XXIV ENCONTRO NACIONAL DE PESQUISA EM CIÊNCIA DA INFORMAÇÃO – XXIV ENANCIB

ISSN 2177-3688

GT Especial

RECUPERAÇÃO DA INFORMAÇÃO E INTELIGÊNCIA ARTIFICIAL GENERATIVA COM LARGE LANGUAGE MODEL E *RETRIEVAL-AUGMENTED GENERATION*

INFORMATION RETRIEVAL AND GENERATIVE ARTIFICIAL INTELLIGENCE WITH LARGE LANGUAGE MODEL AND RETRIEVAL-AUGMENTED GENERATION

Henrique Leal Tavares – Universidade de Marília (UNIMAR), Universidade Estadual Paulista "Júlio de Mesquita Filho" (UNESP)
Caio Saraiva Coneglian – Universidade de Marília (UNIMAR); Universidade Estadual Paulista "Júlio de Mesquita Filho" (UNESP), Instituto Brasileiro de Informação em Ciência e Tecnologia (IBICT)
Emanuelle Torino – Universidade Tecnológica Federal do Paraná (UTFPR)
Silvana Aparecida Borsetti Gregorio Vidotti – Universidade Estadual Paulista "Júlio de Mesquita Filho" (UNESP), Instituto Brasileiro de Informação em Ciência e Tecnologia (IBICT)
José Eduardo Santarem Segundo – Universidade de São Paulo (USP), Universidade Estadual Paulista "Júlio de Mesquita Filho" (UNESP)

Modalidade: Trabalho Completo

Resumo: A Inteligência Artificial Generativa vem impactando toda a sociedade, impulsionando o desenvolvimento de soluções em diversas áreas do conhecimento. No contexto da Ciência da Informação, *Large Language Model* e *Retrieval-Augmented Generation*, amplamente utilizadas em ferramentas como o ChatGPT e Gemini, são técnicas capazes de influenciar e se relacionar de forma ampla com a Recuperação da Informação. Dessa forma, o presente trabalho tem como objetivo apresentar um modelo de Recuperação da Informação utilizando como interface um *chatbot* e aplicando as técnicas de Inteligência Artificial Generativa – *Large Language Model* e *Retrieval-Augmented Generation*. Para tal, realizou-se uma abordagem aplicada junto com a exploração da literatura acerca da temática, o que conduziu à construção do *chatbot*. Enquanto resultados, a solução desenvolvida, o *chatbot*, foi capaz de aplicar as técnicas da Inteligência Artificial Generativa, ao mesmo tempo em que demonstrou como a Recuperação da Informação está presente e se relaciona com a solução. Conclui-se, então, que o desenvolvimento de pesquisas aplicadas de Inteligência Artificial no âmbito da Ciência da Informação pode contribuir para novos estudos e para uma maior compreensão da utilização dessas tecnologias emergentes na área.

Palavras-chave: inteligência artificial generativa; Recuperação da Informação; *Large Language Model; Retrieval-Augmented Generation.*

Abstract: Generative Artificial Intelligence is impacting society as a whole, enabling the development of solutions across many fields of knowledge. In the context of Information Science, techniques such as Large Language Models and Retrieval-Augmented Generation, widely used in tools like ChatGPT and Gemini, have the potential to broadly influence and relate to Information Retrieval. Therefore, this work aims to present an Information Retrieval model using a chatbot as the interface and applying Generative Artificial Intelligence techniques — Large Language Models and Retrieval-Augmented Generation. To achieve this, an applied approach was adopted, along with an exploration of the relevant literature, which led to the construction of the chatbot. As a result, the developed solution, the chatbot, was able to apply Generative Artificial Intelligence techniques while demonstrating how Information Retrieval is present and interconnected with the solution. It is concluded that the development of applied Artificial Intelligence research within the scope of Information Science can contribute to new studies and greater understanding of the use of these emerging technologies in the field.

Keywords: generative artificial intelligence; Information Retrieval; Large Language Model; Retrieval-Augmented Generation.

1 INTRODUÇÃO

A Recuperação da Informação é apontada como uma das principais temáticas da Ciência da Informação (CI), pois é nela que acontecem as relações interdisciplinares, como relatado por Saracevic (1995). Trata-se, portanto, de uma área que, ao longo das décadas, permitiu uma série de pesquisas e inovações.

Nesse contexto, a evolução da Recuperação da Informação (RI) em tempos de Inteligência Artificial Generativa se torna um tema relevante, já que muitos processos de organização, representação e recuperação de dados agora ocorrem de maneira automática, impulsionados por algoritmos de *Machine Learning*. Essa transformação é fundamental para promover avanços teóricos e práticos, ampliando as possibilidades de inovação na área.

Ao analisar os estudos e as práticas desenvolvidas no âmbito da Ciência da Informação, verifica-se a proeminência de soluções com a aplicação de ferramentas como o ChatGPT. No entanto, há a possibilidade de aprofundar tal discussão, ao trazer alguns conceitos e técnicas da Inteligência Artificial (IA) e da Ciência da Computação, como o *Large Language Model* (LLM) e *Retrieval-Augmented Generation* (RAG), para serem discutidos na Ciência da Informação — em especial, na apropriação pela Recuperação da Informação.

Assim, o presente trabalho tem como objetivo apresentar um modelo de Recuperação da Informação, utilizando como interface um *chatbot* e aplicando as técnicas de Inteligência Artificial Generativa — *Large Language Model* (LLM) e *Retrieval-Augmented Generation* (RAG).

O estudo, de caráter teórico-prático, teve como base um estudo exploratório das temáticas de Inteligência Artificial Generativa, em especial de LLM e de RAG, e de Recuperação da Informação, vinculando os conceitos. Além disso, teve um aspecto prático, que gerou uma prova de conceito para validar os conceitos e tecnologias identificados na revisão de literatura.

2 RECUPERAÇÃO DA INFORMAÇÃO NO CONTEXTO DA INTELIGÊNCIA ARTIFICIAL GENERATIVA

Ao longo das últimas décadas, a Recuperação da Informação vem se desenvolvendo e se posicionando como uma das áreas que relaciona e traz os avanços tecnológicos e computacionais para a Ciência da Informação. A definição de Recuperação da Informação que aponta: "[...] a recuperação da informação está diretamente ligada à representação, armazenamento, organização e acesso aos itens de informação. Dizem que a representação e a organização dos itens de informação deveriam prover o uso e o fácil acesso à informação necessária ao usuário" (Baeza-Yates; Ribeiro Neto, 1999, p. 1, tradução nossa). Verifica-se, assim, o enfoque no atendimento às necessidades informacionais dos usuários.

Com o avanço das tecnologias, em especial da Inteligência Artificial, novos modelos e técnicas estão surgindo, permitindo repensar e aprimorar o próprio processo de Recuperação da Informação. Entre as possíveis mudanças, destacam-se a personalização dos resultados de busca, o aumento da precisão na recuperação de conteúdos relevantes e a capacidade de integrar diferentes fontes de dados em tempo real. Além disso, a aplicação de técnicas de IA pode melhorar a eficiência dos sistemas de recuperação, reduzindo o tempo de resposta e proporcionando uma melhor experiência do usuário. Discussões acerca da inserção da Inteligência Artificial na Recuperação da Informação foram realizadas ao longo das últimas décadas, como em Jones (1999), Trotman (2004), Ferneda (2006), Coneglian (2020) Coneglian e Santarem Segundo (2022) e Coneglian, Torino, Santarem Segundo e Vidotti (2023), ainda que com abordagens diferentes da proposta neste estudo. No entanto, as novas abordagens de IA, como os *Large Language Models* e o *Retrieval-Augmented Generation*, trazem a oportunidade de adequar esses processos às demandas atuais, como o tratamento de grandes volumes de dados e a interação mais natural entre usuários e sistemas.

Vale destacar que a Inteligência Artificial é definida como "[...] a ciência e a engenharia de fabricar máquinas inteligentes, especialmente programas de computador inteligentes. Está

relacionada com a tarefa semelhante de utilizar computadores para compreender a inteligência humana, mas a IA não tem de se limitar a métodos que sejam biologicamente observáveis" (McCarthy, 2007, p. 2, tradução nossa).

As discussões sobre a inserção da Inteligência Artificial na Recuperação da Informação se deram especialmente no que se refere a aprimorar o processo da Recuperação da Informação, buscando aprimorar e tornar mais efetivo os resultados alcançados. No entanto, com o advento da Inteligência Artificial Generativa, um novo enfoque teve destaque no contexto da Recuperação da Informação.

A IA Generativa é apontada:

> Com capacidades para gerar conteúdo contextualmente relevante e de alta qualidade, quase indistinguível do trabalho criado por humanos [...]. A IA generativa encontra sua utilidade em diversas modalidades, incluindo a geração de texto, imagem, vídeo, código, som e outros conteúdos produzidos, como moléculas ou renderizações 3D (Banh; Strobel, 2023, tradução nossa).

No contexto da IA Generativa, no que tange à geração automática de conteúdo em textos, destaca-se o *Large Language Model* (LLM), que consiste em "[...] modelos de linguagem *Transformer* que contêm centenas de bilhões (ou mais) de parâmetros, que são treinados em dados de texto massivos, como GPT-3, PaLM, Galactica e LLaMA" (Zhao *et al.*, 2023, tradução nossa).

A partir da criação dos modelos de LLM, com destaque para o ChatGPT, houve um momento de rápido desenvolvimento de soluções em todas as áreas do conhecimento. No contexto da Recuperação da Informação, a inserção de LLM possibilitou uma revolução no processo de busca, transformando o modo como os usuários buscam informações na web, utilizando buscadores como Bing, apoiado no uso de IA, e nos modelos de Gemini.

Em especial, a partir desta transformação do processo de Recuperação da Informação com uso das técnicas de LLM, destacam-se as técnicas de *Retrieval-Augmented Generation* (RAG), que são "[...] modelos que combinam memória paramétrica e não paramétrica pré-treinada para geração de linguagem" (Lewis *et al.*, 2020, tradução nossa).

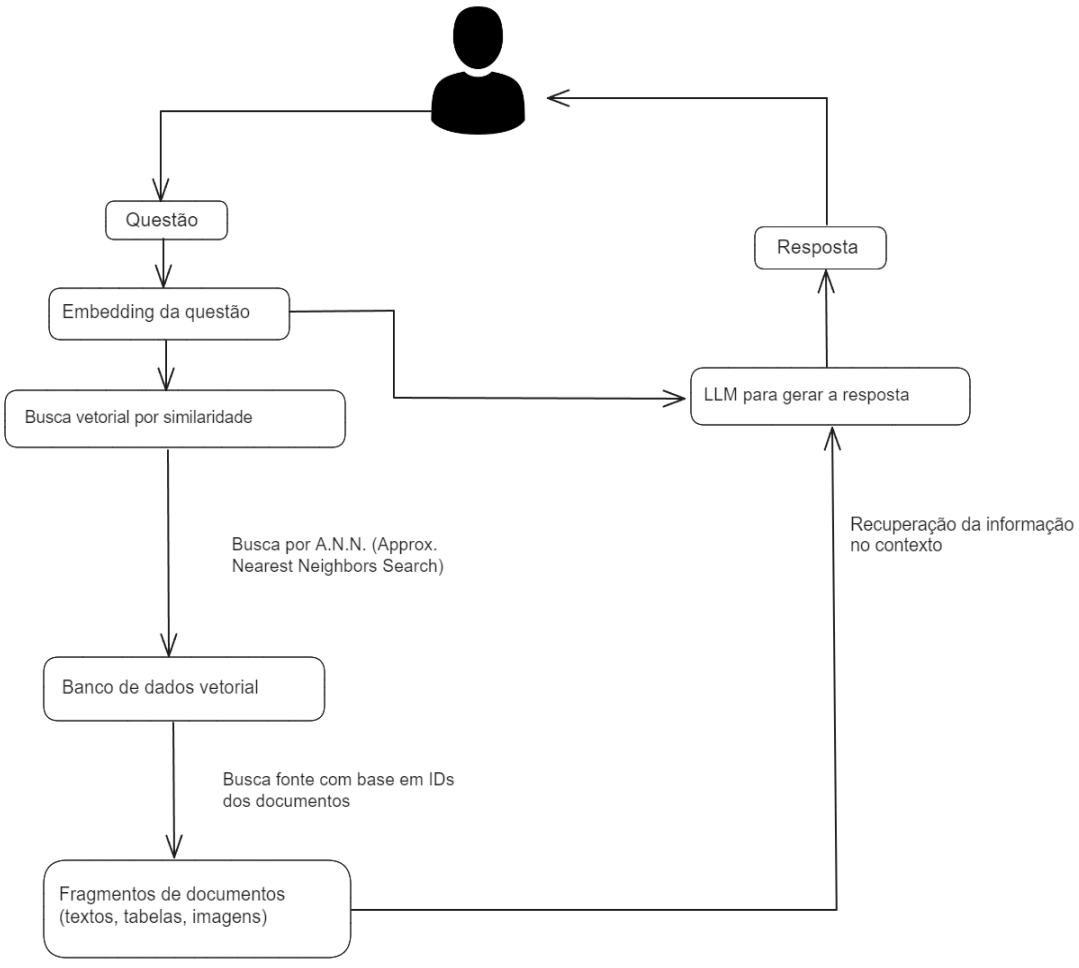
Ou seja, o RAG utiliza informações armazenadas nos parâmetros do modelo de treinamento (memória paramétrica), além de informações de fontes externas que complementam as respostas fornecidas (memória não paramétrica). Desta forma, o RAG é

capaz de gerar respostas de forma natural, utilizando tanto o conhecimento pré-treinado quanto informações recuperadas de fontes externas para criar respostas mais precisas e atualizadas.

3 RESULTADOS: CHATBOT PARA RECUPERAÇÃO DA INFORMAÇÃO UTILIZANDO LLM E RAG

Com o intuito de apresentar como um sistema de Recuperação da Informação pode ser desenvolvido utilizando Inteligência Artificial, em especial *Large Language Model* (LLM) e *Retrieval-Augmented Generation (*RAG), realizou-se a proposta da construção de um *chatbot* considerando as bases da Ciência da Informação, com destaque para a Recuperação da Informação. A Figura 1 demonstra como ocorre o processo de Recuperação da Informação, utilizando uma interface de *chatbot*, a partir do uso de técnicas de LLM e RAG.

Figura 1 – Modelo de Recuperação da Informação a partir de uma interface de *chatbot* utilizando RAG e LLM



Fonte: Elaborado pelos autores (2024).

Inicialmente são abordadas a preparação e a configuração inicial necessárias para a criação do *chatbot*, bem como a sua construção.

Para implementar um sistema RAG eficaz, o primeiro passo envolve a preparação e configuração inicial das bibliotecas necessárias. A *LangChain*[1] é utilizada para construir a cadeia do RAG e conectar com o LLM. A biblioteca *Unstructured* é empregada para extrair informações e elementos de arquivos *Portable Document Format* (PDF), enquanto *Whisper*, da *OpenAI*, é utilizada para transcrever vídeos em textos. Modelos de LLM como *Llama3 70b* (*Llava* para imagens), *Gemini 1.5 Pro* e GPT-4o são integrados ao sistema para processamento de linguagem natural e geração de respostas. O *PGVector* é utilizado como banco de dados vetorial, e o modelo de *embedding* da *OpenAI* é implementado para gerar representações numéricas dos dados. Para a validação do sistema, o *LangSmith* é empregado para rastrear e analisar interações.

No contexto da Recuperação da Informação, tanto na Ciência da Informação quanto na Ciência da Computação, o uso de ferramentas de Inteligência Artificial Generativa, como o Bing e o ChatGPT da OpenAI, oferecem vantagens significativas, cada uma atendendo a diferentes necessidades e etapas do processo. Ambas integram técnicas avançadas de IA generativa, como os *Large Language Models* (LLMs) e *Retrieval-Augmented Generation* (RAG), que proporcionam resultados otimizados na busca e recuperação da informação. No entanto, a escolha entre elas deve ser fundamentada nas características específicas que cada ferramenta oferece.

O Bing, ao incorporar IA generativa, destaca-se por sua capacidade de integrar dados em tempo real com respostas contextualizadas, fornecendo resultados de busca a partir de fontes atualizadas. Isso é especialmente vantajoso quando a recuperação da informação exige dados atuais, ou quando se necessita de uma integração mais direta com a web, como em pesquisas dinâmicas e variadas. Além disso, o Bing facilita a busca de documentos multimodais, ampliando o escopo da recuperação para diferentes formatos, como imagens e

[1] *LangChain* é uma estrutura que facilita a construção de aplicativos que utilizam modelos de linguagem. Ela fornece ferramentas para conectar dados, modelos e fluxos de trabalho, permitindo a criação de sistemas complexos e interativos baseados em linguagem natural.

vídeos, o que pode ser essencial em determinadas etapas da pesquisa ou quando o objetivo for buscar fontes de informação diversificadas.

Por outro lado, o ChatGPT da OpenAI oferece uma abordagem diferenciada no que diz respeito à interação natural com o usuário e à capacidade de geração de respostas contextualizadas e aprofundadas. Por meio de técnicas como o LLM e o RAG, o ChatGPT pode fornecer respostas altamente detalhadas, criando uma interação mais fluida e intuitiva, principalmente em contextos de conversas mais complexas ou consultas que demandam uma compreensão mais elaborada do contexto do usuário. A força do ChatGPT está na personalização e na adaptabilidade da geração de texto, o que o torna ideal para etapas do processo de Recuperação da Informação em que a exploração conceitual ou a geração de conhecimento baseado em grandes volumes de texto é necessária.

Assim, a escolha entre Bing e ChatGPT deve ser guiada pela especificidade de cada etapa da pesquisa. O Bing pode ser mais adequado quando se requer uma recuperação rápida e baseada em dados atualizados e em tempo real, enquanto o ChatGPT se mostra mais eficiente quando o foco está na interação com informações mais conceituais e na geração de insights profundos a partir de grandes conjuntos de dados textuais. Portanto, ao longo do processo de Recuperação da Informação, é importante evidenciar o motivo da escolha de cada ferramenta, considerando os pontos fortes de cada uma e como elas podem complementar as necessidades de cada fase do trabalho.

Após a definição das tecnologias e bases para a construção do *chatbot*, iniciou-se a reflexão acerca da definição dos processos relativos à Recuperação da Informação, considerando que o uso do RAG se mostra como uma nova abordagem para a Recuperação da Informação, tanto no contexto do modo como os resultados são construídos, quanto na forma de apresentação para o usuário, utilizando premissas de *Question Answering*.

Assim, concebeu-se a arquitetura do RAG, que opera em três etapas principais: *embedding*s de vetores, pesquisa de vetores e geração aumentada. Na primeira etapa, documentos e outras fontes de informação são convertidos em representações numéricas (vetores), chamadas *embeddings*, que capturam o significado semântico do conteúdo. Na segunda etapa, uma pergunta é convertida em um *embedding* e uma pesquisa é realizada em

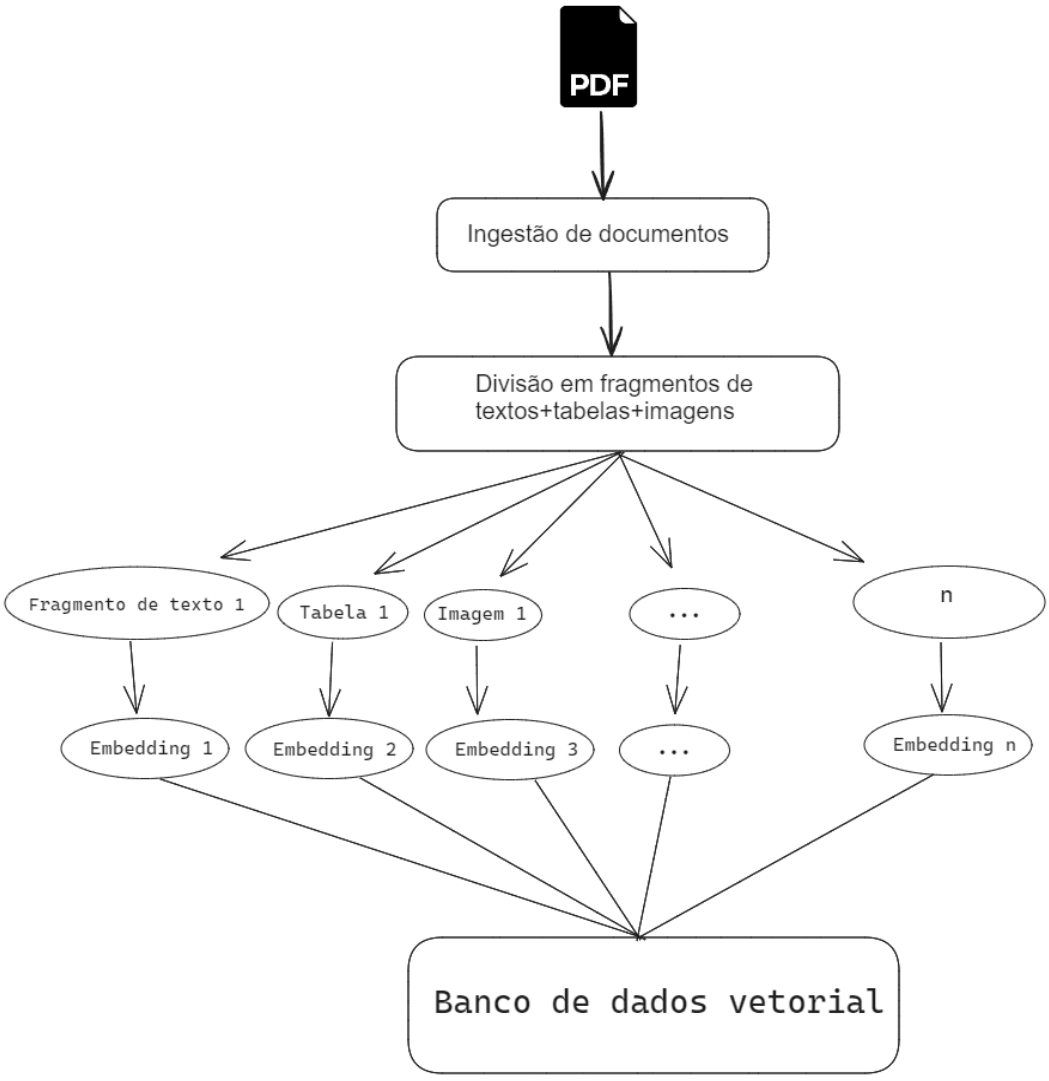
um banco de dados vetorial para encontrar os *embeddings* mais semelhantes, indicando os documentos mais relevantes. Na etapa final, os documentos recuperados são fornecidos como contexto para um modelo de linguagem, que utiliza essas informações adicionais para gerar uma resposta mais completa e contextualizada.

Vale destacar que *embeddings* são representações densas e de alta dimensão de palavras, frases ou outros elementos textuais que capturam informações semânticas e sintáticas em um formato compacto e útil. Essas representações são fundamentais em diversos aplicativos de Processamento de Linguagem Natural (PLN) e aprendizagem de máquina. De acordo com Bengio *et al.* (2003), *embeddings* são gerados projetando vetores de palavras em um espaço contínuo de alta dimensão, no qual a similaridade semântica e sintática entre as palavras pode ser medida. Este processo é baseado em modelos de linguagem neural, que inicialmente produzem essas representações durante o treinamento de grandes redes neurais.

Além dessa etapa de extração de compreensão da pergunta e busca pelos melhores termos, concebeu-se neste *chatbot* a possibilidade de busca por arquivos em formato PDF, que são bases importantes para as organizações. Dessa forma, amplia-se a possibilidade de ter um processo de Recuperação da Informação mais específico, capaz de buscar por informações que estão disponíveis em arquivos próprios, diferenciando-se assim de um processo de busca em um ambiente como o Bing, que utiliza Inteligência Artificial Generativa.

O processamento de documentos em formato PDF começa com o carregamento dos arquivos. A biblioteca *Unstructured* é utilizada para extrair texto, tabelas e imagens dos arquivos PDF. O texto extraído é dividido em pedaços (*chunks*) menores e gerenciáveis, utilizando como guias: títulos de seção, cabeçalhos e tamanho do conteúdo. Essa divisão melhora a precisão da recuperação da informação. Os elementos extraídos são categorizados em texto, tabelas e imagens, permitindo um processamento diferenciado de cada tipo de elemento na cadeia do RAG e no LLM (Figura 2).

Figura 2 – Processo de utilização de um arquivo PDF para a Recuperação da Informação



Fonte: Elaborado pelos autores (2024).

Dessa forma, para que possa ocorrer a sumarização de texto, define-se um *prompt* que instrui o LLM a gerar resumos concisos dos pedaços de texto extraídos. No caso de tabelas, o conteúdo é extraído como um único elemento, preservando sua estrutura, e convertido para *HyperText Markup Language* (HTML) para melhor visualização e processamento. O LLM é então utilizado para gerar um resumo da tabela em linguagem natural. Os dados da tabela são também convertidos para um *DataFrame* do Pandas[2], que é uma representação de um conjunto de dados nesta ferramenta, e que facilita a análise e a exploração. Para imagens, técnicas de *Optical Character Recognition* (OCR) são utilizadas para extrair texto incorporado,

[2] A biblioteca Pandas é uma ferramenta essencial no ecossistema Python para a manipulação e análise de dados.

enquanto bibliotecas de visão computacional analisam as imagens, extraindo informações visuais. As imagens extraídas são convertidas para o formato base64[3] e são sumarizadas por um LLM multimodal.

Para que os dados extraídos possam ser utilizados no processo de recuperação da informação a partir do RAG, é necessário converter o que foi extraído em *embeddings*, o que foi feito utilizando o modelo de *embedding* da OpenAI. Cada pedaço de texto, sumário de tabela e sumário de imagem é convertido em uma representação vetorial. Em seguida, cria-se um índice no banco de dados vetorial para armazenar esses *embeddings* no espaço vetorial. A configuração do banco de dados inclui a definição da dimensionalidade do modelo de *embedding* e parâmetros de busca, como a similaridade de cosseno e o número máximo de documentos retornados com maior similaridade à consulta.

Por fim, cria-se a ferramenta capaz de realizar o processo de Recuperação da Informação, chamado de *MultiVector Retriever*. Por meio desta ferramenta, recuperam-se os pedaços e acessam-se os documentos originais a partir dos *embeddings* armazenados no banco de dados vetorial. Define-se uma função de pesquisa multimodal que recebe uma consulta em linguagem natural, a qual é transformada em *embedding*, e executa uma pesquisa no banco de dados vetorial utilizando a similaridade. Os resultados da busca são usados para recuperar os documentos originais, que são fornecidos como contexto para o LLM. Adicionalmente, é configurada a memória do *chatbot* para retornar resumos de um número predefinido de mensagens anteriores, unindo-os ao *prompt*[4].

Por fim, realizou-se o processo de validação do RAG, no qual o *chatbot* foi testado com perguntas sobre os documentos processados, demonstrando sua capacidade de recuperar informações relevantes. Para isso, utilizou-se o *LangSmith*[5] de forma a rastrear as interações com o *chatbot*, visualizar *prompts*, respostas e tempos de processamento, entre outros detalhes.

---

[3] O Formato base64 é um esquema de codificação que converte dados binários em texto, utilizando um conjunto de 64 caracteres (A-Z, a-z, 0-9, +, /). Essa codificação é comumente usada para transmitir dados de forma segura em sistemas que lidam com dados textuais.

[4] *Prompt* é uma instrução ou entrada fornecida a um sistema de computação, como um comando em um terminal, uma pergunta em um *chatbot*.

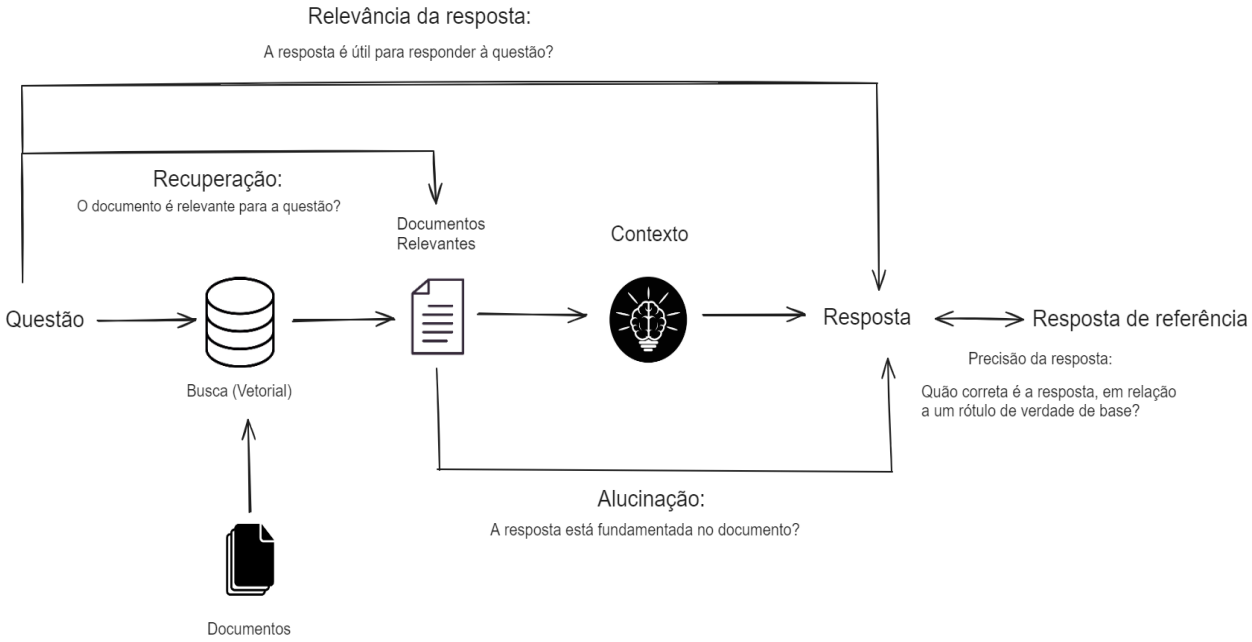
[5] *LangSmith* é uma ferramenta de desenvolvimento de software voltada para a criação de aplicações de linguagem natural,especialmente útil para desenvolver sistemas complexos e interativos que dependem de processamento de linguagem natural (NLP).

Para isso, foram definidas as seguintes métricas de avaliação, baseadas no trabalho de Qiu *et al*. (2024).:

a) correção: mede a precisão da resposta em relação a um rótulo de verdade de base;
b) relevância e utilidade: avalia quão bem a resposta gerada aborda a entrada inicial do usuário;
c) alucinação: mede a concordância da resposta gerada com o contexto recuperado;
d) *score k*, *rank* recíproco médio, NDCG: avaliam a qualidade dos resultados recuperados para uma consulta específica.

A Figura 3 apresenta como ocorreu o processo de validação e avaliação da Recuperação da Informação realizada por meio do *chatbot*.

Figura 3 – Fluxo de validação e avaliação do processo de Recuperação da Informação



Fonte: Elaborado pelos autores (2024).

A Figura 4 apresenta uma tela do ambiente do *chatbot*, na qual o usuário pode realizar o processo de busca e recuperação da informação.

Figura 4 – Tela do protótipo de Recuperação da Informação desenvolvido

Fonte: Elaborado pelos autores (2024).

Como mencionado, este é um protótipo que representa o ambiente de Recuperação da Informação construído a partir da aplicação das técnicas de Inteligência Artificial Generativa, utilizando LLM e RAG.

4 DISCUSSÃO

A introdução da Inteligência Artificial Generativa, em especial do LLM e do RAG, pode ser uma transformação nos modelos de Recuperação da Informação e nas ferramentas de busca. Desta forma, a Ciência da Informação é capaz de contribuir com as bases conceituais e aplicadas da Organização, Representação e da própria Recuperação da Informação para que estes ambientes sejam mais efetivos.

No âmbito da prova de conceito desenvolvida, as técnicas de RAG e LLM foram discutidas e aproximadas à Recuperação da Informação, apresentando o funcionamento das ferramentas e o potencial de aplicação de tais tecnologias para aprimorar os ambientes de Recuperação da Informação.

A partir da prova de conceito, as técnicas utilizadas e seus impactos na Recuperação da Informação e na Ciência da Informação são sintetizados e apresentados no Quadro 1.

Quadro 1 – Síntese das técnicas da Inteligência Artificial Generativa com a Recuperação da Informação e a Ciência da Informação

Técnica/Termo	Definição	Aproximação com a Recuperação da Informação (RI) e Ciência da Informação (CI)
Large Language Model (LLM)	"[...] modelos de linguagem *Transformer* que contêm centenas de bilhões (ou mais) de parâmetros, que são treinados em dados de texto massivos [...] " (Zhao *et al.*, 2023, tradução nossa).	Com o avanço do LLM, em especial do ChatGPT, há a necessidade de refletir como estes modelos são criados e qual a qualidade de tais dados, as fontes utilizadas, entre outras questões. A CI tem potencial para contribuir com modelos de RI mais eficazes e que, apoiados por técnicas de Organização e Representação da Informação, podem aprimorar a efetividade de soluções criadas a partir de LLM.
Retrieval-Augmented Generation (RAG)	"[...] modelos que combinam memória paramétrica e não paramétrica pré-treinada para geração de linguagem" (Lewis *et al.*, 2020, tradução nossa)	No contexto da Inteligência Artificial Generativa, há uma nova abordagem de Recuperação da Informação surgindo. Com isso, os usuários passam a recuperar informação de outra forma, utilizando linguagem natural e a partir de perguntas e respostas. Assim, modelos de RAG combinam a geração de texto utilizado previamente para treinamento com outras fontes. Dessa forma, pensar em modelos e aplicações que utilizam este tipo de tecnologia, mas que ,ao mesmo tempo, partem das premissas da RI, pode gerar soluções mais eficazes e que atendam melhor às necessidades informacionais dos usuários.
Embedding	*Embeddings* são representações densas e de alta dimensão de palavras, frases ou outros elementos textuais que capturam informações semânticas e sintáticas em um formato compacto e útil (Bengio *et al.*, 2003).	A representação de termos, documentos e imagens em vetores é a forma como os algoritmos de Inteligência Artificial utilizam para tratar tais elementos, que são de linguagem natural, na linguagem computacional. Dessa forma, a compreensão destes mecanismos pode ser necessária para a proposição de mecanismos de Recuperação da Informação mais efetivos. Além disso, tais vetores são utilizados para definir a aproximação semântica existente entre os termos, documentos ou imagens.
Banco de Dados Vetorial	De acordo com o estudo *When Large Language Models Meet Vector Databases: A Survey*, os Bancos de Dados Vetoriais emergem como uma solução promissora para otimizar a utilização de recursos computacionais, melhorando a flexibilidade na edição de modelos e reduzindo a ocorrência de alucinações em modelos de linguagem, ao fornecer respostas mais específicas e contextualmente corretas (Jing *et al.*, 2024).	No contexto da conversão dos dados em vetores, a recuperação da informação ocorre de forma distinta da tradicional. Assim, é necessário criar bancos de dados que permitam a busca por similaridade entre os vetores (*embedding*). No âmbito da RI, tal busca se diferencia do modo tradicional e traz um impacto na construção de modelos que se apropriam deste conceito para criar soluções mais eficientes.
Prompt	*Prompt* é o comando dado por	Com o advento da IA Generativa, reflete-se muito

Técnica/Termo	Definição	Aproximação com a Recuperação da Informação (RI) e Ciência da Informação (CI)
	um usuário ao modelo de Inteligência Artificial. "*Prompts* eficazes contribuem para melhorar a precisão, orientando Modelos de IA para gerar conteúdo mais relevante e valioso" (Korzynski *et al.*, 2023, p. 26, tradução nossa).	sobre a criação de *prompts* mais adequados. A discussão sobre *prompts* pode ser importante para a Experiência do Usuário e para a Arquitetura de Dados (Torino, 2022). Pesquisas nesse contexto têm um grande impacto no processo de Recuperação da Informação.

Fonte: Elaborado pelos autores (2024).

O Quadro 1 pode apoiar o desenvolvimento de outras pesquisas e soluções de Inteligência Artificial Generativa a partir das teorias e práticas da Ciência da Informação.

5 CONSIDERAÇÕES FINAIS

A Ciência da Informação e a Recuperação da Informação estão se apropriando cada vez mais da Inteligência Artificial, com diversas pesquisas e discussões acerca do impacto da IA em seus processos e subáreas. No entanto, há a necessidade de construção de aplicações de IA que utilizem as teorias e práticas da Ciência da Informação e que, ao mesmo tempo, dialoguem e tragam para o seu domínio as técnicas de Inteligência Artificial.

Dessa forma, este trabalho apresenta como o processo de Recuperação da Informação ocorre quando alinhado às tecnologias de Inteligência Artificial Generativa, com destaque para LLM e RAG. Tais técnicas, que se destacam atualmente como as principais de soluções como ChatGPT e Gemini, têm um grande potencial de aplicação na área de Ciência da Informação.

O desenvolvimento dessa prova de conceito, aplicando os conceitos e as técnicas de Inteligência Artificial Generativa no contexto da Ciência da Informação, em especial da Recuperação da Informação, mostra-se como um caminho para que esta área possa se apropriar e utilizar as novas técnicas e tecnologias de forma mais efetiva em seus processos.

A partir do proposto, é possível pensar no uso de técnicas de IA Generativa, com destaque para LLM e RAG, em contextos mais aplicados como os repositórios digitais, os serviços de recuperação de dados baseados em catálogos de bibliotecas, nas bases de dados e, ainda, em agregadores de metadados.

Portanto, espera-se que, com esta pesquisa e com a discussão sobre como as técnicas de Inteligência Artificial se relacionam com a Ciência da Informação e a Recuperação da Informação, novas pesquisas aplicadas possam ser desenvolvidas. Espera-se também que, dessa forma, exista uma apropriação ainda maior destes conhecimentos nas discussões e nos trabalhos desenvolvidos.

REFERÊNCIAS

BAEZA-YATES, R.; RIBEIRO-NETO, B. **Modern information retrieval**. New York: ACM Press, 1999.

BANH, L.; STROBEL, G. Generative artificial intelligence. **Electronic markets**, [Berlin], v. 33, art. 63, 2023. Disponível em: https://link.springer.com/article/10.1007/s12525-023-00680-1. Acesso em: 7 jul. 2024.

BENGIO, Y.; DUCHARME, R.; VINCENT, P.; JAUVIN, C. A neural probabilistic language model. **Journal of machine learning research**, [*S. l.*], v. 3, p. 1137-1155, 2003. Disponível em: https:// www.jmlr.org/papers/volume3/bengio03a/bengio03a.pdf. Acesso em: 7 jul. 2024.

CONEGLIAN, C. S. **Recuperação da informação com abordagem semântica utilizando linguagem natural**: a inteligência artificial na ciência da informação. 2020. Tese (Doutorado em Ciência da Informação) - Universidade Estadual Paulista, Marília, 2020. Disponível em: http://hdl.handle.net/11449/193051. Acesso em: 7 jul. 2024.

CONEGLIAN, C. S.; SANTAREM SEGUNDO, J. E. Inteligência artificial e ferramentas da web semântica aplicadas a recuperação da informação: um modelo conceitual com foco na linguagem natural. **Informação & Informação**, Londrina, v. 27, n. 1, p. 625–651, 2022. DOI: https://doi.org/10.5433/1981-8920.2022v27n1p625. Disponível em: https://ojs.uel.br/ revistas/uel/index.php/informacao/article/view/44729. Acesso em: 11 jul. 2024.

CONEGLIAN, C. S.; TORINO, E.; SANTAREM SEGUNDO, J. E.; VIDOTTI, S. A. B. G. Inteligência artificial generativa e recuperação da informação: tendências e oportunidades de pesquisa. In: ENCONTRO NACIONAL DE PESQUISA E PÓS-GRADUAÇÃO EM CIÊNCIA DA INFORMAÇÃO, 23., 2023, Aracaju. Anais [...]. Aracaju: ENANCIB, 2023. Disponível em: https://enancib. ancib.org/index.php/enancib/xxxiiienancib/paper/viewFile/1944/1401. Acesso em: 11 jul. 2024.

FERNEDA, E. Redes neurais e sua aplicação em sistemas de recuperação de informação. **Ciência da Informação**, Brasília, v. 35, p. 25-30, 2006. Disponível em: https://revista.ibict.br/ ciinf/article/view/1149. Acesso em: 11 jul. 2024.

JING, Z. *et al.* **When large language models meet vector databases**: a survey. 2024. Disponível em: https://ar5iv.labs.arxiv.org/html/2402.01763. Acesso em: 7 jul. 2024.

JONES, K. S. Information retrieval and artificial intelligence. **Artificial Intelligence**, [Amsterdam?], v. 114, n. 1/2, p. 257-281, Oct. 1999. Disponível em: https://www.sciencedirect.com/science/article/pii/S0004370299000752. Acesso em: 7 jul. 2024.

KORZYNSKI, P. *et al*. Artificial intelligence prompt engineering as a new digital competence: Analysis of generative AI technologies such as ChatGPT. **Entrepreneurial business and economics review**, Kraków, v. 11, n. 3, p. 25-37, 2023. DOI: 10.15678/eber.2023.110302. Disponível em: https://www.semanticscholar.org/paper/Artificial-intelligence-prompt-engineering-as -a-new-Korzy%C5%84ski-Mazurek/0019e876188f781fdca0c0ed3bca39d0c70c2ad2. Acesso em: 7 jul. 2024.

LEWIS, P. *et al*. Retrieval-augmented generation for knowledge-intensive nlp tasks. **Advances in neural information processing systems**, San Mateo, v. 33, p. 9459-9474, dec. 2020. Disponível em: https://dl.acm.org/doi/abs/10.5555/3495724.3496517. Acesso em: 7 jul. 2024.

McCARTHY, J. *et a*l. What is artificial intelligence. United States: Stanford University, 2007.

QIU, X.; ZHANG, X.; MÜNDLER, J.; KANG, S.; LI, J. **A comprehensive survey of hallucination mitigation techniques in large language models**. 2024. Disponível em: https://ar5iv.org/pdf/2401.01313. Acesso em: 7 jul. 2024.

SARACEVIC, T. A natureza interdisciplinar da Ciência da Informação. **Ciência da informação**, Belo Horizonte, v. 24, n. 1, 1995. DOI: https://doi.org/10.18225/ci.inf.v24i1.608. Disponível em: https://revista.ibict.br/ciinf/article/view/608. Acesso em: 7 jul. 2024.

TORINO, E. **Arquitetura de dados no contexto da Ciência da Informação**. 2022. Tese (Doutorado em Ciência da Informação) – Universidade Estadual Paulista Marília, 2023. Disponível em: https://repositorio.unesp.br/items/b2192b88-8362-488f-9b85-2c173eb66e48. Acesso em: 15 fev. 2024.

TROTMAN, A. An artificial intelligence approach to information retrieval. *In*: ACM SIGIR CONFERENCE ON RESEARCH AND DEVELOPMENT IN INFORMATION RETRIEVAL, 27., 2004, Sheffield. **SIGIR '04**: Proceedings of the 27th annual international [...]. New York: Association for Computing Machinery, p. 603-608. Disponível em: https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=4631952bd13d3a43f7be21d9245185d538d44545. Acesso em: 19 maio 2024.

ZHAO, W. X. *et al*. **A survey of large language models**. 2023. Disponível em: https://arxiv.org/pdf/2303.18223. Acesso em: 15 fev. 2024.