Desenvolvimento e avaliação de um buscador de portarias

Serigne Khassim Mbaye[1], Edimar Manica[1] (orientador), Renata Galante[2] (colaboradora)

[1]Instituto Federal do Rio Grande do Sul - *Campus* Ibirubá

[2]Universidade Federal do Rio Grande do Sul - Instituto de Informática

Resumo. *Portarias são documentos emitidos por órgãos institucionais federais que contém, dentre outras, informações a respeito de servidores de instituições. Este Trabalho de Conclusão de Curso apresenta o desenvolvimento e a avaliação de um buscador de portarias, que coleta os documentos que publicam as portarias nos repositórios das instituições, identifica as portarias nesses documentos e disponibiliza uma interface de busca. Foram realizados experimentos em 5 bases de dados reais, que mostraram uma precisão maior que 90% para a coleta dos documentos e identificação do número, da data e do conteúdo das portarias. Além disso, um experimento com usuários reais mostrou que o buscador é capaz de colocar as portarias relevantes no topo dos resultados.*

Abstract. *Portarias are documents issued by federal institutional organs that contain, among other things, information about the institutions' employees. This paper presents the development and evaluation of a search engine focused on Portarias, which collects the documents that publish the portarias in the institutions' repositories, identifies the portarias in these documents, and provides a search interface. Experiments were carried out on 5 real databases, which showed a precision greater than 90% for crawling of documents and identification of the number, date, and content of the portarias. In addition, an experiment with real users showed that the search engine is able to place the relevant portarias at the top of the results.*

1. Introdução

Portarias são documentos emitidos por órgãos institucionais federais que contém, dentre outras, informações a respeito de servidores de instituições. Por exemplo, o Instituto Federal do Rio Grande do Sul (IFRS) e a Universidade Federal do Rio Grande do Sul (UFRGS) publicam por meio de portarias as progressões dos servidores, bem como designações para comissões, afastamentos, licenças, entre outras informações.

Frequentemente, os servidores precisam encontrar determinadas portarias para solicitar progressões ou para comprovar determinada experiência. No entanto, eles encontram dificuldades para realizar tal busca porque esses documentos estão espalhados em diversas fontes de diferentes formatos, não há uma forma de busca adequada e não existe a opção de busca pelo conteúdo da portaria. Gestores das instituições, frequentemente, precisam fazer consultas, como, por exemplo, todos os professores que se aposentaram em determinado ano ou professores que estão em afastamento.

Nesse contexto, este trabalho tem o objetivo de desenvolver e avaliar um buscador, denominado **Touba**[1], onde os cidadãos possam pesquisar as portarias de uma instituição de forma intuitiva e eficaz. O buscador coleta os documentos que publicam as portarias

[1]Touba é uma cidade senegalesa fundada por Cheikh Ahmadou Bamba Mbacke, que é o sufi mais famoso do Senegal, pois tinha a missão social de resgatar a sociedade da alienação colonial.

nos repositórios das instituições, identifica as portarias nesses documentos e disponibiliza uma interface de busca. Essa busca é realizada por palavras-chave sobre o conteúdo do documento, permitindo, por exemplo, que o usuário encontre as portarias que contenham o nome de um determinado servidor ou comissão. Foram realizados experimentos em 5 bases de dados reais, que mostraram uma precisão maior que 90% para a coleta dos documentos e identificação do número, da data e do conteúdo das portarias. Além disso, um experimento com usuários reais mostrou que o buscador é capaz de colocar as portarias relevantes no topo dos resultados.

O restante deste trabalho está organizado como segue. A Seção 2 apresenta os principais conceitos, ferramentas e bibliotecas utilizadas neste trabalho. A Seção 3 discute os trabalhos relacionados. A Seção 4 descreve o buscador **Touba**. Na Seção 5, são apresentados os experimentos realizados e discutidos os resultados. Por fim, a Seção 6 apresenta as considerações finais e os trabalhos futuros.

2. Fundamentação Teórica

Esta seção descreve os principais conceitos necessários para a compreensão deste Trabalho de Conclusão de Curso, bem como as bibliotecas e ferramentas utilizadas.

2.1. Recuperação de informação

Segundo Baeza-Yates e Ribeiro-Neto (2013), a Recuperação de Informação (RI) "é uma área abrangente da Ciência da Computação que se concentra principalmente em prover aos usuários o acesso fácil às informações de seu interesse". Nesse contexto, é importante diferenciar sistema de recuperação de informação e sistema de recuperação de dados. De acordo com Baeza-Yates e Ribeiro-Neto (2013), a recuperação de dados consiste na identificação de quais documentos da coleção contém as palavras-chave da consulta do usuário, o que, com frequência, não é suficiente para satisfazer a necessidade de informação do usuário. Por outro lado, em um sistema de RI, os objetos recuperados podem ser inexatos e pequenos erros podem passar despercebidos.

O buscador proposto neste trabalho é um sistema de recuperação de informação focado em portarias institucionais. Esse buscador necessita coletar os documentos onde as portarias estão publicadas, os quais estão disponibilizados em diferentes fontes.

2.2. Ferramentas e bibliotecas utilizadas

Esta subseção apresenta as ferramentas e bibliotecas utilizadas no desenvolvimento do buscador **Touba**.

Para coletar os documentos que publicam as portarias nos repositórios, o buscador **Touba** necessita analisar o HTML das páginas dos repositórios a fim de encontrar os links para as portarias. Essa análise foi realizada utilizando a ferramenta JSoup, que é uma biblioteca Java que fornece uma API (*Application Programming Interface* - Interface de Programação de Aplicação) para extrair e manipular dados, usando as tecnologias DOM, CSS e JQuery (JSOUP, 2021).

Como as portarias estão disponíveis em documentos PDF, a ferramenta **Touba** necessita convertê-los para o formato texto. Para realizar essa tarefa, foram testadas três ferramentas: Tabex PDF, itext e PDFbox. A API Tabex PDF é um leitor de PDF que

permite aos desenvolvedores exportar dados de PDF para TXT. A API varre o documento PDF para extrair todo o texto contido no corpo do documento e permite que os desenvolvedores usem o texto para processos adicionais (TABEXPDF, 2019). A Tabex PDF não é gratuita, mas fornece um conjunto de conversões para teste sem custo. A biblioteca iText permite criar e manipular arquivos PDF em Java e .NET. O código fonte dessa biblioteca está disponível como código-aberto (AGPL) bem como há uma licença comercial. PDFbox é uma ferramenta Java de código aberto para trabalhar com documentos PDF. Essa ferramenta permite a criação de novos documentos PDF, manipulação de documentos existentes e a capacidade de extrair conteúdo de documentos (PDFBOX, 2021). Foi escolhida a ferramenta PDFBox, pois apresentou o melhor desempenho.

A ferramenta **Touba** semiestrutura as portarias em documentos XML. A criação dos documentos XML é realizada utilizando a biblioteca JDOM. Essa biblioteca fornece uma maneira de representar esse documento para leitura, manipulação e escrita fáceis e eficientes. Possui uma API simples, leve, rápida e otimizada para a linguagem de programação Java (JDOM, 2021).

Por fim, a ferramenta **Touba** realiza o pré-processamento, a indexação e as consultas por meio da biblioteca Apache Lucene. Essa biblioteca é escrita inteiramente em Java e possibilita a criação de mecanismos de busca de alto desempenho (LUCENE, 2021). Pode-se afirmar que essa tecnologia é adequada para praticamente qualquer buscador que exija pesquisas por palavras-chave.

3. Trabalhos Relacionados

Esta seção descreve os trabalhos relacionados ao escopo do buscador **Touba**. O trabalho de SOBRINHO (2019) tem o objetivo de mostrar como dados de um usuário do Instagram podem ser extraídos, normalizados, enriquecidos e armazenados em uma camada de persistência utilizando um rastreador web não supervisionado. SOBRINHO (2019) aplicaram uma técnica de web *crawling*[2] para baixar automaticamente os dados de uma página web e extrair informações específicas. O buscador **Touba** também aplica uma técnica de web *crawling*, mas usando uma ferramenta diferente. A diferença entres os dois trabalhos é que SOBRINHO (2019) utilizaram os dados extraídos para análises de tendências e padrões, enquanto que este trabalho utiliza os dados para permitir que os usuários encontrem as portarias desejadas.

O trabalho de Ramya e Shreedhara (2017) descreve uma nova metodologia para recuperação de documentos na web. Esse trabalho tem o propósito de reduzir o tempo de resposta do sistema e melhorar a similaridade entre o documento e a consulta, portanto otimizando o processo de recuperação de informações. No trabalho de Ramya e Shreedhara (2017), foram aplicadas técnicas de *crawling*, classificação e indexação, bem como o algoritmo PSO (*Particle Swarm Optimization*). Neste Trabalho de Conclusão de Curso, também são realizadas as etapas de *crawling* e indexação. No entanto, é utilizada a ferramenta Apache Lucene em vez do algoritmo PSO.

O trabalho de Pivetta, Mergen e Kepler (2013) realiza a classificação de documentos do Exército Brasileiro. Os autores propõem formas de realizar essa classificação de maneira automática, utilizando o método Naive Bayes, a fim de identificar quais sentenças

[2]Web *crawler* é um algoritmo usado pelos buscadores para encontrar, ler e indexar páginas de um site.

em um documento são relativas a cada militar, de modo que apenas elas sejam usadas durante o treinamento do classificador. Tanto o trabalho de Pivetta, Mergen e Kepler (2013) quanto o buscador **Touba** utilizam a biblioteca PDFBox para converter documentos PDF em TXT. No entanto, Pivetta, Mergen e Kepler (2013) realizam a classificação de documentos, enquanto este trabalho permite a busca das portarias publicadas nos documentos.

4. Touba: Buscador de portarias

O objetivo da ferramenta Touba é permitir a busca de portarias institucionais por meio de palavras-chave. A Figura 1 apresenta a visão geral do buscador proposto, que contempla seis etapas: coletar, converter, estruturar, pré-processar, indexar e buscar. A etapa **coletar** consiste em identificar e baixar os documentos que publicam as portarias nas diferentes fontes (sites institucionais). A etapa **converter** realiza a conversão dos documentos para o formato de texto. A etapa **estruturar** identifica as portarias nos documentos e as estrutura em documentos XML. A etapa **pré-processar** é responsável pelo pré-processamento dos documentos, incluindo as fases de análise léxica, remoção de *stopwords* e atribuição de pesos. A etapa **indexar** compreende a inclusão dos termos e seus respectivos pesos em uma estrutura de índices que facilita a busca. A etapa **buscar** inclui a disponibilização de uma interface de busca que recebe palavras-chave definidas pelo usuário, consulta o índice e retorna os documentos mais similares a consulta.

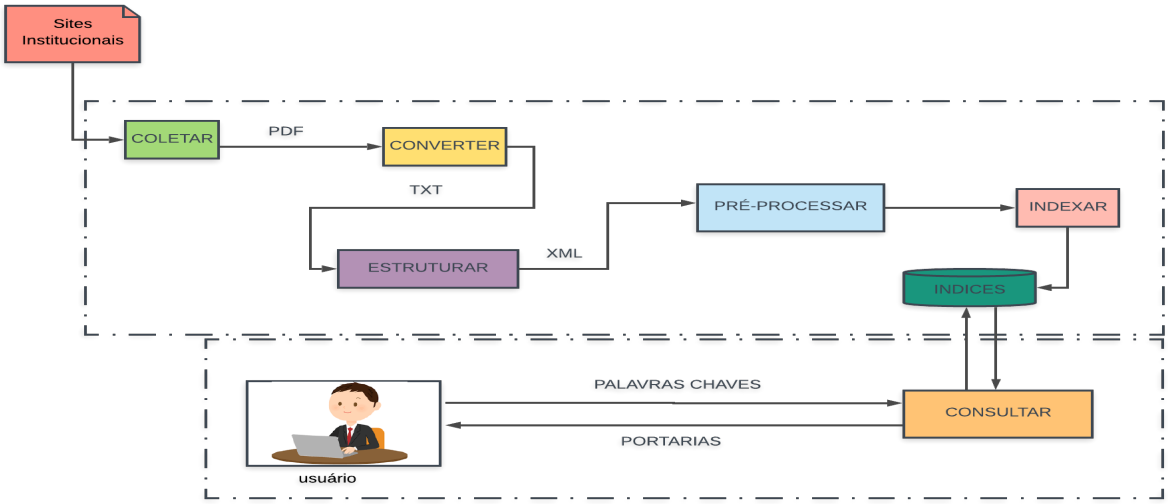


Figura 1. Visão Geral do Touba: Buscador de portarias

A seguir cada etapa é detalhada em uma subseção específica.

4.1. Coletar

O objetivo desta etapa é identificar e baixar os documentos que publicam as portarias nas diferentes fontes. A entrada dessa etapa é composta por sites institucionais. A saída dessa etapa é composta pelos documentos PDF que contém as portarias publicadas nos sites institucionais. Essa etapa é realizada de forma semi-automática uma vez que requer a participação do usuário para definir alguns padrões que auxiliam a coleta dos documentos nos sites institucionais.

Neste trabalho, foram coletadas portarias das seguintes instituições de ensino: Instituto Federal do Rio Grande do Sul (IFRS) e Universidade Federal do Rio Grande do Sul (UFRGS). Foi necessário utilizar técnicas diferentes para cada instituição conforme as restrições e possibilidades que cada repositório possui. A seguir essas técnicas são detalhadas.

4.1.1. Coleta das portarias dos sites do IFRS

O IFRS não possui um repositório central de portarias. Existem portarias publicadas no site oficial do instituto. Outras portarias são disponibilizadas nos sites dos *campi*. Ainda, existem portarias disponibilizadas em sites antigos. Neste trabalho, foram coletadas as portarias do site oficial do IFRS[3], do site atual do *Campus* Ibirubá[4] e do site antigo do *Campus* Ibirubá[5].

Basicamente, a coleta ocorreu pelo caminhamento nas páginas, a identificação dos links para as portarias e o *download* automatizado uma vez que havia uma estrutura hierárquica. Nessa técnica, são criados pequenos trechos de código (scripts) que percorrem as páginas do site que formam a estrutura hierárquica até chegar nos links para as portarias. Esse processo é similar a uma busca em profundidade realizada sobre uma árvore de pesquisa, onde as portarias são os nodos folhas da árvore. Nessa técnica, é necessário analisar a estrutura das páginas HTML, identificando os elementos que contém os links para os documentos de interesse.

Para analisar a estrutura HTML das páginas, foi utilizada a ferramenta JSoup. Inicialmente, utilizou-se a versão 1.12.1 da biblioteca JSoup. No entanto, essa versão não foi capaz de processar os documentos do site antigo do *Campus* Ibirubá, sendo necessário utilizar a versão 1.13.1 para essa tarefa.

4.1.2. Coleta das portarias do IFRS via API

Recentemente, o IFRS também disponibilizou o SIPPAGweb[6], que inclui uma ferramenta de busca de portarias, bem como uma API para consulta de portarias. Essa ferramenta inclui algumas portarias relacionadas a vida funcional dos servidores, como, por exemplo, licença capacitação e progressão. No entanto, não inclui todas as portarias relacionadas aos servidores. Por exemplo, não contempla as portarias de designações para comissões e coordenações.

A coleta nesse repositório foi realizada por meio da API do SIPPAGweb. Os documentos foram obtidos por meio de requisições HTTP, que retornaravam informações das portarias, incluindo a URL dos documentos, no formato JSON.

Ressalta-se ainda que desde 10 de agosto de 2020, o Boletim de Pessoal Diário da Reitoria passou a fazer parte do Boletim de Gestão de Pessoas do Governo Federal

[3]Disponível em: https://ifrs.edu.br/. Último acesso em: 11/02/2021.
[4]Disponível em: https://ibiruba.ifrs.edu.br/. Último acesso em: 11/02/2021.
[5]Disponível em: https://ibiruba.ifrs.edu.br/site/. Último acesso em: 11/02/2021.
[6]Disponível em: https://sippag-web.ifrs.edu.br/. Último acesso em: 26/02/2021.

(BGP)[7]. No entanto, não foram coletados documentos deste repositório.

4.1.3. Coleta das portarias da UFRGS

A UFRGS possui um repositório central para busca das portarias[8]. O acesso a esse repositório é restrito e controlado por uma ferramenta limitadora chamada CAPTCHA. Essa restrição impede o *download* dos arquivos de maneira simples, obrigando o uso de abordagens não convencionais.

A coleta das portarias nesse repositório ocorreu pela inferência de padrões de URL. Essa técnica consiste em gerar URLs candidatas que representam o endereço dos documentos de interesse. A partir da URL, é possível baixar o documento sem passar pela ferramenta CAPTCHA. A inferência do padrão de URLs foi realizada manualmente.

O padrão de URL foi definido pela seguinte expressão regular <https://www1.ufrgs.br/sistemas/sde/gerencia-documentos/index.php/publico/ExibirPDF?documento=[0-9][1,6]>. Na prática, as URLs candidatas são geradas inferindo valores para o campo documento. Iniciou-se com o número 18001. Após, gerou-se URLs candidatas incrementando o número do documento até 105995. Os valores inicial e final foram determinados por meio de submissões manuais na ferramenta de busca da instituição. Destaca-se que algumas URLs candidatas não geraram documentos relevantes, ou seja, documentos que publicavam portarias. Esses foram excluídos. Um documento foi considerado relevante se ele possuía pelo menos uma ocorrência do termo "portaria".

O uso dessa técnica gerou, também, a limitação no número de requisições ao repositório da Universidade. Isso foi percebido devido ao retorno da requisição HTTP que possuía o seguinte erro: *Status Code 429 - Too Many Requests*. Para contornar a limitação, foi inserido um *delay* de 60 segundos a cada 100 requisições ao servidor.

Ressalta-se que a coleta das portarias da UFRGS foi realizada no Trabalho de Conclusão de Curso do Christian Schmitz (SCHMITZ, 2020).

4.2. Converter

O objetivo desta etapa é transformar todos os arquivos PDFs coletados em um formato editável. A entrada dessa etapa é composta pelos documentos PDF que foram coletados nos sites institucionais. A saída dessa etapa é composta pelos respectivos documentos em formato TXT.

Essa etapa foi realizada de forma automática. Foram selecionadas três ferramentas para avaliar sua eficácia nesta tarefa: Tabex-PDF, itextpdf e Apache PDFBox. A ferramenta Tabex-PDF foi descartada uma vez que a versão gratuita possui um limite de conversão de 200 documentos. A ferramenta itextpdf apresentou erros na conversão, principalmente quando havia tabelas no documento. Por fim, a ferramenta Apache PDFBox

[7]Disponível em: https://boletim.sigepe.planejamento.gov.br/publicacao/pesquisa/. Último acesso em: 26/02/2021.

[8]Disponível em: https://www1.ufrgs.br/sistemas/sde/gerencia-documentos/index.php/publico/consultar/. Último acesso em: 26/02/2021.

apresentou os melhores resultados, conseguindo converter adequadamente a maioria dos documentos.

4.3. Estruturar

O objetivo desta etapa é transformar os dados de um formato não estruturado (TXT) para um formato semi-estruturado (XML - *Extensible Markup Language*). A entrada dessa etapa é composta de arquivos TXT obtidos na etapa anterior. A saída dessa etapa é composta de arquivos no formato XML com o conteúdo das portarias individuais.

Essa etapa foi realizada analisando os arquivos de texto de forma a identificar, segmentar e estruturar as múltiplas portarias que estão publicadas em cada documento. Essa não é uma tarefa trivial uma vez que alguns documentos possuem diversas portarias, assim como outros documentos publicam outras informações além de portarias, como, por exemplo, ordens de serviço, férias, aniversariantes, concessão de diárias e passagens e atestados médicos.

A primeira tarefa desta etapa era segmentar as portarias, ou seja, identificar quando uma portaria começava e quando uma portaria acabava. Essa segmentação é realizada através de expressões regulares que capturam os intervalos de caracteres que compõem os padrões que descrevem o formato de uma portaria, seu número e data de publicação, respectivamente. Os documentos analisados neste trabalho possuem estruturas bem definidas que permitem a identificação desses padrões facilmente.

A Tabela 1 apresenta as expressões regulares utilizadas. O início de cada portaria é identificado pela linha que inicia com "PORTARIA". O fim de cada portaria é identificado pela linha que descreve o cargo de quem assinou a portaria. Por exemplo, Diretora-Geral e Reitor. O número das portarias é identificado considerando que sempre é precedido por “PORTARIA No”. A data de publicação das portarias é identificada por uma função que procura um padrão de data no início das portarias, logo após seu número. Essa função preve diferentes formatos de data.

Tabela 1. Expressões regulares utilizadas para estruturação das portarias

FUNÇÃO	EXPRESSÃO REGULARES
Identificar o início de cada portaria	(^)PORTARIA
Identificar o fim de cada portaria	(^)Diretora Geral|Reitor Substituto|Diretora-Geral|Presidente da CPAD|Reitor pro tempore|Vice-Reitora|Pró-Reitora|Pró-Reitor|Diretor Geral|Diretor-Geral|Vice-Pró-Reitora|Vice-Pró-Reitor|Vice-Superintendente|VICE-REITOR|Vice-Superintendente|Reitor|VICE-REITORA|Diretor|Diretora da Faculdade|Diretor-Geral Substituto| Diretor Geral Substituto|(^)End_New_Official
Identificar o número de cada portaria	PORTARIA No *(/d+)|(^)Nº [0-9]+/[0-9]{4}|(^)Port. nº [0-9]+/[0-9]{4}|(^)Portaria [0-9]+/[0-9]{4}

Por fim, é gerado um documento XML para cada documento TXT contendo as portarias publicadas no referido documento. Para essa tarefa, foi utilizada a biblioteca JDOM. A Figura 2 ilustra o formato dos documentos XML resultantes. O documento XML possui um elemento raiz chamado `documento`, que tem como atributos o identificador único do documento (`id`), o nome do arquivo PDF original (`nome_arquivo`)

e o endereço do arquivo no repositório da instituição (`site`). Um documento pode ter um número arbitrário de filhos, denominados `portaria`, que correspondem as portarias que o documento PDF original possui, tendo como atributos o número (`nr`) e a data de publicação (`data`). O conteúdo da portaria é armazenado como um nó texto filho do elemento portaria. A Figura 3 apresenta um exemplo de portaria enquanto que a Figura 4 ilustra a respectiva saída da etapa estruturar.

Figura 2. Formato dos documentos XML resultantes.

```
<documento id="" nome_arquivo="" site="">
    <portaria nr="" data="">
        ConteudodaPortaria
    </portaria>
</documento>
```

Figura 3. Exemplo de Portaria em formato pdf

PORTARIA DO DIA 04 DE MARÇO DE 2020.

O DIRETOR GERAL SUBSTITUTO DO INSTITUTO FEDERAL DE EDUCAÇÃO, CIÊNCIA E TECNOLOGIA DO RIO GRANDE DO SUL - CAMPUS IBIRUBÁ, no uso de suas atribuições legais subdelegadas pela Portaria nº 034, de 28 de fevereiro de 2020, publicado no DOU em 03 de março de 2020, RESOLVE:

Nº 041 – DESIGNAR a servidora [redacted], Assistente de Aluno e Coordenadora de Ensino, matrícula SIAPE nº 2331826, para responder como substituta da Diretoria de Ensino, código CD - 04 do IFRS – Campus de Ibirubá, na ausência do Titular da função.

Edimar Manica
Diretor Geral Substituto
Portaria 034/2020

Figura 4. Exemplo de Portaria em formato XML

```
<Document id="6" nome_arquivo="https_DOISpont__baraduplas_ifrsc.." site="https://ifrs.edu.br/ibiruba/wp-content/uploads/sites/4/2020/06/..">
  <portaria ID="041" data="04 DE MARÇO DE 2020">
    <text>
      O DIRETOR GERAL SUBSTITUTO DO INSTITUTO FEDERAL DE EDUCAÇÃO, CIÊNCIA E TECNOLOGIA DO RIO GRANDE DO SUL - CAMPUS IBIRUBÁ, no uso de suas atribuições legais subdelegadas pela Portaria nº 034, de 28 de fevereiro de 2020, publicado no DOU em 03 de março de 2020, RESOLVE: Nº 041 – DESIGNAR a servidora , Assistente de Aluno e Coordenadora de Ensino, matrícula SIAPE nº 2331826, para responder como substituta da Diretoria de Ensino, código CD - 04 do IFRS – Campus de Ibirubá, na ausência do Titular da função. Edimar Manica Diretor Geral Substituto Portaria 034/2020
    </text>
  </portaria>
</Document>
```

4.4. Pré-processar

O objetivo desta etapa é separar o conteúdo das portarias em termos (*tokens*) indexáveis e pré-processá-los de forma a facilitar a indexação e a busca. Essa etapa foi realizada utilizando a classe `StandardAnalyzer` da ferrramenta Lucene (LUCENE, 2021). Essa classe cria os termos indexáveis usando um algoritmo de segmentação de texto Unicode e aplica os seguintes pré-processamentos: conversão dos termos para minúsculo e remoção de *stopwords* (termos muito frequentes que tem pouca contribuição para busca).

4.5. Indexar

O objetivo desta etapa é indexar os termos relacionando-os às portarias que os contêm de forma a facilitar a busca. Além disso, cada termo recebe um peso para cada documento representando a sua importância naquele documento. O peso é calculado pela medida TF-IDF (*Term Frequency-Inverse Document Frequency* - frequência do termo–inverso da frequência nos documentos), é uma medida estatística que visa indicar a importância de um termo em um documento em relação a uma coleção de documentos ou em um *corpus* linguístico. Essa medida atribui maior peso para os termos que ocorrem várias vezes no documento, mas poucas vezes na coleção, uma vez que esses termos possuem maior poder discriminatório (BAEZA-YATES; RIBEIRO-NETO, 2013). As seguintes fórmulas foram utilizadas : $TF - IDF = TF(T, D) * IDF(T)$ onde $TF(T, D)$ representa a frequência de termo T no documento D, $IDF(T) = LOG\frac{1+N}{1+DF(D,T)}$, onde N representa o total de documentos e $DF(D, T)$ representa o número de documentos que contêm o termo T.

A indexação foi realizada por meio da ferramenta Lucene. A ponderação dos termos foi realizada utilizando a medida TF-IDF. Tradicionalmente, a indexação é realizada seguindo a granularidade documento. No entanto, neste trabalho adotou-se a granularidade portaria. Dessa forma, as portarias foram consideradas como documentos independentes, mesmo que estivessem publicadas no mesmo documento. As seguintes informações foram associadas às portarias: (i) número; (ii) data de publicação; (iii) conteúdo; e (iv) endereço do documento que a publicou.

4.6. Consultar

O objetivo desta etapa é permitir que os usuários encontrem as portarias de seu interesse ou necessidade por meio de uma interface gráfica. A entrada dessa etapa é um conjunto de palavras-chave informadas pelo usuário. A saída dessa etapa é um conjunto composto pelas portarias mais similares à consulta, em ordem decrescente de similaridade. A similaridade entre as palavras-chave da busca e os termos indexados foram calculados por meio da função de similaridade Cossine disponível na ferramenta Lucene. Essa função considera os pesos dos termos em cada portaria. Além disso, é importante destacar que as palavras-chave da consulta passam pelo mesmo pré-processamento realizado nos termos antes da indexação.

A interface gráfica para interação com o usuário foi desenvolvida utilizando a linguagem de programação PHP e o Sistema de Gerenciamento de Banco de Dados (SGBD) Mysql, integrados com a ferramenta Lucene. São armazenadas no SGBD Mysql as informações sobre as consultas realizadas e as portarias clicadas/acessadas. A Figura 5 ilustra a tela inicial de consulta, enquanto a Figura 6 apresenta a tela de exibição de resultados.

5. Avaliação

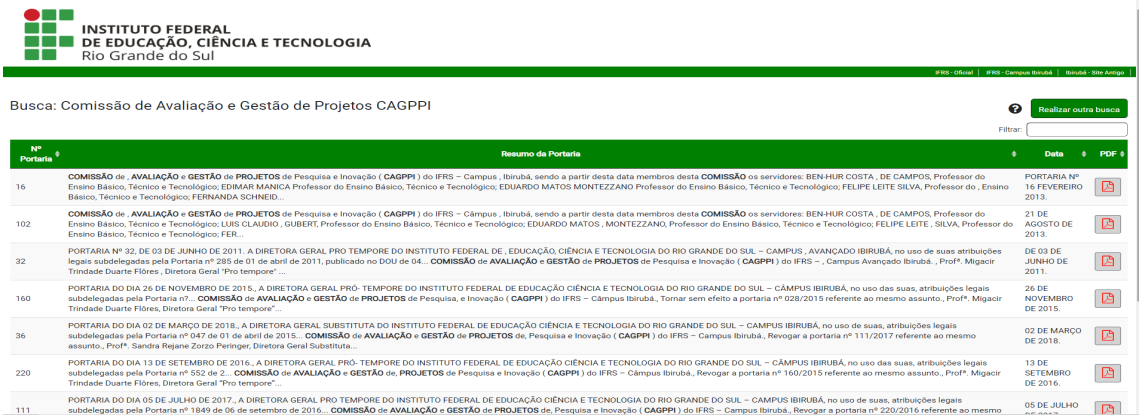
Esta seção tem como objetivo avaliar a eficácia das etapas do buscador de portarias **Touba**, proposto neste Trabalho de Conclusão de Curso.

5.1. Métricas

Foram utilizadas métricas tradicionais da comunidade de recuperação (BAEZA-YATES; RIBEIRO-NETO, 2013), tais como revocação, precisão e precisão@k. A revocação (r), a precisão (p) e a precisão@k ($p@k$) foram calculadas pelas seguintes fórmulas:

Figura 5. Touba : buscador de portarias institucionais antes da consulta

Figura 6. Touba : buscador de portarias institucionais depois da consulta



$p = \frac{VP}{VP+FN}$, $p = \frac{VP}{VP+FP}$ e $p@k = \frac{VP_k}{k}$, onde VP representa os verdadeiros positivos, FN representa os falsos negativos, FP representa os falsos positivos e VP_k representa os verdadeiros positivos entre os primeiros k itens. Essas variáveis são detalhadas em cada experimento.

5.2. Bases de dados

Os experimentos foram realizados em 5 bases de dados:

1. `Ibirubá-Antigo` - inclui as portarias internas do *Campus* Ibirubá publicadas, nos Boletins de Serviço do site antigo[9], entre janeiro de 2011 e fevereiro de 2018;
2. `Ibirubá-Atual` - inclui as portarias internas do *Campus* Ibirubá publicadas, nos Boletins de Serviço do site atual[10], entre março de 2018 e fevereiro de 2021;
3. `SIPPAG` - inclui as portarias publicadas, na ferramenta SIPPAGweb[11], pela Diretoria de Gestão de Pessoas entre 01/01/2018 e 27/02/2021;
4. `IFRS` - inclui as portarias publicadas, no Boletins de Pessoal Mensais[12], pela

[9]Disponível em: https://ibiruba.ifrs.edu.br/site/conteudo.php?cat=50&sub=2385. Último acesso em: 24/03/2021.

[10]Disponível em https://ifrs.edu.br/ibiruba/documentos/boletim-de-servico/. Último acesso em: 24/03/2021.

[11]Disponível em: https://sippag-web.ifrs.edu.br/portarias/. Último acesso em: 27/02/2021.

[12]Disponível em: https://ifrs.edu.br/documentos/tipos/boletim-de-pessoal-mensal/. Último acesso em 24/03/2021.

Reitoria da instituição entre julho de 2017 e agosto de 2020.

5. `UFRGS` - inclui as portarias publicadas, na ferramenta da Pró-Reitoria de Gestão de Pessoas[13], entre 18/02/2016 e 03/03/2020.

5.3. Configuração dos experimentos

Para realizar a hospedagem do buscador de portarias **Touba**, foi utilizada a plataforma *Amazon Web Services* (AWS)[14]. A hospedagem foi realizada em um servidor Ubuntu (86x64) versão 16.04 com 1 GiB de memória RAM e processador Intel Xeon de alta frequência, na opção EC2 que permite 750 horas de uso da instância por mês. Para banco de dados, foi selecionada a opção S3 que permite armazenar até 5 GB, realizar 20.000 solicitações GET e 2.000 solicitações PUT. Os *softwares* instalados foram o Apache 2.4.18, o PHP 7.0.8 e o MySQL 5.7.

5.4. Experimentos

Os experimentos têm como objetivo avaliar a eficácia da ferramenta **Touba** em todas as etapas[15]. A seguir cada experimento é detalhado.

5.4.1. Coleta

Esse experimento tem como objetivo avaliar a eficácia do buscador **Touba** em coletar os documentos PDF que publicam portarias nos sites institucionais. A eficácia foi analisada a partir da precisão da coleta considerando como verdadeiros positivos os documentos coletados que continham pelo menos uma ocorrência do termo "portaria". Os demais documentos coletados foram considerados falsos positivos. A Tabela 2 apresenta o total de documentos PDF coletados em cada base de dados, bem como o total de documentos coletados que contêm o termo "portaria" e a precisão da coleta. Observa-se que a precisão foi superior a 90% em todas as bases de dados. Ressalta-se que a coleta dos documentos da UFRGS foi realizada pelo Christian Schmitz em seu trabalho de conclusão de curso (SCHMITZ, 2020), que foi realizado em colaboração com o presente trabalho de conclusão de curso. Destaca-se que não foi possível calcular a revocação neste experimento uma vez que a informação sobre o total de portarias publicadas em cada site não está disponível.

Tabela 2. Eficácia da etapa coletar.

Base de dados	Ibirubá-Antigo	Ibirubá-Atual	SIPPAG	IFRS	UFRGS
Total de PDFs	74	47	4370	332	45490
Total de PDFs que contém o termo portaria	74	46	4370	312	42309
Precisão	100%	97.8%	100%	93.9%	93%

[13]Disponível em: https://www.ufrgs.br/progesp/?page_id=4965. Último acesso em: 24/03/2021.

[14]Disponível em: https://aws.amazon.com/pt/free/. Último acesso em 24/03/2021.

[15]Os experimentos foram realizados em conjunto com o estudante Iago Ivanir Dalmolin, bolsista de Iniciação Científica.

5.4.2. Conversão

Esse experimento tem como objetivo avaliar a eficácia da ferramenta **Touba** em converter os documentos PDF para o formato TXT. A eficácia foi analisada a partir da revocação e da precisão da conversão considerando como verdadeiros positivos os documentos PDF coletados que foram convertidos para o formato TXT. Os documentos PDF que não foram convertidos para TXT foram considerados como falsos negativos. A Tabela 3 apresenta o total de documentos PDF coletados em cada base de dados e o total de documentos que foram convertidos para o formato TXT, bem como a revocação e a precisão da conversão. Observa-se que a revocação e precisão foram de 100% em todas as bases de dados.

Tabela 3. Eficácia da etapa converter

Base de dados	Ibirubá-Antigo	Ibirubá-Atual	SIPPAG	IFRS	UFRGS
Total de PDFs	74	47	4370	332	45490
Total de TXTs gerados	74	47	4370	332	42309
Precisão	100%	100%	100%	100%	100%

5.4.3. Estruturar

Esse experimento tem como objetivo avaliar a eficácia da ferramenta **Touba** em estruturar as portarias em formato XML, identificando seu número, data de publicação e conteúdo. A eficácia foi analisada a partir da revocação e da precisão da estruturação. A Tabela 4 apresenta a quantidade de portarias reais, a quantidade de portarias identificadas, a revocação da identificação das portarias, a precisão da identificação do número das portarias e a precisão da identificação da data das portarias. A quantidade de portarias reais foi obtida apenas para a base de dados SIPPAG, analisando o arquivo JSON retornado pela ferramenta SIPPAGweb. Essa informação não está disponível para as demais bases de dados. Para calcular a revocação da identificação das portarias, foram consideradas como verdadeiros positivos as portarias que foram identificadas. Na base de dados SIPPAG, a precisão da identificação do número e da data foi calculada comparando o valor identificado pela ferramenta **Touba** nos documentos PDF com o valor disponível em formato JSON pela ferramenta SIPPAGweb, sendo considerados verdadeiros positivos os casos de igualdade ou equivalência (no caso da data). Nas demais bases de dados, a precisão da identificação do número e da data foi calculada considerando como verdadeiros positivos os casos onde o número e a data correspondiam a um padrão de número e data, respectivamente, expresso por meio de expressões regulares.

5.4.4. Consultar

Esse experimento tem como objetivo avaliar a eficácia da ferramenta **Touba** em retornar portarias relevantes para os usuários que realizam buscas. A eficácia foi realizada por meio da precisão@k uma vez que não era possível definir todos os documentos relevantes presentes nas bases de dados. Foram adotados os seguintes valores de k: 1, 5 e 10.

Tabela 4. Eficácia da etapa estruturar

Base de dados	Ibirubá-Antigo	Ibirubá-Atual	SIPPAG	IFRS	UFRGS
Quantidades de portarias real	-	-	4370	-	-
Quantidade Portarias identificadas	1115	836	4370	6446	45272
Revocação da identificação de portarias	-	-	100%	-	-
Precisão de identificação do número	94.5%	99.2%	100%	91.6%	99.9%
Precisão de identificação da data	98.6%	100%	100%	99.5%	100%

Esses valores foram escolhidos uma vez que os usuários de buscadores geralmente olham apenas a primeira página de resultados, que contém geralmente 10 resultados (BAEZA-YATES; RIBEIRO-NETO, 2013). Além disso, quanto mais ao topo os resultados relevantes estiverem, melhor. Foi realizado um experimento com usuários reais utilizando uma adaptação da ferramenta **Touba** com as portarias de todas as bases de dados indexadas. A ferramenta **Touba** foi adaptada de modo que os usuários pudessem anotar as portarias relevantes, bem como explicar o que procuravam e fornecer um *feedback* sobre a ferramenta. O experimento foi realizado entre 29/01/2021 e 08/03/2021. Foi enviado e-mail para os servidores do IFRS - *Campus* Ibirubá com instruções para participação no experimento. A interface gráfica também continha instruções que explicavam o uso da ferramenta (onde digitar as palavras-chave, onde anotar os documentos relevantes, onde fornecer *feedback*), mas não induziam o tipo de consulta. Foram realizadas 30 consultas.

A Tabela 5 apresenta os resultados do experimento. Observa-se que o maior valor foi da precisão@1 e o menor valor da precisão@10. Esse resultado se deve a dois motivos principais: (i) a ferramenta **Touba** coloca as portarias relevantes mais ao topo; (ii) diversas consultas tinham poucas portarias relevantes, em alguns casos, por exemplo, o usuário desejava uma única portaria.

Tabela 5. Eficácia da etapa consulta

Métricas	Quantidade Buscas	Médias
P@1	30	90%
P@5	30	71.3%
P@10	30	65%

A ferramenta **Touba** apresentou os melhores resultados quando o usuário buscava portarias mais gerais. Por exemplo, portarias que continham o nome de determinado servidor. Nesses casos, a ferramenta apresentou 100% de precisão para todos os valores de k. O *feedback* da ferramenta fornecido pelos servidores foi muito positivo. Por exemplo, "Achei excelente. É uma demanda muito importante, seguidamente precisamos de alguma portaria, a qual não guardamos mais a cópia física, por motivos óbvios. Essa busca por esses documentos, muitas vezes é trabalhosa e nem sempre encontramos o que buscamos. Sim atendeu minhas expectativas.", "Parabéns, pessoal! Ficou muito bom! Esta rápido e eficaz. O Layout também está confortável.". Também foram sugeridas algumas melhorias

que poderão ser adicionadas posteriormente: "Sugestão: colocar a possibilidade de busca avançada, onde pode selecionar se quer a expressão exata ou não".

5.5. Casos de falha

Esta seção descreve os principais casos de falha da ferramenta **Touba**. Com relação à coleta, a principal dificuldade da ferramenta **Touba** foi a ferramenta de CAPTCH do repositório da UFRGS. Mas, esse problema foi solucionado coletando os documentos que publicam as portarias por meio da geração de URLs candidatas uma vez que havia um padrão de URLs das portarias. Também, havia a limitação no número de requisições ao repositório da Universidade. Para contornar essa limitação, foi inserido um *delay* de 60 segundos a cada 100 requisições ao servidor.

Com relação à conversão, o principal caso de falha é a presença de documentos escaneados. A ferramenta **Touba** não suporta este tipo de arquivo. Uma possível solução seria a utilização de software para reconhecimento de caracteres ópticos. Esses softwares de reconhecimento óptico de caracteres têm como propósito converter elementos textuais de documentos em textos editáveis e pesquisáveis (KUHN; CERVI; MÂNICA, 2019). Essa tarefa apresenta desafios específicos quando os elementos textuais estão em imagens capturadas por câmeras de *smartfones*. Um desses desafios é a inclinação das linhas do texto que afeta a eficácia e eficiência dos métodos de reconhecimento atuais.

Com relação à estruturação, o principal caso de falha está relacionado a heterogeneidade dos documentos. As bases de dados `Ibirubá - Antigo` e `Ibirubá - Atual` possuíam várias portarias no mesmo documento, às vezes, até na mesma página. Também, as portarias dessas bases de dados são geradas manualmente, logo não possuem um padrão rígido. Embora, quanto mais recentes as portarias, mais rígido é o padrão. A base de dados `IFRS` possuía portarias coletivas, onde todas as portarias do dia estavam juntas com um cabeçalho e assinatura compartilhados. Por fim, a base `UFRGS` continha algumas portarias com o prefixo do número fora do padrão.

Com relação à consulta, a ferramenta apresentou os piores resultados quando o usuário buscava portarias bem específicas. Por exemplo, nas seguintes buscas "edimar manica técnico informática"e "multa biblioteca ifrs", todas as portarias retornadas possuíam os termos buscados, porém não se encaixavam no que o usuário estava procurando. Esse resultado ocorreu porque os termos utilizados na busca estavam contidos em inúmeras portarias. Outro caso de falha ocorria quando o usuário utilizava um sinônimo do termo contido na portaria. Por exemplo, o usuário utilizou o termo "progressão"na busca e a portaria continha o termo "promoção". Nesse caso, a ferramenta **Touba** não obteve resultado satisfatório uma vez que não inclui tratamento para sinônimos, como por exemplo, expansão de consultas (BAEZA-YATES; RIBEIRO-NETO, 2013).

6. Conclusão

Este trabalho apresentou o desenvolvimento e avaliação de um buscador de portarias, denominado **Touba**, onde os cidadãos podem pesquisar as portarias das instituições de forma intuitiva e eficaz. Essa busca é realizada por palavras-chave sobre o conteúdo dos documentos, que foram coletados em fontes de dados com características distintas. Foram realizados experimentos com bases de dados e usuários reais. Os experimentos mostraram que o buscador **Touba** é eficaz em coletar os documentos que publicam as

portarias, transformá-los em TXT e semiestruturá-los em XML. Além disso, o buscador é capaz de colocar as portarias relevantes ao topo dos resultados.

Foram identificados alguns pontos de melhoria e evolução que podem ser considerados possíveis trabalhos futuros. Desses, destacam-se: (i) a classificação do conteúdo das portarias em categorias, como, por exemplo, progressão funcional, afastamento, aposentadoria e substituição de função; (ii) a inclusão da técnica de expansão de consultas (BAEZA-YATES; RIBEIRO-NETO, 2013) para tratamento de sinônimos; e (iii) a adição de opção de busca avançada.

Referências

BAEZA-YATES, R.; RIBEIRO-NETO, B. *Recuperação de informação: conceitos e tecnologia das máquinas de busca*. 2ª. ed. Porto Alegre, RS: Bookman, 2013.

JDOM. *Documentação JDOM*. 2021. Disponível em: <http://www.jdom.org/docs/faq.html#a0000>. Último acesso em: 22/02/2021.

JSOUP. *jsoup Java HTML Parser 1.12.1 API*. 2021. Disponível em: <https://jsoup.org/apidocs/index.html>. Último acesso em: 13/02/2021.

KUHN, D. M.; CERVI, C. R.; MÂNICA, E. Uma abordagem para extração de elementos textuais em imagens com linhas de texto inclinadas. In: SEMINÁRIO INTEGRADO DE SOFTWARE E HARDWARE, 46., 2019, Belém, PA, Brasil. *Anais...* Porto Alegre, Brasil: SBC, 2019. p. 161–172.

LUCENE. *Apache Lucene Core*. 2021. Disponível em: <https://lucene.apache.org/core/>. Último acesso em: 13/02/2021.

PDFBOX, A. *Apache PDFBox Une bibliothèque PDF Java*. 2021. Disponível em: <https://pdfbox.apache.org/>. Último acesso em: 08/03/2021.

PIVETTA, S. P.; MERGEN, S. L. S.; KEPLER, F. N. Uso de aprendizado de máquina para a classificação de documentos do exército brasileiro. In: SIMPÓSIO BRASILEIRO DE SISTEMAS DE INFORMAÇÃO, 9., 2013, João Pessoa, PB, Brasil. *Anais...* Porto Alegre, RS, Brasil: SBC, 2013. p. 768–779.

RAMYA, C.; SHREEDHARA, K. S. A new PSO methodology for web documents retrieval. In: INTERNATIONAL CONFERENCE ON ELECTRICAL, ELECTRONICS, COMMUNICATION, COMPUTER, AND OPTIMIZATION TECHNIQUES, 2017, Mysuru, India. *Proceedings...* [S.l.]: IEEE, 2017. p. 1–5.

SCHMITZ, C. *ACERPI: Uma abordagem para coleta de documentos, extração de informação e resolução de entidades em Portarias institucionais*. 2020. Monografia (Bacharel em Ciência da Computação), UFRGS (Universidade Federal do Rio Grande do Sul), Porto Alegre, Brasil.

SOBRINHO, J. L. V. *Rastreador web não supervisionado para aquisição, enriquecimento e predição de dados de usuários de redes sociais por intermédio de métodos de inteligência computacional*. 2019. Dissertação (Mestrado em Engenharia Elétrica e da Computação) - Universidade Federal de Goiás, Goiânia, Brasil.

TABEXPDF. *Tabex PDF para API TXT*. 2019. Disponível em: <http://pdfextractoronline.com/pdf-to-txt-api-for-developers/>. Último acesso em: 13/11/2019.