ANÁLISE DE COMPORTAMENTO DOS USUÁRIOS DOS WEBSITES DO CPTEC/INPE UTILIZANDO TÉCNICAS DE APRENDIZAGEM DE MÁQUINA

Marcus Vinícius Souza Silva

Relatório de Iniciação Científica do Programa PIBIC, orientado por Dr. Leandro Guarino de Vasconcelos, CPTEC.

INPE
Cachoeira Paulista
2021

ANÁLISE DE COMPORTAMENTO DOS USUÁRIOS DOS WEBSITES DO CPTEC/INPE UTILIZANDO TÉCNICAS DE APRENDIZAGEM DE MÁQUINA

Marcus Vinícius Souza Silva

Relatório de Iniciação Científica do Programa PIBIC, orientado por Dr. Leandro Guarino de Vasconcelos, CPTEC.

INPE
Cachoeira Paulista
2021

RESUMO

Este trabalho, teve como objetivo o desenvolvimento e aplicação da análise de comportamento dos usuários nos web sites do CPTEC/INPE por meio de técnicas de aprendizagem de máquina. Inicialmente, foi necessário entender as características que eram fundamentais para determinar o comportamento do usuário em um web site. Por meio dessas características, é possível determinar como o usuário está usando o web site e desenvolver melhorias para ele, aprimorando a usabilidade da interface. Para isso, é necessária a coleta de informações básicas de navegação, como cliques, envio de formulários, tempo na página, etc. Após a coleta, os dados são enviados para a plataforma ELK, que consiste no conjunto das ferramentas Elasticsearch, Logstash e Kibana. O Elasticsearch é um mecanismo de busca e análise; o Logstash é um pipeline de processamento de dados do lado do servidor que recebe os dados das páginas do INPE; o Kibana é útil para a visualização dos dados com diagramas e gráficos. Por fim, os dados são processados por um algoritmo de detecção de regras de associação, técnica de aprendizagem de máquina, e é possível descobrir padrões de comportamento dos usuários. De acordo com esse funcionamento, foi realizado um estudo de caso com o web site de Previsão Numérica de Tempo do CPTEC/INPE e, por meio dos padrões de comportamento identificados, foi possível propor melhorias de usabilidade da interface.

LISTA DE FIGURAS

Pág.

Figura 1 - Dashboard do Kibana 7

SUMÁRIO

1. INTRODUÇÃO 4
1.1 Objetivos específicos 5
2. MATERIAIS E MÉTODOS 6
2.1 JAVASCRIPT 6
2.2 JQUERY 6
2.3 NODE.JS 6
2.4 ELASTICSEARCH 6
2.5 LOGSTASH 7
2.6 KIBANA 7
2.7 WEKA 7
2.8 método 8
3. IMPLEMENTAÇÃO 9
3.1 DEFINIÇÃO DOS ATRIBUTOS 9
3.2 INSERÇÃO DO SCRIPT 9
3.3 API 9
3.4 KIBANA 9
3.5 Aprendizagem de máquina 9
4. RESULTADOS 11
5. CRONOGRAMA 13
REFERÊNCIAS 14

1. INTRODUÇÃO

As pessoas têm diferentes objetivos quando acessam uma aplicação Web. Em um web site de vídeos, por exemplo, um usuário pode estar interessado em ver um vídeo específico, enquanto outro usuário está interessado em reproduzir sua lista favorita de vídeos. Devido à diversidade de objetivos dos usuários, torna-se essencial entender seus diferentes comportamentos a fim de oferecer a melhor interface de acordo com cada necessidade.

Para aprimorar a usabilidade de uma interface, há diferentes métodos de avaliação de usabilidade, tais como: *thinking-aloud*, testes de usabilidade em laboratório e análise de tarefas. Dentre esses métodos, há a abordagem de avaliar a usabilidade de interfaces a partir do registro das interações dos usuários.

Esses registros podem ser gravados no servidor Web que hospeda a aplicação, chamados de *server logs* (logs de servidor); ou podem ser gravados na interface de aplicação, chamados de *client logs* (logs de cliente), que são mais detalhados devido à possibilidade de coletar ações específicas dos usuários como: passar o mouse sobre um elemento, rolar a tela, digitar em um campo de formulário, etc.

Na literatura, há diferentes ferramentas e métodos para coleta de *client logs* (AZZOPARDI, DOOLAN e GLASSEY, 2012). Uma dessas abordagens foi desenvolvida por VASCONCELOS (2012), chamada de USABILICS, um sistema para coleta de *client logs* e avaliação remota e automática de usabilidade, sem a intervenção humana no processo de avaliação.

Posteriormente, GONCALVES et.al. (2016) implementaram a avaliação de usabilidade em aplicações Web com foco em *smartphones* e outros dispositivos sensíveis ao toque.

Com a coleta das interações dos usuários, é possível processar os logs para detectar padrões de comportamento, que ajudam na tomada de decisão sobre novas funcionalidades e sobre melhorias na interface. Nesse contexto, o assunto é tratado na literatura como mineração de dados da Web (*Web Mining*).

CHANG et.al. (2003) apresentam uma classificação de mineração de dados da Web direcionada para três aspectos de uma interface Web: estrutura, conteúdo e uso. Nesse contexto, VASCONCELOS (2017) criou a abordagem RUM (*Real-time Usage*

Mining) que foca em mineração de uso da Web e permite realizar a análise de padrões de comportamento dos usuários durante a navegação.

Portanto, neste trabalho, o foco foi a mineração de uso da interface, que permite descobrir padrões de comportamento dos usuários.

Para um estudo de caso, foi escolhido o Centro de Previsão de Tempo e Estudos Climáticos do Instituto Nacional de Pesquisas Espaciais (CPTEC/INPE), que possui diversos web sites que são acessados por milhões de brasileiros, ou seja, possui a necessidade de oferecer interfaces com boa usabilidade para diferentes tipos de usuários.

1.1 OBJETIVOS ESPECÍFICOS

- Coletar dados anônimos da interação dos usuários nos web sites do CPTEC/INPE;
- Detectar padrões de comportamento dos usuários dos web sites do CPTEC/INPE a partir do registro da interação nas interfaces;
- Identificar conteúdos de interesse dos usuários nos web sites do CPTEC/INPE a partir do registro da interação nas interfaces.

2. MATERIAIS E MÉTODOS

Ao se propor a construção, desenvolvimento e/ou melhoria de uma ferramenta de coleta de informações das páginas do CPTEC/INPE se faz necessário o auxílio de algumas ferramentas, softwares e outras tecnologias. Também é necessário visualizar e tratar essas informações. Desta forma, foi imprescindível uma pesquisa e estudo de métodos computacionais mais acurados e que atendessem às necessidades.

Portanto, as seguintes tecnologias foram utilizadas para o desenvolvimento da arquitetura necessária para coletar, processar, armazenar e analisar *client logs*: Javascript, JQuery, Node.js, ElasticSearch, Logstash, Kibana e Weka.

2.1 JAVASCRIPT

A linguagem Javascript (MDN Web Docs, 2021), também conhecida como JS, é uma linguagem de programação interpretada, mais conhecida como a linguagem de script para páginas Web, mas usada também em vários outros ambientes sem browser, como o Node.js (Node.JS, 2021). É uma linguagem multi-paradigma e dinâmica, suportando estilos de orientação a objetos, imperativos e declarativos. Por meio da linguagem JavaScript é possível manipular o DOM (MDN Web Docs, 2021), que é um modelo de árvore de nós das páginas web, sendo que cada nó representa uma parte do documento.

2.2 JQUERY

JQuery (JQuery, 2021) é uma biblioteca popular de JavaScript. Ela foi criada por John Resig em 2006 com o propósito de aumentar a produtividade dos desenvolvedores que usam JavaScript nos seus sites. Não é uma linguagem de programação separada, funciona em conjunto com o JavaScript.

2.3 NODE.JS

O Node.js é uma tecnologia usada para executar código JavaScript fora do navegador. Ele torna possível trabalhar no lado do servidor (*server-side*) com JS. Isso é possível graças ao ambiente de execução de código JS do próprio Node.js e o motor de interpretação de execução de Java Script presente no Google Chrome, chamado de V8.

2.4 ELASTICSEARCH

O Elasticsearch (Elastic.co, 2021) é um mecanismo de busca e análise de dados gratuito e aberto para diversos tipos de dados, como texto, número, etc. É o componente central do Elastic Stack, um conjunto de ferramentas utilizadas para ingestão, armazenamento, análise e visualização de dados, chamado ELK Stack (pelas iniciais de Elasticsearch, Logstash e Kibana). Os dados brutos passam pelo Elasticsearch, que indexa-os, para facilitar as consultas.

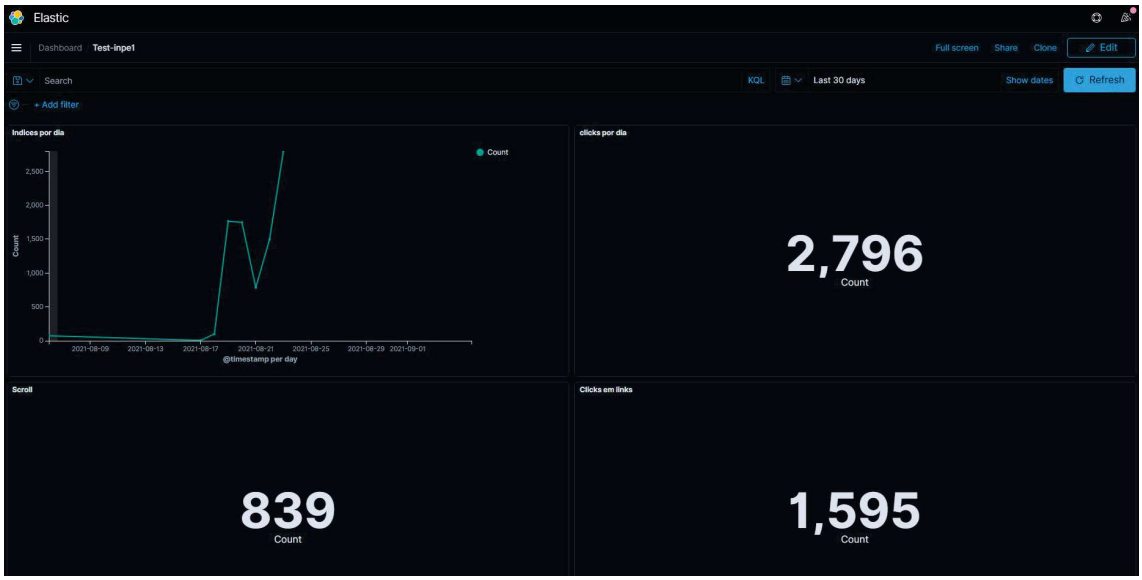
2.5 LOGSTASH

O Logstash (Elastic.co, 2021) é um pipeline capaz de unificar dados vindos de diversas fontes e centralizá-los em uma só fonte. Faz parte da ELK Stack junto do Elasticsearch e Kibana.

2.6 KIBANA

O Kibana (Elastic.co, 2021) é uma interface de usuário gratuita que permite a visualização dos dados do Elasticsearch e navegar no Elastic Stack. Conta com diversos tipos de gráficos e diferentes visualizações de dados.

Figura 1 - Dashboard do Kibana



2.7 WEKA

Weka é uma coleção de algoritmos de aprendizagem de máquina para tarefas de mineração de dados. Ele contém ferramentas para preparação dos dados, classificação, regressão, agrupamento, detecção de regras de associação e visualização.

2.8 MÉTODO

O trabalho realizado baseou-se no processo de descoberta de conhecimento de bases de dados (KDD – *Knowledge Discovery in Databases*) definido por Fayyad et.al. (1996).

As etapas desse processo foram implementadas da seguinte forma:

- Seleção de atributos: a partir de observação das interfaces do web site de Previsão Numérica de Tempo do CPTEC/INPE, foram definidas as ações que seriam coletadas na interface através de um script em Javascript.
- Pré-processamento: uma vez que os logs foram coletados, foi necessário limpar dados duplicados e combinar os dados para transformá-los em ações de mais alto nível, como: "usuário favoritou uma variável do modelo numérico".
- Transformação: após o tratamento dos dados, estes foram transformados em dados nominais (lista pré-definida de valores) e dados binários, a fim de prepará-los para a etapa de aprendizagem de máquina.
- Aprendizagem de máquina: com o uso da técnica de detecção de regras de associação, foi possível processar os logs para identificar padrões de ações combinadas dos usuários.
- Interpretação dos padrões: com os padrões identificados, a equipe de desenvolvimento do web site analisou os padrões, com o objetivo de identificar quais são relevantes e quais são óbvios.
- Conhecimento: o processo foi finalizado com a identificação de padrões relevantes que sugeriram novas funcionalidades para o web site, a fim de melhorar a experiência dos usuários.

3. IMPLEMENTAÇÃO

3.1 DEFINIÇÃO DOS ATRIBUTOS

Antes de começar o desenvolvimento do projeto foi necessário definir quais atributos seriam coletados nas páginas:

- Cliques;
- Clique duplo;
- Rolagem de tela;
- Posicionamento do mouse sobre um elemento (*mouseover*);
- Envio de formulários;
- Detecção de foco em caixas de texto (*focus*);
- Uso de teclado; e
- Redimensionamento da janela do navegador.

3.2 INSERÇÃO DO SCRIPT

Foi desenvolvido um script em JavaScript que é inserido na página para fazer o reconhecimento e tratamento dos eventos que os usuários realizam nas páginas. Por meio dele, é possível definir quais eventos serão coletados e quantos eventos serão enviados ao mesmo tempo para a API. O script deve ser inserido na página através da tag HTML <script> e faz uso da biblioteca JQuery.

3.3 API

A API (*Application Programming Interface*) é uma forma de associar a stack ELK e o script que coleta os dados nas páginas. Foi desenvolvida em JavaScript utilizando o ambiente de execução Node.JS. Possui duas rotas, uma para a criação do identificador da sessão e outra para a recepção dos logs e envio para a stack ELK.

3.4 KIBANA

Para visualização dos logs coletados, foram criados gráficos por tipos de ações dos usuários, com a possiblidade de filtrar por datas.

3.5 APRENDIZAGEM DE MÁQUINA

Os logs foram processados através de um script em JavaScript, automatizando o processo de KDD desde a seleção dos atributos até a geração de padrões.

Para isso, foram mapeadas as funcionalidades da interface do web site de Previsão Numérica de Tempo do CPTEC/INPE:

- Acesso ao modelo WRF 5km;
- Acesso ao modelo WRF 7km;
- Acesso ao modelo BAM;
- Acesso ao modelo ETA;
- Acesso ao modelo BRAMS;
- Uso da lista de categorias de variáveis;
- Usuário adicionou uma variável;
- Usuário removeu uma variável;
- Usuário acessou a combinação de variáveis;
- Usuário usou a linha do tempo;
- Usuário usou a animação;
- Usuário alterou a opacidade de uma imagem; e
- Usuário acessou variáveis favoritas.

Em seguida, foi automatizada a extração dessas características a partir dos logs coletados. O resultado desse script de automatização foi um arquivo texto separado por vírgulas (arquivo CSV).

O arquivo foi importado no Weka para detecção de regras de associação entre as características de comportamento dos usuários. Os padrões mais relevantes são apresentados na seção seguinte.

4. RESULTADOS

Para ambiente de teste foi utilizado o site de previsão numérica de tempo CPTEC/INPE. Nele foram coletados 216 interações e 29.548 ações de usuários. Por meio dos dados coletados foi possível obter algumas informações através de aprendizagem de máquina, com o algoritmo Apriori (AGRAWAL, IMIELINSKI e SWAMI, 1993), que é uma implementação da técnica de detecção de regras associação:

1. 84% dos usuários que visualizaram o Modelo Numérico WRF 7km também visualizaram o Modelo BAM;
2. 54% dos usuários que usam as combinações de variáveis são usuários que navegam entre 1 minuto e 14 minutos;
3. 50% dos usuários que navegam no período noturno interagem menos de 30 segundos;
4. 100% dos usuários que mudaram a opacidade das camadas também usaram o recurso de combinação de variáveis.
5. 100% dos usuários que usaram as variáveis favoritas interagiram no período da manhã (talvez sejam meteorologistas).

A regra 1 mostra que 84% dos usuários acessam diferentes modelos numéricos. Esse padrão foi considerado relevante pela equipe de desenvolvimento porque, atualmente, o site não permite visualizar variáveis de modelos diferentes ao mesmo tempo. Essa regra mostrou que há a necessidade de implementar essa funcionalidade para melhorar a usabilidade.

As regras 2 e 3 foram consideradas relevantes, porém não foi definida uma ação imediata para atendê-las.

A regra 4 caracteriza usuários avançados, possivelmente recorrentes, que já conhecem a interface do web site e usam recursos específicos.

A regra 5 possivelmente caracteriza meteorologistas operacionais, que acessam o web site diariamente para ver variáveis específicas para as suas atividades de trabalho.

Diante dos resultados, observa-se que o método implementado para análise do comportamento dos usuários mostrou-se suficiente. Devido a restrições técnicas de acesso a outros web sites, não foi possível realizar mais experimentos, porém objetiva-se

oficializar essa ferramenta como forma de análise do comportamento dos usuários dos web sites do CPTEC/INPE.

5. CRONOGRAMA

MÊS	ATIVIDADES
AGOSTO	- Pesquisa bibliográfica sobre análise de logs de web sites e técnicas de aprendizagem de máquina
SETEMBRO	- Definição dos atributos da interação do usuário para coleta - Implementação de script para coleta de ações da interação do usuário em web sites
OUTUBRO	- Estudo da linguagem de programação JavaScript e do ambiente de execução Node.js.
NOVEMBRO	- Implementação de script para coleta de ações da interação do usuário em web sites
DEZEMBRO	- Configuração da plataforma ELK para recepção e armazenamento dos logs.
JANEIRO	- Definição dos comportamentos do usuário relevantes para o CPTEC/INPE
FEVEREIRO	- Utilização de softwares e/ou plataformas de aprendizagem de máquina para execução de algoritmos para detecção de padrões de comportamento dos usuários
MARÇO	- Utilização de softwares e/ou plataformas de aprendizagem de máquina para execução de algoritmos para detecção de padrões de comportamento dos usuários
ABRIL	- Análise e interpretação dos padrões detectados
MAIO	- Implementação de gráficos e tabelas para visualização das características dos conteúdos mais visualizados pelos usuários, utilizando os logs coletados na interface
JUNHO	- Implementação de gráficos e tabelas para visualização das características dos conteúdos mais visualizados pelos usuários, utilizando os logs coletados na interface - Escrita de artigo científico para submissão
JULHO	- Escrita de artigo científico para submissão

REFERÊNCIAS

1. AGRAWAL, R.; IMIELINSKI, T.; SWAMI, A. Mining association rules between sets of items in large databases. ACM SIGMOD, v. 22, n. 2, p. 207–216, jun. 1993.

2. AZZOPARDI, L.; DOOLAN, M.; GLASSEY, R. Alf: a client side logger and server for capturing user interactions in web applications. In: INTERNATIONAL ACM SIGIR CONFERENCE ON RESEARCH AND DEVELOPMENT IN INFORMATION RETRIEVAL, 35., 2012, Portland, Oregon, USA. New York, NY, USA: ACM, 2012. p. 1003–1003.

3. CHANG, G.; HEALEY, M.; MCHUGH, J.; WANG, J. Mining the World Wide Web. Dordrecht, the Netherlands: Kluwer Academic Publishers, 2003.

4. Elastic.co. **O que é o Elasticsearch?** Disponível em: < https://www.elastic.co/pt/what-is/elasticsearch >. Acesso em: 29 ABR 2021.

5. Elastic.co. **O que é o Kibana?** Disponível em: < https://www.elastic.co/pt/what-is/kibana >. Acesso em 29 ABR 2021.

6. Elastic.co. **Logstash**. Disponível em < https://www.elastic.co/pt/logstash/ >. Acesso em: 30 ABR 2021.

7. FAYYAD, U.; PIATETSKY-SHAPIRO, G.; SMYTH, P. The KDD process for extracting useful knowledge from volumes of data. Communications ACM, v. 39, n. 11, p. 27–34, nov. 1996.

8. GONCALVES, L. F.; VASCONCELOS, L. G.; MUNSON, E. V.; BALDOCHI, L. A. Supporting adaptation of web applications to the mobile environment with automated usability evaluation. In: ACM SYMPOSIUM ON APPLIED COMPUTING, 2016, New York, NY, USA: ACM, 2016. p. 787–794.

9. JQuery. **What is JQuery**. Disponível em: <https://jquery.com/>. Acesso em: 27 ABR 2021.

10. MDN Web Docs. **What is JavaScript?** Disponível em: < https://developer.mozilla.org/en-US/docs/Learn/JavaScript/First_steps/What_is_JavaScript />. Acesso em: 22 ABR 2021.

11. MDN Web Docs. **Introdução ao DOM**. Disponível em: < https://developer.mozilla.org/pt-BR/docs/Web/API/Document_Object_Model/Introduction />. Acesso em: 22 ABR 2021.

12. Node.JS. **About Node.JS**. Disponível em: < https://nodejs.org/en/about/ >. Acesso em: 6 MAIO 2021.

13. VASCONCELOS, Leandro Guarino de. USABILICS: avaliação remota e automática de usabilidade de aplicações Web baseada em um modelo de interface. 2012. 88 f. Dissertação (Mestrado em Ciência e Tecnologia da Computação) – Universidade Federal de Itajubá, Itajubá, 2012.

14. VASCONCELOS, L. G.; SANTOS, R. D. C.; BALDOCHI, L. A. Classifying user experience of web applications in real time using client logs. In: INTERNATIONAL CONFERENCE WWW/INTERNET, 2014, Porto, Portugal, 2014. p. 11–18.

15. VASCONCELOS, L. G. Uma abordagem para mineração de logs para apoiar a construção de aplicações web adaptativas. Tese (Doutorado em Computação Aplicada). Instituto Nacional de Pesquisas Espaciais, São José dos Campos, 2017. 120 p.

16. WEKA. **WEKA 3: Data Mining with Open Source Machine Learning Software** Disponível em: < https://www.cs.waikato.ac.nz/ml/weka/>. Acesso em: 30 SET 2021.