INOVAÇÃO E AGILIDADE NA CONSULTA DOS DADOS USANDO AS TAGS HTML

INNOVATION AND AGILITY IN DATA QUERYING USING HTML TAGS

José Franklin Miranda Gomes Leite*
Alessandro Viola Pizzoleto**

RESUMO

A disseminação de dados e informações em WebSites vem crescendo gradativamente nas últimas décadas, tornando-se um dos principais meios para consulta e utilização. Apesar de toda a facilidade em obter os dados e as informações, um ponto crítico é a extração destes, por parte dos usuários, para serem utilizados em outros softwares e/ou plataformas. Considerando este fator, busca-se a análise e desenvolvimento de um conjunto de ferramentas que permitam realizar de forma automatizada e simplificada a extração com a possibilidade de armazenagem em SGBDs (Sistemas Gerenciadores de Banco de Dados) devidamente estruturados e implementados para isso. Ainda pensando no usuário final, apresentar um App para consulta e manutenção das tags que serão utilizadas no processo de extração dos dados e das informações contidas no WebSite informado.

Palavras-chave: Parser. HTML. Métodos. Dados. Informações.

ABSTRACT

The dissemination of data and information on Websites has been growing gradually in recent decades, becoming one of the main means of consultation and use. Despite all the ease in obtaining data and information, a critical point is the extraction of these, by users, to be used in other software and/or platforms. Considering this factor, the aim is to analyze and develop a set of tools that allow for automated and simplified extraction with the possibility of storage in DBMSs (Database Management Systems) duly structured and implemented for this purpose. Still thinking about the end user, present an App to consult and maintain the tags that will be used in the process of extracting the data and information contained in the informed website.

Keywords: Parser. HTML. Methods. Data. Information.

Introdução

A internet se tornou uma fonte inestimável de informações, com websites armazenando vastos conjuntos de dados em diferentes formatos. Entre esses formatos, o HTML (HyperText Markup Language) se destaca como a linguagem padrão para

* Discente FATECE. jose.franklin550@gmail.com
** Docente e Pesquisador da FATECE e FAMEESP. alessandro.pizzoleto@fatece.edu.br

estruturação e apresentação de conteúdo na web. As tags HTML, elementos básicos dessa linguagem, servem para organizar e classificar informações dentro de um website, fornecendo contexto e significado ao conteúdo textual.

Nesse contexto, surge a necessidade de um parse específico para identificar e extrair as informações contidas nas tags HTML. Um parse robusto torna-se essencial para automatizar o processo de coleta e organização de dados relevantes a partir de websites, possibilitando diversas aplicações práticas e beneficiando diferentes áreas do conhecimento.

A coleta manual de dados em websites é um processo tedioso e propenso a erros, especialmente quando se trata de grandes volumes de informações. Um parse dedicado à extração de dados em tags HTML permite automatizar essa tarefa, otimizando o tempo e a precisão da coleta. Isso possibilita a atualização frequente de bancos de dados, alimentando sistemas de inteligência artificial e pesquisas em diversas áreas.

Isto porque, as informações extraídas das tags HTML podem ser facilmente estruturadas e armazenadas em formatos compatíveis com ferramentas de análise e manipulação de dados. Isso permite a realização de análises estatísticas, identificação de padrões e tendências, geração de relatórios e visualizações de dados, possibilitando a tomada de decisões mais assertivas em diferentes contextos.

Justificando a necessidade, o parse de tags HTML pode ser integrado a diversos sistemas e aplicações, como bancos de dados, plataformas de e-commerce, ferramentas de CRM (Customer Relationship Management) e sistemas de automação de marketing. Essa integração permite a automatização de tarefas repetitivas, como a atualização de catálogos de produtos, a personalização de campanhas de marketing e a geração de relatórios personalizados para clientes.

Na área de pesquisa e desenvolvimento, o parse de tags HTML é fundamental para coletar dados de experimentos online, analisar o comportamento de usuários em websites e testar a efetividade de interfaces e campanhas. As informações extraídas podem ser utilizadas para aprimorar algoritmos, desenvolver novos produtos e serviços, e validar hipóteses de pesquisa.

O parse de tags HTML pode contribuir para a acessibilidade da informação, permitindo que pessoas com deficiência visual ou outras dificuldades de acesso à internet possam usufruir do conteúdo de websites. Ao extrair e converter o conteúdo textual em formatos compatíveis com leitores de tela e outras ferramentas assistivas, o parse torna a informação mais acessível a um público mais amplo.

Em suma, o desenvolvimento de um parse robusto para identificar e extrair informações em tags HTML se configura como uma ferramenta crucial para otimizar a coleta e o tratamento de dados na web, abrindo um leque de possibilidades para diversas áreas do conhecimento e da sociedade. A automatização da coleta de dados, a análise e manipulação de informações, a integração com sistemas e aplicações, a pesquisa e o desenvolvimento, e a acessibilidade da informação são apenas alguns exemplos dos benefícios que essa ferramenta pode proporcionar.

Nas seções seguintes são apresentadas informações complementares relacionadas e fundamentais para viabilizar e embasar todo o processo de pesquisa, análise e implementação o parse. Na seção 2, encontra-se a fundamentação teórica dos recursos necessários. Já na seção 3

1 Fundamentação Teórica

Analisar e implementar um *parse* com estas finalidades não é trivial, para tanto surge a necessidade de fundamentar e apresentar os conhecimentos necessários, além dos recursos que serão utilizados neste contexto.

1.1 Parse

No universo da computação, o termo "parse" se destaca como um processo fundamental para a análise e interpretação de dados. Imagine um explorador desbravando uma floresta densa, buscando identificar padrões e significados entre as árvores, cipós e animais selvagens. De forma similar, o parse atua como um explorador de dados, desvendando os segredos contidos em sequências de símbolos, como textos, códigos ou arquivos [1]. O parse, com sua capacidade de desvendar a linguagem da web e extrair informações valiosas, se configura como uma ferramenta poderosa para diversos fins. Sua versatilidade e aplicações práticas o tornam um aliado indispensável na era da informação, possibilitando a coleta, análise e utilização eficiente de dados na web [2].

O parse, também conhecido como analisador sintático, atua como um intérprete perspicaz, examinando minuciosamente sequências de símbolos, como código ou texto, e identificando sua estrutura e significado. No contexto da web, o parse se concentra na linguagem HTML (HyperText Markup Language), desbravando a hierarquia de tags e atributos que compõem a anatomia de um website. Com maestria, o parse reconhece as

tags HTML, elementos básicos que organizam e classificam o conteúdo textual em um website [1]. Cada tag possui um nome e, muitas vezes, atributos que fornecem informações adicionais sobre o conteúdo. O parse identifica essas tags e seus atributos, mapeando a estrutura hierárquica do website e desvendando o significado de cada elemento [3].

O poder do parse reside na sua capacidade de extrair informações relevantes das tags HTML. Como um arqueólogo digital, o parse vasculha cada tag e atributo, desenterrando dados valiosos como títulos, textos, imagens, links e metadados. Essas informações, antes ocultas na linguagem da web, se tornam acessíveis e prontas para serem utilizadas. Sua versatilidade o torna um instrumento indispensável em diversas áreas. Na pesquisa, o parse auxilia na coleta de dados de websites para análise e validação de hipóteses. No marketing, automatiza a coleta de informações para personalização de campanhas e otimização de conteúdo. Na área de desenvolvimento, facilita a integração de websites com sistemas e bancos de dados [3].

1.2 Linguagem HTML

Na vasta cena da internet, onde websites se apresentam como palcos virtuais, a linguagem HTML assume o papel de maestro, orquestrando a estrutura e o conteúdo de cada página. Através de tags e atributos, o HTML define a hierarquia, o significado e a aparência dos elementos que compõem um website, transformando ideias em experiências visuais interativas [4].

O HTML (*HyperText Markup Language*) se configura como a linguagem base para a criação de websites. Sua função primordial é estruturar o conteúdo textual, definindo títulos, parágrafos, imagens, links e outros elementos que compõem uma página da web. Através de tags e atributos, o HTML fornece ao navegador instruções sobre como interpretar e exibir o conteúdo, permitindo a criação de interfaces intuitivas e acessíveis [5].

As tags HTML são os blocos de construção fundamentais de um website. Cada tag possui um nome e, muitas vezes, atributos que fornecem informações adicionais sobre o conteúdo. Ao serem interpretadas pelo navegador, as tags definem a estrutura hierárquica da página e o significado de cada elemento [4].

- **<h1>**: Define um título principal na página.
- **<p>**: Define um parágrafo de texto.

- **<img>**: Insere uma imagem na página.
- **<a>**: Cria um link para outra página ou recurso.
- **<table>**: Cria uma tabela para organizar dados.

Os atributos complementam as tags, fornecendo informações adicionais sobre o conteúdo. Por exemplo, o atributo *src* da tag <img> define o endereço da imagem a ser exibida, enquanto o atributo *href* da tag <a> define o destino do link.

O HTML, com sua simplicidade e flexibilidade, se tornou a linguagem universal para a criação de websites. Sua capacidade de estruturar e organizar conteúdo, definir a aparência das páginas e criar links interativos a torna uma ferramenta essencial para qualquer pessoa que deseje navegar ou construir na web. Através das tags e atributos, o HTML permite a criação de websites informativos, interativos e acessíveis, conectando pessoas e ideias em um mundo digital cada vez mais conectado [6].

1.3 Legislação associada a LGPD

Em um mundo cada vez mais digital, onde os dados pessoais se tornaram um ativo valioso, surge a Lei Geral de Proteção de Dados (LGPD) como um farol para garantir a privacidade e segurança dos cidadãos. Promulgada em 2018 e em vigor desde 2020, a LGPD estabelece princípios, direitos e deveres para o tratamento de dados pessoais no Brasil, alinhando o país às melhores práticas internacionais na área [7].

A LGPD define dado pessoal como qualquer informação relacionada a uma pessoa física identificada ou identificável. A lei visa proteger esses dados, garantindo que sejam coletados, armazenados, utilizados e compartilhados de forma ética, transparente e segura. O objetivo principal da LGPD é proteger os direitos dos titulares dos dados, ou seja, as pessoas a quem os dados se referem, concedendo-lhes controle sobre seus dados pessoais e estabelecendo mecanismos para garantir seu uso adequado [7].

A LGPD se baseia em 10 princípios fundamentais que norteiam o tratamento de dados pessoais [7]:

- Legalidade, Legitimidade e Adequação: Os dados devem ser coletados, armazenados e utilizados de forma legal, legítima e adequada aos fins para os quais foram coletados.

- Boa-fé: O tratamento de dados deve ser realizado com boa-fé, de forma transparente e com o objetivo de proteger os direitos dos titulares.
- Finalidade Específica, Explícita e Incompatível: Os dados devem ser coletados para fins específicos, explícitos e legítimos, e não podem ser utilizados para fins incompatíveis com aqueles para os quais foram coletados.
- Necessidade: A coleta de dados deve ser limitada ao mínimo necessário e relevante para alcançar os fins para os quais foram coletados.
- Não Discriminação: O tratamento de dados não pode ser realizado de forma discriminatória, ou seja, com base em raça, cor, religião, sexo, orientação sexual, nacionalidade, idade, entre outros.
- Segurança: Os dados devem ser protegidos contra acessos não autorizados, destruição, perda, alteração, dano ou tratamento inadequado.
- Transparência: Os titulares dos dados têm o direito de saber como seus dados estão sendo tratados, quem os está tratando e para quais fins.
- Livre Acesso: Os titulares dos dados têm o direito de acessar seus dados, obter cópias deles e corrigi-los em caso de erro.
- Apagamento: Os titulares dos dados têm o direito de solicitar o apagamento de seus dados, caso não sejam mais necessários para os fins para os quais foram coletados ou em caso de revogação do consentimento.
- Prevenção: As empresas e órgãos públicos que tratam dados pessoais devem implementar medidas para prevenir e mitigar os riscos à segurança dos dados.

A LGPD garante aos titulares dos dados diversos direitos, incluindo [7]:

- Direito de acesso: O direito de saber como seus dados estão sendo tratados, quem os está tratando e para quais fins.
- Direito de retificação: O direito de corrigir seus dados em caso de erro.
- Direito de apagamento: O direito de solicitar o apagamento de seus dados, caso não sejam mais necessários para os fins para os quais foram coletados ou em caso de revogação do consentimento.
- Direito à portabilidade: O direito de receber seus dados em um formato estruturado, de uso comum e de leitura automática, e de transferir esses dados para outro controlador.
- Direito de oposição: O direito de se opor ao tratamento de seus dados para determinados fins, como para fins de marketing direto.

- Direito à decisão individual automatizada: O direito de não ser submetido a decisões baseadas unicamente em tratamento automatizado de dados, incluindo a criação de perfis.

1.4 JSoup (Java HTML parser)

No vasto universo da web, onde websites se apresentam como ilhas de informação, o Jsoup surge como um navegador perspicaz, explorando o código HTML e extraindo seus segredos com maestria. Essa biblioteca Java oferece ferramentas poderosas para manipular e analisar o conteúdo de páginas da web, tornando-se um aliado indispensável para diversos projetos e tarefas [8].

O Jsoup se configura como um analisador sintático HTML leve e flexível, permitindo aos desenvolvedores navegar, selecionar e modificar elementos HTML com facilidade. Através de uma API intuitiva, o Jsoup torna possível extrair texto, imagens, links e outros dados relevantes de páginas da web, abrindo um leque de possibilidades para diversos fins [9].

O Jsoup se destaca por suas diversas vantagens, tornando-se uma ferramenta valiosa para diversos projetos [10]:

- Facilidade de Uso: A API do Jsoup é intuitiva e fácil de aprender, permitindo que desenvolvedores de todos os níveis de experiência a utilizem com eficiência.
- Flexibilidade: O Jsoup oferece diversas opções para navegar, selecionar e manipular elementos HTML, adaptando-se às necessidades específicas de cada projeto.
- Eficiência: O Jsoup é leve e eficiente, consumindo poucos recursos do sistema e permitindo o processamento rápido de grandes volumes de dados.
- Versatilidade: O Jsoup pode ser utilizado em diversos tipos de aplicações, desde a coleta de dados até a criação de scrapers e crawlers web.

O Jsoup encontra diversas aplicações práticas em diferentes áreas [10]:

- Colagem de Dados: O Jsoup pode ser utilizado para extrair dados relevantes de websites, como textos, imagens, links e metadados.
- Scrapping Web: O Jsoup pode ser utilizado para criar scrapers web que automatizam a coleta de dados de websites específicos.

- Análise de Conteúdo: O Jsoup pode ser utilizado para analisar o conteúdo de páginas da web, identificando padrões e tendências.
- Limpeza de Dados: O Jsoup pode ser utilizado para limpar dados extraídos da web, removendo tags HTML desnecessárias e outros elementos indesejados.
- Criação de Conteúdo: O Jsoup pode ser utilizado para criar conteúdo HTML dinâmico, como tabelas, listas e outros elementos interativos.

O Jsoup se configura como uma ferramenta poderosa e versátil para manipular e analisar o conteúdo de páginas da web. Sua facilidade de uso, flexibilidade e eficiência o tornam um aliado indispensável para diversos projetos, desde a coleta de dados até a criação de *scrapers* e *crawlers* web. Através do Jsoup, os desenvolvedores podem explorar as profundezas do HTML com maestria, extraindo informações valiosas e criando aplicações inovadoras [9].

2 Implementação e Resultados do Protótipo

Através da linguagem Java, embarcamos em uma jornada empolgante para explorar as tags e seus segredos, desvendando o significado oculto por trás da estrutura de cada página da web. Dê início, foi necessário instalar o Jsoup no projeto Java. Pode-se fazê-lo através do Maven ou baixando o JAR manualmente e adicionando-o ao classpath. Com o Jsoup em mãos, podemos criar um objeto *Document* a partir de uma URL, um arquivo HTML ou uma string contendo o código HTML.

O Jsoup apresenta vasta flexibilidade dos seletores **CSS** para navegar pela estrutura hierárquica do HTML. Através de seletores como ***#id, .classe, tag*** e seus combinadores, podemos localizar com precisão os elementos que desejamos analisar. A Figura 1, ilustra um exemplo de código utilizando para acesso a um WebSite especifico e consequentemente, a seleção de uma das tags HTML que poderão ser utilizadas, neste caso, na linha 2, foi definido como elemento de trabalho a tag <p> e suas informação gravadas na variável "*paragraphs*". Este código permite a conexão com qualquer WebSite que se deseja ter acesso às informações.

```
Document doc = Jsoup.parse("https://www.example.com");
Elements paragraphs = doc.select("p");
```

Figura 1 - Código de exemplo

Ao encontrar os elementos desejados, o Jsoup nos permite extrair seu conteúdo textual, atributos e outros dados relevantes. Podemos utilizar métodos como *text(), html(), attr() e data()* para acessar essas informações. Como pode ser observado na Figura 2, após definir o elemento que será trabalhado (Figura 1), pode-se extrair as informações contidas na mesma, neste caso foi extraído o texto e o mesmo impresso para visualização do usuários.

```
for (Element paragraph : paragraphs) {
    String text = paragraph.text();
    System.out.println(text);
}
```

Figura 2 - Acesso as informações da tag

Com o Jsoup em nosso arsenal, podemos desvendar os segredos do HTML, extrair informações valiosas de websites, manipular e transformar o conteúdo das páginas da web e garantir a qualidade e compatibilidade do código HTML. O Jsoup se configura como um aliado indispensável para diversos projetos em Java, desde a coleta de dados até a criação de ferramentas de análise e manipulação de conteúdo web.

Como utilização prática, apresenta-se nas Figuras 3 e 4 o código que fora gerado para acessar um WebSite, identificar todas suas tags e apresentar seus conteúdos. O protótipo apresentado permite que novos usuários possam utilizá-lo e adaptá-lo as reais necessidades de acesso e consequentemente inovar com recursos onde as tags a serem analisadas possam ser definidas facilitando assim a extração das informações.

```
public static void main(String[] args) {
    try {
        String url = "https://devcoffee.com.br";//Trocar URL
        Map<String, String> tagsWithContent = getAllHtmlTagsWithContent(url);

        System.out.println("Tags HTML e conteúdos encontrados:");
        for (Map.Entry<String, String> entry : tagsWithContent.entrySet()) {
            System.out.println("Tag: " + entry.getKey() + " - Conteúdo: " + entry.getValue());
        }
    } catch (IOException e) {
        e.printStackTrace();
    }
}
```

Figura 3 - Conexão com o WebSite e apresentação das tags e informações

```java
public static Map<String, String> getAllHtmlTagsWithContent(String url) throws IOException {
    Map<String, String> tagsWithContent = new HashMap<>();

    // Conecta à URL e obtém o documento HTML
    Document document = Jsoup.connect(url).get();

    // Seleciona todas as tags no documento
    Elements elements = document.getAllElements();

    // Adiciona o nome de cada tag e seu conteúdo ao mapa
    for (Element element : elements) {
        tagsWithContent.put(element.tagName(), element.ownText());
    }

    return tagsWithContent;
}
```

Figura 4 - Identifica todas as tag e adiciona a um array

O Jsoup se configura como uma ferramenta poderosa e versátil para manipular e analisar o conteúdo de páginas da web. Sua facilidade de uso, flexibilidade e eficiência o tornam um aliado indispensável para diversos projetos, desde a coleta de dados até a criação de scrapers e crawlers web. Através do Jsoup, os desenvolvedores podem explorar as profundezas do HTML com maestria, extraindo informações valiosas e criando aplicações inovadoras.

Considerações Finais e Trabalhos Futuros

Este trabalho, explorou profundamente os recursos do Jsoup, uma biblioteca Java poderosa e versátil para navegar, manipular e analisar o conteúdo de páginas da web. Descobrimos como o Jsoup se configura como um aliado indispensável para diversos projetos, desde a coleta de dados até a criação de *scrapers* e *crawlers* web.

O Jsoup se destaca por sua interface intuitiva e amigável, permitindo que desenvolvedores de todos os níveis de experiência o utilizem com facilidade. Sua API bem documentada e exemplos abundantes facilitam o aprendizado e a implementação em diversos projetos. Além disso, a flexibilidade do Jsoup permite a adaptação a diferentes necessidades, desde a extração de dados específicos até a análise complexa de conteúdo HTML. O Jsoup se destaca por ser leve e eficiente, consumindo poucos recursos do sistema e permitindo o processamento rápido de grandes volumes de dados. Essa característica o torna ideal para aplicações que exigem alto desempenho e escalabilidade.

As possibilidades de utilização do Jsoup são vastas e abrem portas para diversos trabalhos futuros, como:

- Desenvolvimento de *scrapers* mais avançados: Implementação de scrapers que lidam com websites complexos, javascript e APIs.
- Criação de ferramentas de análise de dados: Desenvolvimento de ferramentas que extraem e analisam dados de websites para gerar relatórios e insights.
- Exploração de novas aplicações: Integração do Jsoup com outras tecnologias para criar soluções inovadoras na web.

O Jsoup se configura como uma ferramenta essencial para qualquer desenvolvedor que trabalha com websites e dados da web. Sua facilidade de uso, flexibilidade, eficiência e versatilidade o tornam um aliado indispensável para diversos projetos. Através do Jsoup, é possível navegar pelas profundezas do HTML com maestria, extraindo informações valiosas e criando aplicações inovadoras.

Referências

[1] P. Sebesta, A Primer on Data Structures and Algorithms, Addison-Wesley, 2016.

[2] L. Liu, Web Data Mining and Knowledge Discovery, Springer, 2011.

[3] B. Ford e B. Bah, Web Scraping with Python, O'Reilly Media, 2015.

[4] "W3C HTML Standard," 20 Abril 2024. [Online]. Available: https://www.w3.org/html/.

[5] "HTML Tutorial," 10 maio 2024. [Online]. Available: https://www.w3schools.com/html/.

[6] "Dive Into HTML5," 15 abril 2024. [Online]. Available: http://diveintohtml5.info/.

[7] "Brasil," 25 Abril 2024. [Online]. Available: https://www.planalto.gov.br/ccivil_03/_ato2015-2018/2018/lei/l13709.htm.

[8] Jsoup, "Jsoup Tutorial," [Online]. Available: https://www.baeldung.com/java-with-jsoup. [Acesso em 10 Abril 2024].

[9] Jsoup, "Jsoup API Documentation," [Online]. Available: https://jsoup.org/apidocs/. [Acesso em 10 Abril 2024].

[10] Jsoup, "Jsoup Website," [Online]. Available: https://jsoup.org/. [Acesso em 5 maio 2024].