

RECUPERAÇÃO DE INFORMAÇÃO NA WEB: uma experiência com o modelo de linguagem de inteligência artificial ChatGPT

INFORMATION RETRIEVAL ON THE WEB: an experience with the ChatGPT artificial intelligence language model

 Patrícia Nascimento Silva¹

¹Professora Adjunta no Departamento de Organização e Tratamento da Informação na Escola de Ciência da Informação (ECI) da Universidade Federal de Minas Gerais (UFMG). Professora no Programa de Pós-Graduação em Gestão & Organização do Conhecimento (PPGGOC) ECI/UFMG.

E-mail: patricians@ufmg.br



ACESSO ABERTO

Copyright: Esta obra está licenciada com uma Licença Creative Commons Atribuição 4.0 Internacional. 

Conflito de interesses: A autora declara que não há conflito de interesses.

Financiamento: Não há.

Declaração de Disponibilidade dos dados: Todos os dados relevantes estão disponíveis neste artigo.

Recebido em: 3 mar. 2023.

Aceito em: 26 set. 2023.

Publicado em: 18 out. 2023.

Como citar este artigo:

NASCIMENTO SILVA, Patrícia. Recuperação de informação na Web: uma experiência com o modelo de linguagem de inteligência artificial ChatGPT. **Informação em Pauta**, Fortaleza, v. 8, p. 1-19, 2023. DOI: 10.36517/2525-3468.ip.v8i0.2023.83566.1-19.

RESUMO

Recuperação de informação é um termo recente, criado em 1951, contudo seu objetivo é almejado há milênios na sociedade. Com a evolução da informática, na década de 1970, e principalmente a criação da Web, em 1989, o termo teve seu significado potencializado em novos cenários. Os modelos clássicos de recuperação de informação obtiveram contribuições das técnicas de computação para otimizar os resultados e dialogar com os usuários, como proposto pelo modelo de linguagem de inteligência artificial ChatGPT, lançado em 2022. O objetivo deste relato foi analisar a recuperação de informação na Web com o ChatGPT. Para tanto, foi selecionado um domínio: dados abertos, que é originalmente diverso, e realizadas buscas exploratórias para identificar as informações e objetos digitais recuperados, em fevereiro de 2023. O modelo utilizado no ChatGPT demonstrou uma capacidade superior aos motores de busca da Web, ao recuperar informações e recursos de diferentes fontes e apresentá-las em um texto organizado e coeso. No entanto, a falta de referência às fontes utilizadas e de clareza sobre questões éticas são problemas visualizados na versão gratuita. Espera-se que com ajustes o ChatGPT recupere diversos recursos na Web, alcançando interoperabilidade com outros sistemas, contribuindo com o intercâmbio de dados em diversos domínios.

Palavras-chave: recuperação de informação; inteligência artificial; ChatGPT; assistente virtual; dados abertos.

ABSTRACT

Information retrieval is a recent term, created in 1951, but its objective has been desired for millennia in society. With the evolution of information technology, in the 1970s, and especially the creation of the Web, in 1989, the term had its meaning enhanced in new scenarios. The classic information retrieval models obtained contributions from computing techniques to optimize results and dialogue with users, as proposed by the ChatGPT artificial intelligence language model, launched in 2022. The objective of this report was to analyze information retrieval on the Web with ChatGPT. For this purpose, a domain was selected: open data, which is originally diverse, and exploratory searches were carried out to identify the

information and digital objects recovered, in February 2023. The model used in ChatGPT demonstrated a superior capacity to web search engines, by retrieving information and resources from different sources and presenting them in an organized and cohesive text. However, the lack of reference to the sources used and clarity on ethical issues are problems seen in the free version. It is expected that, with adjustments, ChatGPT will recover several resources on the Web, achieving interoperability with other systems, contributing to the exchange of data in different domains.

Keywords: information retrieval; artificial intelligence; ChatGPT; virtual assistant; open data.

1 INTRODUÇÃO

A recuperação de informação é uma área da ciência da computação aplicada a diversos domínios. Na ciência da informação, a temática trata da representação, armazenamento, organização e acesso à informação, sempre considerando o contexto do usuário. Apesar de o termo ter sido criado somente nos anos 1950, por Calvin Moores, há milênios as bibliotecas já organizavam a informação com o objetivo de facilitar sua recuperação.

Durante a década de 1970, vários modelos de recuperação de informação foram propostos, se tornando base para os sistemas de informação da época. Contudo, além dos modelos clássicos, evoluções foram percebidas com a criação dos modelos dinâmicos, especialmente após a criação da Web, em 1989.

Na representação do processo de recuperação, a expressão de busca é o meio que o usuário emprega para comunicar a sua necessidade informacional para o sistema de informação, sendo composta por linguagem natural ou artificial. Ao utilizar uma linguagem artificial é necessário que o usuário tenha conhecimento do tema de interesse e do seu vocabulário de domínio (Ferneda, 2012).

Com isso, pensado em oferecer facilidades ao usuário, são cada vez mais comuns sistemas que utilizam a linguagem natural para interação com usuário e tratam a representação da busca no sistema, traduzindo a necessidade de informação do usuário. Assistentes virtuais, ou *chatbot*, foram desenvolvidos para ter uma conversa com um

humano, interagindo em um formato semelhante ao de mensagens instantâneas (Queiroz; Valls, 2022). O uso de *chatbots*, associados a aplicativos de mensagens, aumentou significativamente nos últimos anos e se popularizou em sistemas de atendimento.

O ChatGPT é uma ferramenta recente, lançada em 30 de novembro de 2022, e tem fomentado várias discussões entre usuários e pesquisadores. A ferramenta é disponibilizada de forma gratuita, inicialmente, e já possui mais de 100 milhões de usuários ativos em dois meses, um recorde que supera o crescimento das redes sociais mais famosas do mundo (Forbes, 2023). A proposta da ferramenta é construir textos, códigos e outros recursos a partir de um modelo de linguagem de inteligência artificial, baseado em uma pergunta do usuário, escrita em linguagem natural.

Esta pesquisa é um relato de experiência com orientação prática/tecnológica que pretende responder a seguinte questão: Quais tipos de informação e recursos são recuperados pela ferramenta ChatGPT? O objetivo do estudo é analisar a recuperação de informação na Web com a ferramenta ChatGPT. Para tanto, foi selecionado um domínio e definidos critérios de busca para analisar a ferramenta. Destaca-se que este relato de experiência não é exaustivo e trata-se um estudo inicial, integrante de um projeto maior, e justifica-se para apresentar a nova ferramenta à comunidade acadêmica e fomentar novos estudos e possíveis aplicações no campo da ciência da informação.

2 MODELOS DINÂMICOS E LINGUAGEM NATURAL

Os modelos dinâmicos são alternativas promissoras e muito estudadas recentemente que possibilitam ao usuário interagir e interferir diretamente no processo de recuperação de informação. Baseado em técnicas oriundas da inteligência artificial, oportuniza a evolução e adaptação da informação de acordo com os interesses de busca do usuário e interações com o sistema (Baeza-Yates; Ribeiro-Neto, 2013)(Ferneda, 2003). Previamente, a inteligência artificial pode ser definida como uma área da ciência da computação dirigida para a criação de soluções inteligentes de tecnologia conectadas com a inteligência humana para resoluções de problemas, aprendizagem, compreensão e raciocínio lógico (Barr; Feigenbaum, 1981).

O Processamento da Linguagem Natural (PLN) é um conjunto de técnicas computacionais para análise de textos em um ou mais níveis linguísticos, com o propósito de simular o processamento humano da língua (Ferneda, 2012). Além disso, é a área da inteligência artificial responsável por estudar a capacidade que uma máquina tem de estender a linguagem dos seres humanos, por meio de diversos algoritmos computacionais.

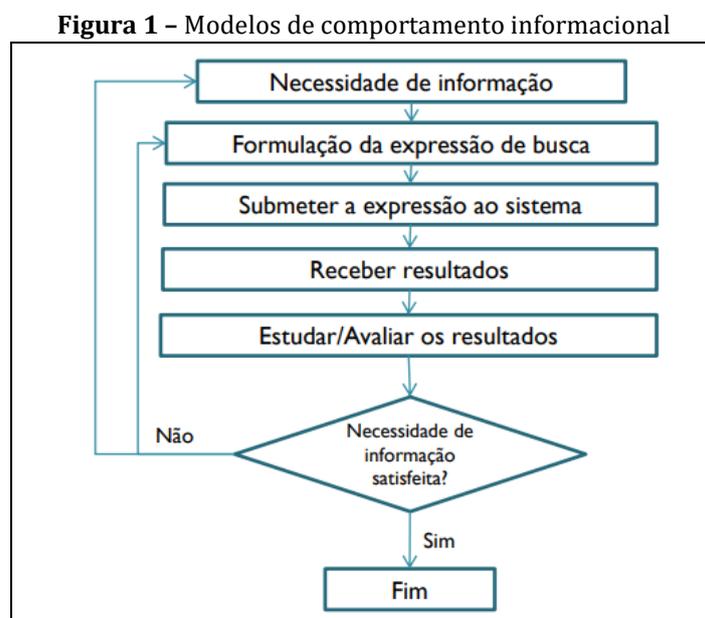
Uma das primeiras aplicações da linguagem natural em sistemas foi por meio das **questionanswering**: sistemas que têm como objetivo fornecer informações precisas e diretas respondendo a uma pergunta construída por um usuário. A ideia estava compreendida no usuário escrever uma pergunta, e buscar em um conjunto de bases, a resposta para tal. Há uma grande complexidade nesse processo, visto que é necessário tratar a pergunta, e encontrar nela, o que o usuário está de fato buscando e, posteriormente, utilizar PLN nos textos que podem conter a resposta, para tratar o que será apresentado como resposta ao usuário. Em seguida foram desenvolvidos **sistemas tutores inteligentes**: sistemas que utilizam diversos conceitos de PLN e constroem uma “rede de conhecimento” com diversos fatos, relações e regras, que permitem ao sistema realizar um diálogo com um indivíduo, **sistemas de automação de tarefas**: sistema que auxiliam em tarefas administrativas e gerenciais de uma empresa, passando por questões como agendamento de reuniões, compras de passagens aéreas e até detecção de erros ortográficos, **assistentes virtuais**: sistemas que auxiliam as pessoas em tarefas cotidianas, que vão desde realizar uma busca, mandar uma mensagem, verificar preços, ou até mesmo realizar uma compra e **sistemas de tradução automática**: sistemas que buscam a correspondência direta entre a língua original e a traduzida. Estas aplicações também aplicam a inteligência artificial e o PLN em sistemas (Coneglian, 2020).

Modelos baseados na Web se configuraram como as primeiras máquinas de busca e eram fundamentalmente sistemas de recuperação de informação cuja estratégia de ranqueamento era baseada em modelos tradicionais. As diferenças estavam nas coleções que eram compostas por páginas Web, e não mais em documentos, que precisavam ser coletadas e pertenciam a coleções muito maiores. Com isso, cada palavra da consulta recuperava muitas páginas com muitos documentos associados a um único resultado (Baeza-Yates; Ribeiro-Neto, 2013). Conforme Ferneda (2003, p.15) “os sistemas de recuperação de informação devem representar o conteúdo dos documentos do corpus e apresentá-los ao usuário de uma maneira que lhe permita uma rápida seleção dos itens

que satisfazem total ou parcialmente à sua necessidade de informação.” No entanto, para otimizar essa rápida seleção, diversas técnicas precisaram ser empregadas.

Grande parte dos mecanismos de busca constrói e armazena um resumo de cada página em sua base de dados. Em muitos casos este resumo é formado por uma quantidade fixa de palavras ou caracteres a partir do início do texto. A indexação antecipada dos documentos (representado por páginas) é essencial e pode ser feita pela **indexação manual**: indexadores profissionais especificam uma hierarquia de assuntos e indexam as páginas Web utilizando as categorias ou pela **indexação automática**: seleção de endereços de páginas, por meio de programas, que rastreiam a estrutura hipertextual da Web, colhendo informação sobre as páginas com técnicas de rastreamento (*crawlers, spiders, robôs*) (Ferneda, 2012).

Conforme apresentado neste breve referencial teórico, modelos, técnicas e aplicações são relevantes na recuperação da informação, no entanto, o sucesso de uma busca dependerá também do comportamento informacional do usuário. A Figura 1 apresenta o modelo básico de comportamento informacional para o usuário. A cada nível de representação pode ocorrer uma perda ou distorção em relação ao nível anterior e caso a necessidade de informação não seja satisfeita, uma nova expressão de busca deverá ser formulada.



Fonte: Ferneda (2020).

Os desafios da recuperação de informação são vastos e aumentam no contexto da Web, onde constantemente se produzem volumes excessivos de informações, por multiplicidade de usuários, de difícil mapeamento e de diferentes modelos de interação, apropriação e representação (Nhacuongue, 2015). Para apoiar a recuperação de informação linguagens e modelos são criados para refletir a representação da informação e otimizar sua recuperação. O ChatGPT é um modelo de linguagem que interage de forma conversacional com os usuários. Ele foi ajustado a partir de um modelo da série GPT-3.5,¹ que terminou o treinamento no início de 2022, e foi lançado em 30 de novembro de 2022. O formato de diálogo permite que o ChatGPT responda a perguntas de acompanhamento, admita seus erros, conteste premissas incorretas e rejeite solicitações inadequadas (OpenAI, 2023).

É importante destacar que o ChatGPT é um protótipo e não está pronto para ser vendido para empresas. Para gerar as informações pedidas pelos usuários, o *software* recorre a um vasto banco de textos disponíveis na internet, o que inclui artigos, notícias e até mesmo publicações do X, antigo Twitter. Atualmente, a ausência de definição sobre questões éticas gera críticas recorrentes à ferramenta (TechTudo, 2022).

3 METODOLOGIA

A pesquisa caracterizou-se como um estudo descritivo e exploratório, com abordagem qualitativa, por meio de uma análise documental para compreender a recuperação de informação na Web com a ferramenta ChatGPT. A pesquisa realizada também pode ser definida como pesquisa aplicada e pesquisa ação, por envolver a participação ativa da autora, e por seu objetivo acadêmico ser a produção de conhecimento para aplicação prática e imediata.

Devido à amplitude da ferramenta, que se propõem a construir um texto ou retornar objetos (códigos-fonte, *links*, vídeos, imagens) sobre qualquer assunto, foi selecionado um domínio específico a ser analisado. Este recorte considerou três critérios: (1) **Conhecimento prévio da autora sobre o assunto** que poderia validar os resultados apresentados pela ferramenta; (2) **Temática de contexto mundial**, para que fosse possível analisar resultados globais e (3) **Resultados que poderiam retornar informações textuais e outros objetos digitais**. Desta forma, a temática selecionada

¹ <https://platform.openai.com/docs/model-index-for-researchers>.

foi: dados abertos, no idioma português. Essa temática é originalmente complexa e diversa, e como resultado pode retornar arquivos e conjuntos de dados de diversas tipologias. É importante salientar que a temática é geral e essa definição foi pensada para não limitar os resultados recuperados pela ferramenta.

Após a definição do domínio foram criadas expressões de busca, ou seja, perguntas em linguagem natural, simulando a interação do usuário com o ChatGPT. Destaca-se que as buscas não foram exaustivas e que foram construídas perguntas na norma culta da língua portuguesa, um padrão diferente do utilizado em motores de buscas conhecidos na internet. Os critérios consideraram pesquisas mais amplas, inicialmente, e elementos mais específicos, no contexto dos dados abertos brasileiros, em seguida, a fim de identificar os principais conceitos e validar os resultados recuperados. O Quadro 1 apresenta a relação do tipo de busca, ou seja, a tipologia esperada/desejada, com a respectiva a necessidade de informação do usuário e a expressão de busca criada, que corresponde a pergunta para o ChatGPT.

Quadro 1 – Intenções e expressões de busca

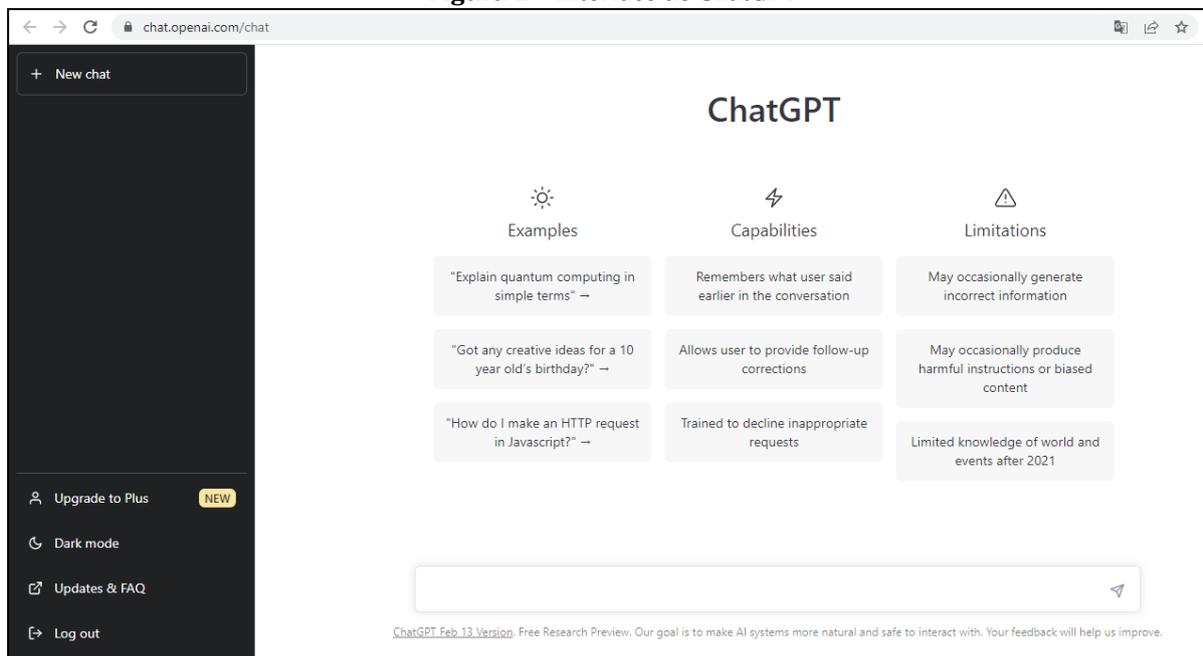
Tipo de busca	Necessidade de Informação	Expressão de busca
Busca textual	Identificar conceitos e informações gerais sobre da temática de dados abertos	(P1) O que são dados abertos? (P2) Apresente exemplos de dados abertos. (P3) Exiba dados abertos disponibilizados no Brasil.
Busca por objetos, links, arquivos.	Identificar conjuntos de dados abertos específicos sobre a vacinação no Brasil	(P4) Quais são os dados abertos sobre vacinação no Brasil? (P5) Onde acessar dados abertos sobre vacinação no Brasil? (P6) Exiba arquivos de dados abertos sobre a vacinação no Brasil.
Busca por recursos: ferramentas, APIs, códigos-fonte	Identificar ferramentas para manipulação de dados abertos e como acessá-las	(P7) Quais ferramentas disponíveis para acesso, uso e reúso dos dados abertos? (P8) Quais são as APIs de dados abertos disponíveis no Brasil? (P9) Como acessar as APIs de dados abertos? (P10) Exiba um código para coleta de dados abertos

Fonte: Próprios autores.

As intenções e expressões de busca estão associadas a resultados textuais e possíveis objetos digitais, uma vez que a ferramenta propõe uma recuperação de diversos recursos. Inclusive, essa diversidade de recursos caracteriza-se como um dos motivos da popularidade da ferramenta entre os usuários.

As 10 expressões de buscas foram executadas em fevereiro de 2023, em três momentos diferentes, por meio de um acesso gratuito. A execução foi feita mais de uma vez, pois a proposta da ferramenta é gerar textos diferentes, sem plágio, para cada resposta. Foi utilizada a interface Web da ferramenta, opção “New chat”, disponível no endereço: <https://chat.openai.com/chat>, com a versão Feb 13,² conforme apresentado na Figura 2. Os resultados obtidos foram disponibilizados integralmente no repositório desta pesquisa, no Mendeley Data, com o identificador: doi: 10.17632/ccdrbyn97g.1 (Silva, 2023).

Figura 2 – Interface do ChatGPT



Fonte: OpenIA (2023).

A análise dos resultados considerou a avaliação da recuperação, analisando a qualidade dos resultados, sem considerar a avaliação do desempenho do sistema, que incluiria, por exemplo, o tempo de processamento das buscas. Na próxima seção, a experiência com a ferramenta e as respostas recuperadas no ChatGPT foram analisadas, classificadas e discutidas sob o olhar da ciência da informação, especialmente no que tange as técnicas de recuperação de informação na Web.

² <https://help.openai.com/en/articles/6825453-chatgpt-release-notes>.

4 RESULTADOS

As respostas obtidas no ChatGPT, para as perguntas formuladas, foram analisadas e classificadas conforme a tipologia esperada e a necessidade de informação do usuário, em quatro categorias:

- Totalmente Pertinente (TP): quando o resultado era totalmente adequado a tipologia esperada e a necessidade de informação;
- Pertinente (P): quando o resultado era adequado a necessidade de informação, mas não apresentava o formato esperado;
- Parcialmente Pertinente (PP): quando o resultado apresentava parcialmente os resultados esperados, tanto na tipologia quanto na necessidade de informação;
- Incorreta (I), quando o resultado não estava adequado a necessidade de informação ou apresenta resultados incorretos.

A avaliação dos resultados pautou-se no atendimento as categorias definidas e foi baseado no conhecimento específico da área de dados abertos e recuperação de informação, pela autora. Independentemente das respostas obtidas na ferramenta, também foram analisados elementos e técnicas de recuperação de informação identificadas no uso da ferramenta. Destaca-se que os resultados foram predominantemente textuais. A ferramenta retornou textos e um único tipo de objeto digital (código-fonte). O Quadro 2 apresenta o resumo das análises das respostas, em cada uma das três execuções.

Quadro2 – Análise das respostas no ChatGPT

Expressão de busca (Pergunta)	Execução 1 (R1)	Execução 2 (R2)	Execução 3 (R3)
(P1) O que são dados abertos?	TP	TP	TP
(P2) Apresente exemplos de dados abertos.	PP	PP	PP
(P3) Exiba dados abertos disponibilizados no Brasil.	PP	P	PP
(P4) Quais são os dados abertos sobre vacinação no Brasil?	PP	PP	P
(P5) Onde acessar dados abertos sobre vacinação no Brasil?	P	P	PP
(P6) Exiba arquivos de dados abertos sobre a	P	P	P

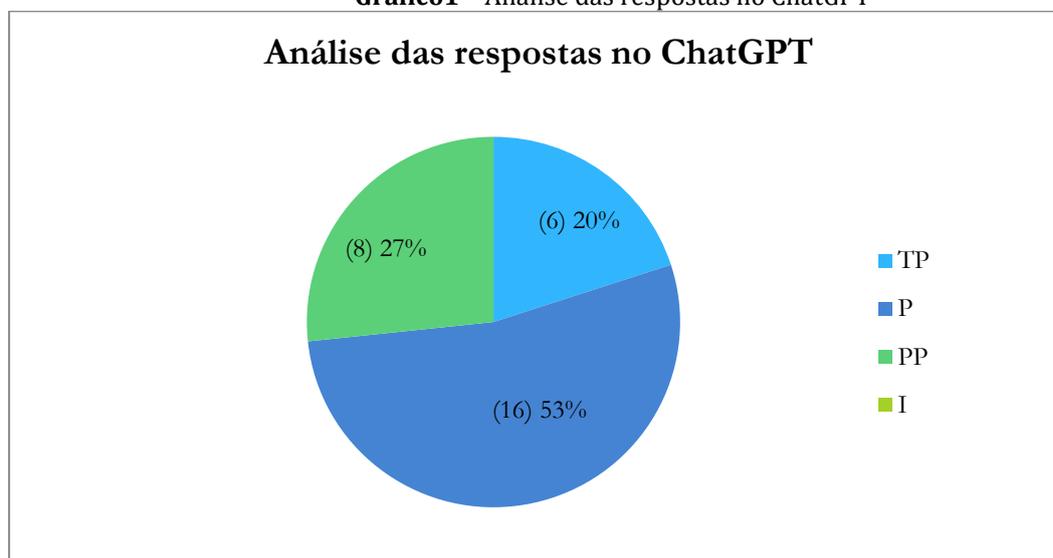
vacinação no Brasil.			
(P7) Quais ferramentas disponíveis para acesso, uso e reúso dos dados abertos?	P	P	P
(P8) Quais são as APIs de dados abertos disponíveis no Brasil?	P	TP	TP
(P9) Como acessar as APIs de dados abertos?	P	P	P
(P10) Exiba um código para coleta de dados abertos	P	P	TP

Fonte: Dados da pesquisa.

Observou-se que as respostas obtidas nas três execuções foram similares, mas não idênticas, cumprindo com a proposta da ferramenta de não configurar plágio. No entanto, a não indicação das fontes de utilizadas impede qualquer tipo de checagem. Assim, a validação realizada foi baseada exclusivamente no conhecimento da autora, fundamentada nas experiências práticas e na literatura da área.

Não foi identificada nenhuma resposta incorreta. Contudo, destaca-se que a maioria das respostas (53%) foi considerada pertinente, ou seja, apresentaram resultados corretos, mas não no formato esperado. Em seguida 27% dos resultados foram classificados como parcialmente pertinentes e 20% como totalmente pertinentes, conforme apresentado no Gráfico 1:

Gráfico1 - Análise das respostas no ChatGPT



Fonte: Dados da pesquisa.

Para pergunta P1, as três repostas (P1 - R1, P1 - R2 e P1 - R3) foram totalmente pertinentes e, além do conceito de dados abertos, os resultados apresentaram no texto

informações sobre o objetivo, formatos e tipologias dos dados abertos. O usuário obteve a informação desejada de forma completa e no formato de um texto coeso, conforme a real necessidade de informação.

Na pergunta P2, os resultados retornados nas respostas do ChatGPT são adequados a necessidade de informação, mas exemplos práticos não foram apresentados, nem em outros formatos. Constatou-se que o conceito foi informado novamente, como apresentado em P1, juntamente com conteúdos e exemplos que representassem alguns tipos de dados abertos:

[...] Dados de saúde - muitos governos disponibilizam dados de saúde, como o Ministério da Saúde no Brasil e o Centers for Disease Control and Prevention (CDC) nos Estados Unidos. Esses dados incluem informações sobre doenças, tratamentos e estatísticas de saúde pública.

Dados de transporte público - muitas cidades disponibilizam dados de transporte público, como horários de ônibus, rotas de metrô e dados de tráfego. Esses dados podem ser usados para desenvolver aplicativos de transporte público e para ajudar as pessoas a planejar suas viagens (P2 - R3).

[...] Dados de pesquisa: Pesquisadores e acadêmicos muitas vezes disponibilizam dados abertos de seus estudos, que podem incluir pesquisas de opinião pública, resultados de testes psicológicos, dados de imagens de satélite e muito mais (P2 - R2).

Apesar de os resultados serem apresentados no formato de uma lista, não é possível identificar o modelo de ordenação utilizado. Os motores de buscas atuais utilizam o ranqueamento e técnicas de ordenação, como, por exemplo, a classificação por popularidade. Nesta técnica encontram-se as páginas que são populares e as classificam antes de outras páginas que contêm as palavras-chave especificadas (Baeza-Yates; Ribeiro-Neto, 2013). Similar a essa classificação estão as páginas e anúncios pagos que são apresentados antes nos buscadores, como os primeiros resultados ranqueados. A definição dos critérios de ordenação pode envolver diversos interesses, inclusive comerciais, prática muito criticada atualmente e que tem impactado nos resultados de grandes motores de busca e plataformas de redes sociais.

Em P3 não foram apresentados excertos de conjuntos de dados abertos, como esperado e modelado na expressão de busca. Mas foram identificados os principais portais e alguns links indicados nas respostas. Destaca-se que das três execuções, somente a segunda execução (R2) apresentou os links dos portais, as outras somente informaram o nome do instrumento e os dados disponibilizados:

[...] DataSUS: o Ministério da Saúde disponibiliza diversos dados abertos sobre saúde, como o Sistema de Informação Hospitalar (SIH), Sistema de Informação da Atenção Básica (SIAB), entre outros (P3 - R1).

[...] Sistemas de informações do Ministério da Saúde: o DataSUS (www.datasus.gov.br) disponibiliza informações sobre a saúde no país, como o número de leitos, atendimentos e internações hospitalares. Já o Sistema Único de Saúde (SUS) oferece informações sobre a rede de saúde pública, como hospitais, postos de saúde e unidades de pronto-atendimento (P3 - R2).

A técnica *page rank* foi introduzida pela empresa Google e pode estar associada a este resultado, uma vez que dados abertos sobre saúde são apresentados com maior frequência nos resultados, no período da pandemia de Covid-19, que foi iniciada em 2020 e ainda estava vigente em fevereiro de 2023. Esta técnica baseia-se na probabilidade de atribuir popularidade ou *page rank* alto às páginas, isto é, parte do princípio de que um usuário pode navegar aleatoriamente em uma página e a partir daí existe a probabilidade de escolher um dos *links* externos da página. Para Silberschatz et al. (2006), a maior desvantagem da medida *page rank* é da recuperação de documentos por popularidade desconsiderando as palavras-chave da consulta. Uma solução para este problema consiste em adotar o uso de palavras-chave no texto de âncora dos *links* para uma página, determinando os tópicos a que a página é relevante, através da *tag* HTML “a href”. A definição de critérios específicos para ordenação destes resultados também não está descartada e uma política clara sobre esses critérios implementados nos algoritmos da ferramenta deveria ser disponibilizado aos usuários.

Ao perguntar sobre um conjunto de dados específico, em P4, não foram indicados os conjuntos de dados e sim os órgãos que os disponibilizam, respostas bem próximas das obtidas em P3. Na terceira execução (P4 - R3) foram indicados alguns *links* diretos para os conjuntos de dados:

[...] OpenDataSus: o portal do Datasus oferece diversas informações sobre o sistema de saúde brasileiro, incluindo dados sobre vacinação. É possível acessar informações sobre a cobertura vacinal por município, estado e país, bem como dados sobre a produção e distribuição das vacinas. Os dados podem ser acessados no endereço eletrônico: <http://opendatasus.saude.gov.br/dataset/covid-19-vacinacao> (P4 - R3).

Na pergunta P5 os repositórios com conjuntos de dados foram indicados diretamente, por meio de *links*, nas primeiras execuções, na terceira foi indicado

somente o instrumento, conforme apresentado em P5-R2. Há um destaque para apresentação de Portais de dados abertos que predominaram entre as respostas da ferramenta:

[...] Localiza SUS: plataforma que reúne informações sobre o sistema de saúde do Brasil, incluindo dados sobre a vacinação contra a COVID-19. Os dados podem ser acessados em: <https://localizasus.saude.gov.br/> (P5 – R2).

Destaca-se que as três execuções de P5 trouxeram elementos complementares e não somente repetiu uma lista. Os elementos novos apresentados, a cada execução da pergunta, podem evidenciar que a ferramenta utiliza técnicas de aprendizado de máquina. O Aprendizado de máquina ou *machine learning* é uma área da inteligência artificial que objetiva o desenvolvimento de sistemas capazes de aprender a partir de dados. Através da análise de problemas anteriores, algoritmos de aprendizado de máquina são capazes de realizar as tomadas de decisões (Monard; Baranauskas, 2003).

Desta forma, o comportamento da ferramenta ChatGPT aponta o uso de algoritmos de aprendizado de máquina, pois a cada execução elementos anteriores são retomados e complementados. No entanto, o fato de a ferramenta não produzir plágio pode influenciar negativamente na apresentação desses resultados, já que nem todos os resultados retornam todos os elementos sobre a temática buscada. Assim, para obter resultados mais completos na ferramenta, recomenda-se que a mesma busca seja feita várias vezes, e em momentos diferentes. É importante destacar que esse comportamento da ferramenta, que apresenta resultados diferentes para uma mesma expressão de busca, também podem incitar discussões sobre a confiabilidade, métodos e técnicas utilizadas e também algum potencial discriminatório do perfil algorítmico da ferramenta. Essas questões técnicas e éticas que não podem ser ignoradas e devem ser esclarecidas aos usuários do ChatGPT.

A pergunta P6, apesar de pertinente, não exibiu arquivos, nem parte do conteúdo de arquivos ou conjuntos de dados abertos. Foram indicados somente portais onde os arquivos poderiam ser acessados, similar às repostas apresentadas em P4. A partir destes resultados, estima-se que a solução de recuperação utilizada pelo ChatGPT deve utilizar consultas baseadas em conceito. Neste tipo de consulta é analisada a ambiguidade de cada palavra no respectivo documento, de modo que a substituição do termo da pesquisa se faça por outro mais próximo e empregue no mesmo documento. Embora a consulta baseada em conceitos tenha a desvantagem na sobrecarga e

processamento de documentos, para retirar a ambiguidade e por ser pouco utilizada na recuperação da Web, ela permite a recuperação de documentos relacionados através da hierarquia de conceitos, como ontologias (Silberschatz *et al.*, 2006).

O termo ‘dados abertos’, quando analisado separadamente, pode ter outros sinônimos e homônimos, em cenários diversificados. No entanto, os resultados apresentados pelo ChatGPT estavam dentro do contexto esperado e os portais indicados eram instrumentos relevantes e utilizados com frequência por pesquisadores da área.

Para a pergunta P7 o ChatGPT indicou o nome e uma descrição das ferramentas populares para análise de dados, conforme a necessidade de informação da busca, contudo foi apresentada uma lista, sem exemplos práticos relacionados:

[...] Google Dataset Search - é um mecanismo de busca que permite a busca de conjuntos de dados em vários repositórios na internet.

OpenRefine - é uma ferramenta gratuita e de código aberto que ajuda a limpar, transformar e analisar dados abertos.

R - é uma linguagem de programação estatística e ambiente de desenvolvimento que permite a análise de dados, visualização e modelagem.

Tableau Public - é uma plataforma gratuita para visualização de dados que permite a criação de gráficos interativos, mapas e dashboards.

QGIS - é um software de Sistema de Informação Geográfica (GIS) de código aberto que permite a visualização, análise e processamento de dados geoespaciais (P7 - R2).

Para a pergunta P8 foi apresentado um breve conceito e indicadas *Application Programming Interface* (API)s que disponibilizam dados abertos, por meio de um *link* direto, na segunda e na terceira execução:

[...] IBGE APIs (<https://servicodados.ibge.gov.br/api/docs/>): o Instituto Brasileiro de Geografia e Estatística (IBGE) também oferece diversas APIs para acesso a dados relacionados a população, produção agrícola, economia, entre outros.

SINESP Cidadão API (<https://github.com/servicosgovbr/api-sinesp-cidadao>): o Sistema Nacional de Informações de Segurança Pública (SINESP) disponibiliza uma API que permite acessar informações sobre veículos roubados, furtados ou clonados.

OpenWeatherMap API (<https://openweathermap.org/api>): essa API permite acessar dados climáticos de diversas cidades brasileiras e de todo o mundo.

ANSAPI (https://www.ans.gov.br/externo/site_novo/informacoes_avaliacoes_oper/list_a_operadoras/todas_operadoras.asp): a Agência Nacional de Saúde

Suplementar (ANS) disponibiliza uma API para acesso a informações sobre operadoras de planos de saúde (P8 - R3).

A indicação do *link* das APIs nas respostas demonstrou que o ChatGPT recuperou a informação no formato esperado. Vale ressaltar que em motores de busca tradicionais essa informação é difícil de ser recuperada, com uma expressão de busca mais ampla, como a utilizada na pergunta P8.

Nas perguntas P9 e P10 eram esperadas respostas com exemplos práticos. No entanto, para P9 foi apresentada uma lista com etapas descritivas sobre como utilizar uma API. Foram indicados recursos comuns para esse tipo de acesso, de forma detalhada e explicativa, para o usuário:

Acesso às APIs de dados abertos varia dependendo da fonte dos dados, mas aqui estão alguns passos gerais que você pode seguir:

Encontre a API: Descubra quais organizações governamentais ou outras entidades oferecem APIs de dados abertos. Uma boa fonte para isso é o portal de dados abertos do seu governo local ou o portal de dados abertos da sua região. Outra opção é fazer uma pesquisa na internet com palavras-chave relevantes, como "API de dados abertos [nome da cidade/estado/país]".

Obtenha uma chave da API: Algumas APIs de dados abertos exigem que você obtenha uma chave de API antes de poder acessá-las. Isso é feito para rastrear o uso da API e limitar o acesso a dados pessoais ou restritos. A maioria das fontes de dados fornece instruções claras sobre como obter uma chave da API.

Entenda a documentação: Depois de encontrar a API e obter a chave, é importante ler a documentação fornecida para entender os endpoints disponíveis, os parâmetros que podem ser usados e os dados que podem ser acessados. Isso ajudará você a construir suas consultas de maneira eficaz e entender como os dados são estruturados.

Construa consultas: Com a chave da API e a compreensão da documentação, você pode começar a construir suas consultas para acessar os dados. Isso geralmente é feito usando uma linguagem de consulta, como o SQL, ou uma linguagem de programação, como o Python ou o Java.

Analise os dados: Uma vez que você tenha feito as consultas e obtido os dados, é hora de analisá-los. Isso pode ser feito com o uso de ferramentas de análise de dados, como o Excel, o R ou o Python.

É importante lembrar que diferentes fontes de dados abertos podem ter requisitos e procedimentos diferentes para acessar suas APIs. Portanto, é importante ler atentamente a documentação e seguir as instruções fornecidas (P9 – R2).

A apresentação dos resultados no formato de listas ou etapas surpreendeu positivamente. Apresentar os dados em etapas ao usuário permite uma melhor estruturação do conhecimento e traz um viés prático e real. O desenvolvimento de

metodologias cada vez mais avançadas no PLN, derivadas do aprendizado de máquina e de outros recursos de inteligência artificial, mais avançados, como *Deep Learning* e Redes neurais, reforçam o papel inevitável do algoritmo e do processamento estatístico da informação, tanto para fins de recuperação como para o uso em visualizações (Wuet *al.*, 2013). Estima-se que essa estruturação possa facilitar a integração com outras ferramentas, otimizando a visualização da informação para os usuários.

Na resposta de P10 foi apresentado um texto e um código-fonte, conforme sugerido na expressão formulada. Apesar de ser representado por um texto, o código pode ser considerado um objeto digital, pois é executável em um interpretador da linguagem Python. Na primeira e na segunda execução foram apresentados de códigos-fonte para uma API de exemplo. Na terceira execução foi apresentado um *endpoint*³ válido sobre dados abertos, conforme solicitado na expressão de busca. Este último código recuperado é simples e corresponde a um acesso à API do Portal Brasileiro de Dados Abertos, sem representar uma coleta específica, conforme apresentado na Figura 3.

Figura 3 – Código-fonte gerado pelo ChatGPT

```
python Copy code  
  
import requests  
  
url = "https://dados.gov.br/api/action/datastore_search?resource_id=abc123&limit=1000"  
  
response = requests.get(url)  
  
if response.status_code == 200:  
    data = response.json()  
    # Faça algo com os dados  
else:  
    print("Erro ao acessar os dados")
```

Fonte: OpenIA (2023).

Quando comparado a motores de buscas tradicionais, o ChatGPT apresenta resultados unificados e estruturados em um texto bem escrito e coeso, no idioma português. A recuperação de objetos ainda é reduzida, mas é um recurso promissor que permitirá a ferramenta “interoperar” com diferentes sistemas e aplicações.

³ https://dados.gov.br/api/action/datastore_search?resource_id=abc123&limit=1000.

A falta de referência à fonte coletada é um problema para o usuário que utiliza a versão gratuita. A falta dessa informação inviabiliza a utilização dos textos em trabalhos acadêmicos e publicações, o que pode ser benéfico frente a discussões de ensino aprendizagem, mas é prejudicial para a ferramenta, que pode estar utilizando fontes duvidosas e contribuindo com a desinformação. Essa “dúvida” sobre as fontes utilizadas pode afastar usuários e findar oportunidades para OpenIA, principalmente no meio acadêmico.

Ao longo da execução das buscas percebeu-se que a ferramenta identifica erros de grafia e concordância ao realizar a busca. Ao digitar uma frase com erro de digitação, por exemplo, o ChatGPT retornou a seguinte resposta:

Peço desculpas, mas preciso de mais informações para entender melhor sua pergunta. Você poderia esclarecer o que você quer dizer com "datos abertos disponibilizados no Brasil"? Existem muitas iniciativas de dados abertos em diferentes áreas no país, e as informações que você procura podem variar de acordo com sua necessidade específica. Quanto mais detalhes você fornecer, melhor poderei ajudá-lo (ChatGPT, 2023, s/p).

A versão gratuita também impõem limitação ao número de consultas em um determinado espaço de tempo. A mensagem: *“Too many requests in 1 hour. Try again later”* foi exibida após realizar aproximadamente 50 consultas em um curto período de tempo. Durante o uso da ferramenta não foram indetificados problemas ou indiponibilidade. Mesmo que o foco da análise não envolvesse o desempenho, contactou-se que o tempo médio das respostas crescia quando as buscas envolviam elementos mais específicos sobre os dados abertos.

5 CONCLUSÃO

O vertiginoso sucesso da ferramenta ChatGPT e incógnitas sobre o funcionamento da ferramenta foi uma das motivações deste estudo que apresentou um breve relato de experiência, com orientação prática/tecnológica, objetivando analisar a recuperação de informação na Web com a ferramenta ChatGPT. Os modelos tradicionais de recuperação de informação ainda são utilizados na maioria dos motores de busca e a utilização de modelos dinâmicos, e o uso da linguagem natural, correspondem a uma evolução natural e muito aguardada pelos usuários.

Apesar de a ferramenta ainda estar em desenvolvimento é perceptível que a proposta é interessante, contribuindo para a recuperação de informação de forma otimizada e em diferentes contextos. A inteligência artificial, juntamente com o PLN, o aprendizado de máquina e outras técnicas da computação poderão contribuir com a recuperação de informação na sociedade contemporânea, onde a quantidade e a diversidade de dados são crescentes e seu significado e contexto são cada vez mais valorizados e “interoperados” entre sistemas.

Os resultados obtidos demonstraram o potencial da ferramenta, ainda que somente 20% dos resultados foram apresentados no formato esperado. No entanto, por ser uma ferramenta em desenvolvimento e muito recente, naturalmente deverá ter correções e ajustes até o lançamento de uma versão *gold*,⁴ principalmente para o serviço de assinatura pago, anunciado em fevereiro de 2023. Fundamentos de modelos tradicionais de recuperação de informação na Web aparentam estar implementados na ferramenta, mas poucos detalhes técnicos foram disponibilizados pela OpenIA até o momento. A não reprodutibilidade das buscas, a fim de evitar o plágio, e questões éticas precisam ser esclarecidas aos usuários, assim como as fontes de informação utilizadas.

Por fim, o ChatGPT demonstrou ser um recurso instigante de recuperação de informação na Web, com vasto potencial de crescimento. Com as devidas informações sobre critérios aplicados, fontes utilizadas e a inclusão de novos recursos de recuperação de informação, em diferentes formatos, o ChatGPT certamente poderá contribuir com o acesso, a recuperação e o reúso dos dados abertos, bem como a recuperação de informação de outros domínios, apoiando os usuários em tarefas de coleta e análise de dados. A integração com outros sistemas, via API, estende ainda mais as possibilidades de intercâmbio de dados com o ChatGPT, cujas aplicações e impactos devem ser acompanhados atentamente pelo profissional da informação.

⁴ Versão final, pronta para ser comercializada.

REFERÊNCIAS

- BAEZA-YATES, R.; RIBEIRO-NETO, B. **Recuperação de informação: conceitos e tecnologia das máquinas de busca**. 2. ed. Porto Alegre: Bookman, 2013.
- BARR, A.; FEIGENBAUM, E. **The Handbook of Artificial Intelligence**. Boston: Addison Wesley, 1981. v. 1.
- CONEGLIAN, C. S. **Recuperação da informação com abordagem semântica utilizando linguagem natural: a inteligência artificial na ciência da informação**. Tese (Doutorado em Ciência da Informação) - Universidade Estadual Paulista, São Paulo, 2020. Disponível em: <http://hdl.handle.net/11449/193051>. Acesso em: 19 fev. 2023.
- FERNEDA, E. **Recuperação de informação: análise sobre a contribuição da Ciência da Computação para a Ciência da Informação**. 2003. 137 f. Tese (Doutorado em Ciências da Comunicação) - Universidade de São Paulo, São Paulo, 2003.
- FERNEDA, E. **Introdução aos Modelos Computacionais de Recuperação de Informação**. Rio de Janeiro: Editora Ciência Moderna Ltda. 2012.
- FERNEDA, E. Recuperação de informação (distância). Revisão. 2020. Disponível em: <https://www.marilia.unesp.br/Home/Instituicao/Docentes/EdbertoFerneda/ri-ead-12-revisao.pdf>. Acesso em: 19 fev. 2023.
- FORBES. ChatGPT tem recorde de crescimento da base de usuários. 2023. Disponível em: <https://forbes.com.br/forbes-tech/2023/02/chatgpt-tem-recorde-de-crescimento-da-base-de-usuarios/>. Acesso em: 19 fev. 2023.
- NHACUONGUE, Januário Albino. **O campo da Ciência da Informação: contribuições, desafios e perspectivas da mineração de dados para o conhecimento pós-moderno**. 2015. 194 f. Tese (Doutorado) - Universidade Estadual Paulista Júlio de Mesquita Filho, Faculdade de Filosofia e Ciências, 2015.
- OpenAI. ChatGPT: Otimizando modelos de linguagem para diálogo. 2023. Disponível em: <https://openai.com/blog/chatgpt/>. Acesso em: 17 fev. 2023.
- QUEIROZ, T. D. S.; VALLS, V. M. O bibliotecário analista de chatbot: as competências desenvolvidas nos cursos presenciais de bacharelado em biblioteconomia da cidade de São Paulo. **Revista Brasileira de Biblioteconomia e Documentação**, v. 18, p. 1-25, 2022. Disponível em: <http://hdl.handle.net/20.500.11959/brapci/193951>. Acesso em: 19 fev. 2023.
- SILBERSCHATZ, A. *et al.* **Sistema de Banco de Dados**. 5. ed. Rio de Janeiro: Elsevier, 2006.
- SILVA, P. Resultados ChatGPT dados abertos - Fevereiro 2023". Mendeley Data, V1, 2023. DOI: 10.17632/ccdrbyn97g.1.
- TECHTUDO. ChatGPT: entenda como funciona o chatbot 'sabe-tudo' da OpenAI. 2022. Disponível em: <https://www.techtudo.com.br/listas/2022/12/chatgpt-saiba-tudo-sobre-o-chatbot-que-usa-ia-para-responder-perguntas.shtml>. Acesso em: 17 fev. 2023.
- WU, X. *et al.* Data mining with big data. **IEEE Transactions on Knowledge and Data Engineering**, v. 26, n. 1, p. 97-107, 2014.