Estudo comparativo de web scraping tradicional e integrado com inteligências artificiais

Amanda de Oliveira Laurindo
UDC
Foz do Iguaçu, Brasil
amanda96laurindo@outlook.com

Luciano Santos Cardoso
UDC
Foz do Iguaçu, Brasil
luciano.cardoso@udc.edu.br

Alessandra Bussador
UDC
Foz do Iguaçu, Brasil
bussador@yahoo.com.br

***Abstract*—The integration of Artificial Intelligence (AI) techniques into web scraping has transformed the way data is collected and analyzed on the web. While traditional methods face limitations due to barriers such as CAPTCHAs, dynamic content, and structural changes in websites, AI-assisted approaches enable greater robustness, adaptability, and accuracy. Large Language Models (LLMs), machine learning, natural language processing (NLP), and computer vision allow tasks ranging from the interpretation of unstructured texts to multimodal analysis and the prediction of changes in web pages. These advances expand the potential of scraping in areas such as research, business, and real-time monitoring, although they raise ethical and legal concerns related to privacy, cultural bias, and the responsible use of information. This study comparatively analyzes traditional and AI-assisted scraping, highlighting benefits, challenges, and social implications.**

***Keywords*—Data; Web bots; LLMs.**

***Resumo*—A integração de técnicas de Inteligência Artificial (*AI*) ao *web scraping* tem transformado a forma como dados são coletados e analisados na *web*. Enquanto métodos tradicionais enfrentam limitações frente a barreiras como *CAPTCHA*, conteúdos dinâmicos e mudanças estruturais em sites, abordagens assistidas por *AI* permitem maior robustez, adaptabilidade e precisão. Modelos de linguagem de larga escala (*LLM*), *machine learning*, *natural language processing* (*NLP*) e visão computacional possibilitam desde a interpretação de textos não estruturados até a análise multimodal e previsão de alterações em páginas. Tais avanços ampliam o potencial do *scraping* em áreas como pesquisa, negócios e monitoramento em tempo real, embora levantem questões éticas e legais relacionadas à privacidade, vieses culturais e uso responsável das informações. Este estudo analisa comparativamente o *scraping* tradicional e o assistido por *AI*, destacando benefícios, desafios e implicações sociais.**

***Palavras-chave*—Dados; *Bots web*; LLM.**

I. Introdução

A informação sempre teve papel decisivo na sociedade, e o avanço tecnológico aumentou a demanda por técnicas como o *web scraping*, que permite coletar grandes volumes de dados para áreas como pesquisa, mercado e treinamento de modelos de linguagem. Apesar disso, medidas de proteção como *CAPTCHA* e bloqueios de *IP* dificultam sua prática, impulsionando métodos mais sofisticados para contornar barreiras. Alternativas como *API* oferecem acesso estruturado, mas com menor flexibilidade que a raspagem.

É importante diferenciar *web crawling* (descoberta e indexação de páginas) de *web scraping* (extração e análise dos dados). Atualmente, busca-se aprimorar o *scraping* com o uso de *AI* e *LLM* (*Large Language Models*), capazes de processar texto em larga escala e interpretar linguagem natural.

A aplicação de *AI* ao *scraping* possibilita avanços como:

- uso de *NLP* para interpretar textos não estruturados;
- redes neurais para reconhecer *layouts* visuais;
- *machine learning* para adaptação automática a diferentes sites;
- técnicas de visão computacional para superar *CAPTCHA*;
- sistemas de autoaprendizagem que mantêm eficiência mesmo diante de mudanças.

O estudo tem como objetivo comparar a raspagem tradicional com a assistida por *AI*, avaliando ganhos em qualidade e quantidade de dados, mas também os riscos envolvidos, como questões éticas, privacidade e vieses culturais. A proposta central é demonstrar como a integração de *AI* pode tornar a raspagem mais robusta, precisa, escalável e representativa, ao mesmo tempo em que enfrenta o desafio de garantir diversidade e alinhamento cultural nos dados coletados.

II. Referencial Teórico

A. Fundamentos da World Wide Web

A *World Wide Web (WWW)* é mais do que um conjunto de páginas interligadas: trata-se de um sistema global de comunicação baseado em universalidade, descentralização e interoperabilidade. Sua evolução levou da *web* estática à *web* semântica, onde metadados estruturados permitem melhor compreensão de conteúdos, facilitando processos como o *web scraping*.

Os protocolos de internet são fundamentais nesse processo:

- *HTTP/HTTPS* regula a transferência de dados via métodos como *GET* e *POST*;
- *TCP/IP* garante confiabilidade e ordem na comunicação;
- *DNS* traduz domínios em endereços *IP*.

A integração de *AI* com esses protocolos permite raspadores inteligentes capazes de ajustar frequência de requisições, prever falhas de rede, gerenciar *proxies* e otimizar coleta de dados.

Os websites modernos são sistemas complexos compostos por múltiplas camadas tecnológicas que trabalham em conjunto para fornecer a experiência do usuário. No front-end, temos a tríade funda- mental: HTML para estruturação do conteúdo, CSS (Cascading Style Sheets) para apresentação visual e JavaScript para adicionar interatividade e comportamento dinâmico. O HTML, em particular, é de extrema importância para o web scraping, pois contém toda a estrutura semântica da página, organizada em elementos hierárquicos que podem ser analisados e percorridos por algoritmos de extração. Técnicas de visões computacionais, por outro lado, analisam a estrutura e o layout das páginas web, facilitando a extração de dados estruturados de documentos HTML não estruturado [1].

A estrutura de dados na *web* envolve:

- Estruturados (*HTML*, *JSON*, *XML*), que podem variar entre sites;
- Não estruturados (textos, imagens, vídeos), que exigem técnicas como *NLP* e *OCR* para interpretação.

Um aspecto particularmente desafiador do HTML moderno é o uso crescente de JavaScript para renderização dinâmica de conteúdo. Muitos sites, especialmente aplicações *web* complexas, dependem pesadamente de *framework* como React, Angular ou Vue.js para construir suas interfaces. Muitas aplicações JavaScript carregam partes do código no tempo de execução, por exemplo, por meio da tag ¡script¿ de uma página HTML, que causa o navegador a buscar o código de um servidor, ou programaticamente por XMLHttpRequest, que carrega o código assincronicamente [2].

Web scraping melhorado com IA garante a possibilidade de juntar dados de websites como páginas web, PDF, vídeos, fotos e meta dados, de uma forma mais rápida e automática. Tecnologias de IA e *web crawling* juntam podem resultar em um processo de raspagem de dados mais produtivo por facilitar extração de dados mais efetiva e eficiente [1].

Com a *AI*, é possível identificar entidades, analisar sentimentos, resumir textos e priorizar páginas com base em relevância. Essa sinergia entre *scraping* e *AI* permite integrar múltiplas fontes, tornando a extração mais efetiva, holística e escalável.

Esses obstáculos tornam necessária a adoção de técnicas avançadas de *AI* e visão computacional para manter a eficiência e precisão da coleta.

B. Fundamentos da Raspagem de Dados

A raspagem de dados (*web scraping*) emergiu como uma das técnicas mais poderosas para coleta e estruturação de informações na era digital, permitindo transformar dados dispersos na *web* em conjuntos organizados e analisáveis. Esse processo automatizado de extração vai desde a simples captura de textos até a coleta complexa de dados estruturados de múltiplas fontes, servindo como base para análises empresariais, pesquisas acadêmicas e desenvolvimento de sistemas inteligentes. Na figura 1 é demonstrado o processo da raspagem de dados que começa pela url do site seguida pelo *fetching* que consiste de capturar a url usando os comandos curl ou wget, com isso o documento HTML é capturado e então a extração dele é feita podendo ser usado uma coleção de ferramentas como Regular Expressions, XPath e bibliotecas de parsing HTML, os dados desejados então são extraídos e são transformados em dados estruturados que então são dirigidos para o banco de dados ou são apresentados na aplicação.

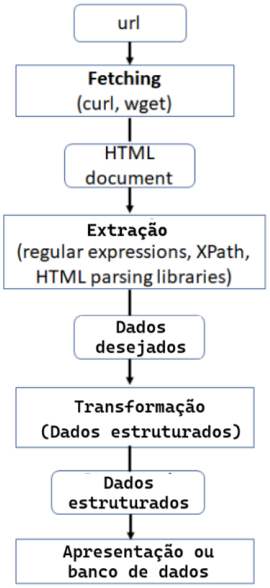

Fig. 1. Diagrama do processo para realizar a raspagem de dados [3]

Web scraping é usado em diferentes indústrias e pode ser usado para determinar preços e o custo de produção por averiguar os dados que foram capturados. Propagandas podem ser criadas para anunciar produtos à diferentes usuários dependendo dos seus dados como localização e configuração de cookies [4].

A ascensão da *web* semântica trouxe novos desafios e oportunidades. Seu objetivo, tornar os dados não apenas legíveis, mas também interpretáveis por máquinas, exige métodos sofisticados para extrair significado contextual. A anotação semântica,

por exemplo, permite marcar elementos com metadados que definem seu papel. *Web mining* faz-se possível localizar resultados permanentes na internet e é usado para separar os dados importantes dos padrões de descoberta armazenados nos servidores [5]. Nesse contexto, a raspagem atua como ponte entre a *web* convencional e a semântica: ao coletar dados brutos, possibilita sua posterior estruturação em ontologias ou grafos de conhecimento, onde relações entre entidades ganham significado.

Na figura 2 a diferença entre a estrutura da *web* superficial e a *web* semântica está sendo demonstrada com a superficial no lado esquerdo com os elementos na web sendo tratados todos iguais e com links ligando um ao outro, e no lado direito demonstra a *web* semântica na qual os recursos possuem metadados definindo o seu conteúdo e eles são interligados com contexto um ao outro.

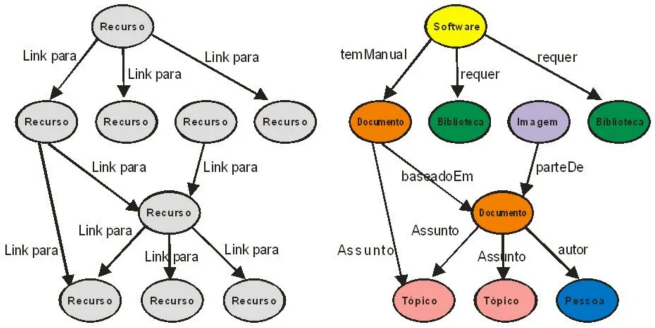

Fig. 2. Comparação entre a *web* atual e a *web* semântica [6]

A *AI* amplia as capacidades do *scraping*, automatizando:

- pré-processamento;
- interpretação multimodal (texto, imagem, vídeo);
- detecção de anomalias;
- previsão de mudanças em páginas.

No mercado, o *scraping* é estratégico em:

- *e-commerce* (monitoramento de preços);
- *digital marketing* (análise de comportamento e sentimentos);
- *SEO* (identificação de tendências).

Métodos tradicionais envolvem ferramentas como *Requests*, *BeautifulSoup*, *Selenium*, *Playwright* e *Scrapy*, enquanto abordagens modernas incluem *scraping* híbrido, distribuído e baseado em *AI*.

- Métodos estáticos: *parsing HTML*, *DOM*, seletores *XPath/CSS*, *regex*;
- Métodos dinâmicos: renderização *headless*, visão computacional;
- Extração via *APIs*;
- Distinções entre *web scraping* e *crawling*.

Para contornar medidas de proteção (*CAPTCHA*, *fingerprinting*, *honeypots*), técnicas de *AI* e *proxies* são usadas, embora muitas operem em zona cinzenta legal.

As limitações do *scraping* tradicional incluem dificuldades com conteúdo dinâmico, fragilidade de seletores e altos custos de manutenção e escalabilidade.

Por fim, a legalidade permanece incerta, regulada por interpretações ligadas a contratos, direitos autorais e legislações de fraude ou abuso digital.

III. Desenvolvimento

A. Inteligência Artificial Aplicado a Raspagem de Dados

A integração de Inteligência Artificial (*AI*) com *web scraping* está transformando a coleta e análise de dados da *web*, superando as limitações dos métodos tradicionais. Técnicas como *Machine Learning (ML)*, *Natural Language Processing (NLP)* e visão computacional permitem não apenas extrair informações, mas também interpretá-las, classificá-las e adaptar estratégias de raspagem de forma autônoma.

As técnicas de Inteligência Artificial (IA), especificamente *Machine Learning (ML)* e *Natural Language Processing (NLP)*, juntamente com o *Natural Language Toolkit (NLTK)*, possibilitam o desenvolvimento de uma ferramenta lógica de resumo de texto, que utiliza a abordagem extrativa para gerar resumos precisos e fluentes. O *NLP* e o *ML* são técnicas de IA que ajudam as máquinas a compreender e analisar textos de entrada. Os algoritmos de *ML* treinam sistemas utilizando dados de amostra, permitindo que façam previsões e tomem decisões de forma autônoma. Já o *NLP* foca nas interações entre máquinas e humanos por meio da linguagem natural, aumentando a produtividade e a precisão. O *Natural Language Toolkit (NLTK)* é um pacote em *Python* que fornece bibliotecas e códigos para o processamento simbólico e estatístico de linguagem natural, facilitando tarefas de pré-processamento de dados, como limpeza, visualização, integração e tokenização [1].

Adicionalmente, técnicas mais avançadas de aprendizado de máquina foram propostas para evadir detecções baseadas nos movimentos de mouse que os bots geram, como o uso de Aprendizado por Reforço (RL) para gerar movimentos de mouse que passam pelo reCAPTCHA v3 do Google, e o uso de Redes Adversárias Generativas (GAN) para gerar movimentos de mouse parecidos com humanos para navegar a web [7].

Já os *LLM* possibilitam a geração automática de raspadores mais escaláveis e adaptáveis, enquanto *frameworks* como *LangChain* e *ScrapeGraphAI* integram esses modelos ao processo de extração.

Os modelos de aprendizado profundo podem ser aplicados para aprender representações a partir de características

de *URL*,como nomes de domínio, subdomínios e estruturas de caminho. Esses modelos conseguem capturar padrões e relacionamentos complexos dentro das *URL*, permitindo uma classificação precisa com base nas características das próprias *URL* [1].

Contudo, desafios persistem, como:

- necessidade de grandes volumes de dados para treinamento;
- alto custo computacional;
- barreiras impostas por mecanismos de defesa (*fingerprinting*, *CAPTCHA*, *rate limiting*).

A aplicação dessas redes neurais vai além da simples localização de dados - elas permitem a classificação inteligente de conteúdo e a extração de informações implícitas. Algoritmos de aprendizado de máquina podem ser treinados para classificar páginas *web* baseados em seu conteúdo, permitindo raspadores de dados focarem em páginas relevantes e extrair dados de uma maneira mais eficiente [1].

Nos últimos anos, o *deep learning* tem demonstrado capacidades extraordinárias em aprender representações expressivas de dados complexos, como dados de alta dimensionalidade, dados temporais, dados espaciais e dados em grafos, ampliando os limites de diferentes tarefas de aprendizado. O *deep learning* para detecção de anomalias, ou simplesmente detecção profunda de anomalias, tem como objetivo aprender representações de características ou pontuações de anomalia por meio de redes neurais, visando a detecção de anomalias [8].

A filtragem de conteúdo em páginas *web* não depende apenas do *focused crawler* para coletar informações relevantes, mas também pode utilizar *RNN*. Essas redes são capazes de classificar textos ao transformar o conteúdo extraído (como elementos *HTML*) em *tokens* numéricos, por meio da tokenização. Esses *tokens* são inseridos na *RNN* e classificados em categorias com base em características subjacentes identificadas pela rede neural. Devido à alta precisão de classificação obtida pela tokenização, as *RNN* se tornaram o padrão da indústria em classificadores de *machine learning*, sendo amplamente utilizadas em conjunto com a classificação direcionada na classificação baseada em texto [9]. Dessa forma, a combinação de *RNN* com *crawlers* permite uma análise mais eficiente e precisa do conteúdo textual coletado.

As *RNNs* são formadas por camadas de nós que processam sequências de dados. A *input layer* fornece os dados iniciais, as camadas *hidden* processam e retêm memória de estados anteriores, capturando dependências temporais ou contextuais, e a *output layer* gera a classificação final. Essa estrutura torna as *RNNs* eficazes para dados sequenciais e contextuais.

Em suma, a *AI* redefine o *web scraping* ao torná-lo mais preciso, adaptativo e estratégico, com aplicações que vão desde análise de sentimentos em avaliações até monitoramento financeiro em tempo real, estabelecendo um novo paradigma de automação inteligente de raspagem.

IV. Conclusão

A análise evidencia que a Inteligência Artificial redefine o *web scraping*, ampliando sua escalabilidade, precisão e capacidade de adaptação frente aos desafios impostos pela *web* dinâmica. O uso de técnicas avançadas como *LLM*, *NLP* e visão computacional transforma a raspagem de uma prática meramente estrutural para um processo inteligente de extração e descoberta de conhecimento. No entanto, os ganhos técnicos vêm acompanhados de dilemas éticos e jurídicos, especialmente em relação ao respeito à privacidade, aos Termos de Serviço e à mitigação de vieses culturais nos dados coletados. Conclui-se, portanto, que o futuro do *web scraping* depende não apenas da evolução tecnológica, mas também da consolidação de práticas responsáveis que equilibrem inovação, eficiência e ética.

Agradecimentos

Agradeço aos meus professores, Alessandra e Luciano, pelo apoio acadêmico e por contribuírem com conhecimento e incentivo para a realização deste trabalho.

Referências

[1] K. Weerasinghe, M. Maduranga, and M. Kawya, "Enhancing web scraping with artificial intelligence: A review," *Department of Information Technology, Faculty of Computing, General Sir John Kotelawala Defence University, Ratmalana, Sri Lanka*, 2024.

[2] E. Andreasen, L. Gong, A. MØller, M. Pradel, M. Selakovic, K. Sen, and C.-A. Staicu, "A survey of dynamic analysis and test generation for javascript," *ACM Computing Surveys, Vol. 50, No. 5, Article 66. Publication date: September 2017.*, 2017.

[3] E. Persson, "Evaluating tools and techniques for web scraping," 2019, acessado em 16 jun. 2025. [Online]. Available: https://www.diva-portal.org/smash/record.jsf?pid=diva2:1415998

[4] M. A. Khder, "Web scraping or web crawling: State of art, techniques, approaches and application," *Int. J. Advance Soft Compu. Appl, Vol. 13, No. 3, November 2021*, 2021.

[5] H. Vijaykumar, "Techniques to prevent attacks caused through web scraping - a review approach," *Journal of Nonlinear Analysis and Optimization Vol. 14, Issue. 2, No. 4: 2023 ISSN : 1906-9685*, 2023.

[6] Henrique, Pedro, "Web semântica: a informação tornando-se conhecimento," 2022, acessado em 16 jun. 2025. [Online]. Available: https://www.deviante.com.br/noticias/web-semantica-a-informacao-tornando-se-conhecimento/

[7] C. Iliou, T. Kostoulas, T. Tsikrika, V. Katos, S. Vrochidis, and I. Kompatsiaris, "Web bot detection evasion using deep reinforcement learning," *Association for Computing Machinery, New York, NY, USA, Article 15, 1–10*, 2022.

[8] G. Pang, C. Shen, L. Cao, and A. V. D. Hengel, "Deep learning for anomaly detection: A review," *ACM Comput. Surv., Vol. 1, No. 1, Article 1.*, 2020.

[9] D. Kanneganti, "Using recurrent neural networks and web crawlers to scrape open data from the internet," *The Young Researcher, 6(1), 60-71*, 2022.