26º ENANCIB
Encontro Nacional de Pesquisa e Pós-Graduação em Ciência da Informação
A Amazônia na sociedade global em rede: assimetrias e potencialidades informacionais, tecnológicas e ambientais
Belém, Pará, 09 a 13 de novembro de 2026
Universidade Federal do Pará

XXVI ENCONTRO NACIONAL DE PESQUISA E PÓS-GRADUAÇÃO EM CIÊNCIA DA INFORMAÇÃO - XXVI ENANCIB

ISSN 2177-3688

GT 2 – Organização e Representação do Conhecimento

LINGUAGENS DOCUMENTÁRIAS E INTELIGÊNCIA ARTIFICIAL PARA RECUPERAÇÃO DA INFORMAÇÃO: PROPOSTA DE UM MODELO BASEADO EM *RETRIEVAL-AUGMENTED GENERATION*

DOCUMENTARY LANGUAGES AND ARTIFICIAL INTELLIGENCE FOR INFORMATION RETRIEVAL: A RETRIEVAL-AUGMENTED GENERATION-BASED APPROACH

Juliana Cristhine Schifino Santos Jardim – Universidade Federal do Rio Grande do Sul (UFRGS)
Rene Faustino Gabriel Junior – Universidade Federal do Rio Grande do Sul (UFRGS)

Modalidade: Trabalho Completo

Resumo: A organização do conhecimento e a recuperação eficiente da informação constituem desafios em ambientes científicos digitais, especialmente em domínios especializados que demandam representação terminológica consistente e mecanismos avançados de descoberta. Nesse contexto, este estudo propõe o modelo SmartRetrieval, que integra a arquitetura de *Retrieval-Augmented Generation* a modelos de linguagem para aprimorar a recuperação da informação em bases especializadas. Trata-se de uma pesquisa aplicada, de abordagem qualitativa e caráter metodológico. Como parte da proposta, foi desenvolvido um microtesauro multilíngue a partir de um recorte de artigos do capítulo brasileiro da *International Society for Knowledge Organization*, destinado à expansão semântica de consultas e à normalização terminológica. O SmartRetrieval foi avaliado em um *corpus* composto por 40 artigos científicos, utilizando as métricas de precisão, revocação e F1-Score. O sistema emprega estratégia de busca orientada por domínio no *Elasticsearch* e utiliza reclassificação baseada em Modelos de Linguagem de Pequena Escala para aumentar a relevância dos resultados recuperados. Os resultados demonstram que a integração entre linguagens documentárias e modelos generativos reduz ruídos, amplia a cobertura semântica e melhora a eficácia da recuperação da informação. Evidenciam, ainda, a relevância da mediação intelectual especializada na construção e validação de estruturas conceituais, uma vez que a qualidade da representação temática influencia diretamente no desempenho dos sistemas de busca. Conclui-se que a abordagem proposta contribui para o aperfeiçoamento dos processos de organização, representação e recuperação da informação, fortalecendo a mediação informacional em ambientes científicos digitais.
Palavras-chave: Retrieval-Augmented Generation; SmartRetrieval; Linguagem documentária.

Abstract: Knowledge organization and effective information retrieval are challenges in digital scientific environments, particularly in specialized domains that require consistent terminological representation and advanced discovery mechanisms. In this context, this study proposes the SmartRetrieval model, which integrates the Retrieval-Augmented Generation architecture with language models to improve information retrieval in specialized databases. This is an applied study with a qualitative approach and a methodological character. As part of the proposal, a multilingual microthesaurus was developed from a selected corpus of the International Society for Knowledge Organization (brazilian chapter) papers to support semantic query expansion and terminological normalization. SmartRetrieval was evaluated using a corpus of 40 scientific articles and assessed through precision, recall, and F1-score metrics. The system employs a domain-oriented search strategy in Elasticsearch and uses Small Language Model-based re-ranking to increase the relevance of retrieved results. The findings demonstrate that the integration of documentary languages and generative models reduces noise, expands semantic coverage, and improves the effectiveness of information retrieval. Furthermore, they highlight the importance of specialized intellectual mediation in the construction and validation of conceptual structures, since the quality of subject representation directly influences the performance of retrieval systems. It is concluded that the proposed approach contributes to the improvement of information organization, representation, and retrieval processes, strengthening information mediation in digital scientific environments.
Keywords: Retrieval-Augmented Generation; SmartRetrieval; Documentary Language.

1 INTRODUÇÃO

A popularização de modelos de Inteligência Artificial Generativa (IAG), no final de 2022, e o crescimento na produção de informações digitais, impuseram desafios à Ciência da Informação (CI), intensificando os debates no campo da Organização do Conhecimento (OC) acerca do uso da IAG na Recuperação da Informação (RI). Muito já se investigava, desde os estudos de Luhn (1958), sobre possibilidades de inovação no apoio à OC. Além de propostas de indexação e recuperação automáticas, foram desenvolvidos vocabulários controlados, técnicas de indexação e sistemas de recuperação cada vez mais sofisticados (Robredo; Ferreira, 1980; Lancaster, 2004). E a evolução tecnológica tem permitido que esses estudos persistam ainda hoje (Manning; Raghavan; Schütze, 2008).

Contudo, ainda não se dispunha de uma solução plenamente funcional para atender às demandas específicas do campo da OC (Baeza-Yates; Ribeiro-Neto, 2013). Apesar dos avanços tecnológicos e do desempenho dos mecanismos de busca modernos, ainda persistem limitações na compreensão semântica (De Luca; Fallucchi; Morelato, 2021). Embora os Modelos de Linguagem de Grande Escala (LLMs) da IAG demonstrem forte capacidade de interpretação e processamento de linguagem natural (PLN), sua aplicação à OC ainda é insatisfatória,

especialmente devido à escassez de dados específicos de domínio em seus conjuntos de treinamento. O resultado: a geração de informações imprecisas ou inadequadas para domínios especializados (Narayanan; Kapoor, 2024).

Nesse contexto, bases de dados especializadas, como a Base de Dados em Ciência da Informação (Brapci), enfrentam desafios relacionados à completude e à qualidade dos metadados bibliográficos. O grande volume de publicações anuais inviabiliza a indexação manual e a validação sistemática dos metadados (como títulos, resumos, autoria e palavras-chave), muitas vezes ausentes ou inadequados, afetando negativamente a precisão na RI.

Para mitigar tais problemas, arquiteturas de Retrieval-Augmented Generation (RAG) surgem como alternativa promissora, ao permitir que modelos de linguagem consultem bases de conhecimento externas e curadas antes de gerar respostas (Lewis et al., 2020). Ao combinar as capacidades dos LLMs com um conhecimento especializado se reduz imprecisões e alucinações (Islam, 2023). Ainda, encontramos entre os LLMs, modelos de linguagem "menores", também chamados de Modelos de Linguagem de Pequena Escala (SLMs), os quais, de acordo com Hoffmann et al. (2022), se bem treinados, superam, em seu desempenho, LLMs com menos treinamento.

Diante desses desafios, apresenta-se a questão: de que forma um sistema baseado em RAG, integrando microtesauros especializados e SLMs, pode aumentar a precisão e eficácia da RI em bases de dados especializadas?

O objetivo geral deste estudo é propor um método para RI em bases de dados, aqui chamado SmartRetrieval. Implementado por meio de arquitetura RAG, o modelo integra microtesauros especializados a SLMs open source.

Para isso, propõe-se especificamente: a) analisar o SmartRetrieval por meio da recuperação automática de uma amostra de artigos do capítulo brasileiro da International Society for Knowledge Organization (ISKO-Brasil) na Brapci, comparando seus resultados com os de um sistema de busca convencional; e b) avaliar a eficácia dessa abordagem.

A justificativa do estudo apoia-se em quatro dimensões principais. A primeira é teórica, ao investigar como alinhar tecnologias de IAG aos princípios da OC (Christian, 2020). A segunda é tecnológica, ao propor uma solução acessível baseada em RAG e modelos open source para superar as limitações da indexação manual e das IAs generalistas (Narayanan; Kapoor, 2024). A terceira é prática e social, pois busca melhorar a qualidade dos metadados e a RI em uma das principais bases de dados da CI no Brasil. Por fim, a pesquisa também é motivada por interesses

profissionais e pessoais dos autores na área de OC, bem como pela busca de soluções que auxiliem indexadores, pesquisadores e usuários na localização de informações relevantes.

2 REFERENCIAL TEÓRICO

A OC compreende os processos e instrumentos voltados à representação e organização do conhecimento (RC e OC) e RI, sendo os Sistemas de Organização do Conhecimento (SOCs), como tesauros, taxonomias e ontologias (Nascimento; Pinho, 2019), os principais tipos a serem considerados para a padronização terminológica e a mediação informacional. Essas estruturas permitem reduzir ambiguidades, estabelecer relações semânticas entre conceitos e aprimorar a RI por meio de vocabulários controlados (Almeida, 2020; Lancaster, 2004), constituindo uma base consolidada para a RC em domínios especializados.

A seleção e a estruturação de termos em SOCs fundamentam-se em princípios clássicos e contemporâneos da teoria do conceito e do controle terminológico (Campos, 2001; Dahlberg, 1978; Herdman; Fox-Rushby; Badia, 1998), onde: a) os termos representam estruturas conceituais estabilizadas no discurso científico; b) se utiliza a análise de frequência dos termos como critério para a seleção de expressões preferenciais; c) os processos de normalização, agrupamento e alinhamento multilíngue são igualmente sustentados pelos princípios de equivalência semântica; d) a adoção de estratégias de normalização; e) a delimitação do corpus e a aplicação de estratégias controladas de busca priorizando a literatura científica recente. Tais procedimentos viabilizam a integração semântica com modelos computacionais, especialmente em ambientes que combinam vocabulários estruturados e técnicas de IA.

Inicialmente baseados em métodos estatísticos, os sistemas automáticos de RI surgiram da necessidade de aprimorar a organização da produção científica, especialmente em resposta ao crescente volume de documentos. Contudo, as contribuições de Lancaster (2004) e Richmond (1974) estabelecem a indexação automática apenas como recurso complementar às práticas manuais, ampliando a precisão, a profundidade e a flexibilidade da RI. A partir da década de 1980, os estudos nessa área passaram a se concentrar em sistemas capazes de identificar termos significativos por meio da análise de conteúdo (Robredo; Ferreira, 1980). Tais estudos vieram a evoluir na direção de incorporar abordagens linguísticas e semânticas capazes de lidar com as ambiguidades da linguagem natural (Andreewski; Ruas, 1983). Ao longo das últimas décadas, diferentes técnicas foram propostas para melhorar a precisão e a revocação

dos sistemas de RI, culminando na adoção de modelos neurais capazes de capturar relações contextuais entre termos e documentos.

Nas décadas de 1990 e 2010, os avanços em IA, Linguística Computacional e RC, ocorridos (Brito, 1992; Correa; Kobashi, 2010, Silva; Milidiú, 1991), contribuíram significativamente para a evolução da RI, impulsionando a adoção de técnicas de PLN, vetores semânticos, modelos probabilísticos e lógicas alternativas para lidar com incertezas e contradições. Esses desenvolvimentos objetivam maior qualidade da RC, aprimorando métricas clássicas da RI, como precisão e revocação (Manning; Raghavan; Schütze, 2008). Além disso, Mamfrim (1991) ainda reforça a importância da RC na construção de sistemas automáticos. A autora avalia as possibilidades técnicas da indexação feita por máquina, destacando limitações nos modelos tradicionais quando aplicados ao conteúdo textual complexo de textos científicos e acadêmicos (Mamfrim, 1991).

Paralelamente, a RI deixou de ser um campo predominantemente acadêmico para tornar-se um mecanismo de acesso à informação (Manning; Raghavan; Schütze, 2008), impulsionado pelos mecanismos de busca na web, pela Web Semântica e pelas infraestruturas de computação em nuvem (Alexander, 2003; Viana; Maimone, 2025). Modelos tradicionais, como TF-IDF (Term Frequency–Inverse Document Frequency) e o modelo vetorial, estabeleceram as bases da busca semântica (Manning; Raghavan; Schütze, 2008), mas permaneceram limitados por sua incapacidade de compreender contextos mais amplos. Nesse meio, Alcaide et al. (2001) destacam a importância da consistência na RI, evidenciando os desafios ainda persistentes na padronização dos descritores e na equivalência semântica dos termos atribuídos automaticamente em ambientes digitais especializados.

Mais recentemente, com os LLMs, baseados em arquitetura Transformer, inaugurou-se um novo paradigma para a RI. Treinados em grandes corpora textuais, LLMs demonstram elevada capacidade de interpretação e geração de linguagem natural (Géron, 2021), representando avanço significativo. Contudo, sua tendência a produzir conteúdos plausíveis, mas nem sempre verificáveis, gera riscos de alucinações, especialmente em contextos acadêmicos e científicos (Narayanan; Kapoor, 2024). Essas limitações, somadas à sensibilidade dos resultados e ao surgimento de comportamentos imprevisíveis em larga escala, evidenciam a necessidade de mecanismos que fundamentem e restrinjam as respostas produzidas.

Como alternativa para mitigar essas limitações, destaca-se a arquitetura RAG, cunhada por Lewis (2020), que combina modelos de IAG com bases de conhecimento externas. Nessa

abordagem, informações relevantes são recuperadas de uma fonte especializada antes da geração da resposta, proporcionando precisão, rastreabilidade e atualização do conhecimento. Desta forma, percebe-se que a eficácia do RAG depende diretamente da qualidade da base consultada (Shang et al., 2025). Lancaster (2004) complementa, ao dizer que a indexação de qualidade, mediada por linguagens controladas, é o que garante a relevância na RI, tornando sua utilização em estratégia promissora para aplicações em RI e indexação automática.

Sistemas baseados em RAG representam um avanço metodológico coerente com a trajetória histórica da RI. Eles oferecem estrutura flexível, auditável e escalável (Islam, 2023) para aprimorar a consistência, a especificidade e a relevância da geração de metadados na OC (Gao et al., 2023).

Além dos aspectos técnicos, a aplicação de IA em sistemas de organização e RI demanda atenção a questões éticas relacionadas a vieses algorítmicos, transparência e responsabilidade (Christian, 2020). Nesse sentido, a curadoria humana e o uso de bases de conhecimento controladas contribuem para reduzir riscos de RC inadequada e para aumentar a confiabilidade dos resultados produzidos pelos sistemas inteligentes. Dessa forma, a abordagem evidencia a relevância dos profissionais da informação, cuja expertise possibilita a produção e a integração de conhecimento estruturado às capacidades dos LLMs.

3 PROCEDIMENTOS METODOLÓGICOS

O presente estudo possui sua natureza aplicada e abordagem qualitativa. Quanto aos seus objetivos, é classificada como exploratória, metodológica e propositiva. A formulação do SmartRetrieval observou a norma da ABNT - NBR 12676 (Associação Brasileira de Normas Técnicas, 1992) e os modelos de RI físicos (Cesarino, 1985) e digitais (Manning; Raghavan; Schütze, 2008). Dessa forma, procurou-se conciliar as distintas características e modelar um sistema que atendesse a bases de dados digitais, mantendo os princípios da OC e da RI.

Este estudo propõe avaliar o modelo integrado denominado SmartRetrieval, capaz de aprimorar sistemas de RI por meio do aumento simultâneo da precisão e da revocação. O corpus da pesquisa foi constituído por 20 artigos relacionados à IA, selecionados com base em sua pertinência temática ao objeto investigado. Como fonte, e por ser uma base especializada na área de CI, optou-se pela realização da busca na Brapci. A seleção não se restringiu à presença explícita do termo nos metadados, contemplando também trabalhos cujo conteúdo abordasse conceitos, aplicações ou discussões associadas à área. Como critérios de delimitação,

estabeleceu-se um recorte contextual correspondente aos artigos publicados em eventos da ISKO Brasil e um recorte temporal, neste caso, limitado às publicações do ano de 2025. Compondo uma amostra de controle, ainda foram selecionados outros 20 artigos que não apresentavam relação direta com a temática da IA. Esses, foram escolhidos dentre os artigos publicados no congresso da ISKO Brasil de 2019, permitindo a comparação entre documentos alinhados e não alinhados ao domínio temático investigado. Assim, o corpus final totalizou 40 artigos, distribuídos igualmente entre o grupo experimental e o grupo de controle (2025 e 2019 respectivamente).

A partir desses 40 documentos, e seguindo a premissa da relevância científica de Dahlberg (1978), foram extraídas palavras-chave para a construção de um microtesauro, a ser utilizado como base externa no RAG. Complementarmente, analisou-se a frequência de ocorrência dos termos (Campos, 2001), contribuindo à definição de termos preferenciais e à organização hierárquica e ontológica do microtesauro. Sua formalização se deu por meio da ferramenta Thesa, onde também foram incluídas relações semânticas (termos genéricos e específicos) e variações linguísticas em português, inglês e espanhol. O conjunto resultante abarca 286 conceitos (120 termos em inglês, 48 em espanhol e 118 em português), constituindo um microtesauro multilíngue.

Para dar início ao SmartRetrieval, foram propostos prompts a partir da metodologia de Manning, Raghavan e Schütze (2008), que aponta a formulação de testes baseados em possíveis necessidades e preferências do público-alvo. O prompt estruturado (ver Figura 1) incorpora o microtesauro dinâmica e automaticamente a cada consulta, permitindo que o modelo fundamente suas respostas em conhecimento especializado do domínio. Para tanto, foram desenvolvidos scripts em python, executados no Jupyter Notebook e aprimorados pelo sistema de “erro e acerto”. Ainda, visando mitigar alucinações e aumentar a confiabilidade das respostas, o modelo foi configurado com temperature=0 e top_p=0,1, restringindo o modelo.

Figura 1 - Prompt usado no modelo SmartRetrieval.

```
prompt = f"""
SYSTEM:
Você é um bibliotecário especializado.
Regras:
• Classifique a pergunta usando APENAS os [TERMOS AUTORIZADOS] mais relevantes.
• Identifique APENAS conceitos centrais.
• NÃO explique suas escolhas.
• NÃO repita os termos extraídos.
• NÃO use frases completas.
• NÃO invente termos.
• TERMOS em plural transforme para singular.
• Selecione apenas termos específicos.
• Retorne APENAS termos conceituais curtos, separados por vírgula.
• NÃO use o termo "indexação" ou "indexação" (ou variações) como resposta, mesmo que seja relevante.

TERMOS AUTORIZADOS:
{terms}

USER:
Extraia os termos relevantes da pergunta: "{question}"
"""
```

Fonte: elaboração própria (2026).

A fim de operacionalizar o uso de LLMs na identificação de termos relevantes, foram selecionados modelos com capacidade de execução local em um servidor equipado com GPU de 16 GB de memória VRAM, garantindo a viabilidade da aplicação em ambiente computacional controlado. Dessa forma, foram avaliados três modelos: *LLaMA 3.2*, *Gemma 3* e *Phi-4* (Quadro 1). Dentre esses, o SLM *LLaMA 3.2:3b* apresentou desempenho superior em termos de interpretação semântica e extração de conceitos, especialmente devido à sua janela de contexto ampliada de até 128 mil *tokens*, o que possibilitou uma análise mais consistente do *corpus*.

Quadro 1 – Modelos testados e suas especificações.

Família	Modelo	Tamanho físico	Parâmetros	Limite de contexto	Input
Llama3.2	Llama3.2:3b	2.0GB	3.21B	128k	texto
Gemma3	Gemma3	3.3GB	4.3B	128K	texto e imagem
Phi4	Phi4	9.1GB	14.7B	16K	texto

Fonte: elaboração própria (2026)

Por fim, cada teste foi composto pela aplicação da metodologia RAG na *BrapciLabs*, por meio do aplicativo *Ollama*, para a adição do conhecimento externo; o SLM foi inserido na

plataforma com a finalidade de ser treinado, customizado e testado para a RI com a aplicação de *prompts*; e o *Elasticsearch*, para formação da estratégia de busca.

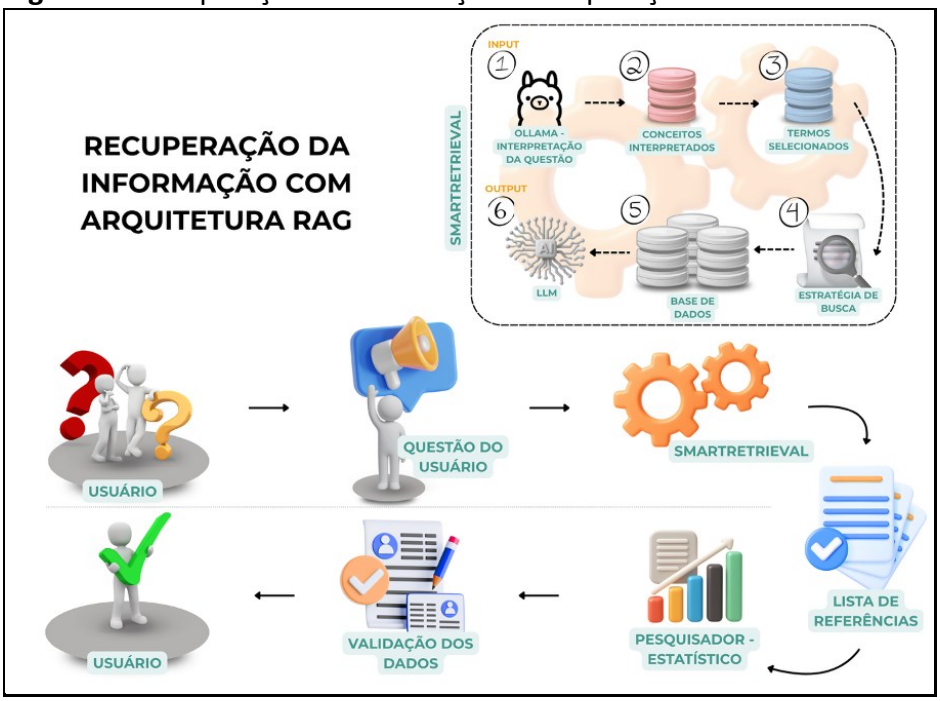
O estudo baseou-se em análise documental de acesso público (Brapci/ISKO-Brasil). Foram respeitadas as licenças aplicáveis aos PDFs, bem como as boas práticas de armazenamento seguro dos arquivos. Não houve tratamento de dados pessoais sensíveis. Artefatos derivados (planilhas, *logs*) não incluem material protegido além do estritamente necessário para fins acadêmicos (Yin, 2016).

4 RESULTADOS

O método SmartRetrieval está exemplificado na Figura 2, mostrando sua aplicação em busca em bases de dados. Por ser parte do sistema, não modifica o *layout* ou a forma de acesso de ferramentas de busca convencionais já existentes. Para o usuário o sistema permanece o mesmo, diferenciando-se apenas no resultado fornecido, mais enriquecido.

O SmartRetrieval, mediador na base de dados, por meio do *Ollama* (etapa 1 do SmartRetrieval), interpreta conceitos a partir das ideias principais da questão do usuário. Na arquitetura RAG (entre as etapas 2 e 3), tais termos são comparados aos conceitos encontrados no microtesauro fornecido. Dessa forma, termos são selecionados (etapa 3), ficando os que encontram correspondência com os termos da busca, acompanhados de suas relações mais próximas (hierárquicas e associativas), enriquecendo o contexto da procura. O *Elasticsearch* (etapa 4) reformula a pergunta em formato de expressão booleana e com valores atribuídos aos termos (quanto mais alto na hierarquia do microtesauro, maior o valor). Na expressão booleana, entre conceitos de uma mesma raiz, usa-se *OR*; para conceitos de raiz distinta, *AND*. Tendo pronta a estratégia de busca, ela é aplicada à base de dados (etapa 5) formando uma lista de referências (etapa 6) que é validada e entregue em resposta ao usuário.

Figura 2 - Recuperação da Informação com aplicação do SmartRetrieval



Fonte: elaboração própria (2026)

No Quadro 2 são apresentados quatro casos presenciados durante os testes, demonstrando reações do sistema convencional de busca e do SmartRetrieval, sob distintas condições. A análise dos dados gerados (Quadro 2), em cada caso apresentado, evidencia que a incorporação do microtesauro amplia a capacidade de recuperação do sistema, aumentando a revocação sem comprometer a qualidade dos resultados. Como resultado, o processo de recuperação consegue diferenciar distintas perspectivas conceituais, preservando suas interconexões. A saída final consiste em uma lista de documentos ordenados em relevância estimada para a necessidade informacional do usuário, proporcionando experiências de RI precisas e contextualizadas, ao passo em que mantém o equilíbrio entre revocação e precisão por meio da combinação de consultas estruturadas e reclassificação semântica.

Quadro 2 – Resumo dos dados gerados nos quatro casos

-	Expressão de busca	Nº de artigos relevantes totais	SmartRetrieval			Sistema convencional		
			Precisão	Revocação	F1 - score	Precisão	Revocação	F1 - score
Caso 1	Sumarização automática	1	1/1 = 100,0%	1/1 = 100,0%	100,0%	1/7 = 14,0%	1/1 = 100,0%	24,6%
	"Sumarização automática"		-	-	-	1/1 = 100,0%	1/1 = 100,0%	100,0%
	Como realizar resumos automáticos		1/1 = 100,0%	1/1 = 100,0%	1/1 = 100,0%	1/13 = 7,7%	1/1 = 100,0%	14,3%
Caso 2	Inteligência artificial	20	20/20 = 100,0%	20/20 = 100,0%	100,0%	18/18 = 100,0%	18/20 = 90,0%	94,7%
	intelligence artificielle@fr		20/20 = 100,0%	20/20 = 100,0%	100,0%	17/17 = 100,0%	17/20 = 85,0%	91,9%
	künstliche intelligenz@gr		20/20 = 100,0%	20/20 = 100,0%	100,0%	0/0	0/20	0,0%
Caso 3	Ferramentas de IA	3	3/3 = 100,0%	3/3 = 100,0%	100,0%	3/19 = 16,0%	3/3 = 100,0%	27,6%
	"Ferramentas de IA"		3/3 = 100,0%	3/3 = 100,0%	100,0%	1/1 = 100,0%	1/3 = 33,0%	49,6%
Caso 4	Organização do Conhecimento	19	19/28 = 68,0%	19/19 = 100,0%	80,9%	18/28 = 64,0%	18/19 = 95,0%	76,5%

Fonte: elaboração própria (2026).

Mais especificamente, no **Caso 1**, o SmartRetrieval apresentou desempenho máximo (precisão, revocação e F1-Score de 100%) tanto para o termo exato quanto para consultas formuladas em linguagem natural utilizando variantes terminológicas. Quando o termo estava presente literalmente no documento, o sistema obteve desempenho equivalente à busca convencional com uso de aspas. Entretanto, ao empregar uma variante linguística ("Como realizar resumos automáticos"), o sistema convencional recuperou documentos irrelevantes, reduzindo sua precisão para 7,7%, enquanto o SmartRetrieval manteve desempenho máximo. Os resultados demonstram que a expansão semântica e a normalização terminológica permitem

ao sistema operar em nível conceitual, reduzindo o ruído informacional e a dependência da formulação exata da consulta.

No segundo caso, o SmartRetrieval recuperou os 20 documentos relevantes do corpus, inclusive aqueles que não continham explicitamente o termo "Inteligência Artificial", mas conceitos correlatos. O sistema também manteve desempenho máximo em consultas realizadas em francês e alemão (línguas não abrangidas pelo microtesauro), alcançando 100% de precisão, revocação e F1-Score. Em contraste, o sistema convencional apresentou perdas de revocação em português e francês e perda completa na consulta em alemão. Dessa forma, os resultados evidenciam a capacidade do SmartRetrieval de realizar alinhamento semântico multilíngue e recuperar documentos com base em relações conceituais, e não apenas em correspondência lexical.

Mantendo desempenho máximo independentemente do uso de aspas, no **Caso 3** o SmartRetrieval recuperou corretamente três documentos relevantes. Dentre esses, identificou dois trabalhos pertinentes que não haviam sido inicialmente classificados sob essa temática, demonstrando potencial como ferramenta de apoio à indexação. Já o sistema convencional apresentou o comportamento de compensação entre precisão e revocação: sem aspas recuperou documentos irrelevantes (precisão de 16%), e com aspas reduziu a revocação (33%). O caso evidencia que a incorporação de mecanismos semânticos reduz a dependência da formulação da consulta e proporciona recuperação mais consistente e contextualizada.

No quarto e último caso, vê-se que o SmartRetrieval alcançou 100% de revocação e F1-Score superior ao sistema convencional, demonstrando maior capacidade de recuperar documentos relevantes. Contudo, o caso revelou uma limitação importante: a qualidade dos resultados depende diretamente da qualidade das relações semânticas estabelecidas no microtesauro. Associações inadequadas, hierarquias imprecisas e equivalências frágeis contribuem para a recuperação de documentos perifericamente relacionados ao tema, reduzindo a precisão. Os resultados indicam que sistemas baseados em expansão semântica e RAG dependem não apenas da arquitetura tecnológica, mas também da qualidade epistemológica e terminológica dos instrumentos de RC que os sustentam, reforçando a importância da curadoria conceitual na integração entre OC e IA.

Os quatro casos demonstram que o SmartRetrieval supera os sistemas convencionais ao integrar RAG, microtesauros e expansão semântica. Mostrando-se menos dependente da

formulação lexical das consultas, é eficiente em cenários multilíngues e capaz de equilibrar simultaneamente precisão e revocação.

Como demonstra o **Caso 2**, o SmartRetrieval é capaz de rever com sucesso os 20 documentos definidos na amostra de teste. De tal forma, percebe-se que a estruturação orientada por domínio possibilita filtragem semântica refinada e apoia a expansão contextualizada das consultas. Essa abordagem contribui para mitigar limitações típicas de sistemas tradicionais de RI baseados em palavras-chave, como incompatibilidade terminológica e ambiguidade semântica, ao explorar as capacidades interpretativas do SLM. Assim sendo, documentos sem correspondências lexicais exatas com a consulta, mas conceitualmente relevantes, são identificados e priorizados.

Vê-se ainda que a configuração dos parâmetros do modelo (*temperature*=0 e *top_p*=0,1) influencia diretamente na qualidade da RI, pois reduziu alucinações e aumentou a estabilidade das respostas. A incorporação do microtesauro ampliou a capacidade de recuperação de documentos relevantes sem introduzir ruído significativo. Além disso, consultas formuladas com termos específicos apresentaram maior precisão, enquanto o uso de variantes terminológicas e equivalentes multilíngues melhorou o alinhamento entre a linguagem do usuário e a representação do sistema.

5 CONSIDERAÇÕES FINAIS

O SmartRetrieval fundamenta-se nos avanços recentes dos LLMs, incorporando suas capacidades de processamento semântico, extração de padrões e apoio à curadoria intelectual realizada por especialistas de domínio, na RI. Entretanto, diferencia-se das abordagens convencionais ao deslocar o foco do conhecimento generalista para o conhecimento especializado de domínio. Essa orientação reconhece que, em contextos relacionados aos SOCs, a precisão conceitual, a consistência terminológica e o controle semântico constituem requisitos de grande importância e que não são plenamente atendidos por modelos de propósito geral.

Nesse sentido, o modelo propõe a articulação entre as capacidades inferenciais dos SLMs e a curadoria intelectual de especialistas de domínio, promovendo uma mediação semântica qualificada. Sem substituir o papel do indexador, as tecnologias de IA são mobilizadas como instrumentos de apoio, capazes de sugerir termos, identificar relações conceituais e ampliar as possibilidades em RI, sempre sob validação humana. Essa perspectiva aproxima-se

de abordagens orientadas pelo conhecimento (*knowledge-driven*), em contraste com modelos exclusivamente orientados por dados (*data-driven*).

Como aplicação prática, o modelo mostrou potencial para apoiar processos de indexação automática, curadoria informacional e recuperação de documentos em bases científicas. Entre as limitações do estudo destacam-se o tamanho reduzido do *corpus* e a dependência da qualidade das relações semânticas estabelecidas. Futuras pesquisas podem ampliar o *corpus* analisado, explorar estratégias mais avançadas de RAG e investigar aspectos relacionados à explicabilidade e mitigação de vieses, contribuindo para o desenvolvimento de sistemas mais robustos, transparentes e alinhados às necessidades informacionais dos usuários.

Ao privilegiar conhecimento especializado, o SmartRetrieval contribui para o desenvolvimento de sistemas mais alinhados às especificidades dos domínios científicos, favorecendo a interoperabilidade semântica, a qualidade da indexação e a confiabilidade dos processos de RI. Essa integração entre IA e expertise humana aponta para um paradigma híbrido, no qual os LLMs atuam como facilitadores da organização e do acesso ao conhecimento, preservando, ao mesmo tempo, os princípios epistemológicos que fundamentam a área.

REFERÊNCIAS

ALCAIDE, Gabriel Santos; GAVA, Roberto Júlio; RODRIGUES, Willame Santos; SANTOS, Débora Ferreira. Análise comparativa e de consistência entre representação automática e manual de informações documentárias. **Transinformação**, Campinas, v. 13, n. 1, 2001. Disponível em: https://brapci.inf.br/v/115276. Acesso em: 16 jun. 2025.

ALEXANDER, Johanna Olson. Library Applications. *In*: BIDGOLI, H. (ed.). **Encyclopedia of Information Systems**. Amsterdam: Elsevier, 2003.

ALMEIDA, Mauricio Barcellos. **Ontologia em Ciência da Informação**: teoria e método. Curitiba, PR: CRV, 2020.

ANDREEWSKI, Alexandre; RUAS, Vitoriano. Indexação automática baseada em métodos linguísticos e estatísticos e sua aplicabilidade à língua portuguesa. **Ciência da Informação**, Brasília, v. 12, n. 1, 1983. Disponível em: https://revista.ibict.br/ciinf/article/view/332. Acesso em: 16 jun. 2025.

ASSOCIAÇÃO BRASILEIRA DE NORMAS TÉCNICAS. **ABNT NBR 12676**: métodos para análise de documentos - determinação de seus assuntos e seleção de termos de indexação - procedimento. Rio de Janeiro, RJ: ABNT, 1992.

BAEZA-YATES, Ricardo; RIBEIRO-NETO, Berthier. **Recuperação da Informação:** conceitos e tecnologias das máquinas de busca. Porto Alegre, RS: Bookman Editora LTDA., 2013.

BRITO, Marcílio de. Sistemas de informação em linguagem natural: em busca de uma indexação automática. **Ciência da Informação**, Brasília, v. 21, n. 3, 1992. Disponível em: https://brapci.inf.br/v/22159. Acesso em: 16 jun. 2025.

CAMPOS, Maria Luiza de Almeida. **Linguagem Documentária:** teorias que fundamentam sua elaboração. Niterói, RJ: EdUFF, 2001.

CESARINO, Maria Augusta da Nóbrega. Sistemas de Recuperação da Informação. **Revista da Escola de Biblioteconomia da UFMG**, Belo Horizonte, v. 14, n. 2, p. 157–168, set. 1985. Disponível em: https://periodicos.ufmg.br/index.php/reb/article/view/36507. Acesso em: 16 jun. 2025.

CHRISTIAN, Brian. **The Alignment Problem:** machine learning and human values. New York, NY: W. W. Norton & Company, 2020.

CORREA, Carlos Alberto; KOBASHI, Nair Yumiko. Estudos de aplicações da lógica paraconsistente em um modelo híbrido de indexação automática. *In*: ENCONTRO NACIONAL DE PESQUISA E PÓS-GRADUAÇÃO EM CIÊNCIA DA INFORMAÇÃO. 11., Rio de Janeiro, 2010. **Anais** [...]. São Paulo: ANCIB, 2010. Disponível em: https://brapci.inf.br/v/177133. Acesso em: 13 fev. 2026.

DAHLBERG, Ingetraut. Teoria do Conceito. **Ciência da Informação**, Brasília, v. 7, n. 2, 1978. Disponível em: https://revista.ibict.br/ciinf/article/view/115. Acesso em: 13 fev. 2026.

DE LUCA, Ernesto William; FALLUCCHI, Francesca; MORELATO, Roberto. Teaching an Algorithm How to Catalog a Book. **Computers**, Basel, v. 10, n. 11, art. 155, 2021. Disponível em: https://www.mdpi.com/2073-431X/10/11/155. Acesso em: 13 fev. 2026.

GAO, Yunfan. *et al.* Retrieval-Augmented Generation for Large Language Models: a survey. **arXiv**, Ithaca, v. 2, 2023. Disponível em: https://arxiv.org/abs/2312.10997. Acesso em: 13 fev. 2026.

GÉRON, Aurélien. **Mãos à obra**: aprendizado de máquina com Scikit-Learn, Keras e TensorFlow. Conceitos, ferramentas e técnicas para a construção de sistemas inteligentes. 2 ed. Rio de Janeiro, RJ: Alta Books, 2021.

HERDMAN, Michael; FOX-RUSHBY, Julia; BADIA, Xavier. A Model of Equivalence in the Cultural Adaptation of HRQoL Instruments: the universalist approach. **Quality of Life Research**, [*s. l.*], v. 7, p. 323–335, 1998. Disponível em: https://link.springer.com/article/10.1023/A:1024985930536. Acesso em: 13 fev. 2026.

HOFFMANN, Jordan. *et al*. Training Compute-Optimal Large Language Models. **arXiv**, Ithaca, 2022. Disponível em: https://arxiv.org/abs/2203.15556. Acesso em: 13 fev. 2026.

ISLAM, Ray. **Retrieval-Augmented Generation (RAG)**: empowering large language models (llms). [*S. l.*]: Dr. Ray Islam, 2023.

LANCASTER, Frederick Wilfrid. **Indexação e resumos**: teoria e prática. 2 ed. Brasília, DF: Briquet de Lemos Livros, 2004. Disponível em: https://bibliotextos.wordpress.com/wpcontent/uploads/2014/07. Acesso em: 3 set. 2025.

LEWIS, Patrick. *et al.* Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. *In*: International Conference on Neural Information Processing Systems. 34., Vancouver, 2020. **Anais** [...]. Vancouver: NeurIPS, 2020. Disponível em: https://proceedings.neurips.cc/paper_files/paper/2020/file/6b493230205f780e1bc26945df7481e5-Paper.pdf. Acesso em: 3 set. 2025.

LUHN, Hans Peter. A Business Intelligence System. **IBM Journal**, [*s. l.*], 1958. Disponível em: http://altaplana.com/ibmrd0204H.pdf. Acesso em: 10 jan. 2025.

MAMFRIM, Flávia Pereira Braga. Representação de conteúdo via indexação automática em textos integrais em língua portuguesa. **Ciência da Informação**, Brasília, v. 20, n. 2, 1991. Disponível em: https://revista.ibict.br/ciinf/article/view/355. Acesso em: 13 dez. 2024.

MANNING, Christopher D.; RAGHAVAN, Prabhakar; SCHÜTZE, Hinrich. **Introduction to Information Retrieval**. Cambridge: Cambridge University Press, 2008.

NARAYANAN, Arvind; KAPOOR, Sayash. **AI Snake Oil**: what artificial intelligence can do, what it can't, and how to tell the difference. Princeton: Princeton University Press, 2024.

NASCIMENTO, Felipe Mozart de Santana; PINHO, Fabio Assis. Sistemas de organização do conhecimento: semelhanças e diferenças. **ConCI**: Convergências mm Ciência da Informação, São Cristóvão, v. 2, n. 3, p. 104–122, set./dez. 2019. Disponível em: https://brapci.inf.br/v/140503. Acesso em: 10 jan. 2025.

RICHMOND, Phyllis A. A Reconsideration of Enumerative Classification for Current Information Needs. **Ciência da Informação**, Brasília, v. 3, n. 1, 1974. Disponível em: https://revista.ibict.br/ciinf/article/view/38. Acesso em: 10 jan. 2025.

ROBREDO, Jaime; FERREIRA, José Adalberto de Paula. Conceituação de um programa para indexação automática de textos. **Revista de Biblioteconomia de Brasília**, Brasília, v. 8, n. 2, 1980. Disponível em: https://brapci.inf.br/v/77346. Acesso em: 5 jul. 2024.

SHANG, Duxin; DUAN, Yufeng; XIE, Jiahong. Exploring Retrieval-Augmented Generation for Multi-Label Discipline Classification of Academic Short Texts. **Scientometrics**, [*s. l.*], 2025. Disponível em: https://link.springer.com/article/10.1007/s11192-025-05472-2. Acesso em: 5 jul. 2024.

SILVA, Wagner Teixeira da; MILIDIÚ, Ruy Luiz. Indexação e recuperação da informação com função de crença. **Ciência da Informação**, Brasília, v. 20, n. 2, 1991. Disponível em: https://revista.ibict.br/index.php/ciinf/article/view/351/0. Acesso em: 5 jul. 2024.

VIANA, Michelângelo Mazzardo Marques; MAIMONE, Giovana Deliberali. **Encyclopedia of Libraries, Librarianship, and Information Science**. 1 ed. Cambridge: Academic Press, 2025. p. 677–688.

YIN, Robert K. **Pesquisa qualitativa do início ao fim**. Porto Alegre, RS: Penso, 2016.

AGRADECIMENTOS

Este trabalho contou com o apoio do Conselho Nacional de Desenvolvimento Científico e Tecnológico – CNPq, por meio da Bolsa de Produtividade em Pesquisa (Processo nº 133119/2025-4), o que contribuiu significativamente para o desenvolvimento desta linha de investigação.