NAEWI - Non-rendering Approach to Extract Web Information

Marcelo C. Nunes [1], **Carina F. Dorneles**[1]

[1]Departamento de Informática e Estatística
Universidade Federal de Santa Catarina (UFSC)
Caixa Postal 476 – 88.040-900 – Florianópolis – SC – Brazil

marcelo.canzian@posgrad.ufsc.br

carina.dorneles@ufsc.br

Abstract. *Extração de informações em páginas da Web é uma tarefa importante que visa facilitar a criação de bases de conhecimento. Levando em consideração que uma página Web é desenvolvida para ser agradável à utilização do usuário, porém é renderizada a partir de uma árvore HTML DOM, identificar e extrair suas informações ainda é um grande desafio. Para superar este desafio, este trabalho propõem uma abordagem que utilizará as informações da árvore DOM em conjunto com as informações visuais extraídas em forma de metadados dos elementos HTML da página para classificar e extrair os conteúdos relevantes de uma página Web. Para isso, será criado um modelo textual que representará a identidade visual do elemento da página, a fim de emular o contexto visual dos elementos e sua hierarquia na página, sem a necessidade de renderização da página por um navegador, para a extração das informações. Para a classificação dos elementos, será utilizado o modelo de linguagem bidirecional ELMo para contextualizar e identificar as características individuais de cada tipo de elemento.*

1. Informações gerais

- **Dissertação de mestrado:** NAEWI - Non-rendering Approach to Extract Web Information.
- **Estudante:** Marcelo Canzian Nunes.
- **Orientadora:** Dra. Carina Friedrich Dorneles.
- **Mês e ano de ingresso:** Junho/2021.
- **Mês e ano previsto para defesa:** Junho/2023.
- **Etapas concluídas:**
  - **Disciplinas Obrigatórias** - Março/2022
  - **Exame de Qualificação** - Julho/2022

2. Introdução

A extração de dados da Web é uma tarefa essencial, pois a Web é vista como o maior banco de dados a que se tem acesso [Tani et al. 2012, Tseng 2014], mas extrair e transformar os dados em informação relevante, ao ponto de se tornarem manipuláveis, é uma tarefa desafiadora [Fayzrakhmanov et al. 2018]. O volume gerado diariamente não para de crescer e todos estes dados são disponibilizados nos mais diversos formatos e em estruturas heterogêneas. As páginas Web são projetadas visando a usabilidade feita por humanos e não por máquinas. Além disso, elas podem ser altamente heterogêneas em estrutura e conteúdo, podendo apresentar estruturas idênticas para conteúdos totalmente diferentes [Park et al. 2015, Crestan and Pantel 2011, Anderson and Hong 2013a].

Como a maioria das páginas Web são escritas em HTML, as soluções que existem hoje são baseadas na análise do DOM HTML. No entanto, com o avanço do próprio HTML e das tecnologias usadas para o desenvolvimento das páginas da Web, novas versões ou novas *tags* são introduzidas, necessitando que os trabalhos anteriores necessitem de alterações nos seus códigos para serem compatíveis com as novas atualizações [Liu et al. 2010a]. Para facilitar a utilização do usuário na página Web, as estruturas são construídas utilizando um mesmo padrão de *layout* para mesmos itens na página e utilizando de artifícios como estilos e classes para padronizar a interface do usuário. Além disso, junto com as informações relevantes da página Web, há muito conteúdo que não trás valor ao conteúdo, poluindo a estrutura do DOM HTML, fazendo com que venham informações irrelevantes durante a extração dos dados, como por exemplo anúncios, menus, etc. [Liu et al. 2010a, Anderson and Hong 2013a].

Várias abordagens já foram utilizadas para tentar sanar esse problema, como a extração via texto bruto [Downey et al. 2004, Weninger and Hsu 2008], a extração via árvore DOM [Mehta and Narvekar 2015], a extração por meio de informações visuais [Anderson and Hong 2013b] e através da remoção de ruído das páginas [Velloso and Dorneles 2020]. Porém todas as abordagens citadas, com exceção da extração por meio de informações visuais, enfrentam dificuldades para extrair dados de páginas Web que apresentam o seu conteúdo de forma dinâmica através da utilização do conceito de *Single Page Application* (SPA). Apesar da abordagem que utiliza informações visuais para extração dos dados conseguir ultrapassar esta barreira, o sucesso da proposta depende da renderização da página Web para extrair as informações, o que torna o método muito custoso para um grande volume de dados [Simon and Lausen 2005, Fayzrakhmanov et al. 2018].

A motivação desse trabalho é propor um algoritmo de extração de dados de ambientes semi-estruturados e dinâmicos, criando um modelo textual para representar a identidade visual do elemento da página, emulando seu contexto visual e sua hierarquia na página, sem a necessidade de renderização por um navegador. Para a classificação dos elementos será utilizado modelo de linguagem bidirecional ELMo para contextualizar e identificar as características individuais de cada tipo de elemento.

3. Trabalhos relacionados

Nesta seção será apresentada uma breve visão geral dos trabalhos existentes na área de extração de dados em páginas Web. A extração de informações de páginas da Web tem sido estudada desde a década de 90, onde eram aplicadas regras em cima de um con-

junto de código HTML ou texto, usando exemplos e contra-exemplos, geralmente rotulados, como os trabalhos de [Soderland 1999] e [Chang and Lui 2001]. Com o passar dos anos, foram implementadas técnicas de extração de informação baseadas nas estruturas das árvores DOM HTML, utilizando o conceito de *Data Regions* e *Data Records* para identificar as informações por meio de técnicas de alinhamento de árvores como em [Liu et al. 2003], [Zhai and Liu 2005], [Pandarge and Chakkarwar 2017].

Trabalhos que buscam realizar a extração de informações através da percepção visual do usuário sobre a página Web, utilizando pistas visuais que são extraídas dos elementos da página, também são abordados por diversos trabalhos como [Liu et al. 2003], [Liu et al. 2010b], [Anderson and Hong 2013b], [Simon and Lausen 2005]. Porém estes trabalhos tendem a limitar o uso das dicas visuais a alguns elementos específicos, baseando-se nas tarefas a serem executadas, dado a complexidade de adotar esta abordagem de forma mais ampla ou pela redução de dimensionalidade na execução dos modelos propostos.

Outra forma de identificar e extrair informações das páginas Web são através de algoritmos que buscam detectar padrões de elementos e remover os ruídos contidos na página, diminuindo assim o universo de busca pelo algoritmo. Trabalhos como os de [Velloso et al. 2014], [Wai et al. 2017], [Guo et al. 2019], [Velloso et al. 2020] utilizam dessa abordagem para fazer a extração das informações contidas nas páginas Web.

Novas abordagens utilizam em suas implementações técnicas de redes neurais artificiais para identificar e classificar padrões dentro de uma página Web. Trabalhos como [Wai et al. 2017], [Xie et al. 2021], [Zhou et al. 2022] utilizam técnicas de classificação, processamento de linguagem natural e transformadores para realizar a extração dos conteúdos das páginas Web.

4. Proposta

O trabalho proposto a ser desenvolvido, chamado de NAEWI, é um algoritmo de extração de dados em ambientes heterogêneos e dinâmicos, sem o uso de renderização dos elementos da página Web, aproveitando a estrutura do DOM HTML, sua sequência de *tags* e seus metadados para inferir o conteúdo relevante dentro de uma página Web, reduzindo ao máximo a captura de ruídos. Para isso, será desenvolvido um algoritmo que irá realizar três fases de execução para executar sua tarefa: pré-processamento, processamento do texto e pós-processamento, conforme mostra a Figura 1.

Na fase de pré-processamento, será realizada uma requisição HTTP para a página Web a fim de obter seu corpo HTML em forma de texto. Através do texto obtido na requisição, o algoritmo original irá inicialmente executar uma limpeza no texto extraído do HTML da página Web, a fim de remover caracteres inválidos e *tags* desnecessárias. Após isso, serão extraídas as *tags* do corpo HTML da página e iniciar a montagem da árvore DOM do HTML limpo. Paralelo a isso, o algoritmo irá buscar extrair metadados da *tag* como: estilos, atributos, execução de *scripts* Javascript, altura do elemento na árvore, posição do elemento na página, etc. Com isso, será criado o Tag Path Sequence (TPS) da *tag* extraída, enriquecida com as informações extraídas dos metadados da *tag*. Por fim, na fase de pré-processamento, a sequência do TPS é invertida, a fim de deixar os dados da folha para o início do texto. Com isso, espera-se que o algoritmo tenha informações suficientes para descrever o elemento sem a necessidade de executar diversas chamadas

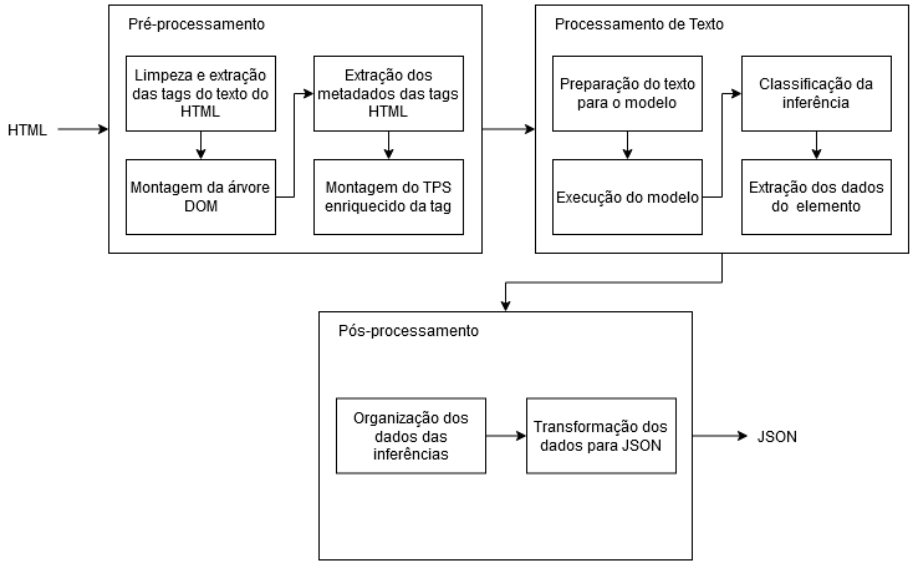

Figura 1. Macro fluxograma do algoritmo NAEWI

HTTP e renderizações utilizando alguma *engine* de navegação. Também acredita-se que invertendo a ordem do TPS enriquecido, o conteúdo mais relevante seja trazido para o início do texto gerado. Como o TPS é a representação da hierarquia do elemento na árvore DOM, invertendo sua sequência colocaria a parte da *string* que mais irá se repetir por todos os elementos, seus pais, para o final da inferência.

Na fase de processamento do texto, o algoritmo irá utilizar o modelo de linguagem bidirecional (biLM) ELMo [Peters et al. 2018] para classificar os elementos da página Web como conteúdo relevante ou ruído. Para isso, o modelo irá utilizar como *input* o texto do TPS enriquecido de cada elemento da página, extraindo e identificando as características individuais de cada tipo de elemento. Como elementos que representam um mesmo conteúdo tendem a ter representações similares dentro de uma mesma página Web e o texto do TPS enriquecido possui todos os metadados do elemento e sua hierarquia dentro da página Web, acredita-se que uma análise do tipo bidirecional conseguirá identificar o tipo de conteúdo de cada elemento e seu contexto dentro da página. Para realizar o treinamento do modelo de classificação, os dados serão divididos em dados de treino, validação e teste, onde os dados de teste não serão utilizados durante o treinamento e serão desconhecidos pelo modelo. Por fim nesta etapa, será utilizado a saída da inferência para classificar o tipo de conteúdo extraído: conteúdo relevante ou ruído. Em caso de ruído, o elemento será descartado.

Na fase de pós-processamento, será realizado a organização dos informações recebidas da etapa anterior para preparar os dados para a transformação em um formato JSON. Por fim, serão utilizados os dados organizados para transforma-los em um formato JSON, retornando-os processados ao usuário.

5. Contribuições e trabalhos futuros

Neste trabalho apresentamos uma proposta de algoritmo de extração de dados da Web, chamado NAEWI, que tem como objetivo criar um modelo textual que representa a identidade visual dos elementos de uma página, emulando seu contexto visual e sua hierarquia

sem necessitar da renderização da página.

Para validar a proposta, será utilizado o *dataset* público *Structured Web Data Extraction Dataset* (SWDE), pois este contém 80 sites distintos, contendo diversas páginas já rotuladas de 8 assuntos diversos: automóveis, livros, câmeras, empregos, filmes, jogadores da NBA, restaurantes e universidades. Com estes assuntos será implementado uma primeira versão do algoritmo. Em uma segunda etapa, serão coletados e rotulados novos sites que implementam os conceitos de SPA para implementar a inferência em páginas onde o conteúdo é dinamicamente adicionado ao corpo da página Web.

Para testar e validar o trabalho proposto, serão realizadas comparações com os trabalhos DEPTA [Zhai and Liu 2005], WebKE [Xie et al. 2021] e LANTERN [Zhou et al. 2022], através das métricas de precisão, *recall* e *F1 score*, utilizando os *datasets* citados acima.

As contribuições deste trabalho atente os seguintes itens:

- Nova abordagem de extração de metadados dos elementos da árvore DOM;
- Nova abordagem quanto à extração de dados em ambientes heterogêneos e dinâmicos;
- Melhoria na performance na extração de dados em ambientes heterogêneos e dinâmicos.

Referências

Anderson, N. and Hong, J. (2013a). Visually extracting data records from the deep web. In Proceedings of the 22nd International Conference on World Wide Web, WWW '13 Companion, page 1233–1238, New York, NY, USA. Association for Computing Machinery.

Anderson, N. and Hong, J. (2013b). Visually extracting data records from the deep web. WWW '13 Companion, page 1233–1238, New York, NY, USA. Association for Computing Machinery.

Chang, C.-H. and Lui, S.-C. (2001). Iepad: Information extraction based on pattern discovery. In Proceedings of the 10th International Conference on World Wide Web, WWW '01, page 681–688, New York, NY, USA. Association for Computing Machinery.

Crestan, E. and Pantel, P. (2011). Web-scale table census and classification. In Proceedings of the Fourth ACM International Conference on Web Search and Data Mining, WSDM '11, page 545–554, New York, NY, USA. Association for Computing Machinery.

Downey, D., Etzioni, O., Soderland, S., and Weld, D. S. (2004). Learning text patterns for web information extraction and assessment. In AAAI-04 workshop on adaptive text extraction and mining, pages 50–55.

Fayzrakhmanov, R. R., Sallinger, E., Spencer, B., Furche, T., and Gottlob, G. (2018). Browserless web data extraction: Challenges and opportunities. In Proceedings of the 2018 World Wide Web Conference, WWW '18, page 1095–1104, Republic and Canton of Geneva, CHE. International World Wide Web Conferences Steering Committee.

Guo, J., Crescenzi, V., Furche, T., Grasso, G., and Gottlob, G. (2019). Red: Redundancy-driven data extraction from result pages? In The World Wide Web Conference, WWW '19, page 605–615, New York, NY, USA. Association for Computing Machinery.

Liu, B., Grossman, R., and Zhai, Y. (2003). Mining data records in web pages. In Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '03, page 601–606, New York, NY, USA. Association for Computing Machinery.

Liu, W., Meng, X., and Meng, W. (2010a). Vide: A vision-based approach for deep web data extraction. IEEE Transactions on Knowledge and Data Engineering, 22(3):447–460.

Liu, W., Meng, X., and Meng, W. (2010b). Vide: A vision-based approach for deep web data extraction. IEEE Transactions on Knowledge and Data Engineering, 22(3):447–460.

Mehta, B. and Narvekar, M. (2015). Dom tree based approach for web content extraction. In 2015 International Conference on Communication, Information & Computing Technology (ICCICT), pages 1–6.

Pandarge, S. S. and Chakkarwar, V. A. (2017). Automatic web information extraction and alignment using ctvs technique. In 2017 International conference of Electronics, Communication and Aerospace Technology (ICECA), volume 2, pages 94–99.

Park, K., Nguyen, M. C., and Won, H. (2015). Web-based collaborative big data analytics on big data as a service platform. In 2015 17th International Conference on Advanced Communication Technology (ICACT), pages 564–567.

Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., and Zettlemoyer, L. (2018). Deep contextualized word representations. In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), pages 2227–2237, New Orleans, Louisiana. Association for Computational Linguistics.

Simon, K. and Lausen, G. (2005). Viper: Augmenting automatic information extraction with visual perceptions. In Proceedings of the 14th ACM International Conference on Information and Knowledge Management, CIKM '05, page 381–388, New York, NY, USA. Association for Computing Machinery.

Soderland, S. (1999). Learning information extraction rules for semi-structured and free text. Machine learning, 34(1):233–272.

Tani, F. Y., Farid, D. M., and Rahman, M. Z. (2012). Ensemble of decision tree classifiers for mining web data streams. International Journal of Applied Information Systems, 1(2):30–36.

Tseng, C.-H. (2014). Crowd aided web search. In 2014 6th International Conference on Knowledge and Smart Technology (KST), pages 1–6.

Velloso, R. P. and Dorneles, C. F. (2020). Optimized extraction of records from the web using signal processing and machine learning. In SBBD, pages 109–120.

Velloso, R. P. et al. (2014). Algoritmo não supervisionado para segmentação e remoção de ruído de páginas web utilizando tag paths.

Velloso, R. P. et al. (2020). Optimized record extraction from web pages using signal processing and machine learning.

Wai, F. K., Yong, L. W., Thing, V. L. L., and Pomponiu, V. (2017). Cmdr: Classifying nodes for mining data records with different html structures. In TENCON 2017 - 2017 IEEE Region 10 Conference, pages 1862–1862.

Weninger, T. and Hsu, W. H. (2008). Text extraction from the web via text-to-tag ratio. In 2008 19th International Workshop on Database and Expert Systems Applications, pages 23–28.

Xie, C., Huang, W., Liang, J., Huang, C., and Xiao, Y. (2021). WebKE: Knowledge Extraction from Semi-Structured Web with Pre-Trained Markup Language Model, page 2211–2220. Association for Computing Machinery, New York, NY, USA.

Zhai, Y. and Liu, B. (2005). Web data extraction based on partial tree alignment. In Proceedings of the 14th International Conference on World Wide Web, WWW '05, page 76–85, New York, NY, USA. Association for Computing Machinery.

Zhou, Y., Sheng, Y., Vo, N., Edmonds, N., and Tata, S. (2022). Learning transferable node representations for attribute extraction from web documents. In Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining, pages 1479–1487.