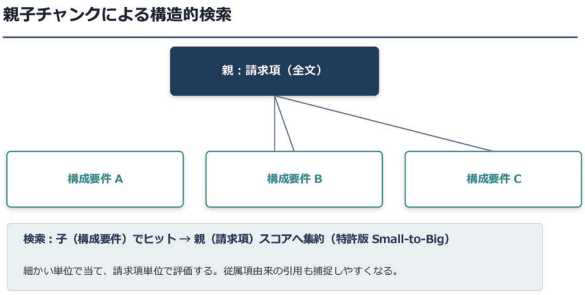

図 13　親子チャンクの構造

子から親への集約には、最大値、上位平均、親との線形結合、順位融合などの方式がある。最適な方式は手法に依存する。

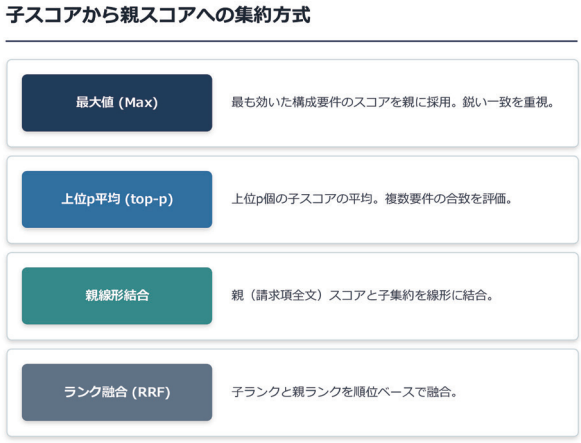

図 14　子から親へのスコア集約

親子検索により、BM25 は MRR が 6.2 パーセント、密検索は 14.3 パーセント改善した。最適な子の粒度は手法によって異なり、密検索は構成要件単位、BM25 は請求項単位が適していた。なお組み合わせ引用の層は標本数が 98 と少なく、検出力が不足するため、層に固有の効果は主張せず全体の再現率改善として報告する。

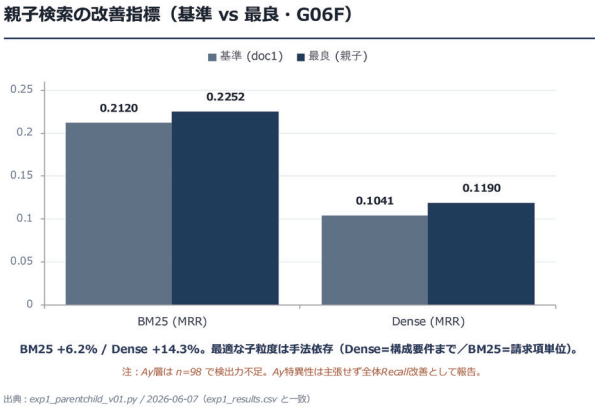

図 15　親子検索の改善

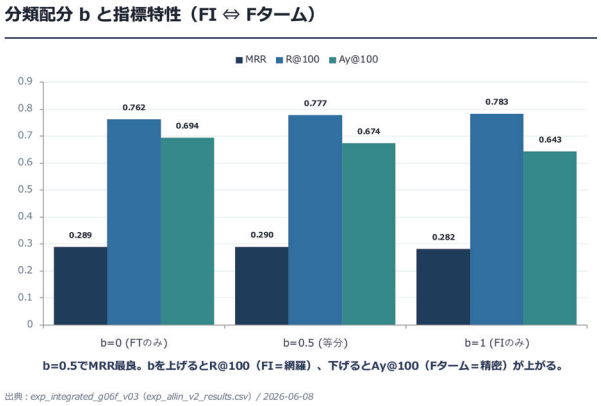

図 17　分類配分 b と指標特性

6 分類メタデータの融合

特許固有のメタデータである分類情報を検索に活用した。国際特許分類（IPC: International Patent Classification）を細分化した日本独自のファイルインデックス（FI: File Index）と、観点別に付与されるFターム（File Forming Term）を用いる。これらは審査官が文献に付与した構造化情報であり、一般的な文書検索にはない、特許固有の強力な手がかりである。クエリと文書の分類の共有度をスコアに加算するソフト融合方式を採った。

図 16　分類メタデータのソフト融合

分類の配分を、Fターム寄りからFI寄りまで変化させると、指標の特性が変わる。等分のときMRRが最良となり、FI寄りにすると網羅性の再現率が上がり、Fターム寄りにすると組み合わせ引用の捕捉率が上がる。FIは網羅、Fタームは精密という役割分担が観察された。

ここで、特定分野で高い精度が得られる要因を分離する実験を行った。全件に同一分類のフィルタを後段で適用した場合と、分野サブコーパスに限定した場合を比較すると、精度向上のうちフィルタそのものの効果は約2パーセントにすぎず、母集団の縮小による効果が約98パーセントを占めた。分野を限定する事前フィルタは、検索漏れを招く危険があるため、本研究ではハードな絞り込みを避け、ソフト融合を基本とした。

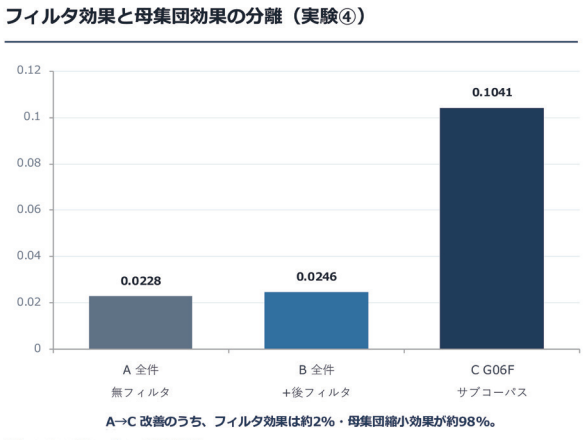

図 18　フィルタと母集団効果の分離

7 自前特徴語生成による商用 DB 非依存化

従来、検索精度の向上には商用データベースが提供する特徴語が用いられてきた。本研究では、公開データのみから特徴語を自動生成する仕組みを構築した。形態素解析[10]により語を切り出し、語の重要度を単語頻度・逆文書頻度（TF-IDF: Term Frequency–Inverse Document Frequency）で評価して特徴語を抽出する。

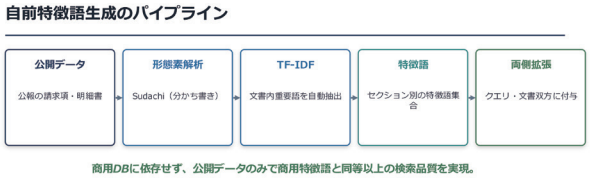

図 19　自前特徴語生成パイプライン

自前生成した特徴語を組み込んだ全部入り構成の MRR は 0.2928 であり、商用特徴語を用いた 0.2904 を上回った。公開データのみで商用水準を超えたことは、ライセンスに依存しない全件展開と顧客環境への配置を可能にする点で意義が大きい。

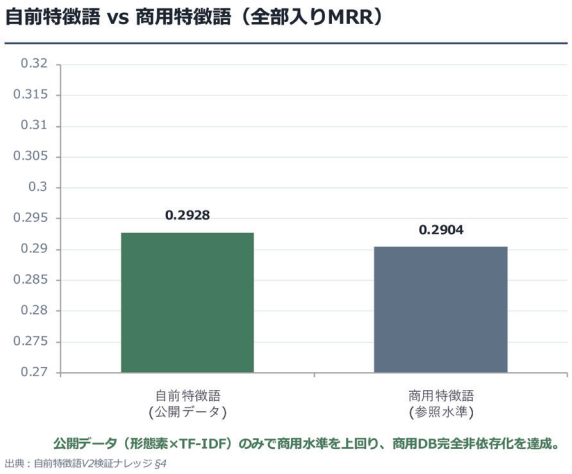

図 20　自前特徴語と商用の比較

生成した特徴語は、請求項・背景技術・課題など 7 つのセクションを横断し、総計 8 万 4 千行あまりに及ぶ。低頻度で識別性の高い語に集中しており、TF-IDF の性質と整合する。

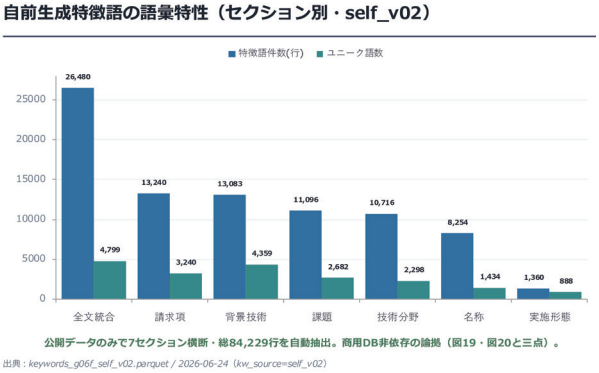

図 21　自前特徴語の語彙特性

8 ハイブリッド検索の統合と信号の直交性

以上の信号を統合したハイブリッド検索を構成した。BM25 を基底とし、親子・分類・特徴語を順に加えると、MRR は基底の 0.2120 から、最終的に 0.2904 へ、すなわち 37 パーセント向上した。自前特徴語版では 0.2928、38 パーセント向上となる。密検索単独の 0.1041 と比べると約 2.8 倍である。

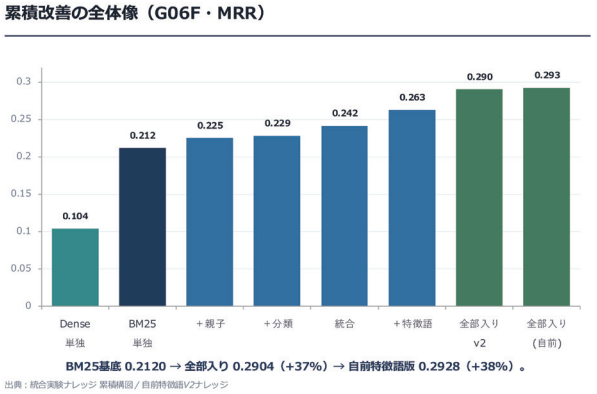

図 22　累積改善の全体像

注目すべきは、各信号がほぼ独立に効くことである。親子で 6.2 パーセント、分類で 7.8 パーセント、特徴語で 24.1 パーセントの寄与があり、その単純な合計は 38.1 パーセントとなる。実測の 37.0 パーセントとの差はわずか 1 ポイントであり、信号はほぼ直交し加法的に作用していると言える。

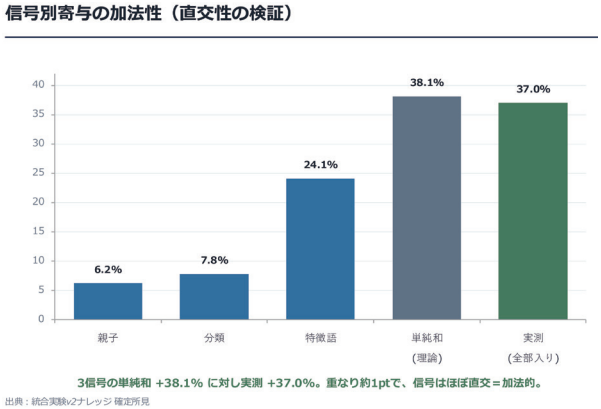

図 23　信号別寄与の加法性

この直交性は偶然ではない。各信号が依拠する情報源が異なるためである。BM25 は語の表層、密検索は意味、親子チャンクは請求項の構造、分類メタデータは審査官が付与した区分、特徴語は語の統計的な重要度に、それぞれ基づく。依拠する情報が重複しないため、信号は互

いに干渉せず加法的に作用する。この性質は実装上も大きな利点をもつ。各信号を独立に設計・改良でき、一つの信号の変更が他に波及しないため、システムの保守性と拡張性が高まる。統合の重みについては、最良条件の近傍で広い平坦域が観察された。上位 10 条件の MRR の差は 0.0026 以内であり、重みの選択に対して頑健である。順位に基づく融合手法 [8] も含め、融合の設計には複数の選択肢があるが、本研究では信号ごとのスコアを線形に統合する方式を中心に検討した。

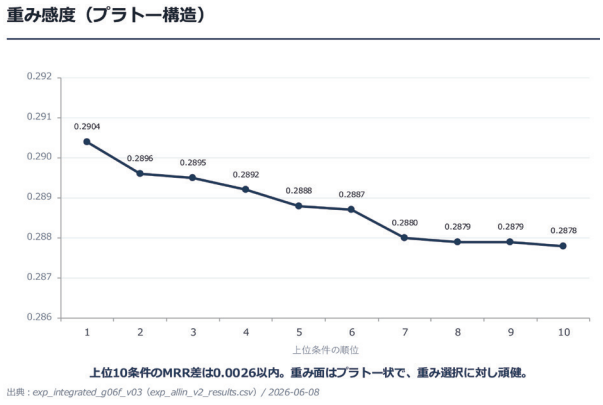

図 24　重み感度のプラトー

9 生成 AI による観点重み付けの検証と限界

ここまでの検索精度の向上に対し、生成 AI を用いてさらなる改善を試みた。具体的には、発明ごとに効く F タームや FI の観点を生成 AI に判定させ、その重要度を重みとして検索ランキングに反映する方式を検証した。

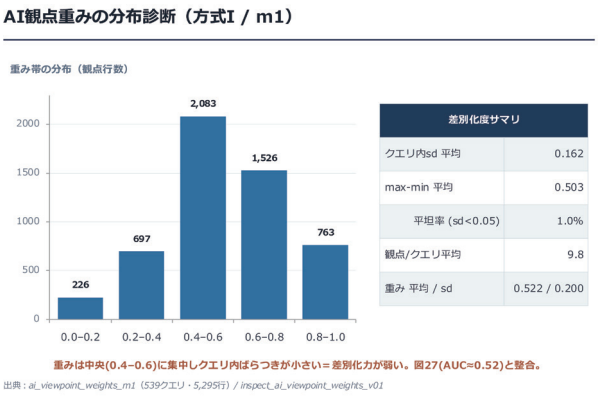

図 25　AI 観点重みの分布診断

しかし結果は、仮説に反するものであった。AI による観点重みは、検証したすべての条件で、素朴な共有数に基づく基準値 0.0827 を下回った。診断のため、観点ごとの重みが正解の共有を弁別できるかを曲線下面積（AUC: Area Under the Curve）で測ると、約 0.52 であった。これは無相関の水準であり、連続的な重み付けは正解の弁別に有効な情報をほとんど持たないことを示す。

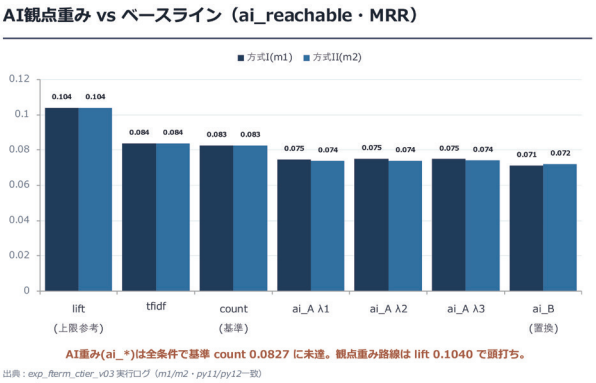

図 26　AI 重みとベースライン比較

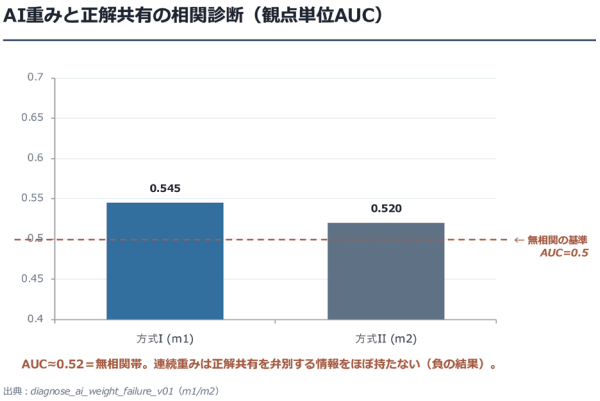

図 27　AI 重みと正解共有の相関

なぜ連続的な重み付けが効かなかったのか。観点の重要度は、調査対象の発明や文脈に応じて動的に変わるものであり、生成 AI が付与する静的な重みでは、その文脈依存性を捉えきれなかったと考えられる。この負の結果は、重要な示唆を与える。生成 AI の判断を、ランキングの連続的な重み付けに用いるのは適切でない。AI の判断は、むしろ調査範囲を決めるという、より上流の離散的な意思決定に向けるべきである。

10 審査官引用の大規模分析による観点共通性の実証

前章の負の結果は、生成 AI が観点の重要度を静的な重みとして当てられないことを示した。しかし、これは観点という軸そのものが無効であることを意味しない。観点が検索に有効かどうかと、その重要度を請求項の文面から予測できるかどうかは、別の問いである。この切り分けのため、審査官が実際に組み合わせ引用を行った

大規模な事例を用いて、組み合わせ引用がどのような共通性によって動機づけられるかを直接分析した。

分析には、特許庁が公開する拒絶理由条文コードを新たなデータ源として用いた。進歩性を否定する条文が通知された出願は、複数文献の組み合わせによって拒絶された事例であり、本研究の組み合わせ引用に対応する。この条文を持つ出願を約２万７千件同定し、それぞれの審査官引用文献と対にすることで、組み合わせ引用のペアを約５万７千組得た。これは、従来の評価枠組みにおける組み合わせ引用の正解 98 組に対し、約 585 倍の規模である。

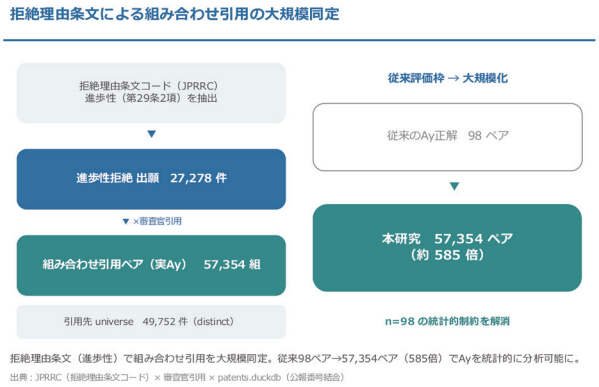

図 28　拒絶理由条文による組み合わせ引用の大規模同定

このペアについて、クエリと引用先が共有する分類観点の強さを、無関係なペアと比較した。結果は明瞭であった。F タームのテーマで約 62 倍、観点で約 72 倍、観点の重なりの割合では約 146 倍という、際立った共有の偏りが認められた。組み合わせ引用は、分類観点という統制された軸の共通性によって強く特徴づけられる。一方、請求項や背景技術といった本文の語彙の共有では、組み合わせ引用に固有の偏りは弱く、分類観点ほどの識別性を示さなかった。組み合わせ引用を駆動するのは、文面の表層的な語彙ではなく、統制された分類観点の共通性であると解釈できる。

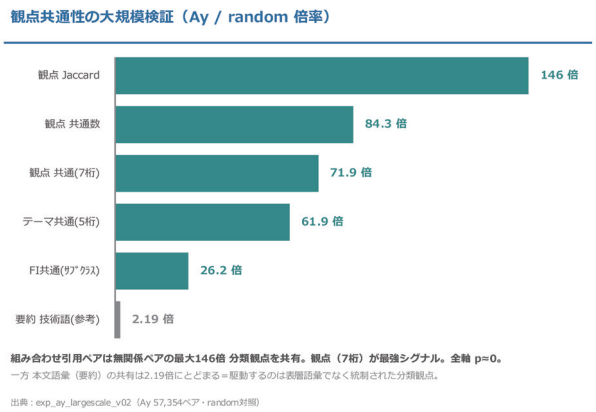

図 29　観点共通性の大規模検証

この知見は、前章の負の結果に整合的な説明を与える。観点共通性は確かに存在し、組み合わせ引用を強く特徴づける。しかしそれは、クエリと引用先という二つの文書の関係として現れるものであり、クエリの請求項の文面のみから静的に予測できるものではない。生成 AI に個々の観点の重みを当てさせる方式が有効でなかったのは、このためである。観点は、個別の重みとしてランキングに反映するのではなく、二文書の関係性をふまえた調査範囲の決定という、より上流の判断に活かすべきである。次章では、この方針に基づき、クエリの分類を予測して調査範囲を絞り込む構想を述べる。

11 分類予測による調査範囲の決定と実装展望

前章の知見を踏まえ、AI の判断を「読むか読まないか」という調査範囲の決定に用いる方向へ転換した。クエリの請求項から分類を予測し、その分類で調査範囲を絞り込む構想である。予測器は、既存の検索基盤で類似文献を集め、その分類を多数決する検索ベースの方式とした。モデルの学習を要さず、中央演算処理装置（CPU: Central Processing Unit）のみで完結する。

予備的な検証として、分類を予測する粒度を変えながら、予測精度と絞り込み効果の関係を測った。粗いセクションの粒度では予測精度の再現率が 0.998 と高い一方、絞り込み効果は 0.778 にとどまる。細かいサブグループの粒度では、絞り込み効果が 0.999 に達する一方、予測精度は 0.612 へ低下する。両者はトレードオフの関係にあり、サブクラスの粒度が予測精度 0.969、絞り込み効果 0.955 と、実用上の最適点を与えた。

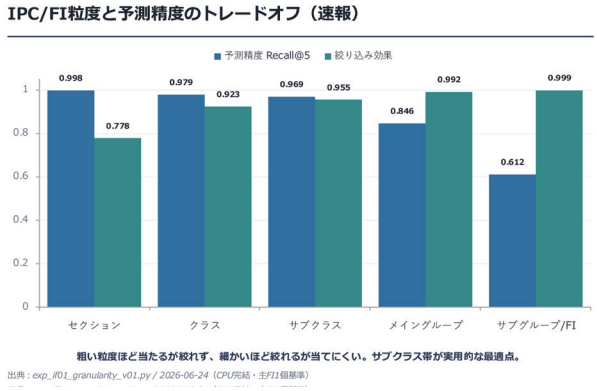

図 30　IPC/FI 粒度のトレードオフ

粒度が粗いほど予測は当たるが範囲を絞れず、細かいほど範囲は絞れるが予測が外れやすい。サブクラスがこの均衡点となるのは、特許の技術分野がこの粒度でおおむね一つの調査単位に対応するためと考えられる。実務における「この分野は読む、この分野は読まない」という判断の単位とも符合する。なお本検証は主分類 1 個を正解とする保守的な見積もりであり、複数分類を考慮すれば予測精度はさらに高まる。この検証は、別手法による分類予測の結果とも整合し、粒度ごとの予測可能性が手法によらず安定して現れることを確認した。重要なのは、絞り込みにおいて検索漏れを最小化することである。調査範囲を狭めれば効率は上がるが、関連文献を範囲外に取りこぼせば調査の信頼性が損なわれる。したがって絞り込みは、適合率よりも再現率を優先して設計する。

実装の観点では、3 層の配置を設計した。ベンダーの演算資源で密検索や索引構築を担う層、顧客のサーバーで疎検索や分類融合を担う層、そして顧客端末で CPU のみで軽量な検索と絞り込みを担う層である。疎検索・分類・特徴語はいずれも CPU で完結するため、端末層での運用が可能である。さらに、約 425 万件・約 316 ギガバイトの全件を、組み込み型の分析データベース [11] を用いて単一のファイルとして外付け記憶装置に格納すれば、演算資源をもたない携帯型計算機でも全件の検索と集計が行える。これにより、未公開の調査意図を外部に出すことなく手元で扱える機密性、ネットワークを介さない即応性、専用サーバーを要さない低コスト性が同時に得られる。

図 31　3 層デプロイとポータブル DB

12 特許庁一括データによる自律的調査基盤への統合

ここまでの検索・融合の枠組みは、手法の側では商用データベースからの独立を達成した一方、評価の正解データは商用データベース経由の審査官引用に依存したままであった。この最後の依存を解消する基盤として、特許庁が提供する公報一括データを導入した。一括データは総容量 13.8 テラバイト・9 千余ファイルのアーカイブであり、形式は 4 系統に分かれる。2022 年以降の新形式 XML、2004 年から 2022 年の旧形式 XML、1993 年から 2003 年の旧 SGML 形式、そして電子公報以前の公報をテキスト化した高精度テキストデータである。

最古の二系統は、当初は本文が画像で提供され文字認識を要すると見込んでいた。しかし全数の走査により、いずれも書誌・要約・請求の範囲・発明の詳細な説明が文字コード EUC-JP のテキストとして格納されており、画像は図面に限られることが判明した。テキストとして本文を取得できた公報の割合は、1993 年以降の系統で 99.5 パーセントを超え、高精度テキストデータでは誤認識の痕跡が一件も検出されなかった。文字認識を経ずに全期間の本文を扱えることが、全年代コーパスの構築を現実的なものとした。判定を実データによらず形式や容量から推測していた当初の見込みは、四つの系統すべてで覆された。

JPOデータ版コーパスの規模

項目	値	内容
一括データ 総容量	**13.8 TB**	9,319アーカイブファイル
形式系統	**4系統**	ST96／旧XML／旧SGML／高精度テキスト
収録年代	**1971〜2026年**	全期間を自前で網羅(1971〜1992一部収録)
取込済 総公報数	**16,335,543**	公開A・登録B・再公表S・特表T
ST96（2022〜）	1,937,559	特表T 214,174を含む
旧XML（2004〜2022）	9,310,561	再公表S・特表Tとも収録
旧SGML（1993〜2003）	4,810,316	本文テキスト99.9%・OCR不要
高精度テキスト（1971〜1995）	277,107	OCRノイズ0・PCT邦訳91.9%
審査官引用GT　※1	**25,671,558**	登録公報（INIDコード56欄）から自前構築
実効GTペア（評価可能）※2	**17,166,553**	citing・cited双方が在籍
拒絶理由条文コード	**988,203**	ユニーク出願・従来比 約26倍
登録公報（種別B）	**4,337,759**	従来コーパス未収録の種別

※1 審査官引用GT：登録公報の引用文献欄（INIDコード56欄）に記載された審査官引用を、引用元・引用先の公報番号ペアとして抽出した検索評価用の正解データ（GT：Ground Truth）。
※2 実効GTペア（評価可能）：※1のうち、引用元・引用先の双方の公報が本コーパス約1,634万件に在籍し、検索評価にそのまま使用できるペア。
出典：4系統横断棚卸し・#7取込後 実測（2026-07-19）

図 32　JPO データ版コーパスの規模（4 系統統合・実測）

第 2 章のコーパスが公開系の公報のみで構成されていたのに対し、一括データには、審査官が引用した文献を記載する登録公報が全期間含まれる。4 系統すべてを取り込み、公開·登録·再公表·公表の全種別で約 1,634 万件を単一の枠組みに構築した。収録年代は 1971 年から 2026 年におよぶ。構築にあたっては、第 2 章で述べたデータ品質の教訓を踏まえ、取り込み対象を文書種別の明示的な列挙で判定すること、公開・登録の別と引用の出所（審査官引用か明細書中の引用か）を第一級の属性として保持することを設計の柱とした。正解データの定義が「登録公報に記載された審査官引用」という 2 属性の組で機械的に定まるため、評価セットの構築が恣意を挟まず再現可能になる。

この設計のもと、登録公報の引用欄から審査官引用を約 2,567 万ペア抽出した。ただし正解データは、存在するだけでは評価に用いることができない。引用する側とされる側の双方が母集団に在籍して初めて、その事例は検索の評価に使える。この条件を満たす実効的な正解は 1,716 万ペアであった。注目すべきは、最古の系統を加える前の実効的な正解が約827万ペアにとどまっていた点である。1971 年から 2003 年の公報を母集団に加えたことで、実効的な正解は 2 倍を超えて増加した。増加分はすべて 2004 年より前の被引用文献であり、従来は正解として記録されていながら母集団に存在しないために評価に使えなかった事例が、母集団の拡張によって評価可能になったことを意味する。旧年代の公報の取り込みは、単に古い文献を追加したのではない。検索の評価という営みそのものの土台を、年代の面で完備させた。

一方で、母集団には依然として到達できない範囲が残る。審査官引用のうち、引用先が国内移行の PCT 公表公報である事例の一部と、実用新案である事例の大半は、本稿の母集団に収録できていない。その主たる理由は、当該年代の全文データが特許庁の提供データに含まれていないことにあり、追加の取り込みでは解消しない構造的な制約である。この限界は自前基盤に固有のものではなく、同じ年代の全文を保持しない限りいかなる評価基盤にも生じうる。母集団の範囲を明示することは、評価結果の解釈可能性を担保するうえで不可欠であり、本稿では到達可能な範囲に限って評価を行っている。

自前の正解が商用データベース由来の正解を代替できるかを、突合により検証した。第 3 章の評価セットの正解 1,434 ペアに対し、全体の 74.4 パーセントが自前の正解と一致した。不一致の主因は、クエリ出願の登録公報がまだ発行されていない（審査係属中である）ことによる構造的なものである。登録公報が存在する範囲に限れば一致率は 97.1 パーセントに達し、X 引用で 97.0 パーセント、組み合わせ引用で 98.4 パーセントであった。引用の実体が異なる不一致は全体の 2.2 パーセントにすぎない。さらに、別の商用データベースから得た約 9.3 万件の引用ペアとの突合でも、同じ条件で 98.2 パーセントが一致した。すなわち自前の正解は、登録公報が存在する範囲において商用の正解を実質的に置換できる。ここに、コーパス・特徴語・正解データの三位一体で、公開データのみで完結する評価・検索基盤が成立した。評価の正解までを公開データで賄えることは、本稿の定量的な主張のすべてを、ライセンスに依存せず第三者が追試できることを意味する。

一括データは、第 10 章の分析基盤も拡張する。拒絶理由条文コードの全期間版は 988,203 出願（2015 年から 2026 年）を含み、従来の約 26 倍の規模である。約 2,567 万ペアの審査官引用は、評価の正解にとどまらず、組み合わせ引用の構造分析や、主引例と副引例の関係を学習するための大規模な素材となる。

とりわけ有望と考えるのは、条文コードと審査官引用の結合が開く進歩性の構造分析である（図 33）。全期間版の条文コードには、進歩性の欠如（特許法第 29 条第 2 項）を理由とする拒絶が約 77 万出願含まれる。これを出願番号を介して審査官引用と結合すれば、進歩性の否定に用いられた組み合わせ引用の実例を、従来の百件規模から数万規模へ拡大して観察できる。本研究の予備的な分析では、組み合わせ引用の文献は、単一文献で新規性を否定する引用に比べて技術分野を表す語の共通性が相対的に強い傾向が観察されているが、事例数の制約から統計的な確証には至っていない。結合後の大規模な実例集合を用い、分野分類（FI・F ターム）の共通性がどの階層まで成立するか、また課題・効果と構成・成分とで観点別の共通性に非対称があるかを検証する。組み合わせ引用の計算的な構造が明らかになれば、無効化調査における副引例の探索を、単一文献の類似度が高い順という従来の原理とは異なる軸で支援する道が開け、拒絶理由の分析を検索機能そのものへ還流できる。

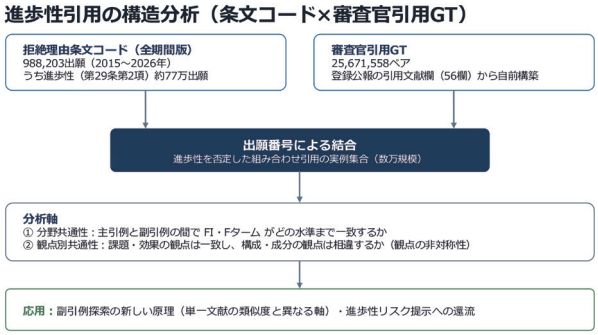

図 33　条文コードと審査官引用の結合による進歩性引用の構造分析

図 34　自律的調査基盤への統合ロードマップ（RM-0 〜 RM-4）

実装の面では、第 11 章で述べた 3 層の配置が具体化する（図 31）。顧客サーバーの層には自前の正解を搭載した本体データベースを置き、顧客端末の層には本文を列指向ファイルに分離した軽量構成を外付け記憶装置で配置する。ここに、2026 年の人工知能の潮流が接続する。生成 AI は対話的な応答の段階から、目標に向けて計画し行動するエージェントとしての実運用段階へ移行しつつある。同時に、小型のオープンウェイトモデルが手元の計算機で実用水準の推論を行えるようになり、機密データを外部に送信せずに言語モデルを運用する構成は、もはや特殊な選択ではなくなった。本研究の設計原則——生成 AI の判断は調査範囲の決定という離散的な意思決定に限定し（第 9 章）、検索と評価は公開データで完結させる——を端末上の小型モデルと組み合わせれば、未公開の調査意図を一切外部に出さないまま、「判断する調査」を手元で完結させる道筋が描ける。

その具体像として、検索上位の公報を入力に、調査観点への回答生成、請求項の構成要件への分解、構成要件と候補文献の対比といった検索後の工程を、16 ギガバイト級の民生用 GPU で動作する小型モデルへ担わせる構成の検証を開始している。回答には公報番号・請求項番号との紐付けを必須とし、根拠の検証可能性を評価の軸とする。定型性の高い構成要件分解については、大規模モデルの出力を教師とする蒸留で小型モデルへ移管できる見込みがあり、成立すれば、外部への通信も運用費用も要しない調査支援が顧客端末の層で完結する。

図 34 に、三位一体の独立から全期間コーパスの統合、分類予測による範囲決定、エージェント型の調査実装、そしてローカル完結の AI 調査へ至るロードマップを示す。

13 おわりに

本稿は、大規模特許コーパスの構築から、ハイブリッド検索の統合、生成 AI による重み付けの限界、そして分類予測による調査範囲決定までを、実装と評価の観点から総説した。ハイブリッド検索は基底に対し 37 パーセントの精度向上をもたらし、公開データのみで商用水準を超える特徴語生成も実現した。生成 AI による連続的な重み付けは有効でなかった一方、審査官引用の大規模分析により、組み合わせ引用が分類観点の共通性によって強く特徴づけられることを実証した。これらの知見は、AI の判断を個々の重みではなく、調査範囲の決定という離散的な役割に向けるべきことを示している。

さらに、特許庁一括データから登録公報を含む約 1,634 万件・1971 年から 2026 年におよぶコーパスと、約 2,567 万ペアの審査官引用を自前構築し、既存評価セットとの突合で登録公報が存在する範囲の 97.1 パーセントの一致を確認した。引用元と引用先の双方が母集団に在籍する実効的な正解は 1,716 万ペアに達し、旧年代の公報を加えたことでその数は 2 倍を超えた。コーパス・特徴語・正解データの三位一体で、公開データのみで完結する自律的な調査基盤が成立した。

「読む調査から判断する調査へ」という移行は、AI に何をさせ、人間が何に集中するかという役割分担の再設計にほかならない。検索の精度を高め、AI に調査範囲を判断させ、人間が最終的な評価に集中する。この分担を実務システムとして実装し、その有効性を検証することが、今後の課題である。

14 謝辞

本報告は 2026 年度の「アジア特許情報研究会」のワーキングの一環として報告するものである。研究会のメンバーの皆様には様々な協力をしていただきました。ここに改めて感謝申し上げます。

参考文献

[1]Robertson, S.E., Walker, S.: Some Simple Effective Approximations to the 2-Poisson Model for Probabilistic Weighted Retrieval. Proc. SIGIR 199　4, pp.232-241 (1994).

[2]Robertson, S., Zaragoza, H.: The Probabilistic Relevance Framework: BM25 and Beyond. Foundations and Trends in Information Retrieval, 3(4), pp.333-389 (2009).

[3]Tsukagoshi, H., Sasano, R.: Ruri: Japanese General Text Embeddings. arXiv:2409.07737 (2024).

[4]Lewis, P., et al.: Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. Advances in Neural Information Processing Systems 33 (NeurIPS 2020), pp.9459-9474 (2020).

[5] Karpukhin, V., et al.: Dense Passage Retrieval for Open-Domain Question Answering. Proc. EMNLP 2020, pp.6769-6781 (2020).

[6]Reimers, N., Gurevych, I.: Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. Proc. EMNLP-IJCNLP 2019, pp.3982-3992 (2019).

[7] Warner, B., et al.: Smarter, Better, Faster, Longer: A Modern Bidirectional Encoder for Fast, Memory Efficient, and Long Context Finetuning and Inference. arXiv:2412.13663 (2024).

[8]Cormack, G.V., Clarke, C.L.A., Buttcher, S.: Reciprocal Rank Fusion Outperforms Condorcet and Individual Rank Learning Methods. Proc. SIGIR 2009, pp.758-759 (2009).

[9]Manning, C.D., Raghavan, P., Schutze, H.: Introduction to Information Retrieval. Cambridge University Press (2008).

[10]Takaoka, K., et al.: Sudachi: a Japanese Tokenizer for Business. Proc. LREC 2018, pp.2　Idt, M., Muhleisen, H.: DuckDB: an Embeddable Analytical Database. Proc. SIGMOD 2019, pp.1981-1984 (2019).