UNIVERSIDADE FEDERAL DE CAMPINA GRANDE

PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

José Rafael Feitosa Remígio

AVALIAÇÃO DO USO DE CRAWLERS FOCADOS NA IDENTIFICAÇÃO DE CRITÉRIOS FISCAIS EM PORTAIS DE TRANSPARÊNCIA PÚBLICA

Campina Grande

2020

Universidade Federal de Campina Grande
Centro de Engenharia Elétrica e Informática
Coordenação de Pós-Graduação em Ciência da Computação

Avaliação do uso de Crawlers focados na identificação de critérios fiscais em portais de transparência pública

José Rafael Feitosa Remígio

Dissertação submetida à Coordenação do Curso de Pós-Graduação em Ciência da Computação da Universidade Federal de Campina Grande - Campus I como parte dos requisitos necessários para obtenção do grau de Mestre em Ciência da Computação.

Área de Concentração: Ciência da Computação
Linha de Pesquisa: Recuperação da Informação

Nazareno Ferreira de Andrade

Campina Grande, Paraíba, Brasil

R387a Remígio, José Rafael Feitosa.
Avaliação do uso de crawlers focados na identificação de critérios fiscais em portais de transparência pública / José Rafael Feitosa Remígio. - Campina Grande, 2020.
71 f. : il. Color.

Dissertação (Mestrado em Ciência da Computação) - Universidade Federal de Campina Grande, Centro de Engenharia Elétrica e Informática, 2020.
"Orientação: Prof. Dr. Nazareno Ferreira de Andrade.
Referências.

1. Recuperação da Informação. 2. Crawlers Focados. 3. Algoritmos de Busca. I. Andrade, Nazareno Ferreira de. II. Título.

CDU 004.42(043)

FICHA CATALOGRÁFICA ELABORADA PELO BIBLIOTECÁRIO GUSTAVO DINIZ DO NASCIMENTO CRB-15/515

AVALIAÇÃO DO USO DE CRAWLERS FOCADOS NA IDENTIFICAÇÃO DE CRITÉRIOS FISCAIS EM PORTAIS DE TRANSPARÊNCIA PÚBLICA

JOSÉ RAFAEL FEITOSA REMÍGIO

DISSERTAÇÃO APROVADA EM 27/02/2020

NAZARENO FERREIRA DE ANDRADE, Dr., UFCG
Orientador(a)

JOÃO ARTHUR BRUNET MONTEIRO, Dr., UFCG
Examinador(a)

EUDISLEY GOMES DOS ANJOS, Dr., UFPB
Examinador(a)

CAMPINA GRANDE - PB

Resumo

O avanço das tecnologias da informação e comunicação (TICs) em conjunto com a popularização da internet, proporcionaram modificações nos processos de transparência nas gestões da máquina pública, impulsionando mudanças na legislação que transformaram a maneira de registrar, organizar e divulgar informações à população, por meio da propagação dos portais de transparência, sites da internet que oferecem o acesso à informações individuais sobre arrecadações e execuções orçamentárias das gestões públicas. Por lei, os sites de transparência devem promover o controle social, viabilizando à população o acompanhamento das contas públicas. Porém, na literatura são encontrados trabalhos que apontam ineficiências em referência ao nível de acessibilidade do site e a qualidade das informações divulgadas. Na tentativa de manter o acesso à informações fiscais atualizadas e coerentes, órgãos como o Tribunal de Contas do Estado da Paraíba (TCE-PB), fiscalizam periodicamente municípios e estados. No entanto, devido ao grande número de portais de transparência existentes entre os estados e municípios brasileiros, fiscalizar o cumprimento da legislação requer um alto custo humano e financeiro, tornando em alguns cenários esta atividade inviável. Este estudo experimentou o uso de um Web Crawler focado com diferentes algoritmos de busca na avaliação de itens fiscais exigidos por lei nos portais de transparência municipais da Paraíba. Os resultados indicaram níveis de eficácia e eficiência satisfatórios em comparação com a ferramenta Turmalina, sendo o *E-greedy + BFS* o melhor algoritmo de busca avaliado. Entretanto, foi possível identificar alguns casos que más práticas de desenvolvimento presentes nos portais de transparência que afetaram diretamente o desempenho do Crawler. Com base nestes resultados, é esperado o aprimoramento dos mecanismos de fiscalização, dos portais de transparência e o incentivo da participação cidadã na administração pública.

Abstract

The advancement of information and communication technologies (ICTs) together with the popularization of the Internet, provides changes in the transparency processes in the public management machine, driving changes in legislation that transforms a method of registering, organizing and disseminating information to the population, through the propagation of the transparency portal, Internet sites that share information on public administration budget executions. By law, transparency sites promote social control, enabling the population or monitoring public tasks. However, in the literature, studies are found that point out the inefficiencies about the level of accessibility of the website and the quality of the information disclosed. Native to maintain or access updated and consistent tax information, bodies such as the Court of Auditors of the State of Paraíba (TCE-PB), periodically inspect municipalities. However, due to a large number of opening portals between Brazilian states and municipalities, inspecting or complying with the legislation requires a human and financial right, showing in some scenarios this unfeasible activity. This study experimented with the use of a Web crawler focused on different search algorithms in the assessment of tax items required by law in Paraíba's municipal transparency portals. The results indicate satisfactory levels of efficiency and efficiency compared to the Tourmaline tool, with itshape E-greedy + BFS or the best-evaluated search algorithm. However, it was possible to identify some cases with more development practices present in the portal directly affected by the performance of the crawler. Based on these results, it is expected or improved the inspection mechanisms, transparency portals and incentives for citizens to participate in public administration.

Agradecimentos

Gostaria de agradecer a Deus, aos meus familiares, em especial, aos meus pais Maria Das Graças e Manoel Ferreira, minha namorada Dara Carneiro pelo apoio durante a realização deste trabalho e na minha vida em geral.

Ao meu orientador, Prof. Dr. Nazareno Andrade, pela paciência, ensinamentos, inspiração, dedicação e generosidade no processo de orientação deste trabalho.

Agradeço ao time do Laboratório Analytics e Laboratório de Sistemas Distribuídos pelo ótimo ambiente de trabalho, pelas discussões construtivas e momentos de descontração.

Agradeço ao Tribunal de Contas do Estado da Paraíba (TCE-PB) e também ao time Turmalina: Fanny Vieira, Lorena Santos, Jefferson Emanuel, Júlio e Daniyel Rocha pela competência, paciência, companhia e dedicação durante nosso projeto.

Aos amigos, que perto ou longe, envolvidos academicamente ou não, estão presentes durante os bons momentos e também durante os desanimadores.

Aos professores e funcionários da Copin e do Departamento de Sistemas e Computação. Finalmente, agradeço ao Governo Brasileiro, que através da CAPES, apoiou financeiramente a execução desta pesquisa.

Conteúdo

Lista de Símbolos

LAI - *Lei de Acesso à Informação*

LC131 - *Lei Complementar número 131*

TCE - *Tribunal de Conta Estadual*

TCE-PB - *Tribunal de Conta do Estado da Paraíba*

Lista de Figuras

Lista de Tabelas

Capítulo 1

Introdução

A aprovação da Lei de Acesso à Informação paralelamente com o amadurecimento das tecnologias da informação e comunicação (TICs) ajudaram a redefinir as formas de relacionamento entre governos e cidadãos, promovendo a criação de novos mecanismos de controle social, como os portais de transparência, que permitiram o acesso pela população à informações fiscais sobre gastos e arrecadações das entidades públicas. [15].

A disponibilização de portais de transparência pelos entes públicos modificou a maneira de divulgação e acompanhamento da transparência pública, propiciando uma aproximação entre a população e suas gestões públicas. No entanto, comprova-se em alguns casos a ineficiência destas instituições na criação de sites que comportem o acesso fácil em relação a navegação, interatividade, desempenho, atualização e coerência das informações fiscais [3].

Na tentativa de manter o acesso às informações fiscais atualizadas e coerentes, órgãos como o Tribunal Contas do Estado da Paraíba (TCE-PB) fiscalizam periodicamente municípios e estados, analisando a disponibilização de informações fiscais referentes às contas públicas destas entidades em seus portais de transparência[1]. Estas iniciativas já resultaram na aplicação de multas e representações de supostas irregularidades sobre os municípios.

Este trabalho tem origem nesse contexto, visando contribuir com a construção de soluções computacionais para situações como a enfrentada pelo TCE-PB ao monitorar um grande número de portais de transparência. A seguir são expostas a motivação por trás deste trabalho, o histórico desta pesquisa, seus objetivos e contribuições para o campo de recuperação da informação com uso de Web Crawlers focados no contexto da avaliação de portais de

[1] Índice de transparência Municipal - http://tce.pb.gov.br/indice-de-transparencia-publica

transparência fiscal públicos, bem como a descrição da estrutura deste documento.

1.1 Motivação

O direito de acesso à informação e a transparência pública podem ser considerados direitos humanos fundamentais em sociedades democráticas. Neste aspecto, gestões transparentes possuem como principais características a garantia do acesso à informações fiscais compressíveis a qualquer cidadão, e a abertura para sua participação no governo (Controle social). Estes critérios promovem uma maior autonomia aos cidadãos e aos órgãos fiscalizadores, no acompanhamento e fiscalização das finanças e execuções orçamentárias, realizadas pelos gestores das entidades públicas [6].

No Brasil a lei de Acesso a Informação (LAI) garante um papel de destaque à população no que diz respeito ao acesso e fiscalização das contas públicas, em contrapartida, os portais de transparência fiscais surgem em alguns casos apenas como instrumento para o cumprimento da legalidade, deixando em segundo plano aspectos ligados a utilização de informações coerentes, acessíveis e compreensíveis a qualquer público [3].

Auditar manualmente esses sites de transparência requer um alto custo tanto de recursos humanos quanto de recursos financeiros. Isso acontece devido ao grande número de portais de transparência existentes entre os estados e municípios brasileiros. Somente o estado da Paraíba possui 223 municípios, e a lei determina que ao menos a prefeitura e a câmara de vereadores de cada cidade possua um portal. Além disso, como há liberdade para a maneira como cada município pode implementar seu portal, fatores como a complexidade na estrutura de alguns portais e a diferenciação na forma de acessar e navegar em diferentes portais agravam ainda mais a realização desta atividade, tornando-a, em alguns cenários, impraticável.

A partir da literatura, foi constatada uma ausência de técnicas computacionais experimentadas com os desafios presentes no contexto de auditoria de portais de transparência pública. Porém, foi percebido a existência de pesquisas na área de recuperação da informação que objetivam a identificação e extração de conteúdos da Web em contextos como, por exemplo, sites de notícias, e-commerces e fóruns de perguntas e respostas, as quais são relevantes na tentativa de desenvolver e experimentar técnicas no que se refere a automatização

da avaliação de portais de transparência pública fiscal [4, 12, 18]. Também há uma lacuna no que diz respeito a capacidade destas técnicas em apresentarem abordagens generalistas para o acesso a conteúdos e a busca por páginas entre a grande diversidade de tipos de sites na Web.

1.2 Histórico

O presente trabalho foi desenvolvido como um desdobramento de um projeto realizado em uma parceria entre o Laboratório Analytics da Universidade Federal de Campina Grande (UFCG) e o Tribunal de Contas do Estado da Paraíba (TCE-PB), cujo objetivo foi implementar um sistema que transformasse as atualizações do Índice de transparência Municipal[2], que por meio de auditores fiscais, periodicamente fiscalizava os portais de transparência dos 223 municípios do estado, em um processo automatizado.

Esse projeto deu origem à ferramenta Turmalina [3], que regularmente acessa cada portal de transparência municipal da Paraíba e avalia a disponibilização das informações fiscais contidas nas diretrizes adotadas pelo Índice do TCE-PB. Neste cenário, um dos grandes obstáculos na implementação da ferramenta foi na criação de uma solução eficaz e eficiente em relação a identificação dos critérios fiscais, considerando a variabilidade dos portais de transparência na forma de navegar e dispor as informações em suas páginas.

Para validar a eficácia da solução foi selecionada uma amostra contendo os portais de transparência dos 10 municípios mais populosos e do estado da Paraíba. Os resultados apresentaram uma acurácia média de 0.7 em referência à capacidade da Turmalina em identificar corretamente os itens fiscais. Ademais, no tocante à eficiência, o Web Crawler leva cerca de 15 dias para avaliar todos os 223 portais considerando a avaliação de 3 portais simultâneos. Apesar destes resultados comportarem a demanda proposta, é importante destacar a necessidade de aprimoramento da eficácia e eficiência como forma de aumentar a confiabilidade das avaliações de transparência e reduzir o tempo gasto entre as execuções.

Nesta pesquisa, propomos melhorias na Turmalina em relação à aplicação de otimizações nos termos de buscas e identificação dos critérios, e a implementação e comparação de novos

[2]Índice de transparência Municipal - http://tce.pb.gov.br/indice-de-transparencia-publica

[3]Turmalina - http://turmalina.tce.pb.gov.br/

algoritmos de busca, possibilitando novos modos de buscar na árvore de Nós.

1.3 Objetivos e Contribuições

O objetivo geral deste trabalho é avançar o estado da arte no uso de Web Crawlers focados na avaliação de portais de transparência fiscal. Com este fim, foram desenvolvidas as seguintes atividades no contexto dos municípios do Estado da Paraíba:

- Identificação de abordagens algorítmicas promissoras para a seleção e priorização do acesso às páginas web baseadas em outros cenários para serem experimentadas na automatização da avaliação de transparência municipal;
- Avaliação e comparação da eficácia e eficiência entre os algoritmos Bfs, Dfs e Epilson-greedy na busca por páginas consideradas relevantes para o contexto da avaliação de transparência;
- Caracterização das adversidades encontradas na automatização da avaliação de transparência.

Com base nas atividades descritas, será possível avaliar a viabilidade na utilização de Web Crawlers na avaliação de portais de transparência fiscal públicos e propor melhorias no que tange a manutenção da transparência entre órgãos fiscalizadores, população e municípios.

1.4 Estrutura do Documento

O restante deste documento está estruturado da seguinte maneira:

No Capítulo 2 são expostos os fundamentos teóricos necessários para a compreensão deste trabalho e o contexto no qual ele está inserido. No Capítulo 3 são descritos e comentados os trabalhos encontrados na literatura relacionados ao contexto de acesso, navegação e extração de dados estruturados em sites da web. Já no Capítulo 4, são mostradas as características do Web Crawler focado proposto neste trabalho. Os materiais e métodos utilizados no planejamento e desenvolvimento deste estudo são descritos no Capítulo 5. No Capítulo

6 são expostos os resultados obtidos e discussões geradas. Por fim, no Capítulo 7 apresentamos as conclusões deste trabalho de dissertação, limitações e perspectivas para atividades futuras.

Capítulo 2

Fundamentação teórica

Este capítulo apresenta os fundamentos teóricos utilizados no trabalho desenvolvido. A Seção 2.1 expõe as mudanças na legislação que modificaram a maneira de gerir as entidades públicas em relação a transparência. O tópico 2.2 apresenta a inserção dos portais de transparência como ferramentas de controle cidadão. A seção 2.3 mostra trabalho do TCE-PB na avaliação de transparência das gestões da Paraíba. Na seção 2.4 é apresentado os conceitos e características dos Web Crawlers. Por fim, o item 2.5 discorre sobre algoritmos de busca utilizados em Web Crawlers para priorizar e selecionar o acesso aos links e componentes clicáveis nas páginas web.

2.1 A Transparência Fiscal Pública

Em 27 de maio de 2009, foi sancionada a Lei Complementar nº 131 (LC131), também conhecida como Lei da Transparência, que alterou Lei de Responsabilidade Fiscal (LRF) no que se refere à transparência da gestão fiscal, determinando a disponibilização, em tempo real, de informações detalhadas sobre a execução orçamentária e financeira da União, dos Estados, do Distrito Federal e dos Municípios [1]. Neste contexto, o processo de transparência pública é complementado e ampliado com o sancionamento da Lei de Acesso à informação (LAI), em 18 de novembro de 2011, que dentre suas disposições estabelece que toda informação referente às atividades do Estado é pública, salvo exceções previstas na legislação [2, 3]. De modo geral as mudanças na legislação garantiram aos cidadãos o acesso à:

- Dados institucionais dos órgãos e entidades do Poder Executivo Federal;
- Dados para o acompanhamento de programas, ações, projetos e obras de órgãos e entidades;
- Inspeções, auditorias, prestações e tomadas de contas realizadas pelos órgãos de controle interno e externo;
- Informações sobre inspeções, auditorias, prestações e tomadas de contas realizadas pelos órgãos de controle interno e externo;
- Todos os registros de repasses ou transferências de recursos financeiros;
- Registros das despesas;
- Procedimentos licitatórios, bem como as informações de todos contratos celebrados;
- Formas de Solicitação de Informações (Serviço de Informações ao Cidadão (SIC) físico e eletrônico).

O dever fiscal das instituições públicas instituiu dois tipos de transparência, a *ativa*, a qual consiste na divulgação espontânea das informações fiscais, e a *passiva*, que equivale ao fornecimento de meios para solicitação de informações mediante a realização de requerimentos [11]. Neste âmbito, surgiram os portais de transparência, sites da web que reúnem informações fiscais gratuitas e acessíveis a qualquer público como instrumento que comportam ambas modalidades de transparência.

2.2 A Democratização dos Portais de Transparência

A evolução das tecnologias da informação e comunicação (TICs) simultaneamente com a popularização da internet, tiveram influência direta na modificação dos processos de transparência nas gestões da máquina pública, provocando mudanças na legislação que transformaram a maneira de registrar, organizar e divulgar informações [3]. Diante deste contexto, houve uma propagação de ferramentas de software com propósito de apoiar o controle social, como é o caso da oferta de portais do governo e portais de transparência.

Os portais de transparência são compreendidos como sites da internet que centralizam uma série de conteúdos e serviços comumente agrupados por critérios fiscais, como por exemplo, despesas, receitas, licitações, quadro pessoal e etc., de modo a simplificar o acesso do usuário por áreas de seu interesse [15]. Um dos principais objetivos destes sites é o de traduzir os conteúdos disponibilizados pelos representantes públicos de forma compreensível e acessível a toda a população [3]. Atualmente os sites de transparência são fornecidos pela União, Estados, Distrito Federal e Municípios.

Contudo, apesar desses sites serem introduzidos como o principal mecanismo de controle social, com intuito de atender as demandas legais criadas pela LAI, é possível identificar na literatura trabalhos que apresentam deficiências em níveis de portais estaduais e municipais no que diz respeito a usabilidade, clareza nas informações e mesmo na disponibilização completa e correta dos critérios fiscais [3, 7, 15]. Neste contexto, além de prejudicar o uso desta ferramenta na fiscalização cidadã das contas públicas, estes problemas evidenciam o comprometimento das gestões apenas pelo cumprimento da legalidade, descartando critérios que tornem os sites mais acessíveis.

2.3 O Índice de Transparência Pública do TCE-PB

Na Paraíba, o Tribunal de Contas do Estado (TCE-PB) apoiado pela LAI criou o Índice de Transparência pública[1] que regularmente avalia os portais de transparência dos 223 municípios, suas câmaras e o estado na Paraíba, analisando a disponibilização de 123 itens divididos entre os temas:

- **Conteúdo:** instrumentos de planejamento (Plano Plurianual, Lei de diretrizes Orçamentárias e Lei Orçamentária Anual), Procedimentos licitatórios, Contratos, Convênios e Execução orçamentária e financeira que representa checagem de informações relacionadas a despesas, receitas, folha de pagamento e etc.;
- **Série Histórica e Frequência de Atualização:** tempo em dias para ocorrência de atualizações no site e a manutenção do histórico das informações fiscais ao longo dos anos;

[1] Índice de transparência Municipal - http://tce.pb.gov.br/indice-de-transparencia-publica

- **Usabilidade:** interatividade, a possibilidade do uso de filtros nas consultas e disponibilização dos dados, filtrados ou não, em formatos como xls, pdf ou txt;

- **Outros:** refere-se à avaliação de itens não classificados nos temas anteriores como a presença de SIC físicos e eletrônicos, a disponibilização de busca no site, disponibilidade do Relatório Resumido da Execução Orçamentária (RREO) e entre outros.

A avaliação é realizada individualmente em cada portal de transparência, com o acesso ao endereço do site na internet e a validação da presença ou ausência de cada um dos itens. Inicialmente este procedimento era feito por auditores fiscais do próprio TCE-PB, mas passou, como descrito no Capítulo 01 deste documento na seção Histórico, a ser executado pela ferramenta Turmalina[2], reduzindo assim o custo de recursos humanos e financeiros envolvidos. Ainda assim, é importante enfatizar que as avaliações ainda necessitam de auditores que validem os resultados obtidos, como forma de manter uma alta confiabilidade.

De acordo com o resultado obtido através das checagens dos itens em cada tema, é atribuída uma nota ao município responsável pelo portal de transparência, a qual pode variar entre 0 até 1000 pontos. Esta pontuação é empregada na criação de um ranque de transparência, o qual quanto mais próximo a nota estiver do valor máximo, melhor é considerado o nível de transparência da gestão.

Essa iniciativa tem apresentado efeitos positivos no que tange a melhoria dos portais de transparência como ferramentas de controle social. Isto deve-se ao uso do Índice como termômetro da transparência entre as gestões municipais e estaduais e a população, seja como mecanismo de autopromoção ou para cobranças de melhorias. Segundo TCE-PB, tais ações já resultaram na aplicação de cerca de quatrocentos mil reais em multas e diversas representações à Controladoria Geral do Estado, Controladoria Geral da União e Procuradoria Geral de Justiça.

2.4 Web Crawlers: Conceitos e Características

A enorme e constante produção de dados na web cria novas perspectivas no que diz respeito a aplicação de inteligência sobre os dados, seja por exemplo para o desenvolvimento de fer-

[2]Turmalina - http://turmalina.tce.pb.gov.br/

ramentas para análises de preços de produtos em e-commerces, na indexação e classificação de conteúdos ou mesmo na avaliação do cumprimento da legislação, no caso dos portais de transparência [10, 17, 23]. Entretanto, reunir estes dados torna-se uma tarefa não trivial, em virtude da falta de mecanismos que os disponibilizem de forma centralizada em formatos adequados a estes tipos de análise.

Desse modo, os Web Crawlers surgem como instrumentos úteis para coletar dados da web, definidos como programas que por meio de um processo designado *web crawling* são capazes de rastrear e extrair conteúdos de páginas da web [8]. Neste cenário, conforme é mostrado na Figura 2.1, o método básico de *web crawling* é iniciado a partir da URL principal do site, que posteriormente é requisitada e se obtém o acesso à página contendo seu código HTML, neste caso, são extraídos os dados de interesse no documento e novos links que mais tarde serão acessados, sendo repetido todo o ciclo até o último link descoberto ser requisitado.

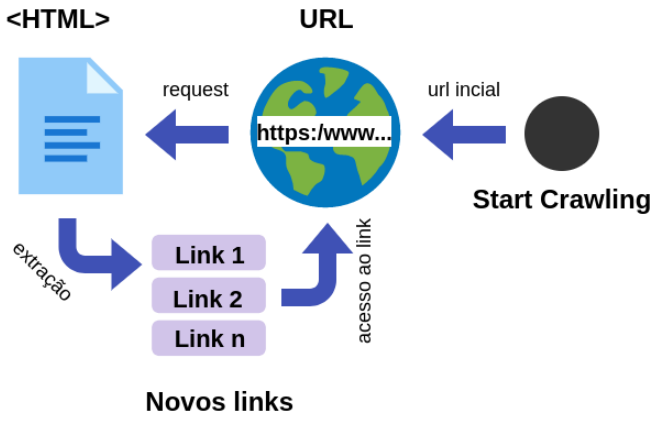

Figura 2.1: Processo de Web Crawling.

O modo de buscar e extrair os dados das páginas nessas ferramentas costumam ser diversificados de acordo com o objetivo proposto. Segundo Du et al. [8], os Web Crawlers podem ser divididos entre as categorias:

- **Gerais:** tendem a percorrer todas as páginas de um determinado site, sem nenhum direcionamento ou contexto analisado, sendo frequentemente aplicados na criação de bases de dados contendo páginas HTML estáticas e na indexação de conteúdos por motores de busca;

- **Focados:** acessam os sites embasado em um contexto como por exemplo: notícias, perguntas e respostas em fóruns, produtos em e-commerces e etc., objetivando exclusivamente a extração de dados significativos, evitando links duplicados ou que não se adéquam à finalidade buscada, ou seja, links que já foram acessados anteriormente ou que direcionam para páginas com conteúdos não relevantes para a cenário procurado.

Um dos grandes desafios encontrados na implementação de Web Crawlers eficazes e eficientes de maneira geral diz respeito a criação de soluções que conseguem lidar de forma equilibrada com a priorização das páginas a serem acessadas, as estruturas contidas nas páginas web e o carregamento de páginas com conteúdos estáticos e dinâmicos. No que corresponde ao meio priorização e seleção do acesso às páginas web, a próxima seção apresenta as definições dos algoritmos utilizados nesta pesquisa.

2.5 Algoritmos de Busca em Crawlers

Durante o processo de Web Crawling é necessário estabelecer regras que determinem a ordem de acesso de cada link encontrado. Para isto, são aplicados algoritmos que auxiliam na seleção apropriada do link a ser acessado em cada ciclo de execução. Este procedimento pode estar diretamente relacionado a eficiência da ferramenta em percorrer e navegar entre as páginas dos sites. Assim, a escolha da técnica computacional deve ser especificada com a análise dos sites aos quais o Crawler será executado, observando, por exemplo, quantas páginas são acessadas para que as informações essenciais estejam disponíveis. Nas subseções 2.5.1 e 2.5.2 são apresentados três algoritmos que podem ser empregado com este objetivo.

2.5.1 Breadth-First Search (BFS) e Depth-First Search (DFS)

O *Breadth First Search* (BFS) é um dos algoritmos mais simples usados nas operações de busca em grafos. A partir de um grafo $G = (V, E)$ – formado por um conjunto de vértices V e um conjunto de arestas E – e um vértice s, o BFS explora as arestas de G para descobrir todos os vértices alcançáveis por s. Isto produz uma árvore BFS com a raiz s contendo todos os vértices atingíveis [5]. Desta maneira, quanto menor o nível do vértice mais próximo ele estará do vértice inicial (Raiz) e mais rápido ele será acessado.

Outro algoritmo de busca em grafos é o *Depth-First Search* que diferentemente do BFS prioriza o acesso aos vértices pertencentes a níveis mais profundos em um grafo com estrutura de dados do tipo árvore, por meio da realização de uma busca em profundidade [13]. Esta comparação pode ser identificada na Figura 2.2 que expõe a ordem de acesso aos vértices também denominados Nós em uma árvore para ambas as abordagens.

No ponto de vista dos Web Crawlers, o uso destes algoritmos visa determinar a ordem de priorização dos links a serem acessados durante o web crawling, seja dando preferência de acesso a links encontrados próximos da página principal do site no uso do BFS ou concedendo a precedência de acesso a links contidos em níveis de navegação mais profundos e distantes da página inicial com o DFS.

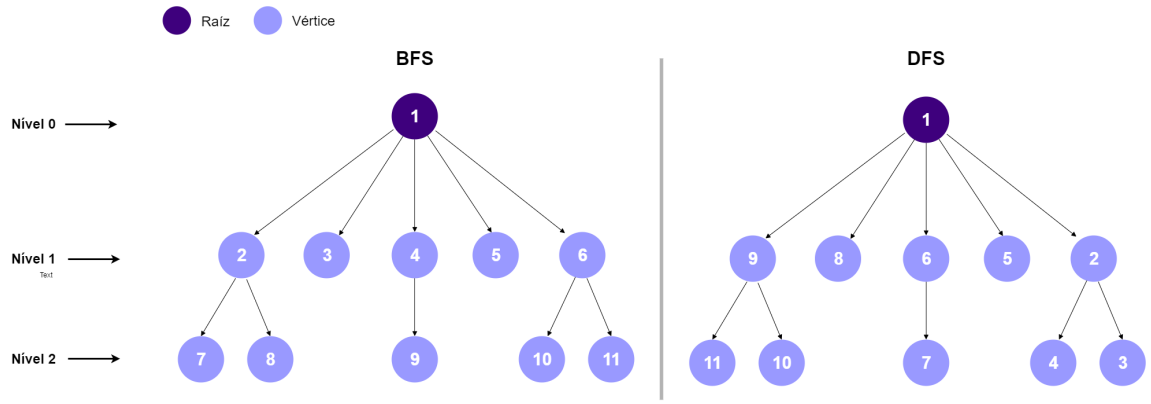

Figura 2.2: Representação dos algoritmos BFS e DFS.

2.5.2 Epsilon-greedy

O *epsilon-greedy* é o mais popular e simples algoritmo de aprendizado por reforço no tratamento do problema entre *exploration* e *exploitation* [19]. No âmbito da forma de um Web Crawler varrer as páginas de um site, é possível classificar a etapa de *exploration* como a tarefa de procurar aleatoriamente páginas web que resultem na identificação de conteúdos importantes para o contexto executado. Já a etapa de *exploitation* compete a investida do algoritmo em links que foram achados próximos a páginas relevantes, partindo da premissa que páginas relevantes estão localizadas próximas das outras.

A escolha do modo de funcionamento dessas etapas é dada pela constante *epsilon* (ε), a qual representa a probabilidade do nível de exploração que deve ser utilizado em cada exe-

cução [19, 22]. Por exemplo, caso se tenha $\varepsilon = 0.1$ o algoritmo em 10% do tempo fará com que o Crawler (Agente) escolha o link a ser acessado de forma aleatória (exploration) e em 90% do tempo irá explorar (exploitation) a melhor opção, ou seja, os caminhos que obtiveram páginas relevantes. Assim, a solução adapta-se a opção que oferece maior performance, baseando-se nas descobertas que alcançaram um maior ganho ao longo de sua execução.

Capítulo 3

Trabalhos relacionados e estado da arte

Neste capítulo é descrita uma visão geral dos trabalhos prévios que abordaram problemas de Web Crawlers relacionados a priorização e identificação de links no processo de web crawling e a extração de dados estruturados em páginas da web. Estes foram associados respectivamente a Navegabilidade (3.1) e Extração (3.2).

3.1 Navegação

De acordo com Lee et al. [14] um Web Crawler Focado deve possuir métodos que determinem a relevância de páginas web baseando-se no tópico buscado. No entanto, as numerosas e diversificadas formas de navegar e disponibilizar informações entre sites de um mesmo domínio torna complexo este tipo de classificação.

Meusel, Mika e Blanco [16] propõem um Web Crawler Focado que combina um classificador online e uma abordagem Bandit com o intuito de maximizar o número de páginas encontradas relacionadas ao tópico buscado. A solução experimentou os modelos de aprendizado de máquina *Naive Bayes* e *Hoeffding Trees* para classificar os links não acessados como relevantes e o algoritmo *epsilon-greedy* (Abordagem Bandit) para priorizar e estabelecer a ordem de acesso aos sites, fundamentando-se em escores obtidos sobre as classificações de links já acessados. Os resultados mostraram um aumento de 26% no número de páginas relevantes com o classificador online em comparação com o uso do mesmo manualmente treinado, sendo o *Naive Bayes* o modelo com maior eficácia.

Com o mesmo objetivo Du et al. [8] apresenta um Web Crawler denominado *Semantic*

Similarity Vector Space Model (SSVSM) que utiliza da similaridade do cosseno e a similaridade semântica para nesta ordem definir a prioridade de acesso entre os links descobertos e não visitados ao longo do processo de crawling e aumentar a cobertura de links significativos. Para validar a técnica foram comparados Web Crawlers apoiados em *Breadth-First Search* (BFS), *Vector Space Model* (VSM), *Semantic Similarity e Retrieval Model (SSRM)* aplicados em sites divididos entre 10 diferentes tópicos, resultando em uma superioridade do SSVSM em 0.65% na similaridade média do nível de páginas encontradas em comparação ao SSRM que apresentou o segundo melhor resultado.

Apesar desses trabalhos exporem soluções que suportam a extração de conteúdos de diferentes tópicos, são encontradas abordagens na literatura que apontam ineficiências nas buscas por páginas relevantes em ferramentas mais generalistas como, por exemplo, fóruns de perguntas e respostas que possuem links em níveis profundos na árvore contida Document Object Model (DOM) [12].

Nesse cenário Cai et al. [4] sugere um Crawler focado chamado IRobot especializado na extração de dados em fóruns, que consiste no aprendizado do mapeamento do site com base em páginas similares agrupadas em clusters de acordo com seus leiautes e formatos das URLs. A técnica foi testada em uma comparação a um Web Crawler genérico usando BFS sobre 7 fóruns distintos. O IRobot expôs uma melhor eficácia em extrair páginas úteis, com uma proporção média de 91% das páginas relevantes extraídas contra 69% da abordagem genérica.

Por sua vez, Jian [12] apresenta a ferramenta Focus que demonstrou superar o IRobot em relação a eficácia na extração de dados em fóruns, indicando 100% de cobertura. Para isto, o problema da variabilidade de leiautes entre os fóruns durante a extração foi reduzido para reconhecimento de URLs (links), pois apesar de serem diferentes, os sites possuem caminhos de navegação semelhantes. Neste sentido, é proposto a utilização do modelo de aprendizado de máquina *Support Vector Machine (SVM)* para classificar os tipos de páginas existentes nestes ambientes, baseando-se em features que representam os leiautes como o número de itens identificados na página, se existe timestamp nos dados, se cada item identificado possui um link para um perfil de usuário e entre outras. Os experimentos foram executados em 200 fóruns diferentes, onde 40 deles foram empregues como treino e 160 como teste.

3.2 Extração

Ainda que a World Wide Web Consortium (WC3)[1] disponibilize guias de boas práticas que incentivam a criação de web sites consistentes no que tange a compreensão, processabilidade, reúso, confiança, acessibilidade e interoperabilidade no compartilhamento de dados, segundo a Opera Software Company apenas 4.13% dos sites possuem suas estruturas nas conformidades com o HTML padrão definido [23]. Nesta circunstância, a tarefa de identificar e extrair porções de dados relevantes em páginas web converte-se em um processo difícil de ser realizado, dando margem a proposição de estudos que visam realização deste processo de modo simplificado.

A pesquisa realizada por Furche et al. [9] expõe um web extrator identificado como OXPath que estende o *XML Path Language (Xpath) 2.0*, oferecendo suporte a criação de consultas robustas e acessíveis ao usuário para a extração de conteúdos, permitindo a execução de eventos de cliques, preenchimentos de formulários, rolagens nas páginas e entre outras operações. A proposta mostrou-se mais eficaz em conferência aos web extratores Web Content Extrator, Lixto, Visual Web e Ripper no que diz respeito ao carregamento das páginas e uso de memória.

Vidal et al. [21] busca extrair os dados das páginas com o uso da similaridade entre estruturas do DOM, apoiando-se sobre exemplos predefinidos para comparar as estruturas como forma de possibilitar a coleta de conteúdos e a descoberta dos padrões de navegação do site. Para esta finalidade, é aplicado o conceito de *Tree edit distance* (TED) que condiz ao custo associado em transformar uma árvore DOM em outra, determinando o quão similares elas são. A validação da solução foi feita através de 11 sites com estruturas HTML semelhantes, resultando em valores de Recall e Precisão entre 95% a 100%. Ainda assim, esta técnica foca na extração de estruturas bastante similares, o que acaba não refletindo na realidade detectada entre alguns sites com o mesmo contexto.

Wu, et al. [24] aborda o mesmo problema com o uso do modelo de aprendizado de máquina *Regressão Logística*, que classifica cada Nó (tag HTML) da árvore DOM como relevante ou não por intermédio de features relacionadas à *posição da tag no DOM*, *à fonte e cor do texto* e ao *tamanho do texto visível*. Neste processo, a etapa de treino é realizada pela

[1] Data on the Web Best Practices W3C - https://www.w3.org/TR/dwbp/audience

inserção de árvores DOM previamente rotuladas com Nós considerados relevantes, posteriormente as páginas são classificadas e cada Nó é separado em grupos, respeitando os níveis hierárquicos da árvore, por fim, é calculado o escore de cada grupo de acordo com probabilidade identificada entre os Nós durante a classificação, em seguida é feita seleção do grupo com maior score que deve representar as tags com dados importantes na página. Para os experimentos foram usados 805 sites, contendo 2000 páginas separadas entre treino e teste, ocasionando em um de 0.8 para o Recall e 0.9 para a Precisão.

Omari et al. [17] indica um framework para construção de Crawlers supervisionados sintetizados para tópicos predeterminados que propõem aprenderem as estruturas das páginas, adotando exemplos de outros sites com o mesmo contexto de informação, por exemplo, livrarias online possuem livros com informações semelhantes como título, autor e preço. Deste modo, ocorre uma generalização para criação de um Crawler sintetizado para extrair dados de sites com esta categoria. A proposta mostrou valores de Recall entre 0.85 a 1 e 0.74 a 1 para a Precisão em relação a extração correta dos conteúdos das páginas. Estes resultados foram concebidos por meio do teste de 30 Crawlers sintetizados para extrair conteúdos de 9 tópicos diferente como livros, tvs, música, filme e entre outros.

Velloso et al. [20] experimentou o uso de uma estratégia de processamento de sinais com o *power spectrum density* (PSD) para distinguir áreas com dados estruturados relevantes de prováveis ruídos (menus, propagandas, footer e etc.), mediante ao reconhecimento de padrões de repetição nas estruturas da árvore DOM contidos no sinal espectral. A abordagem foi testada e validada comparando as técnicas do estado da arte MDR, TCP e ClustVX no processo de extração sobre três datasets que incluía cerca de 107 sites estáticos, resultando em valores de f1-scores entre 99.80% a 99.66% para o ClustVX contra f1-scores na faixa de valores entre 93.05% a 98.83% para a técnica apresentada. No entanto, foi constatado que a solução pode ser até seis vezes mais rápida na extração dos dados quando confrontada com o ClustVX.

Wu et al. [23] desenvolveu um algoritmo (CEDS) para a extração de dados em sites de notícias usando *Dempster-Shafer evidence thoery*, que em conjunto a um processo de extração tags HTML (Nós), reconhece regiões do DOM que possuem partes das matérias das notícias. Neste sentido, a DS theory é adotada como mecanismo para estimar a probabilidade de cada tag path extraída ser relevante, desejando evitar tags paths com ruídos. Os

resultados mostraram valores de f1-score entre 87.63% a 97.47% para o CEDS, superando as técnicas CETR e CEPR no que compete a cobertura e acertabilidade na extração. Este trabalho coletou aleatoriamente páginas de sites de notícias como NYPost, NYTimes, BBC e entre outros.

3.3 Considerações finais

Os trabalhos citados correspondem a contribuições primordiais para estado da arte de recuperação da informação, mais precisamente em um avanço significativo quando se refere a criação de ferramentas de extração de dados da web eficazes e eficientes para os mais variados e diversificados cenários e objetivos.

Os trabalhos encontrados demonstram a aplicabilidade de Web Crawlers a diversas finalidades de extração desde de notícias e threads em fóruns de perguntas e respostas até a coleta de arquivos de Microsoft PowerPoint de instituições acadêmicas [4, 12, 14, 23]. Contudo, nesta pesquisa é proposto o uso de Web Crawlers focados em um ambiente ainda não explorado pela literatura retratada, cujo objetivo principal modifica-se para checagens da disponibilidade de informações fiscais obrigatórias estabelecidas pela LAI em portais de transparência municipal, ou seja, existe um interesse maior em identificar quais informações estão inseridas entre cada estrutura da página.

Apesar desta pesquisa tratar de um diferente contexto, através do estado da arte foi encontrado e aplicado o algoritmo *Epsilon-greedy* proposto por Meusel, Mika e Blanco [16] como uma possível alternativa para priorizar a ordem de acesso aos links. Além disso, foi usado o *Xpath* combinado a termos chaves para criação de consultas que visam identificar e classificar a existência de 64 itens fiscais nos portais. Neste sentido, trabalhos como Omari et al. [17] & Furche et al. [9] também adotaram o *Xpath* para buscas em páginas.

Capítulo 4

Web Crawler focado

Neste capítulo são retratados os procedimentos e características principais do Web Crawler focado proposto nesta pesquisa.

4.1 Estruturas e conceitos

Duas características na provisão marcam o cenário em que nosso Crawler deve operar. Primeiro há um número considerável de empresas e combinações de empresas fornecendo portais de transparência aos municípios paraibanos, gerando bastante diversidade nas estruturas de site encontradas. Em segundo lugar, existe uma ampla presença de más práticas de desenvolvimento Web nos portais que observamos.

Um exemplo prático deste tipo de situação é o que ocorre entre os municípios de Ouro Velho e Alcantil, que possuem a mesma empresa gerenciando seus portais de transparência. O portal implementado pela empresa distingue os usuários interessados nos portais das diferentes cidades por meio de tokens contidos na URL inicial acessível pelo site da prefeitura. Ou seja, o mesmo site em um mesmo servidor gerencia os dados de mais de um município diferenciando-os através da criação de sessões com a identificação do token.

Note que essa implementação não permite o acesso à página inicial dos portais de dados fiscais dos dois municípios através de URLs únicas. É necessário que um navegador visite o portal da prefeitura, e obtenha um link contendo o token de acesso, e só então seja redirecionado para a URL do portal. Administrar essas sessões de modo automatizado é uma atividade complexa, devido ao término da sessão após passado um determinado limite de

tempo, o que causa a perda de todo o conteúdo apresentado no site, mesmo após o recarregamento da página. Sob esta perspectiva, é necessário acessar novamente o link encontrado na prefeitura do município para criação de uma nova sessão no portal. Neste ponto de vista, a ferramenta de Crawler após realizar o acesso à página, executa os procedimentos que não obrigatoriamente requerem interações diretas na página, podendo em alguns casos ultrapassar este tempo limite.

Para tornar viável o bom funcionamento do Web Crawler também nesses contextos, implementamos estruturas que possibilitam mapear todo o histórico de relacionamento entre as páginas, de modo semelhante ao apresentado em grafos bidirecionais com estruturas de árvore, propondo os seguintes conceitos:

- **Nó**: é uma representação unificada de *URLs* ou *componentes dinâmicos clicáveis (button, input, span e etc.)* que consideram o mapeamento de sites em forma de árvores;
- **Nó Pai**: é um Nó que a partir dele foi identificado ao menos um novo Nó (filho), a URL inicial do portal é o Nó pai de todos os Nós;
- **Nós Filhos**: são Nós que foram identificados por meio de um outro Nó (pai);
- **Nível do Nó**: corresponde quantidade de Nós que devem ser acessados para localizar o Nó, por exemplo, o Nó contendo a URL inicial possui o nível 0.

Na seção a seguir mostramos os passos desempenhados pela ferramenta antes e depois do processo de *web crawling*.

4.2 Pré-processamento e pós-processamento

Para o melhor entendimento das técnicas utilizadas neste estudo, é necessário compreender alguns passos que precedem e procedem a varredura dos portais de transparência. Desta forma, a Figura 4.1 e as próximas seções apresentam com detalhes cada fase deste fluxo.

4.2.1 Metadados

Os metadados são conjuntos de diferentes coleções de dados que agregam ao Web Crawler o conhecimento necessário do contexto ao qual será inserido. Todos os metadados usados

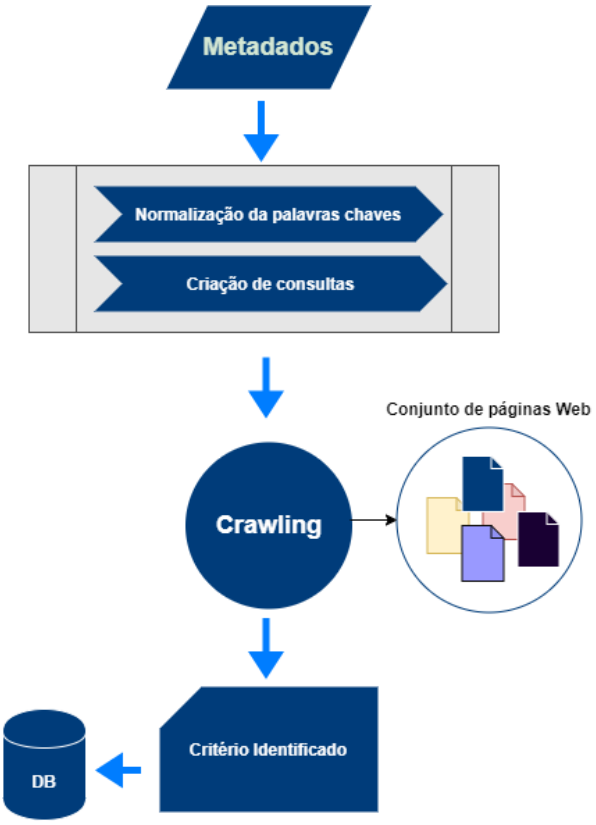

Figura 4.1: Pré e pós processamento do Web Crawler.

neste trabalho foram criados pela equipe de desenvolvimento da Turmalina de forma manual. Abaixo são expostas as características de cada uma delas:

- **Palavras chaves de busca**: são mapeamentos de diferentes terminologias descobertas nos sites fiscais que dão acesso a um mesmo critério fiscal, por exemplo, é possível encontrar os termos *folha de pagamento* e *servidores* em diferentes sites para obter acesso à área contendo critério Quadro Pessoal. Atualmente a lista de busca contém em média entre 8 até 16 diferentes termos considerando cada critério fiscal avaliado;

- **Palavras chaves de identificação**: refere-se a distintos termos usados para identificar um mesmo item fiscal. Em cada página é checada todas as variações mapeadas de cada item para classificar sua presença ou ausência, oferecendo suporte a portais de transparência com diferentes terminologias. Cada item procurado possui em nossos dados média de 1 até 10 termos diferentes para referenciá-lo;

- **Metadados dos municípios**: diz respeito às informações básicas dos municípios como

a URL da prefeitura, a URL do portal da transparência e a(s) empresa(s) fornecedora(s) do portal transparência;

- **Criação e salvamento do critério avaliado**: condiz a etapa de pós-processamento ao qual o Web Crawler finalizou todas as operações, neste momento, são organizadas e salvas no banco de dados todas as informações referentes ao critério buscado como quais itens foram identificados, o tempo de duração final da operação e a data da avaliação.

4.2.2 Normalização das palavras chaves

Na etapa de normalização das palavras chaves todos os termos de busca e identificação são normalizados, sendo removidos acentos, pontos, traços, espaços em branco e convertendo todas as palavras em letras minúsculas (lowercase). A título de exemplo, o termo *Despesa extra-Orçamentária* é transformado em *despesa extra orçamentaria*. Este processo visa expandir a cobertura dos termos durante as buscas nos sites de transparência.

4.2.3 Criação das consultas

Após a obtenção do acesso às palavras chaves normalizadas, o Web Crawler cria automaticamente as consultas utilizando a linguagem *XML Path Language* (XPath) que propiciarão a busca por novos Nós e checagens da presença de itens fiscais nas páginas HTML acessadas. Estes meios de buscas são desenvolvidos para serem reutilizáveis em portais de transparência com estruturas distintas em relação à árvore DOM. Com este intuito, são construídas consultas com o foco principal nos termos chaves, sem o uso de atributos específicos de cada site como classes Css e ids.

Se por um lado as consultas possuem palavras normalizadas, por outro, as páginas web às quais elas serão aplicadas não. Assim, a sintaxe do Xpath deve conseguir interpretar as palavras normalizadas para identificar termos que contenham o mesmo sentido e possuam diferentes grafias, por exemplo, palavras com letras maiúsculas ou acentuadas. Considerando as novas versões do Xpath (2 e 3), é possível descobrir funções úteis para normalização que permitem transformações dos textos das páginas em lowercase e uppercase de forma nativa. Entretanto, a maioria das linguagens de programação e ferramentas como o próprio

Puppeteer[1] ainda adotam o Xpath 1.0 como a principal especificação de implementação, na qual não há suporte a qualquer tipo de normalização nativa.

Como forma de reduzir o impacto da falta de suporte aos termos normalizados pelas consultas, foram criadas manualmente sequências de caracteres que traduzem um caractere para outro, estas foram aplicadas à função translate, nativa do Xpath 1.0. A Figura 4.2 mostra como os caracteres foram traduzidos, por exemplo, ã foi traduzido para a, isto é, se na página existir um termo com o caractere ã com um Xpath contendo a é possível identificá-lo.

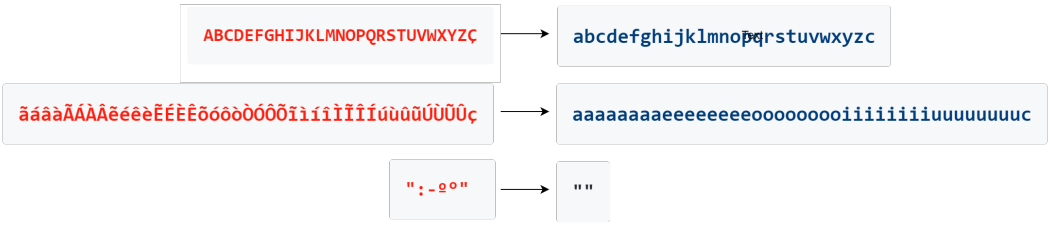

Figura 4.2: Cadeia de caracteres para normalização no Xpath.

4.3 Fluxo completo

Durante a execução do Web Crawler, optamos pela separação do processo de busca e identificação no nível de critério fiscal, visando facilitar uma maior distinção das páginas acessadas por meio dos termos utilizados em cada critério, evitando problemas como a identificação de itens semelhantes em locais pertencentes a outros critérios, visto que critérios como despesa extraorçamentária e despesa orçamentária possuem itens similares. Além disso, a independência entre os processos de avaliação dos critérios permite uma paralelização entre eles, resultando em um melhor aproveitamento dos recursos disponíveis, tornando as execuções mais eficientes. Para ilustrar todo o fluxo de avaliar um critério fiscal, o diagrama apresentado na Figura 4.3 foi proposto. A seguir são detalhadas cada um dos passos presentes no diagrama:

- **Start Crawler**: diz respeito ao processo de inicialização do Crawler, onde são passados todos metadados necessários para sua execução. Nesta etapa, a varredura de cada critério é iniciada de forma paralela, usando operações assíncronas;

[1]https://github.com/puppeteer/puppeteer

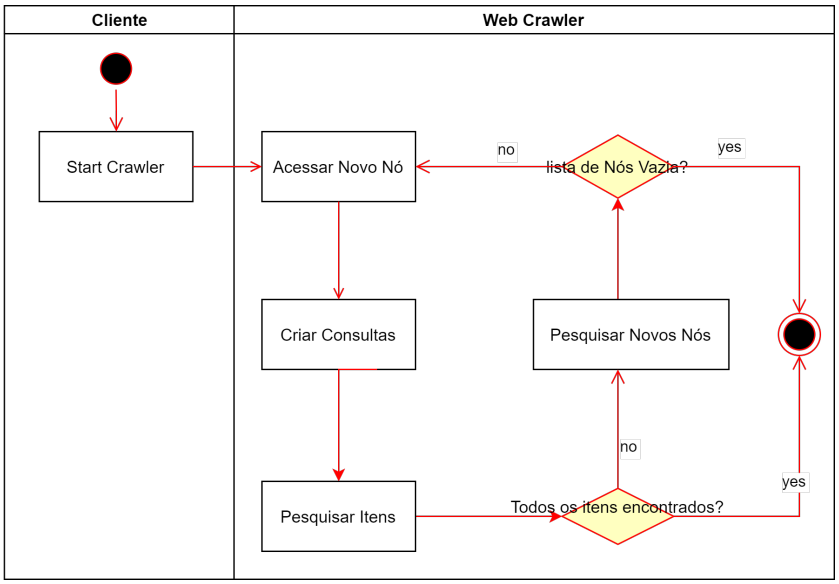

Figura 4.3: Fluxo de avaliação de um critério fiscal pelo Crawler.

- **Acessar Novo Nó**: esta atividade refere-se ao acesso do Crawler a um Nó previamente selecionado pelo algoritmo de busca adotado. Vale enfatizar que neste pesquisa o algoritmo de busca tem o papel exclusivo de priorizar e selecionar os Nós considerando uma representação de grafos bidirecionais em formatos de árvore;

- **Criar consultas**: são criadas consultas no formato de Xpaths pelo Crawler para identificar os itens e encontrar novos Nós (filhos) relevantes para o critério avaliado;

- **Pesquisar itens**: com as consultas criadas, os itens do critério avaliado são buscados. Neste processo algumas validações são aplicadas, como por exemplo verificar se o item encontrado está contido em alguma tabela ou em uma lista. Caso todos itens buscados forem encontrados o processo de avaliação do critério é finalizado.

- **Pesquisar novos Nós**: caso todos os itens não sejam identificados, o Crawler prossegue com a atividade de procurar novos Nós para serem acessados (Nós filhos do Nó atual) podendo ser uma URL ou um elemento HTML clicável. No processo são aplicadas validações aos novos elementos encontrados, verificando a duplicidade de elementos e se eles são relevantes para o critério buscado, evitando possíveis ruídos nas buscas como elementos que dão acesso a páginas de critérios semelhantes como entre Receita Orçamentária e Receita Extra-Orçamentária. Por fim, caso novos Nós

filhos não sejam encontrados e todos os Nós já tenham sido percorridos o processo de avaliação do critério é finalizado. É importante lembrar que os algoritmos de busca propostos neste trabalho não possuem o papel de buscar conteúdos ou páginas, mas em priorizar e acessar a árvore de Nós criada pelo Crawler;

Com base na análise do procedimento de busca e validação dos critérios fiscais é visto que existem dois caminhos possíveis para finalização do processo de *web crawling*: o esgotamento de todos os Nós considerados relevantes e a identificação completa de todos os itens do critério. Neste contexto, fica evidente que quanto mais rápido a ferramenta encontrar as áreas da árvore de Nós com páginas que possuem os itens fiscais procurados, mais rápido a avaliação do critério é terminada, constatando a enorme importância na escolha de algoritmos de busca que consigam mapear Nós que dão acesso a conteúdos relevantes. A Seção 4.4 discorre sobre os algoritmos usados neste trabalho e a motivação por trás de cada um deles.

4.4 Algoritmos de busca

No Web Crawler a maneira de selecionar e acessar os Nós está estritamente relacionado ao algoritmo de busca utilizado. Neste âmbito, como é visto na seção 4.3 deste documento, a forma de percorrer os Nós pode determinar a eficiência da ferramenta em encontrar com um menor custo todos os itens fiscais buscados, pois dependendo da estratégia de acesso aos Nós utilizada, é provável a ocorrência de casos onde os Nós com conteúdos relevantes sejam os últimos acessados.

Diante do contexto de portais de transparência de municípios da Paraíba, foram efetuadas análises empíricas pelo autor deste trabalho, que constataram que itens de um mesmo critério estão habitualmente concentrados em páginas próximas umas das outras e em níveis de profundidade entre 3 e 4 em média, considerando uma hierarquia de árvore. Por esta razão, foi optado pelo uso do algoritmo *BFS* como base de busca no Web Crawler, pois sua busca em largura proporciona o acesso mais rápido às páginas próximas que possuem baixos níveis de profundidade quando comparado a algoritmos como o *DFS*.

Na literatura, foi encontrada a técnica algorítmica *Epsilon-greedy* para a seleção de hosts com a finalidade de maximizem o número de páginas relevantes encontradas [16]. Por este ângulo, nesta pesquisa experimentamos adaptá-lo para a seleção de Nós de forma a viabilizar

o maior ganho em referência à localização de páginas com os itens fiscais buscados, partindo da premissa que páginas (Nós) que contém itens fiscais relevantes estão próximas umas das outras. Durante os experimentos foi testado o desempenho do Crawler com o algoritmo diante de um processo de *exploration* convencional e outra abordagem trocando a aleatoriedade padrão por uma busca em largura (BFS). Além disso, considerando os melhores parâmetros adotados pelo trabalho de Meusel, Mika e Blanco [16], foi adotado o valor de $\varepsilon = 0.1$ em nossos experimentos, representando 10% para *exploration* e 90% para *exploitation*.

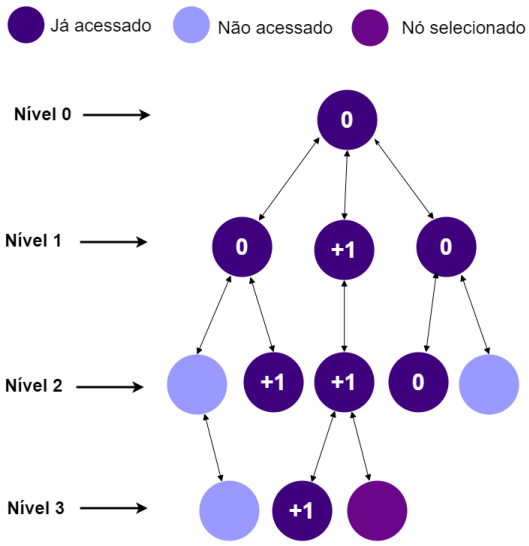

Figura 4.4: Seleção de um Nó com o algoritmo Epsilon-greedy baseado no ganho.

No que diz respeito ao correto funcionamento do *Epsilon-greedy*, é preciso atribuir ganhos aos Nós não acessados para que o algoritmo consiga escolher as opções que possivelmente darão acesso às páginas que possuem itens fiscais (exploitation). Nesta circunstância, para determinar o ganho de um Nó não acessado, é utilizado o seu histórico de vizinhança, identificando Nós pais e irmãos já acessados. Cada Nó classificado como relevante oferece *+1 de ganho* para o Nó não acessado próximo. Conforme é apresentado na Figura 4.4, o Nó não acessado com mais Nós relevantes (+1) na vizinhança é escolhido pelo algoritmo.

4.5 Diferenças entre a Turmalina

A versão do Web Crawler apresentada neste trabalho é parte do conhecimento adquirido pela Turmalina, mas foi reimplementada aproveitando todos os conceitos e metadados do projeto

original. Ainda assim, foram realizadas otimizações que visaram a redução de palavras chaves de busca, e consequentemente melhorias na eficiência durante as execuções, reduzindo o número de consultas realizadas. Ademais, a motivação principal em construir uma nova implementação do Web Crawler deu-se pela imprescindibilidade de uma solução mais escalável, baseando-se em uma ferramenta denominada Puppeteer[2], robusta a sites estáticos e dinâmicos que fazem o uso extremo de operações em Javascript, tendo em vista que a versão da Turmalina emprega os Frameworks Scrapy[3] e Selenium[4] para ser capaz de acessar ambos tipos de sites.

Nesse âmbito, a ferramenta de Web Crawler apresentada nesta pesquisa complementa a Turmalina com a adição de otimizações e a implementação e uso de novos algoritmos de busca *Epsilon-greedy*, *DFS* e *E-greedy + BFS* .

[2]https://github.com/puppeteer/puppeteer

[3]https://scrapy.org/

[4]https://selenium.dev/

Capítulo 5

Materiais e Método

Neste capítulo é descrito de forma detalhada os dados e métodos aplicados durante os experimentos.

5.1 Fontes de Dados

Nesta pesquisa foram produzidas e usadas específicas coleções de dados com a finalidade de validar o projeto de experimentos proposto. A seguir são apresentadas as estratégias adotadas e as descrições a respeito das bases de dados criadas.

5.1.1 Portais de Transparência Municipais da Paraíba

Diante dos trabalhos retratados neste estudo, é demonstrado que conhecer o ambiente ao qual o Web Crawler será inserido pode estar pontualmente relacionado a sua efetividade. Sob esta perspectiva, a equipe de desenvolvedores da Turmalina, analisou os portais de transparência dos municípios da Paraíba a fim de encontrar padrões em seus designs que pudesse classificá-los em grupos homogêneos, na tentativa de medir o quão parecidos eles eram.

Nesse âmbito, os 223 portais municipais foram individualmente categorizados por intermédio do reconhecimento da empresa fornecedora e da semelhança entre seus leiautes principais. Este processo ocasionou no descobrimento de 24 empresas atuantes no fornecimento de portais de transparência no estado (Apêndice B). Porém, durante esse processo também percebemos a presença de uma característica peculiar neste tipo de site: a aparição

de portais fiscais com múltiplas empresas gerenciando diferentes critérios, isto é, um mesmo site pode ter leiautes distintos no acesso a diferentes categorias de informação.

Como forma de considerar as mudanças entre os portais que possuíam uma ou mais empresas divergentes, foram criados grupos que representavam todas as combinações de fornecedores observadas nos portais paraibanos. Este processo deu origem a 49 grupos contendo até no máximo 3 empresas, entretanto, foi notado que apenas 15 delas possuíam mais de uma recorrência dentre todos portais fiscais.

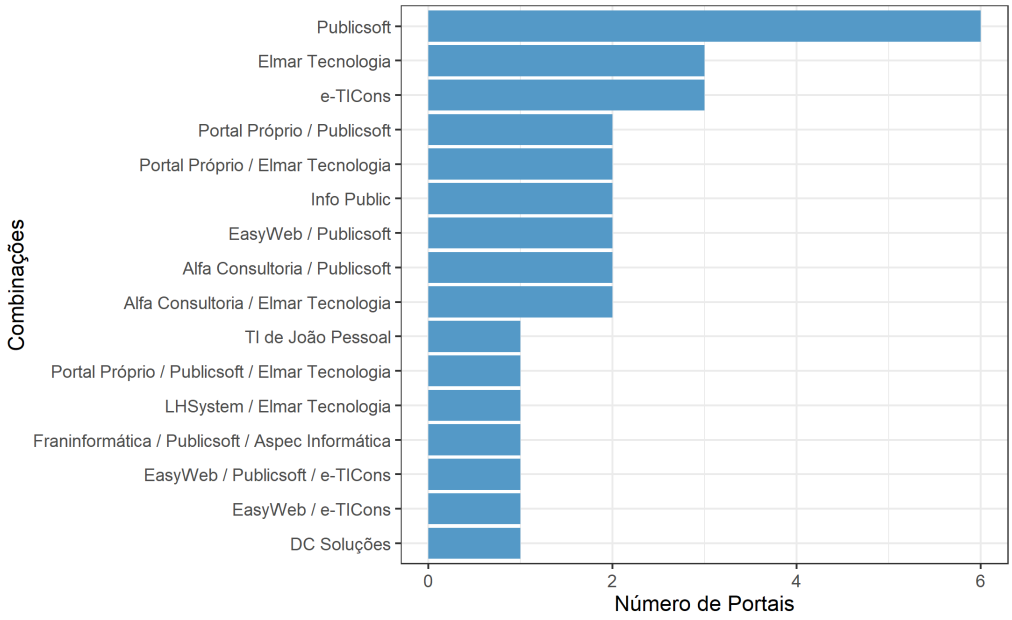

Figura 5.1: Distribuição da amostra de portais municipais na Paraiba por Combinação.

O estudo produzido pelo time da Turmalina serviu como apoio para a geração da amostra aplicada durante os experimentos deste trabalho. Diante deste cenário, embasando-se nas 15 combinações de portais mais utilizadas pelos municípios da Paraíba, foram selecionados aleatoriamente um mínimo de 10% dos portais de transparência de cada grupo, resultando em uma amostra com 30 portais, distribuída proporcionalmente no número de recorrência das combinações. A Figura 5.1 apresenta a divisão do número de portais da amostra por combinação. A lista completa dos municípios da amostra e suas combinações é exposto no Apêndice C.

5.1.2 Critérios e Itens fiscais avaliados

De acordo com o que foi mostrado na seção 2.3, o Índice de Transparência pública do TCE-PB é composto pela avaliação da presença de 123 itens nos portais de transparência de cada município e do estado. Nesta perspectiva, para o melhor entendimento da amostra utilizada, abaixo é apresentado a explanação de alguns conceitos importantes relacionados a granularidade das informações fiscais examinadas:

- **Item**: é compreendido como um atributo que agrega uma única informação fiscal. Este, quando agrupado em grau de semelhança com outros itens dão forma a um critério, por exemplo, os itens nome, salário, CPF e cargo dão origem ao critério quadro pessoal;

- **Critério**: é definido como a caracterização de um agrupamento de itens fiscais semelhantes, que representam uma categoria de transparência como despesa orçamentária, licitação, receita orçamentária e entre outros;

Em função da numerosa quantidade de itens e do custo para criar gabaritos manualmente descrevendo quais portais atendem cada um dos itens, optou-se pela sua redução durante os experimentos, priorizando critérios considerados essenciais no que tange a fiscalização das despesas e receitas nas gestões. Assim, foram utilizados 64 itens divididos entre os critérios despesa orçamentária, despesa extraorçamentária, receita orçamentária, receita extraorçamentária, licitação e quadro pessoal. No Apêndice A deste documento é apresentado todos os itens selecionados e seus respectivos critérios.

5.1.3 Gabaritos

Com o objetivo de medir a eficácia das avaliações de transparência feitas durante os experimentos, o autor desta dissertação desenvolveu gabaritos que registram a presença ou ausência dos 64 itens fiscais nas páginas web de cada portal de transparência entre os municípios contidos na amostra. A Tabela 5.1 apresenta um exemplo do registro do gabarito para os itens do critério Receita extraorçamentária no portal do município de Santa Rita.

Durante a elaboração dos gabaritos, cada um dos itens foi classificado manualmente, sendo atribuído à coluna *encontrado* o valor *TRUE*, caso a presença da informação no portal

município	critério	item	encontrado	local encontrado
Santa Rita	Despesa Extra Orc.	valor	TRUE	http://siteseticons...
Santa Rita	Despesa Extra Orc.	codigo	TRUE	http://siteseticons...
Santa Rita	Despesa Extra Orc.	nomenclatura	TRUE	http://siteseticons...

Tabela 5.1: Exemplo de um gabarito para itens de Despesa extraorçamentária.

fosse confirmada e *FALSE* caso contrário. Além disso, foi registrado o local de identificação de cada item, assegurando a corretude da informação durante a comparação entre o gabarito e os resultados do Web Crawler. Todos os gabaritos criados podem ser encontrados no *link*[1].

5.1.4 Avaliações

As estruturas de dados criadas para representar uma avaliação de transparência efetuada pelo Crawler foram pensadas de maneira similar aos gabaritos, pretendendo gravar dados úteis para avaliar a eficácia e eficiência. Para este fim, em cada avaliação é guardada as seguintes informações:

- **Município**: o nome do município responsável pelo portal de transparência;
- **Item**: objeto que agrega informações relacionadas a um único item fiscal, por exemplo, se ele foi encontrado, o texto identificado e onde foi localizado no site (link);
- **Critério**: objeto que representa um critério fiscal real, agregando itens relacionados;
- **Quantidade de Nós acessados**: compreende ao registro do número total de Nós (páginas, componentes clicáveis) acessados durante a execução da avaliação. Por meio desta métrica, podemos quantificar a eficiência do Crawler nas avaliações, pois quanto menor o número de Nós acessados menor será o consumo de recursos e o tempo de execução;

O pleno funcionamento do Web Crawler no contexto dos sites de transparência está fortemente relacionado a inserção de metadados envolvendo informações como a URL inicial do portal de transparência de cada município e palavras chaves de busca e de identificação.

[1]Gabaritos da amostra - https://bit.ly/2vsUqQ5

Os termos chaves de busca foram concebidos por uma avaliação empírica nos portais de transparência que congregou distintas terminologias que davam acesso a uma mesma área do site, por exemplo, *consultar Despesas* e *detalhamentos das despesas* são termos que refletem a busca pelo critério Despesa orçamentária. Da mesma maneira, os termos de identificação são um conjunto de nomenclaturas diferentes para identificar um mesmo item fiscal entre as páginas web do site. Um exemplo destas variações é o ocorrido entre os termos *codigo* e *cod. despesa* que condizem com a identificação do item código pertencente ao critério Despesa extraorçamentária.

As avaliações de transparência coletadas no decorrer dos experimentos foram executadas em tempo real com o acesso ao portal de transparência oficial de cada município. A escolha deu-se pelo dever da avaliação de transparência retratar precisamente o ambiente real considerando as adversidades como o carregamento de conteúdos dinâmicos, tempo de resposta a uma requisição, instabilidades e entre outros.

5.2 Métricas

A fim de quantificar o nível de eficácia da solução proposta neste trabalho com os diferentes algoritmos de busca testados, foi estabelecido o uso das métricas Recall e Precisão. Estas foram individualmente calculadas em cada avaliação, através de um confrontamento entre os resultados achados no gabarito do portal com a mesma avaliação realizada pelo Crawler. Para ilustrar isto, abaixo são mostrados as formas e definições como esta operação procedeu.

$$Recall = \frac{TP}{TP+FN} \tag{5.1}$$

$$Precisao = \frac{TP}{TP+FP} \tag{5.2}$$

- TP: são os verdadeiros positivos (true positive) que equivalem ao número de itens identificados corretamente como presentes e localizados em locais válidos pelo gabarito;

- FN: são os falsos negativos (false negative) que correspondem ao número de itens identificados incorretamente como ausentes na avaliação do Crawler.

- FP: são os falsos positivos (false positive) que representam o número de itens identificados incorretamente como presente durante avaliação do Crawler.

Avaliar essas métricas separadamente possibilita a extração de características próprias do contexto estudado, por exemplo, os valores da Precisão trazem uma noção da qualidade das avaliações de transparência, respaldando-se no nível de acertos para os itens em relação ao texto detectado e sua localização no site. Já o Recall tem a intenção de mensurar a capacidade do Web Crawler em identificar todos os itens presentes nos portais de transparência.

No que diz respeito à avaliação da eficiência do Web Crawler foi optado pelo uso da mediana do *número de Nós acessados* em cada avaliação, calculada a partir do número de Nós acessados durante em cada critério. Esta escolha foi motivada pelo o número de Nós acessados ser uma medida mais confiável, visto que as execuções foram realizadas em um ambiente não controlado e medidas como tempo de execução podem absorver ruídos relacionados a instabilidades da rede de internet usada.

5.3 Projeto dos Experimentos

Durante o desenvolvimento da pesquisa foram realizados três experimentos. O primeiro experimento avalia e compara novas formas de priorizar e acessar os Nós descobertos ao longo do processo de *web crawling*, visando tornar o Web Crawler mais acertivo no que tange a seleção de páginas relevantes e consequentemente mais eficiente em relação a redução do número de Nós acessados. Sob este contexto, implementamos e testamos a eficiência e eficácia do Crawler com o uso dos algoritmos de busca *Epsilon-greedy*, *BFS* e *DFS*.

Diante dos resultados do primeiro experimento, foi percebido a existência de itens que nunca são encontrados ou mesmo que possuem uma frequência baixa de aparições dentre todas as avaliações. Nestas circunstâncias, é essencial enfatizar que existem duas situações que sinalizam a conclusão da avaliação de transparência pelo Web Crawler, a identificação de todos os itens fiscais (melhor caso) ou o acesso a todos os Nós disponíveis (pior caso). Levando estes dois aspectos em consideração, é provável que a falta de identificação de alguns itens entre as avaliações está diretamente relacionado ao número de Nós acessados, afetando no resultado final das técnicas algorítmicas de busca.

Dessa maneira, surgiram novas perguntas para entender e avaliar o desempenho da ferramenta proposta em situações mais específicas. Assim, o segundo experimento altera dois pontos do primeiro para ajudar a identificar se os resultados do primeiro acontecem devido a características de alguns dos itens avaliados nos portais. Para isso, nosso segundo experimento descarta itens que foram identificados com uma frequência inferior a 12 aparições dentre todas as avaliações do primeiro experimento, restando 48 itens avaliados. Além disso, nesse experimento focamos nos dois algoritmos que tiveram melhor desempenho no primeiro experimento (*Epsilon-greedy* e *BFS*) e em uma combinação entre ambos. Esta combinação refere-se ao uso do algoritmo *Epsilon-greedy* com uma alteração em sua aleatoriedade padrão durante a etapa de *exploration* por uma busca em largura (BFS), dado que o *BFS* obteve bons resultados no primeiro experimento, e portanto deverá selecionar os Nós de um maneira mais eficiente em comparação a uma seleção aleatória.

O terceiro experimento foi elaborado com a finalidade de testar a eficácia e a eficiência do Web Crawler com o algoritmo de busca o BFS, porém usando termos de busca extraídos do próprio nome do critério em lugar da lista de palavras-chave criada manualmente pelo time da Turmalina durante seu desenvolvimento. Por exemplo, para procurar páginas referentes à despesa orçamentária seriam construídas consultas com os termos: *despesa orcamentaria*, *despesa* e *orcamentaria*.

5.4 Infraestrutura e Ambiente de Execução

A ferramenta de Web Crawler apresentada neste trabalho, foi implementada sobre a linguagem de programação Javascript utilizando como ambiente de execução o *Node.js*[2] na versão 12.14. Para acessar e percorrer as páginas dos sites foi aplicado a biblioteca *Puppeteer* [3] 1.7, permitindo automatização destas operações. Já o gravamento dos dados foi feito por um modelo de dados construído na ferramenta *Mongoose* versão 5.7.6, adotando o banco de dados *MongoDB*[4]. As execuções ocorreram em um computador com um processador core i7-7700HQ com 16GB de memória Ram usando um sistema operacional Linux. O código completo do projeto pode ser encontrado em *https://github.com/joserafael97/auditor-crawler*.

[2]https://nodejs.org/

[3]https://github.com/puppeteer/puppeteer

[4]https://www.mongodb.com/

Capítulo 6

Resultados

Este capítulo mostra os resultados alcançados durante os experimentos propostos neste trabalho. A seção 6.1 exibe os resultados sobre o uso da técnica algorítmica *Epsilon-greedy* em comparação com os algoritmos *BFS* e *DFS*. Já a seção 6.2 expõe os resultados sobre o segundo experimento, retratando a eficácia e eficiência de uma combinação do algoritmo *Epsilon-greedy* com o *BFS* em confrontamento com ambos algoritmos separadamente. Por fim, a seção 6.4 aponta os resultados obtidos com testes realizados com o Web Crawler utilizando o algoritmo *BFS* com redução dos termos chaves de busca a termos extraídos do nome dos critérios fiscais buscados.

Com o objetivo de tentar validar todos os resultados referentes a eficácia e eficiência dos algoritmos em nossa amostra, foram aplicados intervalos de confiança calculados sobre *bootstraps* com 4 mil reamostras com um nível de confiança de 95%. O método utilizado para computar o intervalo a partir dos *bootstraps* foi o *BCa*.

6.1 Avaliação dos algoritmos Epsilon-greedy, BFS e DFS

Neste experimento avaliamos a eficácia e eficiência dos algoritmos Epsilon-greedy, BFS e DFS. Para quantificação dos resultados foram coletadas 261 avaliações divididas entre os 29 portais de transparência presentes em nossa amostra, sendo desconsideradas as avaliações do município de Curral de Cima que estava com seu portal fora do ar no momento das execuções. Neste enquadramento, para cada site existe proporcionalmente três avaliações diferentes para cada algoritmo testado.

6.1.1 Recall entre os algoritmos

Inicialmente comparamos os valores do Recall das avaliações entre cada um dos algoritmos. Diante do que é mostrado na Figura 6.1, é possível observar uma semelhança entre as distribuições do Recall nas avaliações para cada um dos algoritmos, estando a maior parte das avaliações entre as faixas de valores de 0.7 até 1. Além disso, comparando as medianas entre os algoritmos é quase imperceptível a diferença entre os métodos de busca, variando entre 0.90 (*Epsilon-greedy*) e 0.89 (*DFS* e *BFS*).

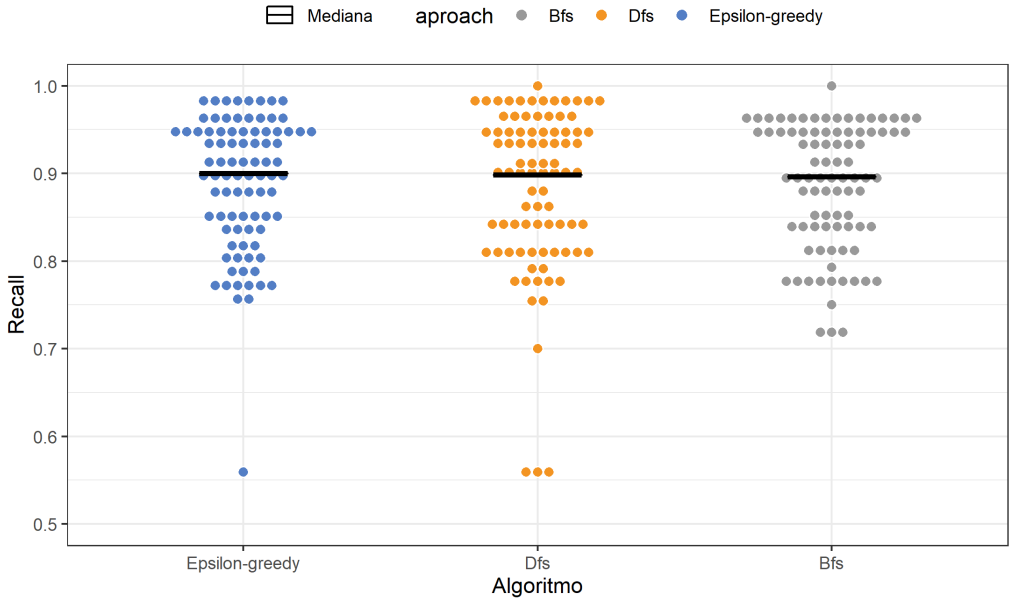

Figura 6.1: Distribuição dos valores de Recall das avaliações entre os algoritmos de busca.

Apesar disso, é notado a existência de quatro outliers situados abaixo dessas faixas de valores, localizados nos resultados obtidos com os algoritmos *Epsilon-greedy* e *DFS*. Dada esta observação, foi conferido individualmente cada um deles, verificando que todos representam avaliações pertencentes ao portal de transparência do município de Remígio[1]. Com algumas novas execuções neste site, foi visto que diante do uso dos algoritmos *Epsilon-greedy* e *DFS*, o Web Crawler não consegue lidar com as sessões temporárias dentro de *iframes* de forma efetiva, devido o acesso não sequencial aos Nós, no uso de *iframes* embutidos nas páginas deste portal fiscal.

[1]https://www.remigio.pb.gov.br/portal/transparencia-fiscal

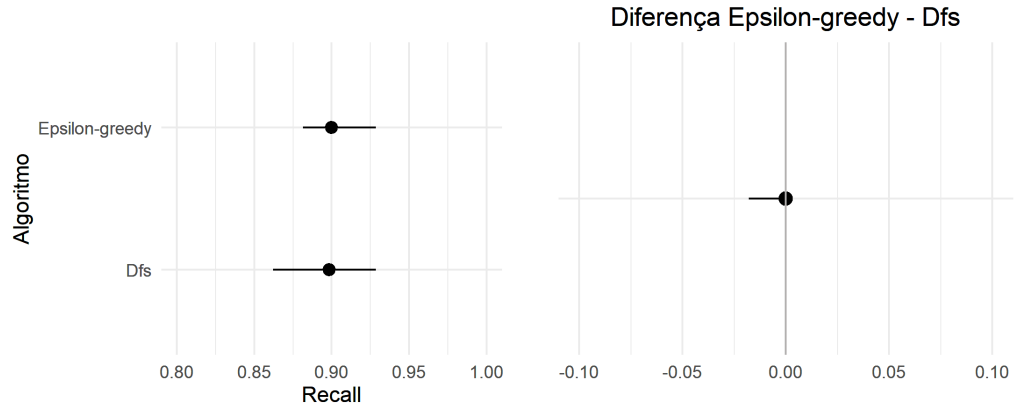

Figura 6.2: Intervalo de confiança da diferença entre a mediana do Recall para o *Epsilon-greedy* e *DFS*.

Considerando nossa amostra e estimando o valor da mediana do Recall na população de avaliações de transparência, estabelecemos como é exibido na Figura 6.2 que a mediana do Recall estaria situada entre 0.86 até 0.92 no *DFS* e 0.88 até 0.92 no *Epsilon-greedy*. Já no que está relacionado a diferença nos algoritmos, as barras de erros do intervalo de confiança variam de -0.017 até 0. O fato de o intervalo englobar valores negativos expressa que é aceitável a existência de uma vantagem de até 0.017 na mediana do Recall no emprego do algoritmo *DFS* em conferência ao *Epsilon-greedy*. Ainda assim, é capaz que esta diferença se mantenha na maior do tempo próxima ao zero, podendo ser desprezada.

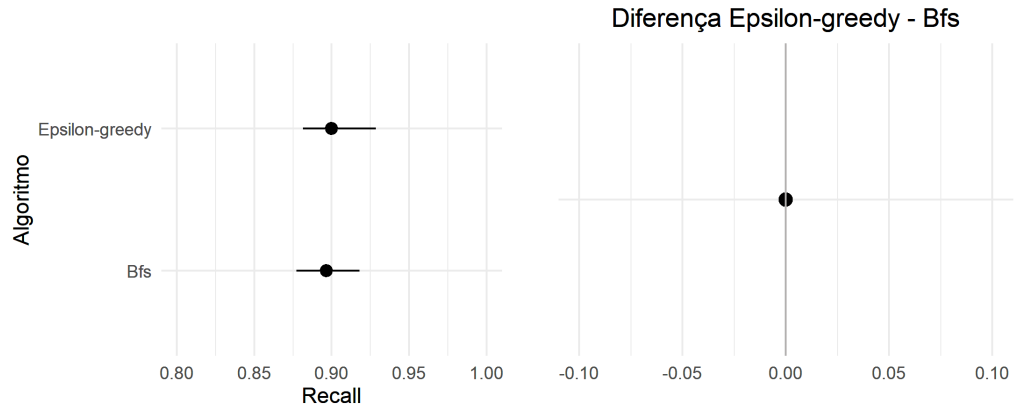

Figura 6.3: Intervalo de confiança da diferença entre a mediana do Recall para o *Epsilon-greedy* e *BFS*.

O intervalo de confiança da diferença entre os algoritmos *Epsilon-greedy* e *BFS* ostentado

na Figura 6.3, expõe os mesmos limites inferior e superior com o valor -0.001, consistindo em uma minúscula ascendência do algoritmo *BFS* no comparativo com o *Epsilon-greedy*, podendo ser desconsiderada no que tange a população. Ademais, é identificado que as medianas do Recall apresentadas nos intervalos do *BFS* estariam contidas entre 0.87 e 0.91 na população.

6.1.2 Precisão entre os algoritmos

Com base na Figura 6.4, atesta-se que as faixas de valores da Precisão das avaliações entre os algoritmos são similares, variando entre 0.85 até 1 e apresentando medianas nos valores de 0.95 no *Epsilon-greedy* e *DFS* e de 0.96 no *BFS*, ocorrendo uma diferença de cerca de 0.01 entre elas.

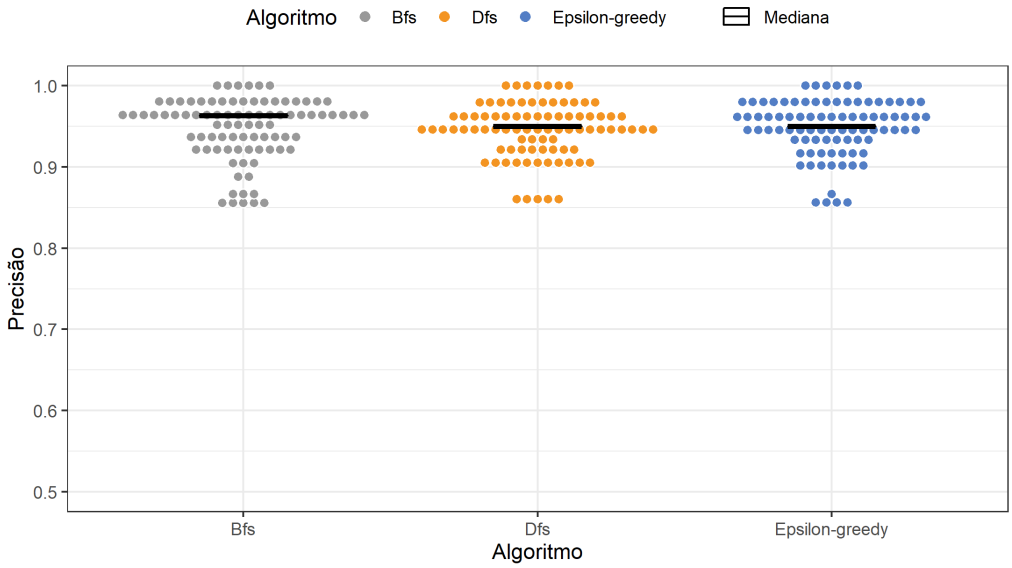

Figura 6.4: Distribuição dos valores de Precisão das avaliações entre os algoritmos de busca.

Embasado pelo intervalo da diferença entre os algoritmos *BFS* e *DFS* que varia de -0.003 até -0.001 incluído na Figura 6.5, é plausível que há uma supremacia no valor da mediana da Precisão no *DFS* em comparação ao *BFS*. Todavia esta distinção é insignificante no nosso cenário, visto que se trata de valores vizinhos ao zero. No mais, é visto que os intervalos da mediana da Precisão estão contidos na população com 95% de confiança entre os valores de 0.94 até 0.95 no *DFS* e de 0.94 até 0.96 no *BFS*.

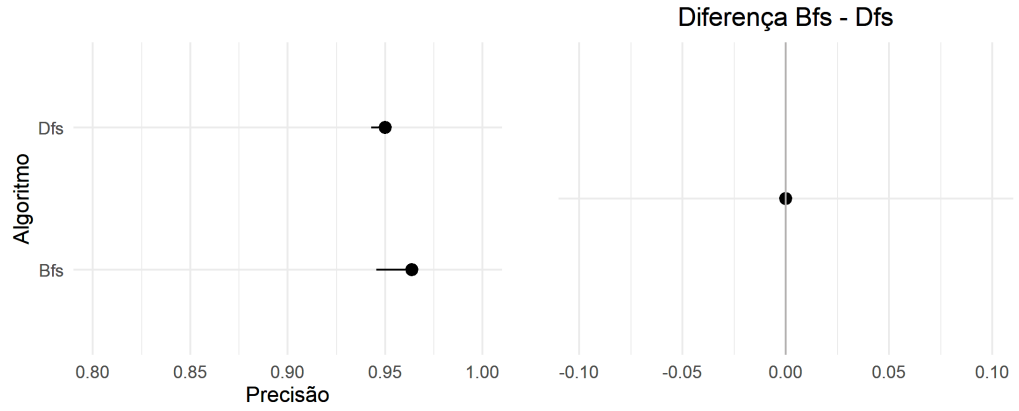

Figura 6.5: Intervalo de confiança da diferença entre a mediana da Precisão para o *BFS* e *DFS*.

Ao analisar a amostra é presenciado que o algoritmo de busca *BFS* possui a melhor mediana da Precisão. Neste sentido, como é exposto na Figura 6.6, comparamos por intermédio do intervalo de confiança a diferença das medianas da Precisão entre o *BFS* e o *Epsilon-greedy*, resultando em um intervalo com os limites inferior e superior iguais, indicando o valor de -0.001. Esta situação sugere uma hegemonia de 0.001 para a mediana da Precisão no algoritmo *Epsilon-greedy* em um confronto com o *BFS*. Entretanto, a diferença é irrisória e pode ser desconsiderada.

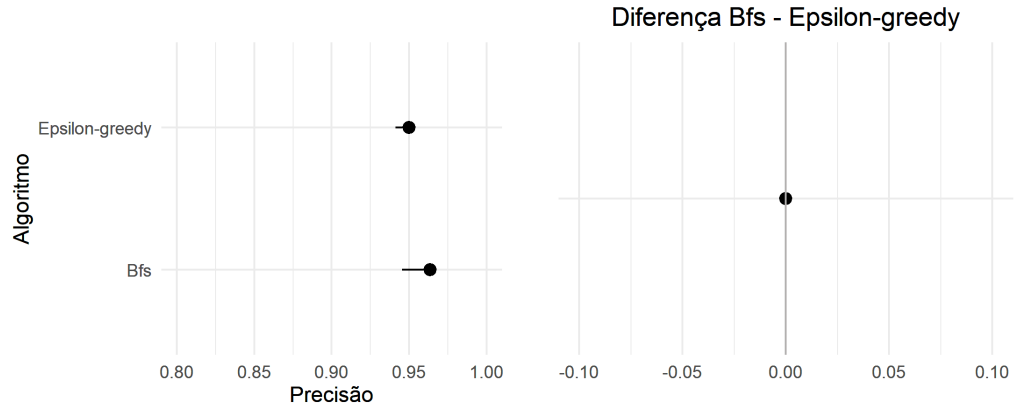

Figura 6.6: Intervalo de confiança da diferença entre a mediana da Precisão para o *BFS* e *Epsilon-greedy*.

A partir do intervalo de confiança da diferença variando de -0.016 até 0 para os algoritmos *DFS* e *Epsilon-greedy* mostrado na 6.7, é plausível que exista uma melhora na mediana

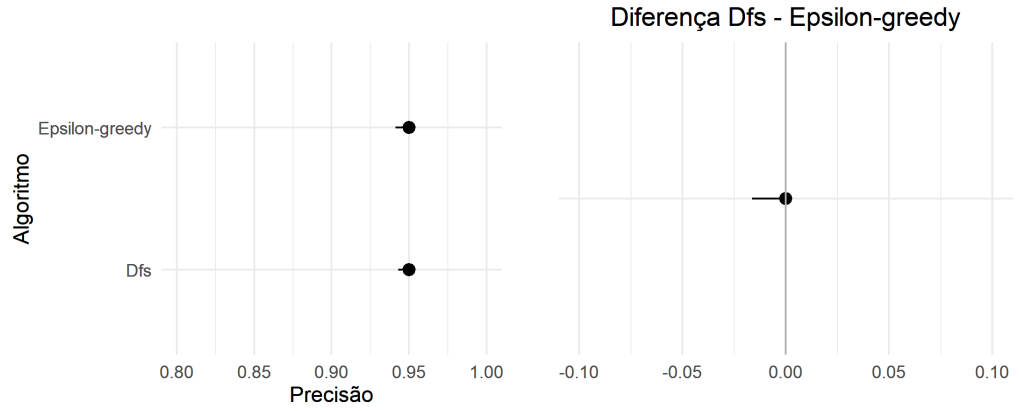

Figura 6.7: Intervalo de confiança da diferença entre a mediana do Recall para o *BFS* e *Epsilon-greedy*.

da Precisão de no máximo 0.016 na adoção do *Epsilon-greedy* como o algoritmo de busca do Web Crawler no comparativo com o *DFS*. Não obstante, esta diferença é relativamente pequena considerando o universo ao qual a população está contida e é provável que na maioria dos casos fique próxima ao zero.

6.1.3 Análise dos resultados das métricas Recall e Precisão

As modestas diferenças identificadas nos resultados entre os algoritmos refletem a ocorrência de instabilidades no link de internet utilizado, no site de transparência ou no processamento das avaliações durante as execuções entre os distintos algoritmos, dado que todas as avaliações foram realizadas online e em tempo real, com o acesso às URLs oficiais dos portais de transparência.

As semelhanças nos resultados dessas métricas entre os três algoritmos testados estão relacionadas inalterabilidade dos termos de busca e identificação, e de todo o processo desempenhado pelo o Web Crawler em gerar e aplicar consultas via Xpath no HTML. Desta forma, o algoritmo não interfere no comportamento da ferramenta em relação aos critérios utilizados para descobrir novos Nós a serem acessados, mas na maneira de selecionar e priorizar cada um deles.

6.1.4 Número de Nós acessados

Para avaliar a eficiência do Web Crawler focado no uso dos diferentes algoritmos de busca, foi empregado a mediana do número de Nós acessados entre os critérios fiscais pesquisados em cada avaliação. Esta métrica foi escolhida em virtude de sua maior estabilidade em relação a possíveis instabilidades ocorridas durante as avaliações, pois independentemente da duração da avaliação, o número de Nós acessados deve permanecer similar tendo em vista um mesmo portal. Neste cenário, a Figura 6.8 mostra a distribuição da mediana do número de Nós acessados entre os critérios fiscais nas avaliações entre os três algoritmos testados.

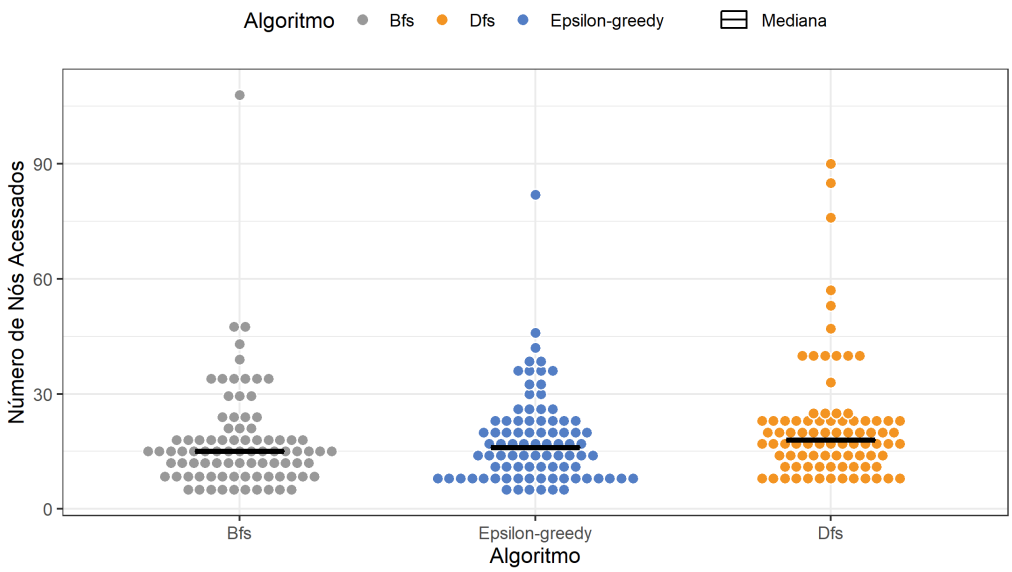

Figura 6.8: Distribuição dos valores da mediana do número de Nós acessados das avaliações entre os diferentes algoritmos busca.

Diante desse quadro, a maioria das avaliações possuem a mediana do número de Nós acessados entre as faixas de valores de 4 até 60 Nós. No entanto, existem algumas exceções como as avaliações de Itabaiana para os algoritmos *Epsilon-greedy* e *DFS*, que apresentam os valores das medianas de Nós acessados entre as faixas de 76 até 90 Nós. Sob este contexto, é detectado também algumas medianas de Nós acessados pertencentes às avaliações de Bayeux nos algoritmos *DFS* e *BFS* que destoam da fração principal de suas avaliações, com a mediana de Nós acessados em respectivamente 68 e 108 Nós. Isto pode ter ocorrido devido a oscilações na internet ou no próprio site de transparência que possam ter afetado tais

avaliações. Esta afirmação embasa-se na não ocorrência deste efeito nas demais avaliações para o mesmo município.

No nosso contexto quanto menor a mediana do número de Nós acessados dos critérios procurados, melhor será a eficiência do algoritmo, pois significa dizer que o Web Crawler precisou acessar menos Nós para ter acesso às informações buscadas. Neste sentido, é apurado que o *BFS* foi o algoritmo de busca que denota a menor mediana de Nós acessados, com um valor de 14 Nós acessados seguido pelos algoritmos *Epsilon-greedy* e *DFS* com respectivamente 16 e 18 Nós. Como usamos a mediana do número de Nós de todos os critérios em cada avaliação uma diferença de 2 Nós entre os algoritmos é considerada representativa.

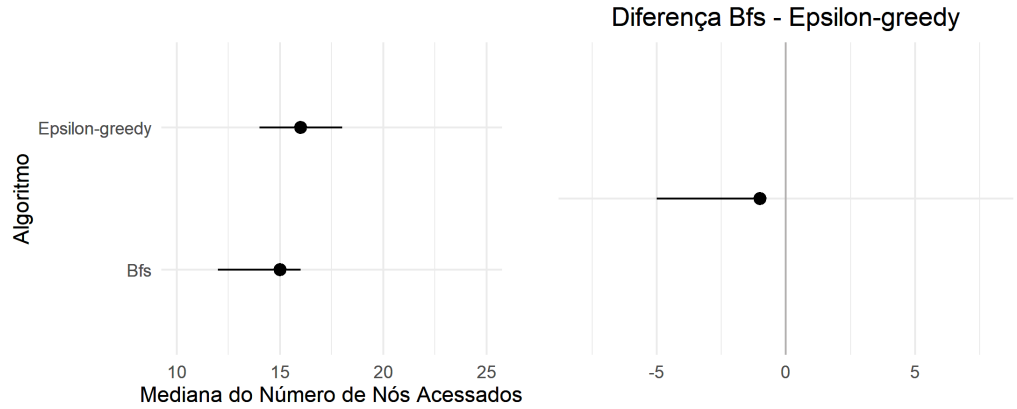

Figura 6.9: Intervalo de confiança da diferença entre a mediana do número de Nós acessados para o *BFS* e *Epsilon-greedy*.

Diante de uma possível superioridade do algoritmo de busca *BFS*, é estimado pelo intervalo de confiança, apresentado na Figura 6.9, que os seus valores para a mediana do número de Nós na população podem estar contidos entre 12 e 16 Nós. Ademais, por meio do intervalo de confiança para *Epsilon-greedy*, é observado que a mediana do número de Nós acessados deve estar contido na população entre os valores de 14 até 19 Nós.

Com relação a diferença entre ambos os algoritmos, mediante ao intervalo de confiança da diferença que exibiu a faixa de valores entre -5 e -1, é constatado que pode haver uma distinção positiva de 1 até 5 Nós na mediana do número de Nós acessados nas avaliações no *BFS* quando comparado ao *DFS*. Contudo, em média esta diferença tende a ser de 1 Nó.

Na Figura 6.10 mostramos o intervalo de confiança da diferença entre *Epsilon-greedy* menos o *DFS*, nele está contido a faixa de valores de -8 até -1. Este resultado evidencia uma

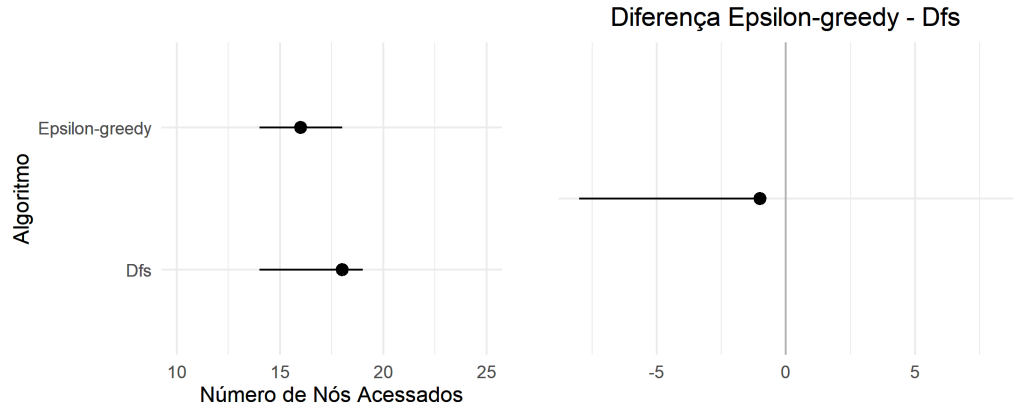

Figura 6.10: Intervalo de confiança da diferença entre a mediana do número de Nós acessados para o *Epsilon-greedy DFS*.

diferença positiva na diminuição de até 8 Nós sobre a mediana do número de Nós acessados no uso do algoritmo *Epsilon-greedy* em comparação ao *DFS*. Porém é provável que em grande parte das situações a diferença esteja situada em 1 Nó. .

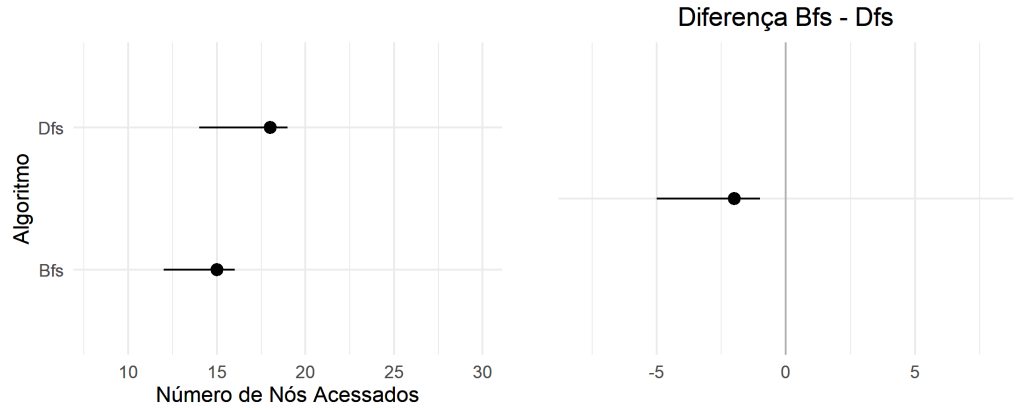

Figura 6.11: Intervalo de confiança da diferença entre a mediana do número de Nós acessados para o *BFS DFS*.

Como foi percebido, é provável que o Crawler em alguns cenários usando o algoritmo *BFS* se sobressaia em relação ao *Epsilon-greedy* no que diz respeito a uma menor mediana do número de Nós acessados. Neste sentido, resolvemos também comparar com base em um intervalo de confiança da diferença os algoritmos *BFS* e *DFS*. De acordo com o que é exposto na Figura 6.11, o intervalo ficou entre -5 e -1, revelando que é plausível que o algoritmo *BFS* tenha uma mediana de Nós acessados menor, havendo uma propensão de se identificar na

população uma diferença de 2 Nós entre eles.

Os resultados dos intervalos de confiança são análogos ao exposto na amostra, indicando uma superioridade do algoritmo *BFS* quanto a mediana do número de Nós acessados e portanto na eficiência durante as execuções das avaliações de transparência. Além disso, foi comprovado que o algoritmo de busca *DFS* não é a melhor opção para o contexto de sites de transparência, considerando um ambiente que não contém Nós relevantes em níveis profundos.

6.2 Avaliação dos algoritmos E-greedy + BFS, BFS e Epsilon-greedy

Neste experimento adaptamos o algoritmo *Epsilon-greedy*, alterando sua aleatoriedade padrão durante a seleção de Nós na etapa de *exploration* por uma seleção baseada no *BFS*, denominada neste trabalho como *E-greedy + BFS*. O seu desempenho em referência a eficácia e eficiência foi comparado com os mesmos algoritmos separadamente (*Epsilon-greedy* e *BFS*). Além disso, desconsideramos o uso do algoritmos DFS, pois sua mediana do número de nós acessados entre os critérios fiscais das avaliações apresentou valores inferiores em comparação aos demais algoritmos. Para quantificação dos resultados foram avaliados 47 itens fiscais sobre 174 avaliações divididas entre os 29 portais de transparência da amostra, com proporcionalmente duas avaliações diferentes para cada algoritmo testado.

Diante da mudança no número de itens fiscais avaliados, foi novamente analisado os valores de Recall e Precisão das avaliações de transparência entre os algoritmos de busca propostos neste experimento. Contudo, é importante destacar que os resultados vistos no experimento anterior apontaram que pode não haver distinção entre as métricas relacionadas a eficiência, pois a principal variação entre algoritmos está na maneira de priorizar a ordem de seleção dos Nós.

6.2.1 Recall entre os algoritmos

Conforme é exposto na Figura 6.12 os valores de Recall estão na maior parte concentrados entre 0.75 e 1 entre os três algoritmos, apresentando medianas entre 0.94 no *Epsilon-greedy*

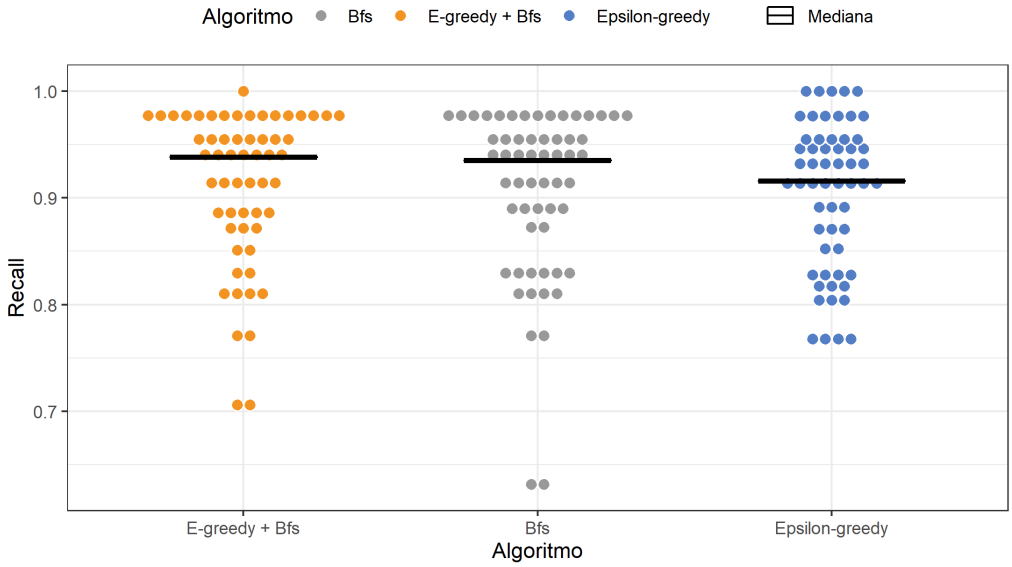

Figura 6.12: Distribuição dos valores de Recall entre os diferentes algoritmos.

e 0.97 no *BFS* e *E-greedy + BFS*. Além disso, são encontradas quatro avaliações fora destas faixas de valores, sendo duas com um Recall de 0.70 no algoritmo *E-greedy + BFS* e as outras duas com um Recall de 0.63 no *BFS*.

Para tentar entender o motivo dessa discrepância, foi feita uma análise individual entre essas avaliações, descobrindo que todos os casos checados tratam-se do portal de transparência do município de Itabaiana. A nível de critério fiscal, em ambos os experimentos, despesa orçamentária e quadro pessoal, comportam a maior proporção de itens não identificados no site de transparência do município.

Por meio dessas informações, acessamos manualmente o site e verificamos que apesar da página principal ser provida pela empresa HC multimídia, ambos os critérios destacados são gerenciados por uma mesma empresa, a Elmar Tecnologia. Ao acessar o ambiente tanto do critério de despesa orçamentária, quanto de quadro pessoal, nos deparamos com um carregamento consideravelmente lento da página, o que provavelmente teve efeito direto nos resultados obtidos na avaliação, causando possíveis timeouts na ferramenta durante o procedimento de *web crawling*.

Como forma quantificar a lentidão no carregamento do site, adotamos a ferramenta do Google o *Lighthouse*[2], que avalia os sites da internet baseando-se em critérios como per-

[2]https://developers.google.com/web/tools/lighthouse?hl=pt-br

formance, acessibilidade e boas práticas de desenvolvimento. Os resultados reportados pela ferramenta atribuíram ao site um nível de performance 1 em uma escala que varia de 1 até 100, mais detalhadamente, o site demora cerca de 19 segundos para estar completamente pronto para interação e 15.6 segundos para exibir o primeiro componente da página. O Apêndice D mostra o resumo do relatório gerado.

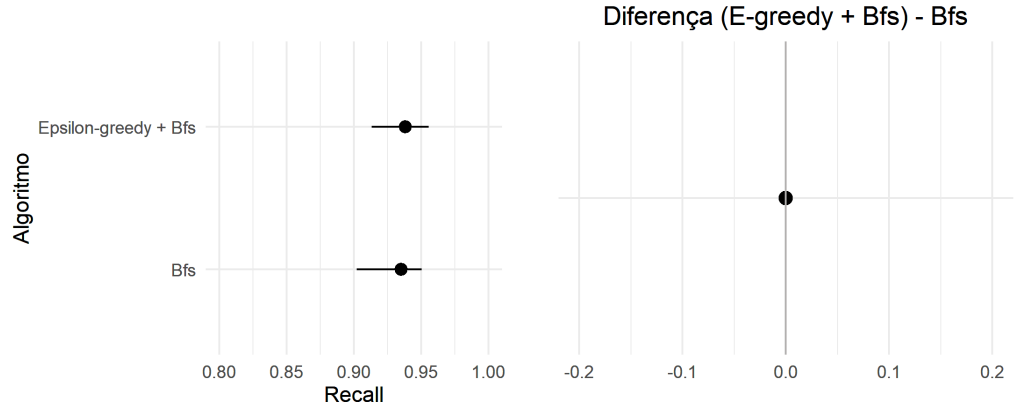

Figura 6.13: Intervalo de confiança da diferença entre a mediana do Recall para o *E-greedy + BFS* e *BFS*.

A princípio conferimos os intervalos de confiança dos dois algoritmos de busca com os maiores valores da mediana para o Recall na amostra, o *E-greedy + BFS* e *BFS*. Na Figura 6.13 são exibidos os intervalos de 0.91 até 0.95 no *E-greedy + BFS* e 0.90 até 0.95 para o *BFS*, podendo refletir com 95% de confiança as medianas encontradas na população de avaliações de transparência. Em outra perspectiva, tendo em vista o intervalo da diferença entre eles com os limites superior e inferior iguais a zero, não é possível detectar distinções diante desta métrica.

No mesmo cenário, a Figura 6.14 mostra o intervalo do algoritmo *Epsilon-greedy*, estando concentrado em referência a população entre 0.89 e 0.93. Ademais, também é exibido o intervalo da diferença entre os algoritmos *E-greedy + BFS* e *Epsilon-greedy*, assumindo o valor de -0.011 para ambos os limites do intervalo. Através deste intervalo, há evidência que pode ocorrer uma superioridade no uso do *Epsilon-greedy* de 0.011 em comparação ao *E-greedy + BFS* no tocante ao valor da mediana do Recall resultante. Entretanto, esta diferença é pequena e pode assumir um valor próximo ao zero.

Com a redução dos termos de busca, foi investigado novamente a presença de qualquer

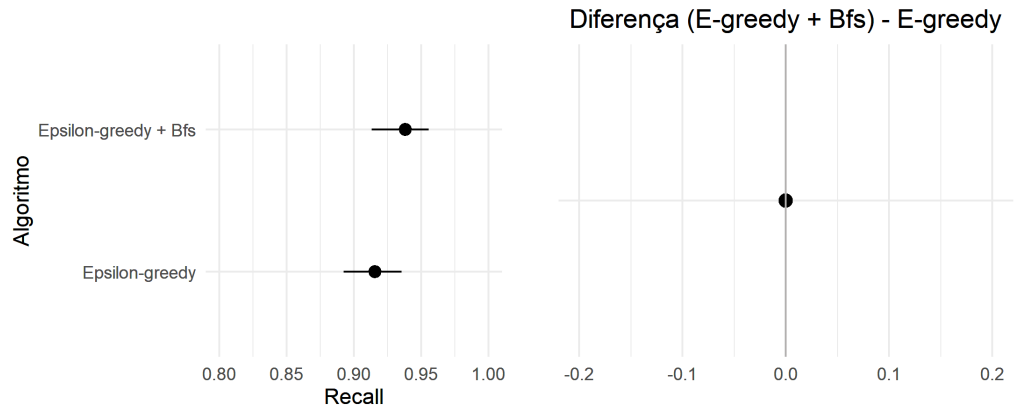

Figura 6.14: Intervalo de confiança da diferença entre a mediana do Recall para o *E-greedy + BFS* e *Epsilon-greedy*.

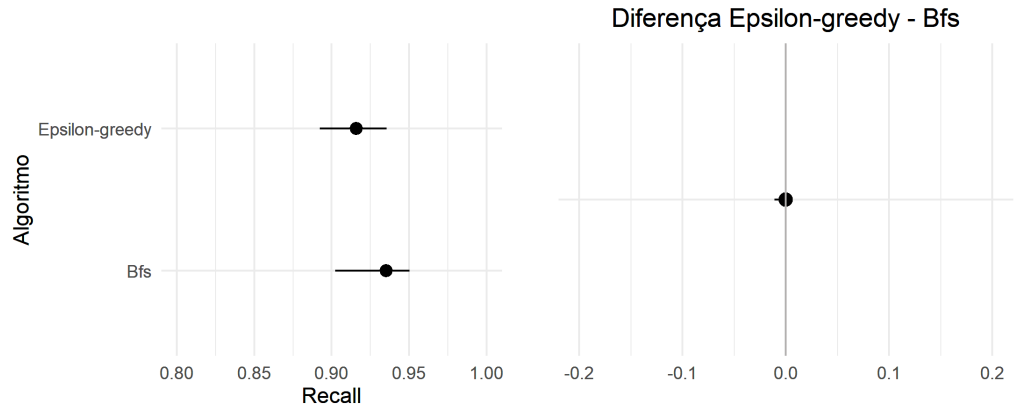

Figura 6.15: Intervalo de confiança da diferença entre a mediana do Recall para o *Epsilon-greedy* e *BFS*.

distinção entre *Epsilon-greedy* e *BFS* quanto a seus valores das medianas do Recall. Assim, de acordo com o intervalo da diferença entre *Epsilon-greedy* e *BFS* presente na Figura 6.15, constata-se uma vantagem do algoritmo *BFS* em referência aos valores das medianas do Recall, ficando entre -0.011 até 0, existindo uma tendência de grande parte dos casos a diferença ficar em valores próximos ao zero.

6.2.2 Precisão entre os algoritmos

Em conformidade com a Figura 6.16, é possível observar que os valores da Precisão entre os distintos algoritmos de busca estão em sua maioria nos intervalos de 0.85 até 1. Já as

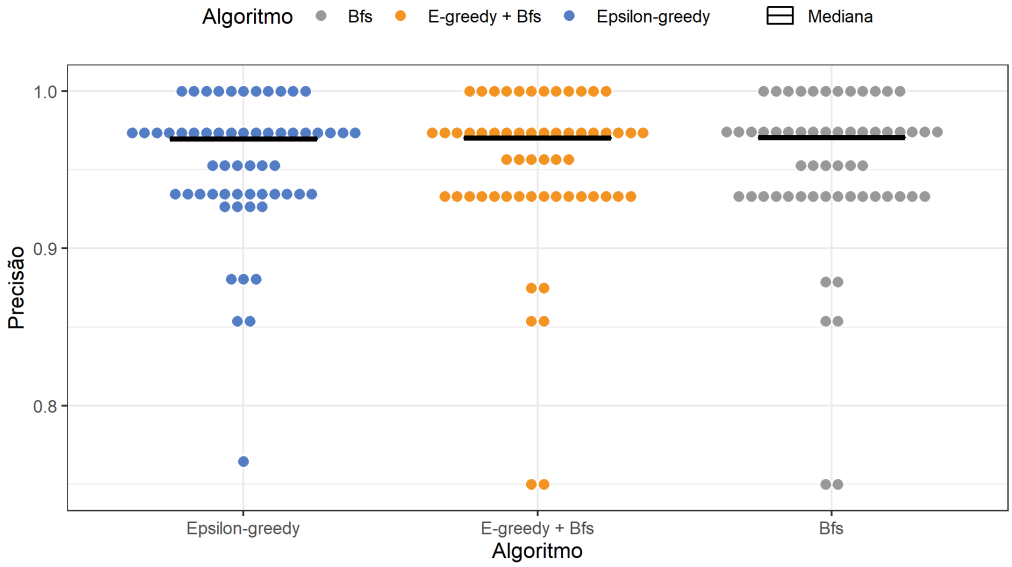

Figura 6.16: Distribuição dos valores de Precisão entre os diferentes algoritmos.

medianas da Precisão estão entre 0.96 no *Epsilon-greedy* e 0.97 no *BFS* e *E-greedy + BFS*. Além disso, da mesma maneira do que acontece nos valores do Recall, existem 5 avaliações relacionados a portal de fiscal de Itabaiana que possuem valores da Precisão abaixo de 0.85.

Os intervalos de confiança presentes na Figura 6.17 são compostos por valores entre 0.93 até 0.97 nos algoritmos *Epsilon-greedy* e *E-greedy + BFS*. Já em referência ao intervalo da diferença entre eles, podemos identificar os limites inferior e superior iguais a 0, não sendo possível confirmar a presença de uma vantagem no uso do algoritmo *E-greedy + BFS* quanto a mediana da Precisão nas avaliações.

Comparando os algoritmos *E-greedy + BFS* e *BFS* através do intervalo de confiança da diferença, presente na Figura 6.18, não é comprovada a existência de qualquer superioridade entre eles em relação as medianas da Precisão, uma vez que o intervalo assume uma relação de equilíbrio entre eles, englobando somente o valor zero.

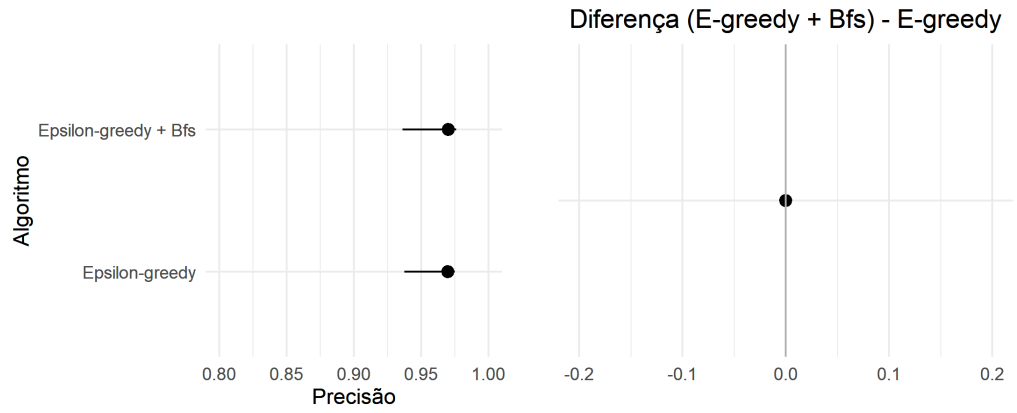

Figura 6.17: Intervalo de confiança da diferença entre a mediana da Precisão para o *E-greedy + BFS* e *Epsilon-greedy*.

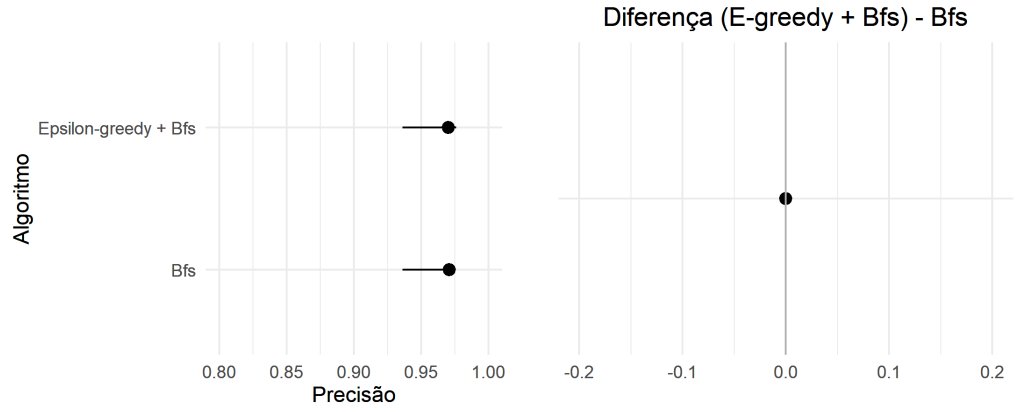

Figura 6.18: Intervalo de confiança da diferença entre a mediana da Precisão para o *E-greedy + BFS* e *BFS*.

De acordo com o intervalo de confiança da diferença, averiguamos a aparição de uma distinção entre os algoritmos *Epsilon-greedy* e *BFS*, avaliando se a redução do número de itens fiscais buscados pode ter interferido nos valores das medianas da Precisão entre eles. Por meio da Figura 6.19, encontramos apenas o valor zero no intervalo, o que aponta somente um possível equilíbrio entre os dois. Isto pode confirmar que a diferença encontrada entre eles no primeiro experimento está relacionada a instabilidades durante as execuções das avaliações.

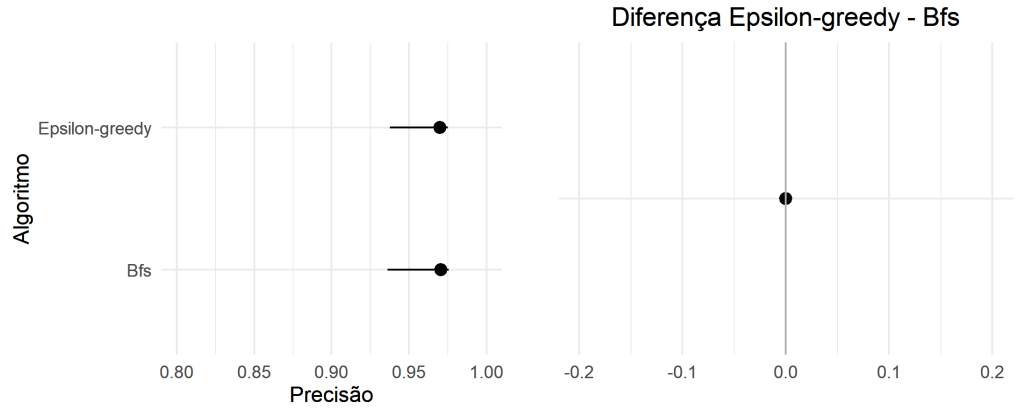

Figura 6.19: Intervalo de confiança da diferença entre a mediana da Precisão para o *Epsilon-greedy* e *BFS*.

6.2.3 Análise dos resultados das métricas Recall e Precisão

Com base nos resultados retratados na amostra para os valores de Recall durante o segundo experimento, foi comprovado uma melhora nos valores de sua mediana variando entre os algoritmos em 0.94 até 0.97 contra 0.89 até 0.90 presentes no primeiro experimento. Este resultado é esperado devido a retirada dos itens fiscais com uma frequência de aparição inferior a 12 vezes. Contudo, na comparação da mediana da Precisão entre os experimentos esta diferença é discreta, ficando entre 0.95 e 0.96 no primeiro experimento e entre 0.96 e 0.97 neste experimento.

De modo similar ao primeiro experimento, os intervalos de confiança da diferença apontaram diferenças pequenas, podendo ser desprezadas no que tange as avaliações de transparência. Diante desta simetria nas medianas de Recall e Precisão entre ambos os experimentos, é possível confirmar a hipótese criada no primeiro experimento que expõe a ideia de não haver diferenças na eficácia com as mudanças dos algoritmos de busca, pois a maneira de identificar novos links e componentes clicáveis não é alterada neste processo, isto é, entre os distintos algoritmos busca serão encontrados os mesmos Nós, porém a ordem que estes serão acessados mudará de abordagem para outra.

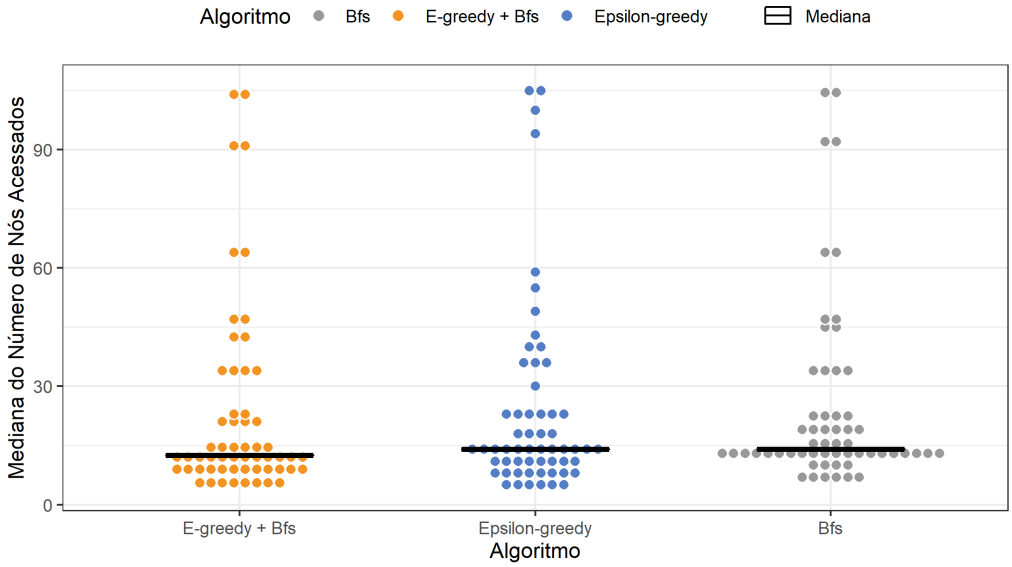

Figura 6.20: Distribuição dos valores da mediana do número de Nós acessados das avaliações entre os diferentes algoritmos busca.

6.2.4 Número de Nós acessados

Com relação a mediana dos Nós acessados durante o experimento dois, podemos identificar através da Figura 6.20 uma semelhança na mediana de Nós acessados entre os três algoritmos, estando fixos entre os valores de 4 até 106 Nós. Já em referência a mediana central de cada abordagem algorítmica, o *E-greedy + BFS* mostrou um melhor resultado com 12.5 Nós seguidos pelo *Epsilon-greedy* e o *BFS* ambos com 14 Nós.

Ao explorar as avaliações com a mediana do número de Nós acessados superior a 90 Nós, identificamos que estas pertencem aos portais de transparência dos municípios de Bom Jesus e Arara, com a maioria dos seus critérios sendo gerenciados por uma mesma empresa. Surpreendentemente ao examinar as avaliações realizadas no primeiro experimento, notamos que em ambos os municípios as medianas do número de Nós atingiram no máximo 27 Nós. Esta mudança no número de Nós acessados está relacionado a modificações nos portais de transparência que fez o Web Crawler localizar um maior número de Nós classificados como relevantes, dado que as avaliações entre os experimentos foram executadas em momentos distintos. Contudo, é importante enfatizar que as avaliações não sofreram alterações em relação a eficácia, mantendo níveis de Recall e Precisão proporcionais.

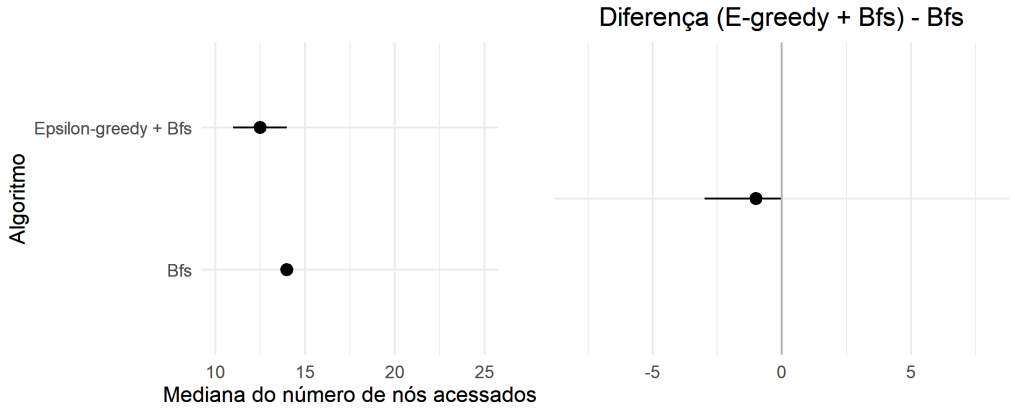

Figura 6.21: Intervalo de confiança da diferença entre a mediana do número de Nós acessados nos algoritmos *E-greedy + BFS* e *BFS*.

Baseando-se na amostra é perceptível que o *E-greedy + BFS* apresenta uma redução de cerca de 1.5 na mediana do número de Nós acessados em comparação as demais técnicas algorítmicas testadas. Assim, apoiando-se na amostra, calculamos o valor de sua mediana do Número de Nós acessados para a população de avaliações de portais de transparência. Na Figura 6.21, é verificado que a mediana para os dados da população pode assumir valores entre 11 até 14 Nós.

Ao aplicar intervalos de confiança para verificar a faixas de valores que incluem a mediana do número de Nós acessados para o *BFS* na população, percebemos que o intervalo possui como limites inferior e superior o valor de 12 Nós. No mais, a Figura 6.21 ostenta também o intervalo da diferença para mediana do número de Nós acessados entre *E-greedy + BFS* e o *BFS*, resultando em faixas de valores de -3 até 0. Este intervalo indica que é plausível que uso do *E-greedy + BFS* durante as avaliações apresente uma mediana do número de Nós acessados de até 3 Nós a menos em um confronto com o *BFS*. Apesar disto, pode existir uma tendência desta diferença ficar em 1 Nó ou ainda que ela não exista.

Na comparação entre os algoritmos *E-greedy + BFS* e *Epsilon-greedy*, a Figura 6.22 exibe o intervalo de confiança da diferença que varia entre -2.5 até -1. Com base nestes valores, é provável que exista uma superioridade do *E-greedy + BFS* em relação a mediana do número de Nós acessados quando comparado com o *Epsilon-greedy*, podendo apresentar uma melhora de até menos 2 Nós na mediana ou mais provavelmente de menos 1 Nó.

Por fim, diante do intervalo da diferença -4.5 até -1.5 entre os algoritmos *BFS* e *Epsilon-*

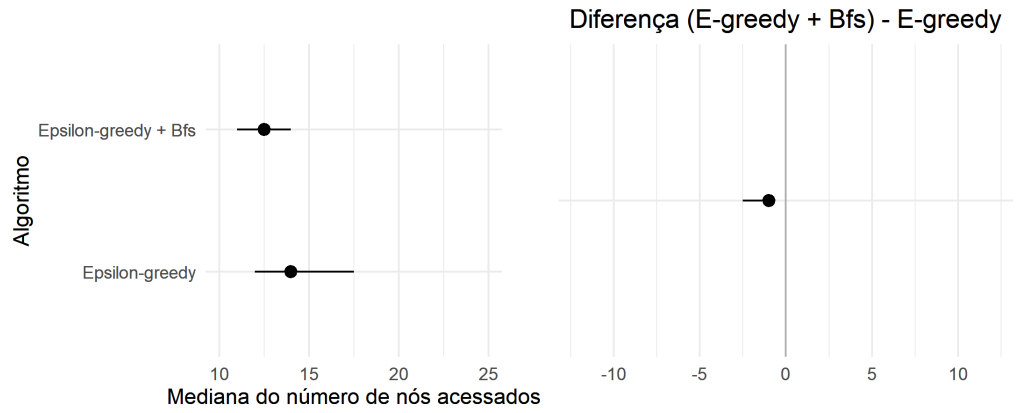

Figura 6.22: Intervalo de confiança da diferença entre a mediana do número de Nós acessados nos algoritmos *E-greedy + BFS* e *Epsilon-greedy*.

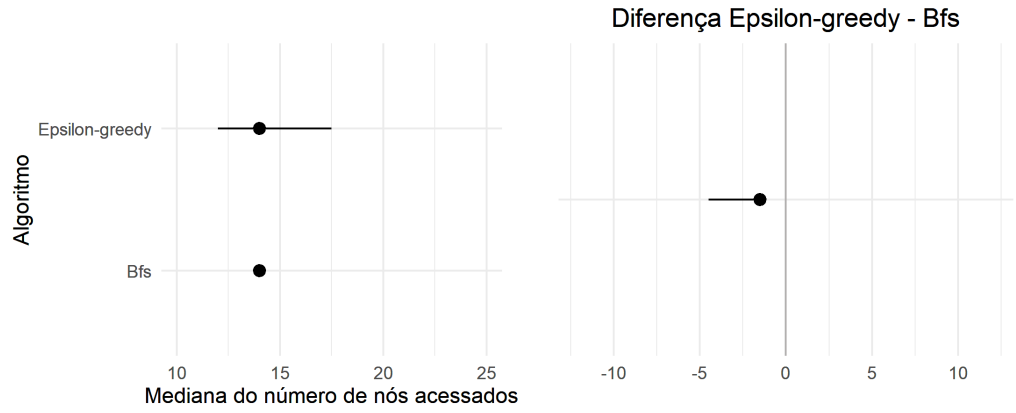

Figura 6.23: Intervalo de confiança da diferença entre a mediana do número de Nós acessados nos algoritmos *BFS* e *Epsilon-greedy*.

greedy presente na Figura 6.23, é plausível que o algoritmo *Epsilon-greedy* tenha um desempenho melhor com a diminuição de até 4.5 no valor da mediana do número de Nós em um confronto com o *BFS*. No entanto, é plausível que na maioria das situações esta diferença fique em torno de 1.5. Este resultado mostrou-se o inverso do experimento um, onde a BFS se sobressaiu nesta comparação. Sob este âmbito, diante dos critérios que determinam a conclusão da avaliação de transparência pelo Crawler, é visto que o *Epsilon-greedy* é recomendado na priorização de Nós em sites de transparência com o maior número de critérios fiscais completos, ou seja, com todos os itens procurados sendo identificados.

A adaptação proposta neste experimento avalia a utilização do algoritmo com os melhores resultados no primeiro experimento o *BFS* para substituir a aleatoriedade do algoritmo

Epsilon-greedy em momentos que os Nós acessados ainda não resultaram em ganho, isto é, não foram acessadas páginas ou componentes clicáveis que contenham itens fiscais identificados. Neste contexto, por meio dos resultados retratados pela amostra e os intervalos de confiança, é plausível que o algoritmo de busca *E-greedy + BFS* proporcione a melhor eficiência com relação a obtenção do menor número de Nós acessados entre os algoritmos testados, consequentemente propiciando um melhor aproveitamento dos recursos e uma diminuição do tempo gasto durante cada avaliação.

6.3 Considerações Finais dos Experimentos 01 e 02

Diante dos resultados obtidos com o primeiro experimento, foi identificado uma superioridade do algoritmo *BFS* em comparação aos algoritmos *Epsilon-greedy* e *DFS*. Em suma, é plausível que o *BFS* possua o valor da mediana do número de Nós acessados para os critérios fiscais avaliados inferior em no mínimo 1 Nó. Isto resulta em uma melhor eficiência em relação aos demais algoritmos de busca, pois reflete a existência de uma redução de 1 Nó na mediana do número de Nós acessados considerando todos os critério avaliados.

O segundo experimento avaliou novamente os algoritmos *BFS* e *Epsilon-greedy*, verificando o desempenho das etapas *exploration* e *exploitation* do algoritmo *Epsilon-greedy* em relação a avaliação de critérios com itens fiscais mais presentes nos portais de transparência. Assim, analisando os resultados foi possível comprovar um melhor aproveitamento dos ganhos no *Epsilon-greedy* diante do cenário com itens mais acertivos, ocorrendo um superação em termos de eficiência quando comparado ao algoritmo *BFS* com uma redução média 1.5 na mediana do número de Nós acessados. Isto evidência que *Epsilon-greedy* consegue um melhor desempenho em ambientes com o maior número de itens fiscais identificados, onde algoritmo tem a possibilidade usufruir dos ganhos.

Como os portais de transparência não possuem sempre o melhor cenário para o *Epsilon-greedy*, foi implementada uma versão adaptada do mesmo, que troca o modo de busca aleatório na etapa *exploration* por uma busca em largura (*BFS*). Este apresentou os melhores resultados quando comparado ao *Epsilon-greedy* e o *BFS*, com uma mediana do número de Nós acessados de 12.5 e uma diminuição da mediana em média de 1 Nós.

6.4 Redução do número de termos chaves de busca

Nos experimentos anteriores foram usados termos de busca identificados pela equipe desenvolvimento do projeto Turmalina. Neste contexto, a nível de critério o número de termos diferentes adotados variam entre 8 até 16 termos chaves para serem buscados entre páginas acessados, objetivando reconhecer novos links ou componentes clicáveis relevantes, baseando-se no critério procurado. Porém, quanto mais termos de busca o critério utiliza mais consultas nas páginas são realizadas e provavelmente mais páginas são acessadas. Ainda assim, isto não garante sua efetividade, pois a relevância do termo pode variar de um portal para outro.

Neste experimento, avaliamos a restrição dos termos de busca para palavras extraídas do nome do critério. Por exemplo, para procurar páginas relacionadas a quadro pessoal, usamos os termos: quadro, pessoal e quadro pessoal. Neste âmbito, foi empregue o Web Crawler com o algoritmo de busca BFS, sendo avaliado igualmente ao experimento dois 47 itens fiscais, resultando na coletada 58 avaliações divididas entre os 29 portais de transparência contidos na amostra, com duas avaliações para cada site.

6.4.1 Precisão e Recall

Ao analisar o contexto de procurar páginas que possuem conteúdos relevantes no que tange a avaliação de transparência, é compreendido que o impacto da redução de termos de busca está diretamente relacionado ao valor do Recall, pois pode afetar a capacidade de Crawler encontrar todos os itens buscados (cobertura). Neste sentido, a Figura 6.24 mostra que o intervalo com os valores do Recall entre as avaliações aumentou, estando concentrados entre as faixas de 0.55 até 0.97.

Comparando com os resultados alcançados no experimento dois que utilizou a mesma quantidade de itens fiscais, é constatado perdas nos valores do Recall nas avaliações, ocorrendo um atenuamento do valor da mediana de 0.97 para 0.91 considerando o mesmo algoritmo de busca.

Com relação a Precisão é notável que as faixas de valores das avaliações estão localizadas entre 0.73 e 1. Novamente é percebido uma diminuição de 0.02 no valor da mediana entre os dois experimentos indo de 0.97 para 0.95.

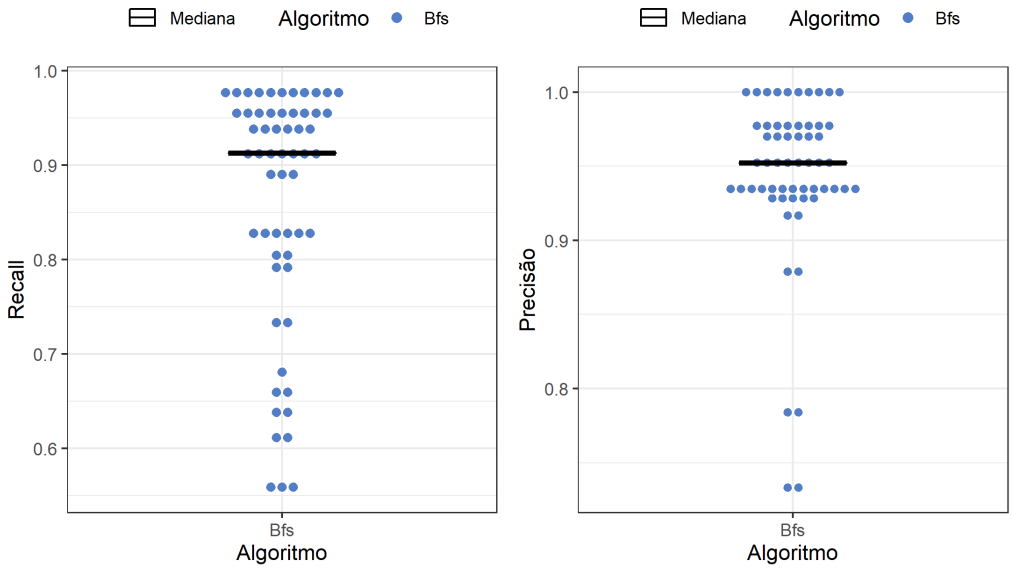

Figura 6.24: Distribuição dos valores de Recall entre os diferentes algoritmos.

Apesar da atenuação dos valores de Recall e Precisão em relação a suas medianas, é importante enfatizar que o custo de extrair os termos de busca baseando-se no nome do critério é mínimo em comparação a elicitação manual de uma lista de termos de busca para os distintos portais de transparência e critérios fiscais. Além disso, é importante destacar que independentemente da redução nas métricas os resultados expuseram níveis satisfatórios de eficácia em referência a capacidade da ferramenta identificar os itens nos lugares corretos em um processo automatizado.

6.4.2 Mediana do número de Nós acessados

Em concordância com a Figura 6.25, é observado que a mediana do número de Nós acessado está entre 1 até 166 Nós com a mediana central em 16 Nós. Estes resultados apontam uma desigualdade nas faixas de valores para o número de Nós acessados em comparação ao segundo experimento. Porém, quanto ao valor da mediana central, é visto que ambos os experimentos resultaram no número de 16 Nós.

O aumento do número de Nós acessados deve corresponder a adoção de termos mais genéricos em comparação aos pré-definidos e usados nos demais experimentos, por exemplo, os termos despesa e orcamentaria extraídos do critério fiscal despesa orçamentária são capa-

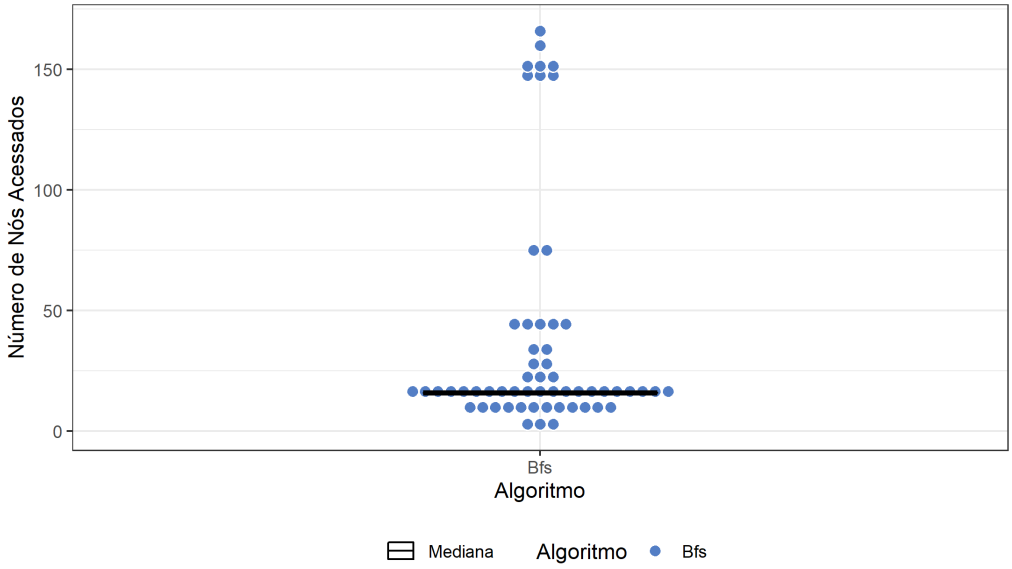

Figura 6.25: Distribuição dos valores de Recall entre os diferentes algoritmos.

zes de classificarem como relevantes páginas do critério despesa extraorçamentária, dado que estes termos estão contidos em ambos os critérios. No entanto, levando em consideração as medianas iguais, é notável que este efeito não está presente na maioria das avaliações. Além disso, no que diz respeito ao custo de produção dos termos de busca, a solução proposta neste experimento leva vantagem quando confrontado com o segundo experimento.

6.5 Considerações Finais

Com base nos resultados do terceiro experimento é identificado a possibilidade da utilização de termos de busca retirados do próprio nome do critério, reduzindo o custo de elicitar novos termos em portais de transparência distintos. Porém, este procedimento resulta em uma diminuição eficiência e eficácia em alguns portais de transparência mais específicos, isto é, sites onde os termos chaves contidos no nome do critério fiscal buscado não propiciam o acesso às páginas que possuem os itens procurados.

Capítulo 7

Conclusões

7.1 Discussão

O objetivo deste trabalho foi investigar o contexto e propor melhorias na avaliação de portais de transparência municipais da Paraíba, visando promover o controle social com um acompanhamento mais frequente das gestões públicas municipais e incentivando o desenvolvimento de sites fiscais mais acessíveis e inclusivos.

No contexto da ferramenta, buscamos no estado da arte soluções de Crawlers focados aplicados a outras situações, mas que poderiam ser úteis no aprimoramento do processo de seleção e priorização de páginas relevantes no cenário dos portais de transparência municipais. Diante desta perspectiva, consideramos experimentar o algoritmo *Epsilon-greedy* tendo em vista o aumento da eficiência na ferramenta com o aperfeiçoamento da seleção de Nós contendo links ou componentes clicáveis que proporcionam o acesso às páginas web que possuem itens fiscais avaliados.

A efetividade do algoritmo no ambiente de transparência fiscal pública foi determinada através de sua comparação com o nosso algoritmo de busca base *BFS*, e outro algoritmo tradicional no cenário de Crawlers, o *DFS*. Apoiado sobre este cenário, foi possível inferir que não existe uma diferença clara entre os algoritmos no que tange a eficácia das avaliações. Entretanto, no que diz respeito a eficiência o BFS foi algoritmo melhor avaliado.

Ao perceber que os critérios determinantes para a conclusão da avaliação de transparência estão respaldados sobre a identificação de todos os itens fiscais ou o esgotamento de todos Nós acessíveis, avaliamos novamente os dois melhores algoritmos resultantes do primeiro

experimento com a remoção de itens fiscais com baixos níveis de aparições dentre todas as avaliações. Além disso, foi proposto o uso de uma nova abordagem com uma adaptação do *Epsilon-greedy*, denominada neste trabalho como *E-greedy + BFS*.

Diante dos resultados foi possível inferir uma superioridade entre o algoritmo Epsilon-greedy em confrontamento com o BFS em relação ao número de Nós acessados, confirmando que em cenários onde o portal de transparência possui todos os itens fiscais disponíveis, isto é, que o Crawler consegue atingir o critério de parada *'encontrou todos os itens'* o algoritmo consegue uma melhor eficiência. No entanto, entre todos as técnicas algorítmicas testadas o melhor desempenho no tocante a eficiência foi com o uso do *E-greedy + BFS*.

Apoiados sobre os dois experimentos confirmamos a hipótese que o uso de um algoritmo de busca diferente não altera a eficácia do Web Crawler em referência as avaliações de transparência realizadas, visto que o processo de descoberta de novos Nós utiliza os mesmos termos de busca e identificação. Sob este quadro, um terceiro experimento foi proposto com o propósito de verificar o comportamento do Web Crawler em referência novamente a eficácia e eficiência em um restringimento de palavras de busca a inclusivamente termos retirados do próprio nome do critério avaliado.

A partir dessas comparações do Web Crawler focado no uso de diferentes algoritmos de busca, podemos inferir também sua efetividade em relação a processo de automatização da avaliação de transparência. Além disso, baseado nos testes realizados durante o terceiro experimento, apontamos novos caminhos para a diminuição do trabalho extra na manutenção de termos de busca.

7.2 Limitações

Compreendemos que os resultados trazidos por esta pesquisa são um ponto de partida para o desenvolvimento de instrumentos eficientes e eficazes na avaliação de transparência, e de modo geral na fiscalização da legislação pública vigente, no que corresponde a garantia de acesso à informação a qualquer público.

Contudo, é importante destacar que a ferramenta apesar de conseguir bons resultados durante as avaliações, não descarta a necessidade de auditores fiscais validarem as informações identificadas, a fim de manter uma alta confiabilidade nas avaliações dos portais de

transparência municipais.

Com as execuções dos portais de transparência sendo realizadas online em tempo real, não é possível controlar a ocorrências de instabilidades que possam ter afetado as execuções durante os experimentos. Porém, neste trabalho tentamos mitigar este risco baseando-se na coleta de diferentes execuções para um mesmo portal de transparência, e removendo dos dados avaliações que englobam reduções drásticas nas métricas em um mesmo site.

7.3 Trabalhos Futuros

Ainda considerando o contexto de portais de transparência podem ser desenvolvidas a partir do presente estudo, pesquisas relacionadas a: experimentação de novas técnicas de extração e identificação de páginas relevantes, a caracterização dos ambientes de transparência fiscal e a criação de Frameworks para auxiliar na implementação de portais de transparência eficazes e eficientes perante métricas relacionadas a performance, acessibilidade e boas práticas de desenvolvimento.

Bibliografia

[1] Lei complementar nº 131, de 27 de maio de 2009.

[2] Lei nº 12.527, de 18 de novembro de 2011.

[3] Paulo Ricardo Zilio Abdala and Carlos Marcos Souza de Oliveira Torres. A transparência como espetáculo: uma análise dos portais de transparência de estados brasileiros. *Administração Pública e Gestão Social. APGS. 8. Viçosa, Universidade Federal de Viçosa-PPGAdm-, 2016. p. 136-200*, 2016.

[4] Rui Cai, Jiang-Ming Yang, Wei Lai, Yida Wang, and Lei Zhang. irobot: An intelligent crawler for web forums. In *Proceedings of the 17th international conference on World Wide Web*, pages 447–456. ACM, 2008.

[5] Thomas H Cormen, Charles E Leiserson, Ronald L Rivest, and Clifford Stein. *Introduction to algorithms second edition*. 2nd ed. MIT press, 2001. Bibliografia: p. 531–532.

[6] Cláudia Ferreira Cruz, Aracéli Cristina de Souza Ferreira, Lino Martins da Silva, and Marcelo Álvaro da Silva Macedo. Transparência da gestão pública municipal: um estudo a partir dos portais eletrônicos dos maiores municípios brasileiros. *Revista de Administração Pública*, 46(1):153–176, 2012.

[7] Walisson da Costa Resende and Mônica Erichsen Nassif. Aplicação da lei de acesso à informação em portais de transparência governamentais brasileiros. *Encontros Bibli: revista eletrônica de biblioteconomia e ciência da informação*, 20(42):1–16, 2015.

[8] Yajun Du, Wenjun Liu, Xianjing Lv, and Guoli Peng. An improved focused crawler

based on semantic similarity vector space model. *Applied Soft Computing*, 36:392–407, 2015.

[9] Tim Furche, Georg Gottlob, Giovanni Grasso, Christian Schallhart, and Andrew Sellers. Oxpath: A language for scalable data extraction, automation, and crawling on the deep web. *The VLDB Journal—The International Journal on Very Large Data Bases*, 22(1):47–72, 2013.

[10] Thamme Gowda and Chris A Mattmann. Clustering web pages based on structure and style similarity (application paper). In *2016 IEEE 17th International conference on information reuse and integration (IRI)*, pages 175–180. IEEE, 2016.

[11] Patrícia Adriani Hoch, Lucas Martins Rigui, and Rosane Leal da Silva. Desafios à concretização da transparência ativa na internet, à luz da lei de acesso à informação pública: análise dos portais dos tribunais regionais federais. *Revista direitos emergentes na sociedade global*, 1(2):257–286, 2012.

[12] Jingtian Jiang, Xinying Song, Nenghai Yu, and Chin-Yew Lin. Focus: learning to crawl web forums. *IEEE Transactions on Knowledge and Data Engineering*, 25(6):1293–1306, 2012.

[13] Dexter C Kozen. *The design and analysis of algorithms*. Springer Science & Business Media, 1992. Bibliografia: p. 19–21.

[14] Jae-Gil Lee, Donghwan Bae, Sansung Kim, Jungeun Kim, and Mun Yong Yi. An effective approach to enhancing a focused crawler using google. *The Journal of Supercomputing*, pages 1–18, 2019.

[15] Nyalle Barboza Matos, Maurício Corrêa da Silva, José Dionísio Gomes da Silva, and Lincoln Moraes de Souza. Avaliação de portais de transparência dos 30 municípios mais populosos da região nordeste. *Registro Contábil*, 4(2):17–35, 2013.

[16] Robert Meusel, Peter Mika, and Roi Blanco. Focused crawling for structured data. In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management*, pages 1039–1048. ACM, 2014.

[17] Adi Omari, Sharon Shoham, and Eran Yahav. Cross-supervised synthesis of web-crawlers. In *Proceedings of the 38th International Conference on Software Engineering*, pages 368–379. ACM, 2016.

[18] Davi de Castro Reis, Paulo Braz Golgher, Altigran Soares Silva, and AlbertoF Laender. Automatic web news extraction using tree edit distance. In *Proceedings of the 13th international conference on World Wide Web*, pages 502–511. ACM, 2004.

[19] Mohit Sewak. *Deep Reinforcement Learning: Frontiers of Artificial Intelligence*. Springer, 2019. Bibliografia: p. 60–61.

[20] Roberto Panerai Velloso and Carina F Dorneles. Extracting records from the web using a signal processing approach. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, pages 197–206, 2017.

[21] Márcio LA Vidal, Altigran S da Silva, Edleno S de Moura, and João Cavalcanti. Structure-driven crawler generation by example. In *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 292–299. ACM, 2006.

[22] John White. *Bandit algorithms for website optimization*. "O'Reilly Media, Inc.", 2012. Bibliografia: p. 11–13.

[23] Gong-Qing Wu, Lei Li, Li Li, and Xindong Wu. Web news extraction via tag path feature fusion using ds theory. *Journal of Computer Science and Technology*, 31(4):661–672, 2016.

[24] Shanchan Wu, Jerry Liu, and Jian Fan. Automatic web content extraction by combination of learning and grouping. In *Proceedings of the 24th international conference on World Wide Web*, pages 1264–1274, 2015.

Apêndice A

Itens Avaliados pelo Web Crawler

A Tabela abaixo expõe todos os itens com seus respectivos critérios aos quais foram considerados nos experimentos realizados por esta pesquisa.

	Critério	Item
1	Despesa Orçamentária	acao
2	Despesa Orçamentária	categoria economica
3	Despesa Orçamentária	elemento
4	Despesa Orçamentária	emp cnpj cpf
5	Despesa Orçamentária	emp data
6	Despesa Orçamentária	emp indicacao licitacao
7	Despesa Orçamentária	emp numero
8	Despesa Orçamentária	emp valor
9	Despesa Orçamentária	funcao
10	Despesa Orçamentária	lic data pag ultima
11	Despesa Orçamentária	lic modalidade
12	Despesa Orçamentária	lic num contrato
13	Despesa Orçamentária	lic numero
14	Despesa Orçamentária	lic obj servico
15	Despesa Orçamentária	modalidade
16	Despesa Orçamentária	natureza despesa
17	Despesa Orçamentária	orgao or uni orcamentaria

18	Despesa Orçamentária	programa
19	Despesa Orçamentária	sub elemento
20	Despesa Orçamentária	sub funcao
21	Despesa Orçamentária	valor fixado
22	Despesa Orçamentária	valor liquidado
23	Despesa Orçamentária	valor pago
24	Despesa Orçamentária	nome
25	Receita Orçamentária	uni gestora
26	Receita Orçamentária	alinea
27	Receita Orçamentária	arrecadacao
28	Receita Orçamentária	categoria economica
29	Receita Orçamentária	especie
30	Receita Orçamentária	lancado
31	Receita Orçamentária	origem
32	Receita Orçamentária	previsao
33	Receita Orçamentária	rubrica
34	Receita Orçamentária	sub alinea
35	Despesa Extra Orçamentária	valor
36	Despesa Extra Orçamentária	codigo
37	Despesa Extra Orçamentária	nomenclatura
38	Receita Extra Orçamentária	valor
39	Receita Extra Orçamentária	codigo
40	Receita Extra Orçamentária	nomenclatura
41	Quadro Pessoal	nome
42	Quadro Pessoal	cpf
43	Quadro Pessoal	cargo
44	Quadro Pessoal	tipo cargo
45	Quadro Pessoal	salario cargo
46	Licitação	numero licitacao

47	Licitação	objeto licitacao
48	Licitação	modalidade licitacao
49	Licitação	data publicacao licitacao
50	Licitação	valor
51	Licitação	cnpj cpf
52	Licitação	edital
53	Licitação	nome vencedores
54	Licitação	nome perdedores
55	Licitação	data realizacao
56	Licitação	setor interessado
57	Licitação	integra
58	Licitação	termo ratificacao
59	Licitação	pregao
60	Licitação	aviso
61	Licitação	licitado

Apêndice B

Empresas atuantes no fornecimento de Portais de transparência na Paraíba

	Empresa
1	Alfa Consultoria
2	Aspec Informática
3	Betha Sistemas
4	Connecta - Tecnologia da Informação
5	Conteúdo Design
6	DC Soluções
7	e-TICons
8	EasyWeb
9	Elmar Tecnologia
10	F5 Tech
11	Franinformática
12	GJSM
13	Grupo Assesi
14	IMAP
15	Info Public
16	Jader Formiga
17	LHSystem

18	Newsites
19	Portal Próprio
20	Publicsoft
21	RedeNet Soluções
22	SD NetWord
23	Softgov
24	WebCreativ

Apêndice C

Municípios presentes na amostra e suas combinações

	Município	Combinação
1	Belém do Brejo do Cruz	Publicsoft
2	Ouro Velho	Publicsoft
3	Cruz do Espírito Santo	Publicsoft
4	Pocinhos	Publicsoft
5	Santa Cecília	Publicsoft
6	Alcantil	Publicsoft
7	Bom Jesus	Elmar Tecnologia
8	Arara	Elmar Tecnologia
9	Carrapateira	Elmar Tecnologia
10	Santa Rita	e-TICons
11	Bayeux	e-TICons
12	Coremas	e-TICons
13	Esperança	Info Public
14	Umbuzeiro	Info Public
15	Serraria	Alfa Consultoria / Elmar Tecnologia
16	Curral de Cima	Alfa Consultoria / Elmar Tecnologia
17	Ingá	Portal Próprio / Publicsoft

18	Remígio	Portal Próprio / Publicsoft
19	Campina Grande	Alfa Consultoria / Publicsoft
20	Areia	Alfa Consultoria / Publicsoft
21	Quixaba	EasyWeb / Publicsoft
22	Serra Grande	EasyWeb / Publicsoft
23	Itabaiana	Portal Próprio / Elmar Tecnologia
24	Barra de Santa Rosa	Portal Próprio / Elmar Tecnologia
25	Conceição	EasyWeb / e-TICons
26	Pedra Lavrada	LHSystem / Elmar Tecnologia
27	Bom Sucesso	DC Soluções
28	Areia de Baraúnas	EasyWeb / Publicsoft / e-TICons
29	Gado Bravo	Franinformática / Publicsoft / Aspec.
30	João Pessoal	TI de João Pessoal

Apêndice D

Relatório Lighthouse Resumido do portal de Itabaiana-PB

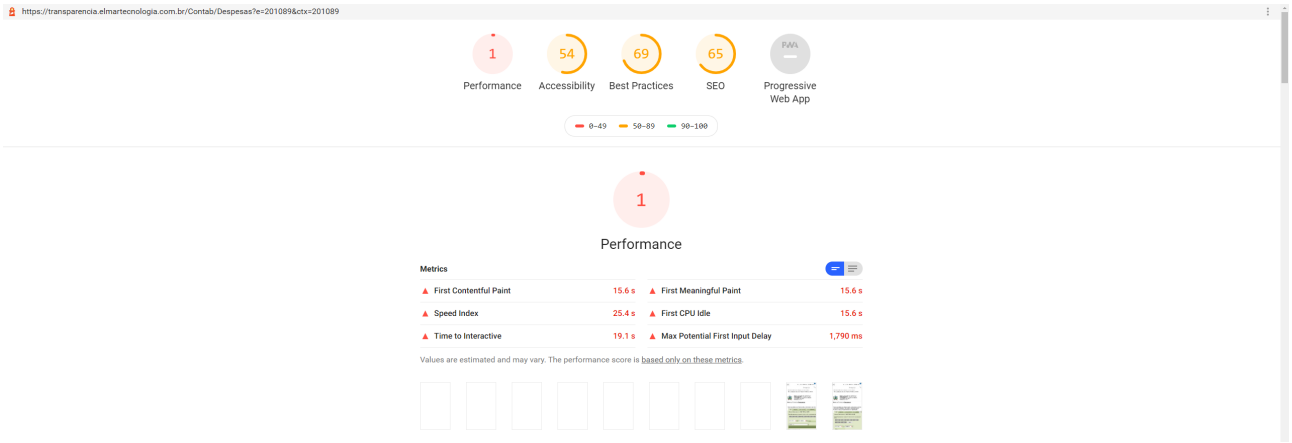

Figura D.1: Relatório Lighthouse Resumido do portal de Itabaiana-PB