UNIVERSIDADE FEDERAL DA PARAÍBA

CENTRO DE CIÊNCIAS APLICADAS A EDUCAÇÃO

DEPARTAMENTO DE CIÊNCIAS EXATAS

BACHARELADO EM SISTEMAS DE INFORMAÇÃO

MONITORAMENTO DE PUBLICAÇÕES NO DIÁRIO OFICIAL DA UNIÃO

FRANCISCO CELESTINO DOS SANTOS NETO

Orientador: Prof. Dr. Alexandre Nóbrega Duarte

RIO TINTO - PB

2013

FRANCISCO CELESTINO DOS SANTOS NETO

MONITORAMENTO DE PUBLICAÇÕES NO DIÁRIO OFICIAL DA UNIÃO

Trabalho de conclusão de curso apresentado para obtenção do título de Bacharel à banca examinadora no Curso de Bacharelado em Sistemas de Informação do Centro de Ciências Aplicadas e Educação (CCAE), Campus IV da Universidade Federal da Paraíba.
Orientador: Prof. Dr. Alexandre Nóbrega Duarte.

RIO TINTO - PB
2013

S237m Santos Neto, Francisco Celestino dos.

1.1.1.1 Monitoramento de publicações no Diário Oficial da União / Francisco Celestino dos

1.1.1.2 Santos Neto. – Rio Tinto: [s.n.], 2013.

55f.: il. –

Orientador: Alexandre Nóbrega Duarte.
Monografia (Graduação) – UFPB/CCAE.

1.Sistema de recuperação da informação. 2.Necessidades tecnológicas. 3.Diário Oficial da União. I. Título.

FRANCISCO CELESTINO DOS SANTOS NETO

MONITORAMENTO DE PUBLICAÇÕES NO DIÁRIO OFICIAL DA UNIÃO

Trabalho de Conclusão de Curso submetido ao Curso de Bacharelado em Sistemas de Informação da Universidade Federal da Paraíba, Campus IV, como parte dos requisitos necessários para obtenção do grau de BACHAREL EM SISTEMAS DE INFORMAÇÃO.

Assinatura do autor: Francisco Celestino dos Santos Neto

APROVADO POR:

Orientador: Prof. Dr. Alexandre Nóbrega Duarte
Universidade Federal da Paraíba – Campus I

Prof. Msc. Carlos Eduardo Silveira Dias
Universidade Federal da Paraíba – Campus IV

Prof. Msc. Rodrigo de Almeida Vilar de Miranda
Universidade Federal da Paraíba – Campus IV

RIO TINTO - PB
2013

Aos amigos, colegas e professores, minha eterna gratidão por compartilhar comigo seus conhecimentos.

AGRADECIMENTOS

Gostaria de gradecer primeiramente a Deus por estar onde estou hoje e por estar realizando este sonho.

Agradeço a todos os professores que tive a honra de conhecer e de ser aluno, que compartilharam seu conhecimento e que contribuíram com minha formação durante estes anos na universidade.

A meus pais, José Alves e Maria Suely, e a minha irmã Suênia Karla, que me deram força, acreditaram em mim e batalharam para que eu concluísse meu curso, e mesmo distantes, foram essenciais para minha chegada até aqui.

Aos meus tios, Maria Verônica e Auristênio Lucena, que acompanharam de perto minha caminhada na universidade.

Agradeço a todos os meus colegas de turma, desde os que concluíram, ainda estão cursando, e aos os que abandonaram o curso, pois, foi uma honra conhecer a todos, e hoje posso dizer que fazem parte da minha história e serão lembrados sempre.

A minha namorada por todo o carinho e compreensão, principalmente nos últimos meses em que estive mais ausente devido ao estágio e aos trabalhos da universidade.

Ao orientador deste trabalho, Alexandre Duarte, por ter acreditado em mim e ter aceitado ser meu orientador mesmo estando lecionando em outro campos da universidade.

A todos os funcionários que fazem parte do DCE da Universidade Federal da Paraíba.

RESUMO

Este trabalho de conclusão de curso teve como objetivo propor uma solução para um problema levantado ao analisar a forma de recuperação da informação das publicações do Diário Oficial da União. O sistema atual impõe a necessidade de interação do usuário com o sistema de busca disponibilizado no portal da imprensa, o que nem sempre é possível para o usuário por vários motivos como, por exemplo, indisponibilidade para acessar o sistema ou, até mesmo, por se tratar de uma atividade cansativa. Como solução para este problema foi proposto e desenvolvido um sistema capaz de recuperar tais informações sem a necessidade de interação com o usuário. Após o usuário realizar o cadastro no sistema e cadastrar a consulta que deseja, o sistema pesquisa diariamente todas as publicações do Diário Oficial da União, e caso alguma informação relevante seja encontrada, o usuário será notificado. Todo este processo ocorre automaticamente sem a necessidade da interação com o usuário. Para alcançar os objetivos foi necessário a analise do problema, levantamento dos requisitos do sistema proposto como solução, definição da arquitetura do sistema e elaboração do caso de uso. Após o processo de desenvolvimento e realização de testes, o sistema apresenta-se estável e cumprindo com o seu objetivo, tornando-se assim de grande valia para os usuários interessados nas publicações do Diário Oficial da União.

Palavras chave: Diário Oficial da União. Crawler. Sistema de Recuperação da Informação. Indexação.

ABSTRACT

This work presents a solution to help people to obtain information from the official journals published by the Federal Brazilian Government. Nowadays, in order to have access to articles on a given subject the user needs to access the system and run a query providing the related key words. So, if someone needs to be informed as soon as possible when a determined piece of information is published, he/she needs to access and query the system on a daily basis. In this work we provide a solution to help these users. Instead of having to access and query the system every day, the user accesses our solution and provides the query he/she is interested into. Our system is then responsible for executing the queries provided by the users and as soon as a query retrieve some results the system notifies the user about the results.

Keywords: Official Gazette. Crawler. Information Retrieval System. Indexing.

LISTA DE FIGURAS

LISTA DE SIGLAS

D.O	Diário Oficial
RI	Recuperação da Informação
URL	Uniform Resource Locator
DNS	Domain Name System
IP	Internet Protocol
SRI	Sistema de Recuperação da Informação
MVC	Model View Controller
HTML	HyperText Markup Language

SUMÁRIO

1 INTRODUÇÃO

Este capítulo apresenta a motivação para o desenvolvimento deste, os seus objetivos gerais e específicos, a metodologia utilizada para o alcance dos objetivos e a estrutura do trabalho.

1.1 MOTIVAÇÃO

Este subtópico apresenta uma breve introdução sobre Diários Oficiais e o problema de pesquisa levantado.

1.1.1 INTRODUÇÃO SOBRE DIÁRIOS OFICIAIS

Os Diários Oficiais (D.O.) são o principal meio utilizado por órgãos Municipais, Estaduais e Federais para tornar públicas suas ações e notificar cidadãos sobre assuntos de seu interesse em relação a esses órgãos. O portal da Imprensa Nacional, responsável pela publicação do Diário Oficial da União (D.O.U) descreve sua missão da seguinte forma:

> "A missão fundamental da Impressão Régia era, assim como é atualmente com a Imprensa Nacional, publicar os atos oficiais do Governo que se instalou no Rio de Janeiro em 7 de março de 1808."
> (Em: <http://portal.in.gov.br/ascom/imprensa1/a-imprensa-nacional>. Acesso em: 22 janeiro 2013.)

Os jornais são importantes para que todos tenham acesso às decisões tomadas pelos administradores públicos, tomando-se assim, ciência de todos os atos do governo, servindo como uma ferramenta oficial de transparência pública.

Existem três instâncias de Diários Oficiais, organizados de acordo com a abrangência de suas publicações,

1. **Diário Oficial da União**
   O Diário Oficial da União é conhecido através da sigla DOU e nele se encontram atos realizados e divulgados pelos administradores públicos federais.

2. **Diário Oficial do Estado**

   Possui a sigla DOE, e apresenta os atos realizados pelos administradores públicos estaduais.

3. **Diário Oficial do Município**

   E, por fim, o Diário Oficial do Município, que possui a sigla DOM e onde são divulgados os atos realizados pelos administradores públicos municipais.

1.1.2 PROBLEMA DE PESQUISA LEVANTADO

As informações publicadas nos Diários Oficiais são de interesse público, principalmente as que se encontram no Diário Oficial da União, já que as informações lá publicadas não se restringem ao interesse de população de uma cidade ou estado, mas sim ao interesse de toda a população do país.

No portal da Imprensa Nacional geralmente são realizadas três tipos de publicações, o Diário Oficial da União, os Suplementos, e o jornal do TRF Primeira Região, sendo o Diário Oficial da União normalmente dividido em sessões, que recebem o nome de DOU1, DOU2 e DOU3.

O DOU1 refere-se a leis, decretos, resoluções, instruções normativas, portarias, e outros atos normativos de interesse geral. O DOU2 refere-se a atos de interesse da administração pública federal. Por fim o DOU3 refere-se a contratos, editais e avisos ineditoriais.

A busca por uma informação nestas publicações não é uma tarefa trivial, pois a quantidade de páginas de cada um desses jornais ou sessões varia, e podem chegar a conter mais de 1000 páginas cada.

O site disponibiliza uma ferramenta de busca onde o usuário informa em um campo de texto a informação que deseja que seja pesquisada, a ferramenta, por sua vez, informa em qual página e jornal a informação foi encontrada. Apesar de eficaz para encontrar uma informação nas páginas das publicações, a ferramenta depende da solicitação do usuário. Sendo assim, se um determinado usuário deseja saber, por exemplo, se seu nome foi mencionado em algum

Diário Oficial ele deve acessar diariamente o sistema de busca do referido diário e fazer uma pesquisa por seu nome.

Essa característica acaba dificultando ou atrasando a tomada de ciência dos usuários sobre uma publicação de seu interesse, pois, ele pode, por algum motivo, estar indisponível para realizar a consulta ou esquecer-se de consultar o diário por um período de tempo.

Observando este cenário, torna-se clara a necessidade de um sistema que auxilie a população na recuperação das informações publicadas no Diário Oficial da União. Pretende-se desenvolver um sistema que será capaz de baixar diariamente as publicações do Portal da Imprensa Nacional, e realizar pesquisas nestas publicações. O sistema, que receberá o nome de Observador Diário, deverá disponibilizar um site onde o usuário será capaz de se cadastrar, e ao realizar o login no sistema, o usuário poderá realizar o cadastro dos textos que deseja que sejam pesquisados. De posse desses textos de pesquisa, o sistema poderá realizar buscas diariamente nas publicações baixadas e, caso seja encontrada alguma informação relevante na pesquisa, informar ao usuário através de seu e-mail cadastrado.

1.2 OBJETIVO GERAL

O objetivo deste trabalho é desenvolver um sistema que facilite o acesso da população a publicações de seu interesse no Diário Oficial da União, automatizando as pesquisas e realizando notificações, dessa forma, invertendo o fluxo de informações no acesso ao conteúdo.

1.3 OBJETIVOS ESPECÍFICOS

- Análise de bibliotecas e frameworks para a implementação de sistemas de recuperação da informação.
- Analisar técnicas de inversão de fluxo no acesso à informação.
- Identificar requisitos funcionais e não-funcionais para solução proposta.
- Desenvolvimento um sistema para o monitoramento do Diário Oficial da União.
- Avaliar o sistema desenvolvido através de um experimento prático.

1.4 METODOLOGIA

Para alcançar os objetivos será utilizada a metodologia a seguir:

- Análise do problema identificado no acesso à informação dos Diários Oficiais da União.
- Levantamento de informações sobre indexação e recuperação da informação.
- Definição dos requisitos funcionais e não-funcionais e dos de casos de uso.
- Definição do modelo arquitetural do sistema.
- Desenvolvimento do sistema.

1.5 ORGANIZAÇÃO DO TRABALHO

O restante do trabalho está organizado da seguinte forma:

- Capítulo 2 – Fundamentação Teórica.

  Este capítulo apresenta um breve resumo sobre sistemas de rastreamento web e de sistemas de indexação e recuperação da informação, conceitos estes que servem como base para a criação dos dois principais componentes do sistema proposto como solução para o problema levantado.

- Capítulo 3 – Análise do Cenário Atual e Definição da Solução

  Este capítulo analisa a atual forma de acesso ao Jornal do Diário Oficial da União e define uma solução para resolução do problema levantado.

- Capítulo 4 – Tecnologias

  Neste capítulo serão levantadas as necessidades tecnológicas do sistema para o seu desenvolvimento, levando em consideração que a linguagem escolhida para o desenvolvimento do sistema é a Java, pois e uma linguagem robusta, segura, independente de plataforma e possui frameworks e bibliotecas que atendem a todas as necessidades do sistema proposto como solução do problema de pesquisa levantado.

- Capítulo 5 – Desenvolvimento do Sistema

  Este capítulo apresenta os requisitos funcionais e não funcionais do sistema, o diagrama de caso de uso, a arquitetura completa do sistema e o design de alto nível dos componentes do sistema, acompanhados da descrição de seu funcionamento.

- Capítulo 6 – Avaliação do Sistema

  Este capítulo apresenta uma análise do sistema em desenvolvimento, apresentando seu estado atual e suas pendências, e o resultado apresentado pelo sistema após um experimento prático.

- Capitulo 7 – Conclusão

  Este capítulo apresenta as considerações finais sobre o trabalho, analisando o problema e a solução proposta, se os objetivos foram alcançados, problemas enfrentados durante a elaboração da solução e possíveis melhorias na solução do problema.

2 FUNDAMENTAÇÃO TEÓRICA

Este capítulo está dividido em dois subtópicos e apresenta um breve resumo explicativo sobre dois temas que servirão como base para a criação dos dois principais componentes do sistema proposto como solução.

2.1 SISTEMAS DE RASTREAMENTO WEB

Crawlers, que em inglês significam rastreadores, também conhecidos como Spiders ou Bots, são sistemas capazes de percorrer a rede formada pelos hyperlinks encontrados nas páginas da web. Ao acessar uma página da web, o Crawler, além de coletar o conteúdo considerado relevante, extrai os seus hyperlinks para serem acessados posteriormente, repetindo o mesmo processo para cada página acessada.

Segundo Aires (2005, p.166) um Crawler pode ser compreendido como um programa que visita cada página ou as páginas representativas de cada site da Web que deseja estar disponível para busca, e as "lê" utilizando os hiperlinks para descobrir o endereço de outras páginas.

Segundo (SILVA e MOURA, 2002, p.4) uma representação da coleta de páginas de um Web Crawler e sua arquitetura típica podem observados nas figuras 1 e 2 respectivamente.

O Web Crawler é o nome dado ao Crawler que tem como função suprir bancos de dados dos sistemas de busca da web.

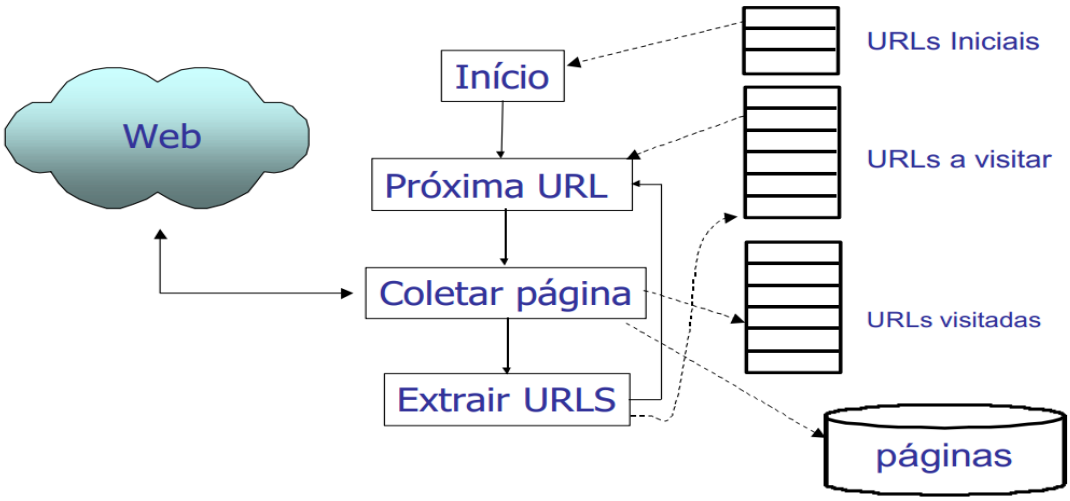


Figura 1 – Representação de um Web Crawler

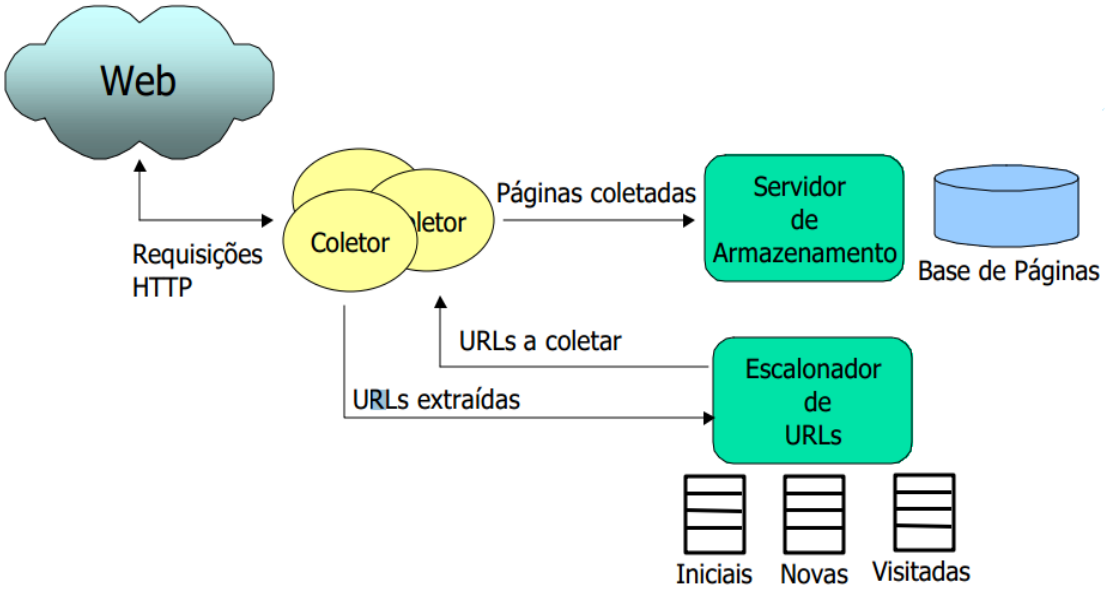


Figura 2 - Arquitetura Típica do Web Crawler

Um Web Crawler típico segue algumas convenções bem definidas. Inicialmente ele possui uma lista de URLs, de posse dessas URLs ele inicia o processo de visitas, ao visitar a primeira URL da lista ele realiza a coleta da página e a salva em um banco de dados, após realizar esse processo ele adiciona a URL na lista de URLs visitadas e em seguida extrai da página visitada as URLs dos hyperlinks para outras páginas. Esse processo é repetido utilizando as próximas URLs da lista de URLs iniciais, e ao chegar ao fim da lista, se inicia a visita das URLs contidas na lista de URLs a visitar.

Segundo Henrique (2011, p. 7), existem três tipos de componentes básicos em um Web Crawler, sendo eles o **Fetcher**, que é o responsável por coletar as páginas da web a partir de um conjunto de URLs iniciais e salvá-las em um repositório de páginas, o **Extrator**, que é responsável por extrair as URLs de dentro das páginas que foram adicionadas no repositório de páginas e por fim o **Verificador de Unicidade**, que realiza a unificação das URLs, ou seja, verificar se as URLs se repetem, e caso se repitam, exclui uma das URLs contidas no repositório de URLs.

Um sistema Crawler não se destina apenas a obtenção de conteúdos de páginas web, como é o caso dos Web Crawlers, podendo também ser projetado para um domínio específico, ou seja, o Crawler pode ser desenvolvido para atender uma necessidade específica de acordo com o propósito da aplicação na qual ele será inserido, como por exemplo, um Crawler pode ser utilizado em uma aplicação para monitorar uma página de vale refeição, e notificar o usuário, informando a data e o valor da disponibilização do próximo saldo, neste

caso o Crawler não teria a estrutura citada anteriormente no exemplo, mas sim apenas uma URL inicial, que seria da página que ele teria que monitorar.

Tomando como base, para um exemplo, o sistema proposto, será necessária a implementação de um Crawler que deverá acessar, durante os dias úteis, uma URL que direcionará o Crawler para a página, do Portal da Imprensa Nacional, que exibe as publicações do dia em que está sendo executado o acesso. Dessa página o Crawler deverá extrair informações que são relevantes para o sistema, como a quantidade de páginas que as publicações possuem e o nome das publicações que foram disponibilizadas neste dia. Estas informações são importantes, pois serão utilizadas, em um momento posterior, para a construção das URLs que servirão para realizar o download dos arquivos publicados. As URLs são comuns para todos os jornais, contendo apenas algumas divergências que são passadas por parâmetros, como o numero do jornal, que são números pré-definidos, como por exemplo, 1, 2, 3, 20, 1000, 1010, onde cada um destes números representa um dos tipos de jornais publicados, e o número das páginas dos jornais, que varia de 1 até o valor de páginas que a publicação possui, sendo assim, sabendo a quantidade de páginas de uma publicação, basta substituir o valor do número da página no parâmetro que será possível ter acesso a página do Portal da Imprensa Nacional que exibe esta página do jornal. Tendo acesso a estas páginas do Portal da Imprensa Nacional, o Crawler só precisa coletar a URL do arquivo PDF que será utilizada para realizar o download do arquivo, e utilizar o nome do arquivo coletado anteriormente para salvar o PDF.

2.2 SISTEMAS DE RECUPERAÇÃO DA INFORMAÇÃO

Estamos vivendo em meio à era da informação. Isso significa que nos dias atuais a informação é o “item” mais importante que podemos ter. Somos bombardeados por informações em todos os momentos, podemos perceber que até em casos específicos, encontrar uma informação relevante é algo difícil, por exemplo, é muito difícil encontrar um trecho específico de um texto que você leu há cerca de um ano em um livro guardado na estante ou encontrar a receita de um bolo em um livro que você leu há alguns meses.

Para resolver problemas como estes citados anteriormente, existem os sistemas de recuperação da informação.

Segundo (AIRES, 2005, p.8) a Recuperação da Informação (RI) é a tarefa de encontrar itens de informação relevantes para uma determinada necessidade de informação expressa pela requisição de um usuário através de consulta e disponibilizá-los de uma forma adequada a esta necessidade.

Ainda Segundo (AIRES, 2005, p.8) um RI pode ser classificado de acordo com o tipo de informação que ele recupera como mostra a tabela abaixo:

CLASSIFICAÇÃO DE UM SISTEMA DE RECUPERAÇÃO DA INFORMAÇÃO	
TIPO DO RI	**INFORMAÇÃO RECUPERADA**
RI Textual ou Documental	Recuperação de Texto
RI Visual	Recuperação de Imagens e Vídeos
RI de Áudio	Recuperação de Áudio
RI Multimídia	Recuperação de Dados Multimídia

Um sistema de recuperação da informação é um sistema capaz de realizar uma busca para um usuário e lhe trazer informações relevantes. Este tipo de sistema trabalha com o armazenamento e recuperação de objetos de dados, geralmente texto.

Segundo Gey, citado por (CARDOSO, 2000, p.1) um Sistema de Recuperação da Informação pode ser estruturado como mostra a figura 3.

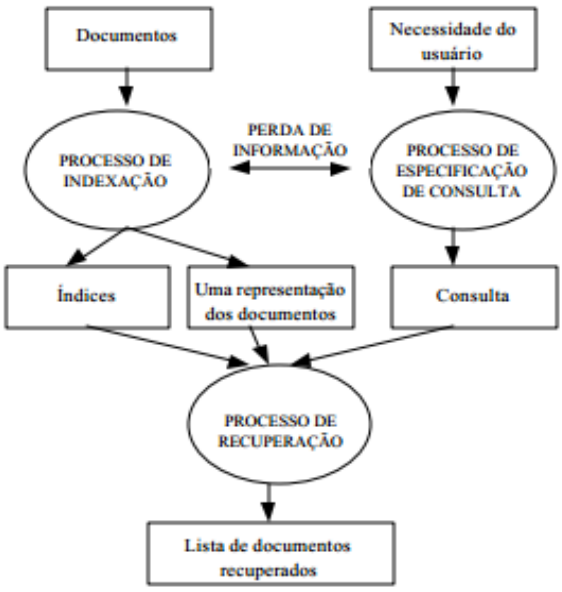


Figura 3 - Estrutura de um Sistema de Recuperação da Informação

Analisando esta estrutura podemos entender o funcionamento de um SRI. Inicialmente o sistema tem acesso a uma série de documentos que passam por um processo chamado de indexação.

A indexação é o processo pelo qual o sistema cria um índice baseado no conteúdo de arquivos de texto. Estes índices são criados através de técnicas como a do "índice invertido" que, segundo AIRES (2005, p.15), é uma técnica baseada em uma lista de palavras-chave onde para cada palavra-chave existem links indicando os documentos onde esta palavra está contida.

Segundo Ferneda (2003, p.16) durante a indexação são extraídos conceitos do documento através da análise de seu conteúdo e traduzidos em termos de uma linguagem de indexação, tais como cabeçalhos de assuntos, tesauros, etc. Esta representação identifica o documento e define seus pontos de acesso para a busca e pode também ser utilizada como seu substituto.

Ferneda (2003, p17) afirma que no processo de indexação automático é possível a utilização de filtros que eliminam palavras de pouca significação (stop words). Como exemplo, podemos utilizar, na língua portuguesa, filtros que ignorem as palavras como "o, a, em, de, para, por", pois são repetidas constantemente durante a criação de frases, e influenciariam negativamente na relevância do texto pesquisado. Os filtros também servem para normalizar os termos reduzindo-os a seus radicais em um processo conhecido como stemming.

Após o processo de indexação de arquivos, torna-se possível a realização de buscas em cima dos arquivos indexados. Para se realizar uma busca é necessário informar um parâmetro de busca, e este parâmetro passará por um processo de especificação da consulta, neste processo a consulta passa por uma filtragem, semelhante a que ocorre no processo de indexação, onde são eliminadas as palavras de pouca significação e ocorre a normalização dos termos.

Após a realização da busca, segundo AIRES(2005, p.12), costuma-se retornar um conjunto de citações de documentos relevantes para a consulta. As citações podem conter, por exemplo, títulos, nome de autores, trechos do texto que contém o termo da consulta, data em que o documento foi publicado, há quanto tempo o documento está disponível no sistema, resumo, localização física ou eletrônica do documento.

Podemos perceber que um sistema de recuperação da informação se encaixa como um componente necessário para o sistema proposto, pois servirá para indexar os documentos

baixados pelo Crawler e para realizar as consultas cadastradas pelo usuário, trazendo como retorno o documento e página onde o dado foi encontrado.

3 ANÁLISE DO CENÁRIO ATUAL E DEFINIÇÃO DA SOLUÇÃO

Este capítulo está dividido em dois subtópicos onde em **Diário Oficial da União** é feita uma análise da atual forma de acesso às publicações dos Diários Oficiais no Portal da Imprensa Nacional e em **Definição de Solução para o Problema Levantado** é proposta uma solução para o problema de pesquisa levantado.

3.1 DIÁRIO OFICIAL DA UNIÃO

Para chegarmos a uma solução viável para o problema apresentado, devemos entender os passos necessários para se ter acesso ao Diário Oficial da União no portal da Imprensa Nacional.

O Diário Oficial da União é publicado durante os dias úteis da semana. Para se ter acesso às publicações, na pagina inicial do portal, existe um formulário de busca chamado Leitura de Jornais, neste formulário existem algumas checkboxes, que são campos em formato de pequenas caixas que podem ser marcados clicando com o mouse em cima deles, estes campos devem ser selecionados para indicar sobre qual jornal o usuário deseja que seja feita a pesquisa, e entre as opções disponíveis estão o checkbox denominado “Todos”, que indica que a busca aconteça para todos os jornais, o “DOU1” que indica que a busca aconteça no Diário Oficial da União Sessão 1, assim como os checkboxes DOU2 e DOU3 indicando a sessão 2 e 3, o checkbox “DJUnico” que indica a busca do Diário da Justiça, porém, este está descontinuado desde 1º de janeiro de 2001, e por fim o checkbox eDJF1, que indica que a busca aconteça no Diário da Justiça Federal da 1ª Região. Também existe no formulário três campos de texto, onde podem ser digitada as datas do período em que deve ser feita a pesquisa, onde o campo “Data Início” indica que a pesquisa deve trazer os jornais publicados desde está “Data Início” informada, o campo de texto “Data Fim” indica que a pesquisa deve trazer jornais publicados até a “Data Fim” informada, e o campo “Ano” indica o ano das publicações que devem ser trazidas ao se realizar a pesquisa.

Ao realizar a consulta, o usuário é direcionado a uma página de resultado. Nesta página o usuário tem acesso a uma tabela informando o nome do jornal, a sua edição, a data de

disponibilização, a data de publicação e, por fim, a quantidade de páginas que o jornal contém.

Ao clicar em qualquer desses campos, o usuário é direcionado para a página de exibição do jornal escolhido. Nesta tela é exibida apenas uma página do jornal por vez, e o jornal pode ser percorrido a gosto do usuário.

3.2 DEFINIÇÃO DE SOLUÇÃO PARA O PROBLEMA LEVANTADO

Analisando esse cenário, percebe-se que a primeira necessidade do sistema é possuir um mecanismo capaz de encontrar e descarregar os arquivos PDFs publicados no portal da Imprensa Nacional, pois, desta forma, o sistema terá acesso a qualquer informação que possa ser considerada relevante ao usuário.

A primeira necessidade do sistema se encaixa no conceito de Crawler, uma ferramenta capaz de percorrer páginas da web e extrair dela dados ou informações relevantes.

No caso do sistema proposto, será necessária a obtenção de informações das publicações do Diário Oficial da União durante os dias úteis da semana, dias estes em que as publicações são realizadas. O Crawler utilizará inicialmente uma URL obtida através da observação da URL gerada pelo formulário de consulta de jornais do Portal da Imprensa Nacional, chamado “Leitura de Jornais. Para a obtenção da URL inicial foi observado a URL gerada quando se selecionava, no formulário, o checkbox “Todos”, que indica que deseja que todos os jornais sejam retornados na consulta, e indicava-se nos campos de “Data Início” e “Data Fim” com o dia e o mês atual, e no campo “Ano” o ano atual. A URL gerada nesta consulta apresenta um padrão, onde os campos selecionados no formulário são passados através de parâmetros, dessa forma, o sistema poderá modificar estes parâmetros quando necessário, como os parâmetros de data, informando sempre a data atual ao acessar esta URL. Ao acessar a URL inicial o Crawler é direcionado para a página que exibe o resultado da consulta, apresentando as publicações realizadas entre a “Data Início” e “Data Fim” informadas. Nesta página o Crawler deve extrair algumas informações relevantes, como a o nome das publicações e a quantidade de páginas de cada uma delas e as URLs para acessar a página de exibição das páginas dos jornais. As URLs são padronizadas contendo parâmetros de número do jornal, que são números fixos e indicam qual jornal deverá ser exibido, e número da página do jornal que varia de 1 até o número de páginas que a publicação possui.

Acessando estas URLs e substituindo o número do jornal e de suas páginas, é possível ter acesso a todas as páginas de exibição das publicações, e assim, extrair delas o link do PDF

que será utilizado posteriormente para salvá-lo, com o nome do jornal que foi obtido anteriormente e o número da página indicado na URL.

A segunda necessidade do sistema é extrair o conteúdo dos PDFs para que seja possível realizar a recuperação da informação desejada pelo usuário.

Esta necessidade se enquadra no conceito de sistema recuperação da informação, onde o sistema, de posse dos documentos, realiza o processo de indexação, e dessa forma torna-se apto a realizar pesquisas sobre informações relevantes sobre os documentos indexados através de parâmetros de consulta.

O sistema proposto deverá ter acesso aos parâmetros de consulta cadastrados pelos usuários. A forma de cadastro destes parâmetros se dará através de um site, onde o usuário poderá se cadastrar informando alguns dados pessoais como e-mail, que será utilizado posteriormente para notificações e para a realização do login, a senha e seu nome. De posse desses parâmetros, o sistema deverá realizar consultas sobre os arquivos de índice, obter os resultados das consultas, gerar respostas, e caso as consultas retornem algum dado relevante, informar os usuários.

A forma de disponibilização da informação aos usuários ocorrerá através de um e-mail de notificação. O usuário também saberá se sua pesquisa foi encontrada acessando sua página inicial no site do sistema. Dessa forma o usuário terá acesso à informação que deseja sem ser necessário acessar diariamente o Portal da Imprensa Nacional para realizar consultas, resolvendo dessa forma o problema de pesquisa levantado, invertendo o fluxo da informação através da automatização das pesquisas.

Obtemos assim as três necessidades básicas do sistema, um Crawler para a obtenção das URLs e realização do download das publicações, um Indexador para indexar o conteúdo dos PDFs das publicações e realizar pesquisas sobre estes índices, e um Cliente de E-mail que servirá para notificar os usuários e para funções básicas do site do sistema, como confirmação de cadastro e recuperação de senhas esquecidas, porém, os três componentes necessitam trabalhar em conjunto para que a necessidade do usuário seja suprida, sendo assim, o sistema necessita de um componente que coordene as atividades principais dos três componentes citados, fazendo com os três trabalhem prestando serviço para este, que será o Gerenciador de Atividades.

Analisando o problema e a solução proposta, chegamos à conclusão de que para a resolução do problema de pesquisa levantado, o sistema deverá conter os seguintes componentes:

- Um componente Gerenciador de Atividades que será responsável por coordenar todas as atividades realizadas pelo Crawler, pelo Indexador e pelo Cliente de E-mail.
- Um Crawler que será responsável por, durante os dias úteis da semana, visitar o portal da Imprensa Nacional e baixar todos os arquivos do Diário Oficial da União publicados.
- Um Indexador que será responsável pela indexação da informação contida nos arquivos baixados pelo Crawler e pela realização pesquisas sobre os arquivos indexados.
- Um Banco de Dados que será responsável por armazenar os dados cadastrados pelos clientes através do site, como suas informações pessoais e as consultas que deseja que o sistema realize, sendo utilizado pelo Gerenciador de Atividades para obtenção das consultas, que serão utilizadas como parâmetro ao chamar o Indexador para realizar as pesquisas sobre os índices.
- Um Cliente de E-mail, que será utilizado pelo Gerenciador de Atividades para notificar os usuários quando uma informação relevante for encontrada, e também pelo site para a notificações como confirmação de cadastro e recuperação de senhas.
- Um Servidor Web para a disponibilização do sistema web.
- Um sistema web para a realização dos cadastros dos clientes e para o cadastro dos parâmetros de busca.

4 TECNOLOGIAS

Este capítulo é dividido em seis subtópicos, onde são apresentadas as necessidades tecnológicas de cada em componente do sistema. A linguagem de programação escolhida para o desenvolvimento do sistema foi a Java, pois já é uma linguagem consolidada no mercado, é considerada segura, robusta, independente de plataforma e apresenta bibliotecas e framworks que atendem a todas as necessidades da aplicação proposta como solução.

4.1 NECESSIDADES TECNOLÓGICAS DO CRAWLER

O Crawler é um sistema ou script que "lê" o conteúdo de uma página web e realiza alguma ação com estes dados obtidos, como armazená-los ou trata-los, porém, em um Crawler típico, a ação geralmente tomada é, a de obtenção de hyperlinks contidos nas páginas para armazená-los em uma lista de URLs que serão visitadas futuramente, e o armazenamento do conteúdo da página em um banco de dados.

No caso do sistema proposto, para que o Crawler funcione de forma satisfatória, ele deverá ler o conteúdo de alguns componentes específicos contidos nas páginas que ele irá visitar, como o conteúdo da tabela de resultados, onde são exibidos os resultados da consulta realizada inicialmente através de uma URL obtida através da observação da URL gerada pelo formulário de pesquisa de jornais, contido na página inicial do Portal da Imprensa Nacional. Outra atividade do Crawler é ler o conteúdo do componente que exibe o PDF na página de exibição das publicações, pois é através deste componente que o Crawler poderá extrair a URL que dará acesso ao PDF, utilizando-a posteriormente para realizar os downloads das publicações.

Para auxiliar nesta atividade, existe uma biblioteca Java chamada de JTidy. Esta biblioteca é capaz de analisar a sintaxe de uma página web e informar a qualidade da formatação de seu conteúdo, tornando assim o processo de correção do conteúdo HTML uma tarefa mais fácil.

Através de uma interface DOM, o JTidy é capaz de capturar e retornar o conteúdo de tags HTML especificas, facilitando assim o trabalho de leitura do conteúdo de uma página, extraindo mais facilmente as informações relevantes para o sistema.

> JTidy é uma versão Java do HTML Tidy, um verificador de sintaxe HTML e de agradável visualização do conteúdo. Como o seu primo não-Java, o JTidy pode ser usado como uma ferramenta de limpeza de HTML mal formatado ou defeituoso. Além disso o JTidy provê uma interface DOM para o documento que está sendo processado, que efetivamente lhe dá a possibilidade de usar o JTidy como um conversor DOM para o HTML do mundo Real.
>
> (Em: < http://jtidy.sourceforge.net/>. Acesso em: 23 janeiro 2013.)

Com a utilização do JTidy o Crawler será capaz de obter o conteúdo de tags específicas das páginas, incluindo as que contem dados específicos do jornais, como, nome e quantidade de páginas, que são informações importantes para a descarga dos arquivos.

4.2 NECESSIDADES TECNOLÓGICAS DO INDEXADOR

A indexação do conteúdo das publicações, obtidas pelo Crawler, é a tarefa mais importante de todo o sistema pois dela depende o seu bom funcionamento e o alcance do objetivo final, que é reconhecer, com segurança, se o conteúdo cadastrado para pesquisa pelo usuário foi encontrado em alguma das publicações, e caso tenha sido, notificar o usuário através de um e-mail e da exibição na página inicial do usuário ao acessar o site e realizar o login no sistema.

A linguagem Java dispõe de uma poderosa biblioteca disponibilizada pela Apache para a indexação e recuperação de informações chamada Lucene.

> "A biblioteca Apache Lucene é um motor de busca de texto de alto desempenho escrito inteiramente em Java. Esta tecnologia é adequada para quase todas as aplicações que precisam de uma completa busca de texto, especialmente multi-plataforma."
> (Em: <http://lucene.apache.org/core/>. Acesso em 24 jan. 2013.)

Com a utilização da biblioteca Lucene o sistema terá a capacidade de realizar a indexação e a busca do conteúdo das publicações baixadas pelo Crawler, porém, para realizar a indexação garantindo que as palavras serão indexadas corretamente, é necessário realizar um tratamento no conteúdo dos arquivos devido a forma como o texto das publicações são dispostos, um exemplo de tratamento é a remoção de palavras hifenizadas por quebras de linha, por exemplo, a palavra "exem- / -plo", também é necessário dispor todo o conteúdo de texto das publicações em uma única linha, o que ajuda no processo de recuperação de resumos do texto ao identificar uma ocorrência. O resumo citado se trata de um trecho do texto onde uma pesquisa cadastrada pelo usuário foi encontrada, as 10 palavras que vierem antes do conteúdo encontrado, e as 10 palavras posteriores ao conteúdo encontrado. Estes resumos serão enviados juntamente com o conteúdo do e-mail de notificação, informando,

desta forma, o arquivo e página onde o texto foi encontrado, e também os resumos encontrados na página citada.

Para realizar esta tarefa primeiramente é necessária a extração do conteúdo dos PDFs para serem posteriormente salvos em arquivos correspondentes no formato TXT.

A Apache também disponibiliza uma biblioteca capaz de extrair o conteúdo de texto de um arquivo PDF, esta biblioteca chama-se PDFBox.

> "A biblioteca Apache PDFBox é uma ferramenta Java de código aberto para se trabalhar com documentos PDF. Este projeto permite a criação de novo documentos PDF, manipulação de documentos existentes e possui a habilidade de extrair o conteúdo destes documentos. O Apache PDFBox também inclui vários utilitários em linha de comando."
> (Em: <http://pdfbox.apache.org/>. Acesso em 24 jan. 2013.)

Desta forma, o PDFBox se torna uma ferramenta perfeita para o trabalho em conjunto com a Lucene, permitindo assim a extração do conteúdo de texto dos arquivos em formato PDF para serem salvos em formato TXT, para que, posteriormente sejam tratados e indexados.

A biblioteca Lucene permite que o desenvolvedor indique o que deve ser salvo no arquivo de índice através da criação de um documento personalizado. Um documento de índice é formado por vários campos, e são nestes campos que o desenvolvedor indica se o dado deve ser mantido ou simplesmente indexado.

O armazenamento de um campo do documento serve para que, ao se realizar uma consulta, o resultado traga o dado que foi armazenado, por exemplo, em um processo de indexação de um livro, deve-se indicar que o nome do livro seja mantido, e que seu conteúdo seja indexado, mas não mantido, dessa forma, ao realizar uma consulta passando como parâmetro um texto contido no livro, o nome do livro será retornado pela busca da Lucene, tornando possível saber em qual livro o conteúdo foi encontrado, se o mesmo fosse feito com o conteúdo do livro, todo o texto do livro seria retornado na busca, o que geralmente não é algo desejado, pois, na maioria dos casos, a busca pelo nome do arquivo que contém o conteúdo buscado já é um resultado satisfatório.

O sistema proposto deverá criar dois campos no documento de indexação, um para o nome do arquivo, que deverá ser mantido, e outro para o conteúdo do documento, que deverá ser indexado. Como o Portal da Imprensa Nacional divide o conteúdo de suas publicações por

página, ao se realizar o download do arquivo PDF, o sistema adiciona o número da página ao nome do arquivo, sendo assim, não há necessidade de criar um campo no documento de índice para manter o número da página.

4.3 NECESSIDADES TECNOLOGICAS DO CLIENTE DE E-MAIL

Para o desenvolvimento do cliente de e-mail não há a necessidade de utilização de nenhuma biblioteca adicional, necessitando apenas da utilização da API JavaMail, já contida na JDK do Java.

4.4 NECESSIDADES TECNOLOGICAS DO GERENCIADOR DE ATIVIDADES

O Gerenciador de Atividades é o componente principal do sistema, tendo como responsabilidade coordenar as atividades do Crawler, Indexador e do Cliente de E-mail.

Além dessas atividades, também será responsável por acessar o banco de dados onde estarão cadastradas as informações dos usuários e de suas consultas.

Para realizar o acesso ao banco de dados torna-se interessante, em questão de desenvolvimento, a utilização do padrão JPA, que define um meio de mapeamento objeto/relacional. Diversos frameworks implementam este padrão, porém para o desenvolvimento do sistema foi escolhido o framework Hibernate, já consolidado no mercado como uma das melhores soluções como serviço de persistência para bancos de dados.

> Hibernate é um serviço de consulta e persistência Objeto/Relacional de alta-performance. A mais flexível e poderosa solução Objeto/Relacional do mercado. O Hibernate toma conta do mapeamento de classes Java para tabelas do banco de dados e de tipos de dados Java para tipos de dados SQL. Provê facilidades na consulta e recuperação de dados que significativamente reduz o tempo do desenvolvimento. Os objetivos do projeto do Hibernate é aliviar em 95% a tarefa do desenvolvedor na programação de dados comuns relacionados à persistência eliminando a necessidade do trabalho artesanal do processamento de dados usando SQL e JDBC. No entanto, diferentemente de muitas outras soluções de persistência, o Hibernate não esconde de você o poder do SQL e

> garante que seu investimento e conhecimento da tecnologia relacional será válido como sempre.
> (Em: <http://www.hibernate.org/about.html>. Acesso em 24 jan. 2013.)

Apesar da definição do hibernate contida na página citada acima, o Hibernate é um framework que realiza o mapeamento Objeto/Relacional, porém não realiza a persistência em banco de dados orientados a objetos, trabalhando apenas com banco de dados relacionais.

A escolha da utilização do Hibernate para o acesso aos dados do banco de dados não tem relacionamento com uma necessidade tecnológica específica do projeto, mais sim como uma alternativa a criação manual do acesso ao banco via JDBC. A utilização do Hibernate é um fator positivo na redução do trabalho de programação do desenvolvedor, e consequentemente do tempo para a conclusão do sistema.

4.5 NECESSIDADES TECNOLOGICAS DA APLICAÇÃO WEB

A aplicação web é uma das partes mais importantes do sistema, pois será através dela que haverá a interação com o usuário. É através da aplicação web que o usuário irá cadastrar os seus dados e também as consultas que serão pesquisadas pelo sistema.

O Java possui um framework bastante interessante para o desenvolvimento de aplicações Web, o Java Server Faces (JSF 2.0). O JSF é um Framework baseado no padrão de projeto MVC (Model View Controller) e possui componentes que auxiliam na criação da interface gráfica com o usuário.

Segundo (GAMMA et al, 1995, p. 20) a abordagem MVC é um padrão de projeto que divide em três tipos de objetos, o modelo, a visão e o controlador. O modelo é o objeto de aplicação, a visão é a apresentação na tela e o controlador é o que define a maneira como a interface reage de acordo com as entradas do mesmo. A MVC separa esses objetos para aumentar a flexibilidade e a reutilização.

O JSF possui duas implementações bem conhecidas, uma desenvolvida pela Sun, chamada de Majoha, e outra desenvolvida pela Apache, chamada de MyFaces, sendo possível desenvolver um projeto utilizando qualquer uma das duas implementações.

O JSF possui um conjunto de componentes visuais básicos na sua implementação de referência. Para acessar estes componentes utilizam-se duas bibliotecas de tags que devem ser

adicionadas na propriedade de namespace do documento HTML (HyperText Markup Language):

- **Biblioteca HTML**
  Responsável por representar vários elementos HTMLs.
  Geralmente utiliza-se o alias "h" na tag, como por exemplo, <h:div>, para acessar os componentes.
- **Biblioteca Core**
  Responsável por tarefas comuns no desenvolvimento de sistemas como internacionalização, validação e conversão de dados de entrada.
  Geralmente utiliza-se o alias "f" na tag, como por exemplo, <f:view>, para acessar os componentes.

O JSF possui uma especificação formal e segura que permite o desenvolvimento de ferramentas e componentes.

Um exemplo de criação de um componente para ser utilizado com o JSF seria a criação de um componente Calendário que poderia ser renderizado em uma página web utilizando uma tag especifica definida pelo desenvolvedor.

Existem diversas bibliotecas de componentes JSF disponíveis, como por exemplo, o Tomahawk, RichFaces e o PrimeFaces, que facilitam ainda mais a vida dos desenvolvedores dispondo de componentes gráficos prontos para serem utilizados na criação da interface com o usuário.

Outro ponto que deve ser observado no sistema proposto como solução é necessidade do usuário se cadastrar para, posteriormente, realizar o login no sistema e ter acesso a página onde poderá cadastrar suas consultas. Analisando esta necessidade percebe-se que não se deve manter o foco apenas no desenvolvimento das páginas web, mas também da implementação de um sistema de autenticação e de controle de acesso.

Para aplicações Java Web existe um poderoso framework que se encarrega em garantir um serviço de autenticação e autorização seguro, chamado de Spring Security.

> O Spring Security é um poderoso e altamente personalizável framework de autenticação e controle de acesso. Ele é, de fato, padrão para a segurança de aplicações baseadas no Spring.

> Em: <http://static.springsource.org/spring-security/site/>. Disponível em 28 jan. 2013.

Com a utilização do Spring Security será possível definir a quais páginas o usuário terá acesso na aplicação através de verificações de autorização do usuário, garantindo segurança para a aplicação, impedindo que usuários não autorizados tenham acesso a páginas restritas ou dados de outros usuários.

4.6 BANCO DE DADOS

Para o funcionamento do sistema proposto, é necessária a utilização de um banco de dados onde serão salvos os dados dos clientes.

Este banco será acessado pelo Gerenciador de Atividades, que tem como parte de suas funções a recuperação de dados dos usuários e suas consultas, que servirão como parâmetros para a consulta realizada pelo Indexador no arquivo de índice. Também será acessado pela Aplicação Web, que será o meio por onde ocorrerão os cadastros dos clientes e de suas consultas.

Para a aplicação proposta foi escolhida a utilização de um dos mais modernos e populares bancos de dados de código aberto disponíveis no mercado, o MySQL.

> O MySQL é o software de banco de dados de código aberto mais popular do mundo, com mais de 100 milhões deste software baixados ou distribuídos em toda a sua história. Com sua velocidade superior, confiabilidade e facilidade de uso, o MySQL se tornou a escolha preferida para a Web, Web 2.0, SaaS, ISV, companhias de telecomunicações e Gerentes de TI de previsões corporativas porque este elimina os maiores problemas associados com inatividade, manutenção e administração para aplicações online modernas.
> Em: <http://www.mysql.com/about/>. Acesso em: 27 jan. 2013.

5 DESENVOLVIMENTO DO SISTEMA

Este capítulo está dividido em três subtópicos, sendo o primeiro **Requisitos Funcionais e Não-Funcionais,** onde são definidos os requisitos funcionais e não funcionais do sistema, o segundo **Diagrama de Casos de Uso,** onde são definidos os casos de uso da aplicação web, e por fim, o terceiro **Arquitetura do Sistema e Design dos Componentes,** onde é apresentada a arquitetura do sistema e o design de alto nível de seus componentes, juntamente com suas descrições.

5.1 REQUISITOS FUNCIONAIS E NÃO-FUNCIONAIS

Este subcapítulo apresenta os requisitos funcionais e não-funcionais do sistema, um requisito funcional define uma função ou serviço do sistema ou de algum de seus componentes, já os requisitos não-funcionais definem restrições ou atributos de qualidade para o software, levando em consideração fatores como, por exemplo, segurança, precisão, usabilidade, performance, manutenabilidade, entre outros.

REQUISITOS FUNCIONAIS DO SISTEMA	
NOME	**DESCRIÇÃO**
Cadastrar Usuário	Cadastrar o usuário no sistema.
Recuperar Senha	Recuperar senha esquecida pelo usuário.
Efetuar Login	Realizar o login para acessar o sistema.
Editar Perfil	Editar as informações cadastradas pelo usuário.
Cadastrar Pesquisa	Cadastra uma pesquisa.
Excluir Pesquisa	Exclui uma pesquisa cadastrada.
Efetuar Logout	Realizar o logout para sair do sistema.
Enviar E-mail	Envia um e-mail para um destinatário.
Baixar PDFs	Baixa arquivos PDFs.
Buscar Texto	Busca um texto nos índices.
Executar Rotina de Busca e Notificação	Inicia o processo de consulta e posteriormente de notificação dos usuários.

REQUISITOS NÃO FUNCIONAIS DO SISTEMA
O usuário deve se autenticar para acessar o sistema.
Um usuário cadastrado só poderá acessar o sistema caso tenha confirmado seu cadastro.
A aplicação irá rodar em um ambiente Linux.
O sistema deverá garantir que os arquivos baixados não estão corrompidos.
O serviço deverá rodar em uma rede com disponibilidade para download de, no mínimo, 10Mbps.
Um e-mail cadastrado no sistema é único e não pode ser alterado, a não ser que seja por um administrador.
O Portal da Imprensa Nacional disponibiliza as publicações às 8 horas da manhã, dessa forma, o sistema deverá iniciar os downloads das publicações as 8:30, dando uma margem de segurança de 30 minutos.

5.2 DIAGRAMA DE CASOS DE USO

Um diagrama de casos de uso tem como objeto auxiliar a comunicação entre os analistas e clientes descrevendo um cenário que mostra as funcionalidades do sistema do ponto de vista do usuário.

Os casos de uso da Aplicação Web podem ser visualizados como mostra a figura 4.

Aqui será apresentado apenas o caso de uso da Aplicação Web devido ao fato de não existir interação do usuário com os outros componentes do sistema.

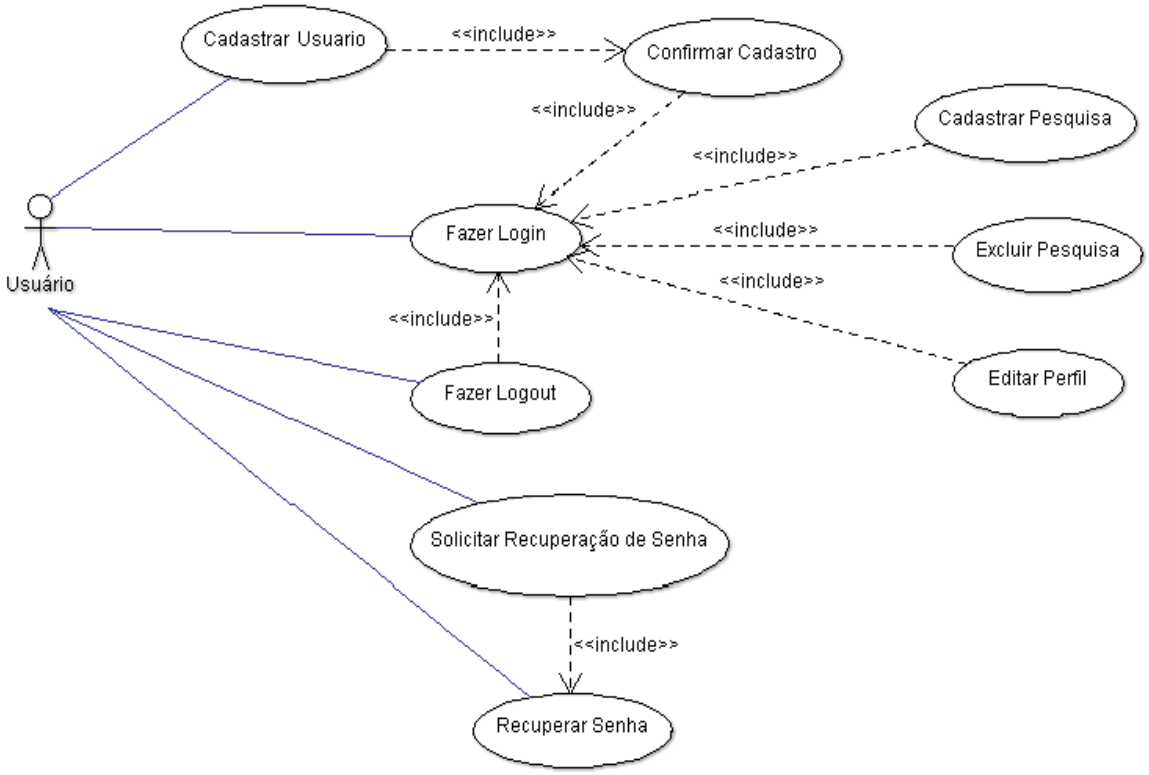


Figura 4 - Diagrama de Caso de Uso da Aplicação Web

5.2.1 DESCRIÇÃO DOS CASOS DE USO

5.2.1.1 CADASTRAR USUÁRIO

Projeto: Observador Diário

Finalidade: Cadastrar o usuário no sistema.

Atores: Usuário

Evento Inicial: O usuário acessa a página inicial do sistema.

Fluxo Principal:

- **a.** O sistema disponibiliza na interface o formulário para a realização do cadastro.
- **b.** O usuário preenche os campos do formulário.
- **c.** O usuário clica no botão cadastrar.
- **d.** O sistema valida os dados informados.
- **e.** O caso de uso é encerrado.

Fluxo Alternativo:

Não contém.

Fluxo de Exceção:

E1 – O usuário não preenche um e-mail válido.

- **a.** O usuário clica no botão cadastrar.
- **b.** O sistema valida os dados informados.
- **c.** O sistema exibe na tela a mensagem de alerta informando que o e-mail está inválido.
- **d.** O caso de uso é encerrado.

E2 – O usuário não preenche um ou todos dos campos.

- **a.** O usuário clica no botão cadastrar.
- **b.** O sistema valida os dados informados.
- **c.** O sistema exibe na tela a mensagem de alerta informando que o campo não informado ou todos os campos devem ser informados.
- **d.** O caso de uso é encerrado.

5.2.1.2 CONFIRMAR CADASTRO

Projeto: Observador Diário
Finalidade: Confirmar o cadastro do usuário no sistema.
Atores: Usuário
Evento Inicial: O usuário recebe um e-mail contendo um código e um link para a página de confirmação de cadastro.

Fluxo Principal:

- **a.** O usuário acessa o link recebido por e-mail.
- **b.** O sistema exibe um formulário com os campos e-mail e código de confirmação.

c. O usuário informa seu e-mail cadastrado e o código recebido por e-mail.

d. O usuário clica no botão confirmar.

e. O sistema valida os dados informados no formulário.

f. O sistema exibe na tela uma mensagem de sucesso na confirmação.

g. O caso de uso é encerrado.

Fluxo Alternativo:

Não contém.

Fluxo de Exceção:

E1 – Código de confirmação errado.

a. O usuário executa até o passo "b" do fluxo principal.

b. O usuário informa seu e-mail e um código de confirmação errado.

c. O sistema valida os dados informados.

d. O sistema exibe uma mensagem informando que o código informado não é válido.

e. O caso de uso é encerrado.

E2 – E-mail errado.

a. O usuário executa até o passo "b" do fluxo principal.

b. O usuário informa um e-mail errado e o código de confirmação.

c. O sistema valida os dados informados.

d. O sistema exibe uma mensagem informando que o e-mail informado não é válido.

e. O caso de uso é encerrado.

E3 – E-mail e código de confirmação errados.

a. O usuário executa até o passo "b" do fluxo principal.

b. O usuário informa um e-mail e um código de confirmação errados.

c. O sistema valida os dados informados.

d. O sistema exibe uma mensagem informando que o e-mail informado não é válido.

e. O caso de uso é encerrado.

E4 – Não preenche todos os campos ou um dos campos em branco.

a. O usuário executa até o passo "b" do fluxo principal.

b. O usuário não informa nenhum dos campos.

c. O sistema valida os dados informados.

d. O sistema exibe uma mensagem informando que todos os devem ser preenchidos.

e. O caso de uso é encerrado.

5.2.1.3 FAZER LOGIN

Projeto: Observador Diário

Finalidade: Realizar o login para acessar o sistema.

Atores: Usuário

Evento Inicial: O usuário acessa a página inicial do sistema.

Fluxo Principal:

a. O sistema disponibiliza na interface os campos de login e senha.

b. O usuário informa o e-mail cadastrado e sua senha.

c. O usuário clica no botão Login**.**

d. O usuário acessa o sistema.

e. O caso de uso acaba.

Fluxo Alternativo:

Não possui.

Fluxo de Exceção:

E1 – Logar sem ter realizado o cadastro ou com e-mail ou senha errados.

a. O usuário informa um e-mail e senha não cadastrados.

b. O usuário clica no botão Login.

c. O sistema valida os dados.

d. O usuário retorna para a página inicial.

e. O caso de uso termina.

E3 – Logar sem preencher os campos de e-mail e senha.

a. O usuário não informa os campos de e-mail e senha.

b. O usuário clica no botão Login.

c. O sistema valida os dados.

d. O usuário retorna para a página inicial.

e. O caso de uso termina.

5.2.1.4 CADASTRAR PESQUISA

Projeto: Observador Diário

Finalidade: Cadastrar uma pesquisa no sistema.

Atores: Usuário

Evento Inicial: O usuário acessa a página inicial do sistema e realiza o login.

Fluxo Principal:

a. O sistema exibe na sua interface o formulário para cadastrar uma pesquisa.

b. O usuário escreve a pesquisa.

c. O usuário clica no botão cadastrar.

d. O sistema valida o campo.

e. O sistema retorna uma mensagem de sucesso.

f. O caso de uso termina.

Fluxo Alternativo:

Não possui.

Fluxo de Exceção:

E1 – Cadastro sem escrever pesquisa.

- **a.** O usuário não escreve uma pesquisa.
- **b.** O usuário clica no botão cadastrar.
- **c.** O sistema valida o campo.
- **d.** O sistema retorna uma mensagem indicando que a pesquisa deve ser informada.
- **e.** O caso de uso termina.

5.2.1.5 EXCLUIR PESQUISA

Projeto: Observador Diário
Finalidade: Excluir uma pesquisa do sistema.
Atores: Usuário
Evento Inicial: O usuário acessa a página inicial do sistema, realiza o login, o sistema exibe na interface o formulário de cadastro de pesquisas, caso não exista uma pesquisa cadastrada, o usuário cadastra uma pesquisa.

Fluxo Principal:

- **a.** O sistema exibe na interface uma tabela informando as pesquisas cadastradas.
- **b.** O usuário clica no botão excluir contido na tabela.
- **c.** O sistema exibe uma mensagem de sucesso.
- **d.** O caso de uso termina.

Fluxo Alternativo:

Não possui.

Fluxo de Exceção:

Não possui.

5.2.1.6 EDITAR PERFIL

Projeto: Observador Diário

Finalidade: Atualizar os dados pessoais cadastrados pelo usuário.

Atores: Usuário

Evento Inicial: O usuário acessa a página inicial do sistema, realiza o login, o sistema exibe na interface o link para a página de edição de perfil, o usuário clica no link Editar Perfil.

Fluxo Principal:

- **a.** O sistema exibe na interface o formulário para a edição dos dados pessoais.
- **b.** O usuário edita seus dados.
- **c.** O usuário clica no botão atualizar.
- **d.** O sistema valida os dados.
- **e.** O sistema exibe na interface uma mensagem de sucesso.
- **f.** O caso de uso termina.

Fluxo Alternativo:

A1 – O usuário não altera os dados.

- **a.** O usuário executa até o passo "a".
- **b.** O usuário não altera seus dados.
- **c.** O usuário clica no link principal.
- **d.** O sistema exibe na interface a página principal.
- **e.** O caso de uso termina.

Fluxo de Exceção:

A2 – Alterar dados deixando um ou mais campos em branco.

- **a.** O usuário executa até o passo "a".
- **b.** O usuário altera um ou mais dados deixando em branco.

c. O usuário clica em editar.

d. O sistema valida os dados.

e. O sistema exibe na interface uma mensagem informando que todos os campos devem ser preenchidos.

f. O caso de uso acaba.

5.2.1.7 SOLICITAR RECUPERAÇÃO DE SENHA

Projeto: Observador Diário

Finalidade: Solicita a recuperação de uma senha perdida ou esquecida pelo usuário.

Atores: Usuário

Evento Inicial: O usuário acessa a página inicial do sistema.

Fluxo Principal:

a. O sistema exibe na interface um link chamado “Esqueceu a Senha?”.

b. O usuário acessa o link “Esqueceu a Senha”.

c. O sistema exibe na interface o formulário de recuperação de senha.

d. O usuário escreve seu e-mail no campo de e-mail.

e. O usuário clica no botão “Recuperar Senha”.

f. O sistema exibe mensagem de sucesso.

g. O caso de uso acaba.

Fluxo Alternativo:

Não possui.

Fluxo de Exceção:

E1 – Não preenche o e-mail no campo do formulário.

a. O usuário segue até o passo “c”.

b. O usuário não preenche o campo de e-mail.

c. O usuário clica no botão “Recuperar Senha”.

d. O sistema valida os dados.

e. O sistema envia uma mensagem de e-mail com um link para a recuperação

f. O sistema exibe mensagem informando que o campo deve ser inserido.

g. O caso de uso acaba.

5.2.1.8 RECUPERAR SENHA

Projeto: Observador Diário

Finalidade: Recuperar uma senha perdida ou esquecida pelo usuário.

Atores: Usuário

Evento Inicial: O usuário acessa o e-mail recebido ao solicitar a recuperação de senha e acessa o link contido no e-mail.

Fluxo Principal:

a. O sistema exibe na interface o formulário de definição de nova senha.

b. O usuário preenche o formulário com seu e-mail e sua nova senha.

c. O usuário clica em "Alterar".

d. O sistema valida os dados.

e. O sistema exibe na interface uma mensagem de sucesso.

f. O caso de uso acaba.

Fluxo Alternativo:

Não possui.

Fluxo de Exceção:

E1 – Não preenche um ou mais campos do formulário.

a. O usuário segue até o passo "a".

b. O usuário não preenche um ou mais campos do formulário.

c. O usuário clica no botão "Alterar".

d. O sistema valida os dados.

e. O sistema exibe, na interface, uma mensagem indicando que todos os campos devem ser preenchidos.

f. O caso de uso acaba.

E2 – E-mail errado.

a. O usuário segue até o passo "a".

b. O usuário informa o e-mail errado.

c. O usuário clica no botão "Alterar".

d. O sistema valida os dados.

e. O sistema exibe, na interface, uma mensagem indicando que o e-mail informado não é valido para a alteração de senha.

E3 – Senhas diferentes.

a. O usuário segue até o passo "a".

b. O usuário informa o campo senha diferente do campo senha de confirmação.

c. O usuário clica no botão "Alterar".

d. O sistema valida os dados.

e. O sistema exibe, na interface, uma mensagem indicando que os campos senha e senha de validação apresentam senhas diferentes.

5.2.1.9 FAZER LOGOUT

Projeto: Observador Diário

Finalidade: Realizar o logout para sair do sistema.

Atores: Usuário

Evento Inicial: O usuário acessa a página inicial do sistema e realiza o login.

Fluxo Principal:

f. O sistema disponibiliza na interface os principal do usuário logado.

g. O usuário clica no link logout.

h. O sistema retorna para a página inicial.

i. O caso de uso acaba.

Fluxo Alternativo:

Não possui.

Fluxo de Exceção:

Não possui.

5.3 ARQUITETURA DO SISTEMA E DESIGN DOS COMPONENTES

Este subcapítulo apresenta a arquitetura do sistema como um todo e os designs de alto nível de seus componentes.

5.3.1 ARQUITETURA DO SISTEMA

A arquitetura completa do sistema proposto pode ser compreendida como mostra a figura 5.

Analisando esta arquitetura podemos perceber que ela é dividida em três camadas, a de cliente, servidor de aplicação e de dados, e o o sistema é composto por cinco componentes principais, o Gerenciador de Atividades, o Cliente de E-mail, o Crawler, o Indexador e a Aplicação Web.

O Gerenciador de Atividades, como o nome já diz, gerencia as atividades que deverão ser realizadas, tanto por ele próprio quanto pelos três componentes que o compõe. O Cliente de E-mail, o Crawler e o Indexador trabalham como prestadores de serviços para o Gerenciador de Atividade.

O Gerenciador de Atividades tem acesso direto ao Banco de Dados de onde obtém os dados dos usuários e de suas consultas cadastradas, que são utilizados para realizar buscas sobre os índices criados pelo Indexador. Ele também possui o papel de gravar dados obtidos através das buscas no banco, como as ocorrências de uma busca e resumos das ocorrências.

O Crawler é utilizado para realizar a coleta de dados e download dos documentos do Diário Oficial da União publicados no portal da Imprensa Nacional.

O Indexador tem o papel de extrair o conteúdo das publicações baixadas pelo Crawler e gravá-los em arquivos de texto correspondentes a cada uma das publicações, realizar a adaptação do conteúdo dos arquivos de texto extraídos colocando o texto em linha e corrigindo palavras hifenizadas por quebras de linha, realizar a indexação dos arquivos de texto adaptados, criando arquivos de índice sobre o conteúdo dos documentos, e por fim utilizar o índice gerado para a realização das consultas cadastradas pelos usuários.

Ao termino de todo o processo de coleta de dados e download, indexação de documentos e recuperação de dados dos usuários, o Gerenciador de Atividades gera uma mensagem de e-mail para cada usuário que teve uma busca realizada trazendo informações relevantes, este e-mail serve para notificar o usuário que sua pesquisa cadastrada foi realizada e que houve sucesso na busca, informando-o qual pesquisa cadastrada foi encontrada, qual a página onde foi encontrado o parâmetro da pesquisa, e um resumo do trecho do texto da página que contém a pesquisa cadastrada.

O Cliente de E-mail é utilizado pelo Gerenciador de Atividades para enviar as mensagens de e-mail geradas a seus destinatários.

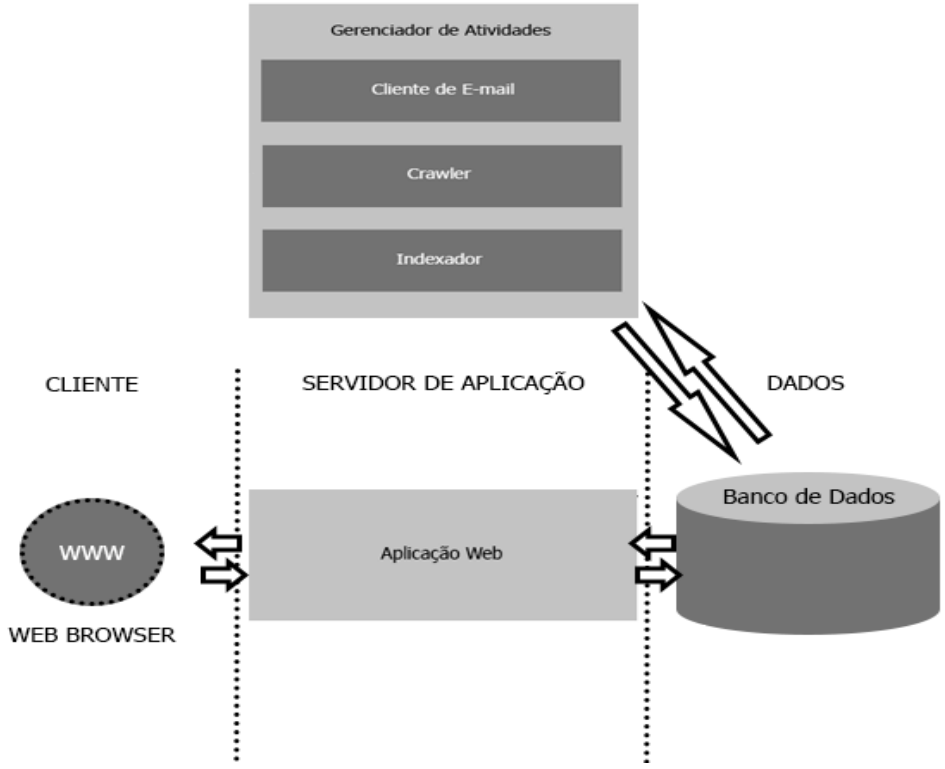


Figura 5 - Arquitetura Completa do Sistema

5.3.2 DESIGN DE ALTO NÍVEL DO CRAWLER

O projeto de alto nível do Crawler para a solução proposta pode ser compreendida como mostra a figura 6.

Para o funcionamento do Crawler proposto, é necessário a informação apenas de uma URL inicial, esta URL foi obtida através da observação da URL gerada pelo formulário de busca disponibilizado pelo portal da Imprensa Nacional para se ter o acesso as publicações do Diário Oficial da União.

De posse desta URL, se dá o início do processamento do Crawler, que por sua vez, modifica os parâmetros da data de início, do data do fim, e do ano da consulta na URL informada de modo que a consulta traga os dados de acordo com o dia em que ele está sendo executado.

Após acessar a URL e ser direcionado para a página de resultados, o Crawler realiza uma coleta de informações sobre as publicações, como número de páginas e nome da publicação, e após esta coleta obtém as URLs referentes às páginas de exibição do conteúdo das publicações. Ao obter o conteúdo desta página, o Crawler terá acesso a URL de do arquivo PDF da página, que servirá para a realização do download. Este processo na imagem está representado apenas por um quadro chamado de “Coleta de Dados e Obter URL Download”.

Após a obtenção das URLs de Download das publicações, o Crawler gera uma lista de URLs de download, trocando na URL coletada o parâmetro da página, gerando uma URL para cada página da publicação.

O download da lista de URLs de download geradas ocorre através da utilização de um Pool de Threads, onde para cada download é gerada uma Thread especifica.

A quantidade de Threads do Pool de Threads é fixa em 10, para evitar a sobrecarga nos acessos ao sistema do Portal da Imprensa Nacional.

A utilização de Threads para a realização do download torna o processo de download mais rápido, devido ao fato de ocorrerem 10 dawnloads simultâneos em vez de apenas um.

No processo download, os documentos pdfs são salvos temporariamente no disco rígido, e após o termino deste processo a atividade do Crawler chega ao fim.

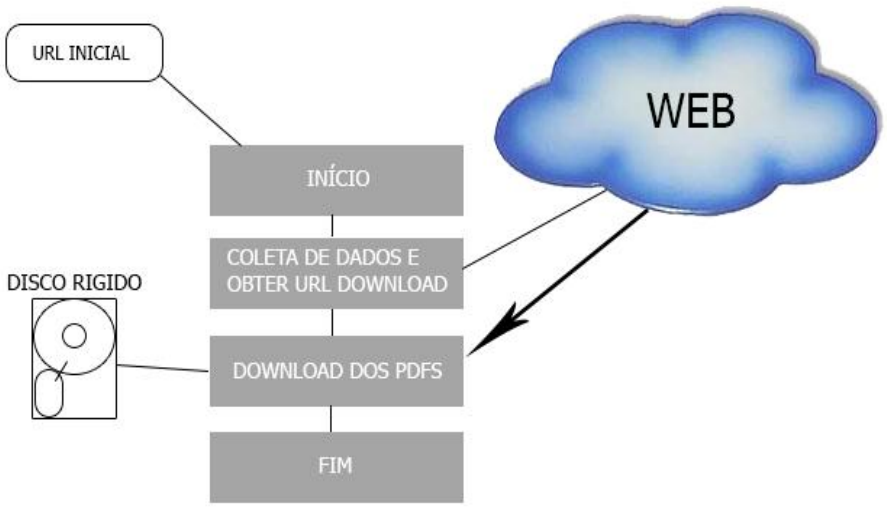


Figura 6 - Design de Alto Nível do Crawler

5.3.3 DESIGN DE ALTO NÍVEL DO INDEXADOR

A arquitetura do Indexador para a solução proposta pode ser compreendida de acordo com a figura 7.

Pode-se considerar que o Indexador proposto é subdividido em três componentes:

- **Extrator TXT:**

  Tem a responsabilidade de extrair dos documentos PDFs o seu conteúdo e salvar em documentos correspondentes em formato TXT.

- **Adaptador TXT:**

  Tem como responsabilidade tratar os textos extraído do PDFs, onde todo o texto é colocado em apenas uma linha e as palavras hifenizadas por quebra de linha são tratadas, removendo o hífen e unindo as duas partes das palavras. Este processo é útil para a realização da indexação e recuperação de resumos do texto, que serão enviados no e-mail de notificação do usuário, quando uma pesquisa retornar dados relevantes.

- **Indexador Textual:**

  Tem como responsabilidade indexar o conteúdo do texto contido nos arquivos de texto extraídos pelo extrator e salvar um arquivo de índice.

- **Buscador Textual:**

  Tem como responsabilidade realizar as buscas de acordo com as consultas cadastradas pelos usuários. O Gerenciador de Atividade é o responsável por passa as consultas por parâmetro para o buscador textual.

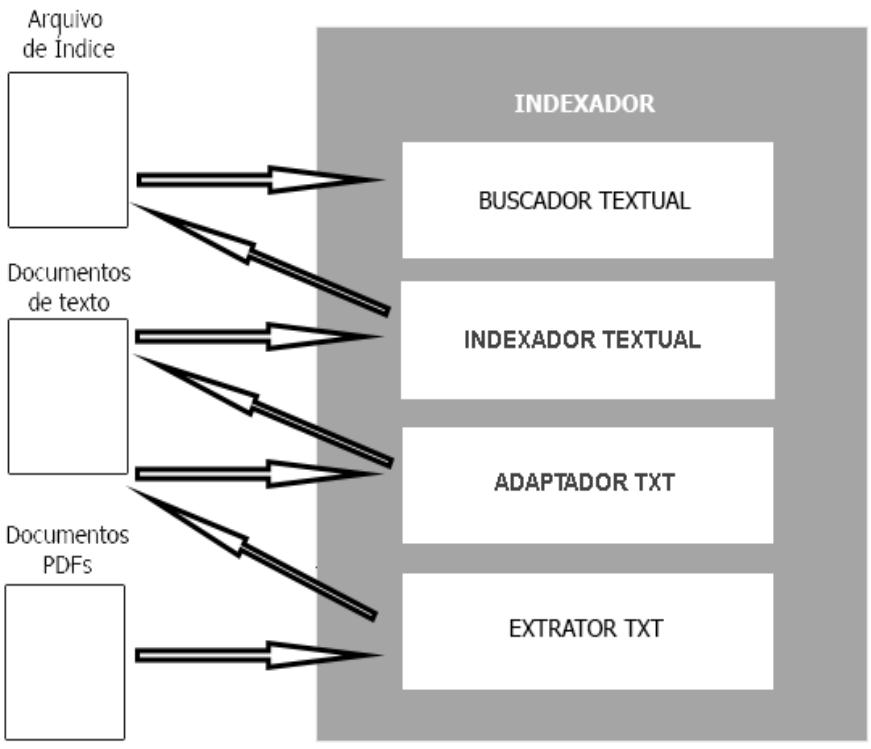


Figura 7 - Design de Alto Nível do Indexador

5.3.4 DESIGN DE ALTO NÍVEL DO CLIENTE DE E-MAIL

A arquitetura do Cliente de E-mail para a solução proposta pode ser compreendida de acordo com a figura 8.

O Cliente de E-mail inicialmente tem acesso a um arquivo de configuração, neste arquivo contém todos os dados necessários para que o cliente se conecte ao servidor de e-mail e possa realizar o envio das mensagens geradas pelas aplicações para quem ele serve.

O Cliente de E-mail presta serviço para a Aplicação Web e para o Gerenciador de Atividades.

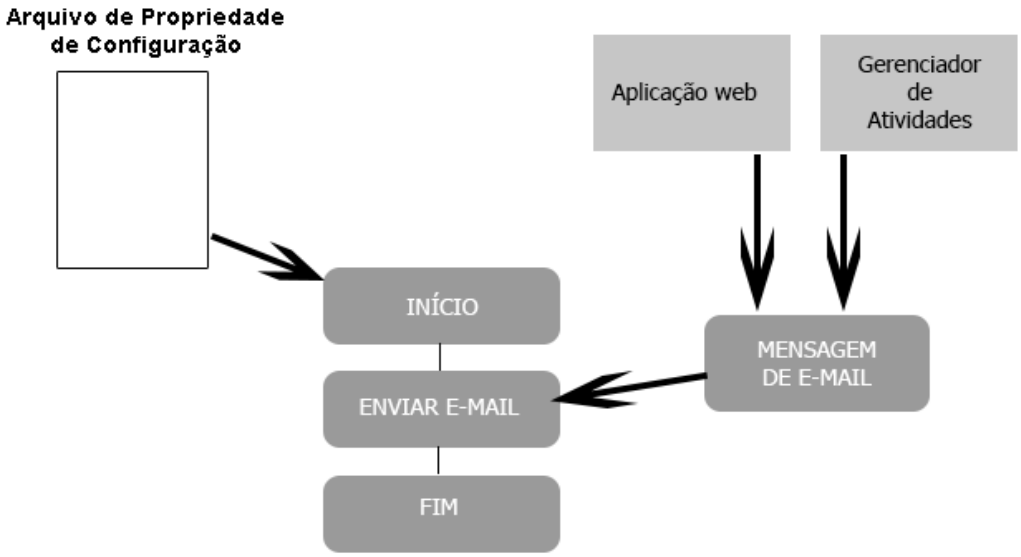


Figura 8 - Design de Alto Nível do Cliente de E-mail

5.3.5 O SISTEMA WEB

O Sistema Web será construído em cima da arquitetura do JSF, dessa forma, a Aplicação Web segue todo o padrão imposto pela arquitetura do JSF como, separação clara dos elementos de lógica e visão, não utilizando scriptlets, elemento comum no jsp que possibilita a inserção de código java dentro das páginas html, uso do modelo MVC2 com um servlet controlador que intercepta as requisições, e baseia-se no ciclo de vida do JSF.

A Aplicação Web utiliza o componente Cliente de E-mail, pois ele é necessário para o envio do e-mail de confirmação de cadastro, recuperação de senha esquecida.

6 AVALIAÇÃO DO SISTEMA

Este capítulo apresenta uma análise do sistema em desenvolvimento, apresentando seu estado atual e suas pendências, e o resultado apresentado pelo sistema após um experimento prático.

6.1 ESTADO ATUAL E PENDÊNCIAS DO SISTEMA PROPOSTO

O sistema proposto está atualmente em fase de conclusão. Neste momento todos os componentes do sistema já foram desenvolvidos e estão totalmente funcionais.

O sistema também necessita de um refatoramento de código no componente de Gerenciamento de Atividades, pois apesar de estar funcionando como esperado o código pode ser melhorado tornando-se melhor manutenível e inteligível.

Quanto a Aplicação Web foram criadas as páginas de acesso público e de acesso restrito, ou seja, acesso apenas por usuários cadastrados, faltando assim a criação das páginas administrativas do sistema, que serão acessadas pelos usuários administradores.

Ainda é necessária à criação de algumas páginas estáticas, como as que compõem os links da barra inferior do site, também é necessária a criação de um sistema de feedback para que o usuário tenha um meio de entrar em contato com os administradores do sistema.

Para tornar ainda mais prático a utilização do sistema, ainda será desenvolvido um componente que tornará possível o cadastro de usuário e consultas através da rede social Twitter. Através deste componente, qualquer usuário da rede social que siga o perfil do sistema será capaz de, através de um twitter, realizar seu cadastro e de sua consulta, sendo avisado posteriormente se sua consulta foi encontrada através de uma mensagem direta (DM).

6.2 EXPERIMENTO PRÁTICO

Neste subcapítulo será apresentado um experimento prático do sistema e, para uma melhor visualização serão apresentados printscreens mostrando o fluxo de teste.

Como experimento prático será percorrido todo o fluxo do sistema, desde o cadastro do usuário até a execução do Gerenciador de Atividades que será responsável por coordenar o as atividades do Crawler, Indexador e realiza a notificação utilizando o Cliente de E-mail caso alguma consulta gere retorno.

A imagem abaixo representa a tela inicial da Aplicação Web. Através dela será cadastrado um usuário de teste que receberá como nome “Usuário” e sobrenome “de Teste do Sistema”.

Figura 9 - Tela Inicial da Aplicação Web

Após a realização do cadastro é exibida uma tela que informa ao usuário que o cadastro foi realizado com sucesso e que um e-mail de confirmação foi enviado para o e-mail informado no cadastro, posteriormente a página é redirecionada novamente para a página principal.

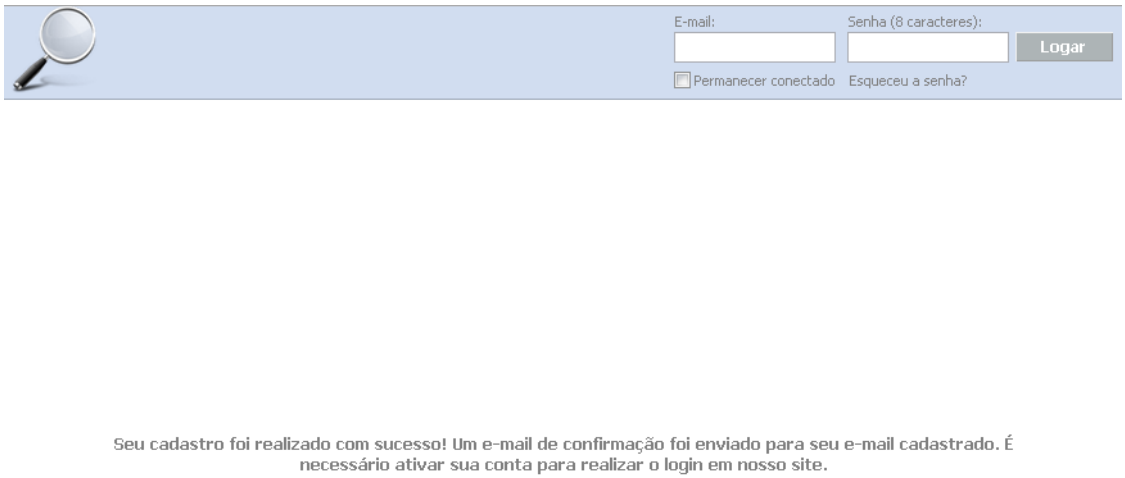


Figura 10 - Tela de Confirmação de Cadastro da Aplicação Web

Após a realização do cadastro é necessário acessar o e-mail informado para ter acesso ao link de confirmação de cadastro e ao código de confirmação. A imagem abaixo exibe o conteúdo do e-mail enviado ao usuário.

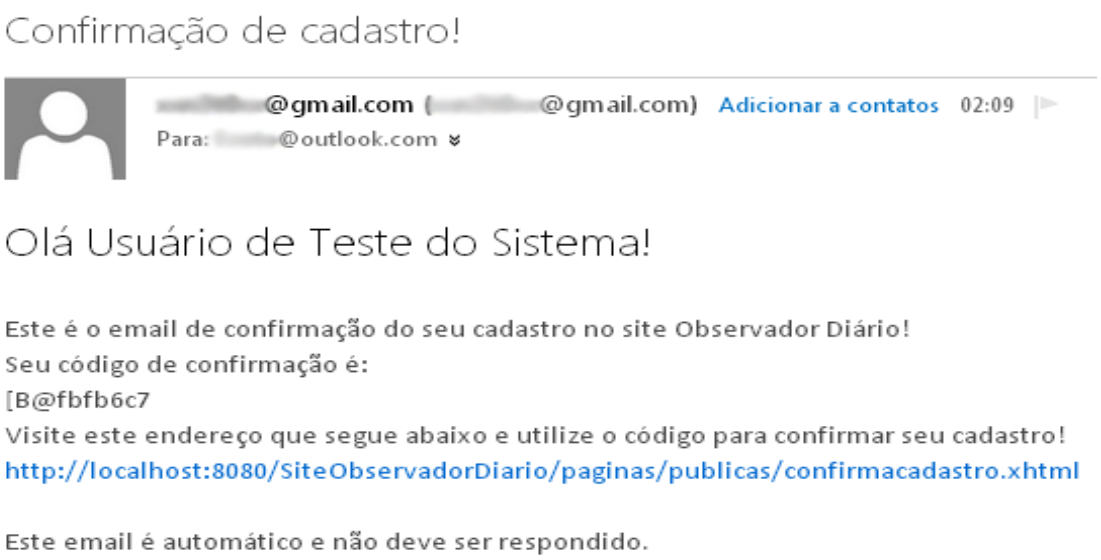


Figura 11 – Conteúdo do E-mail de Confirmação de Cadastro

Ao acessar a tela de informação de cadastro, deve-se passar o e-mail e o código de confirmação informado.

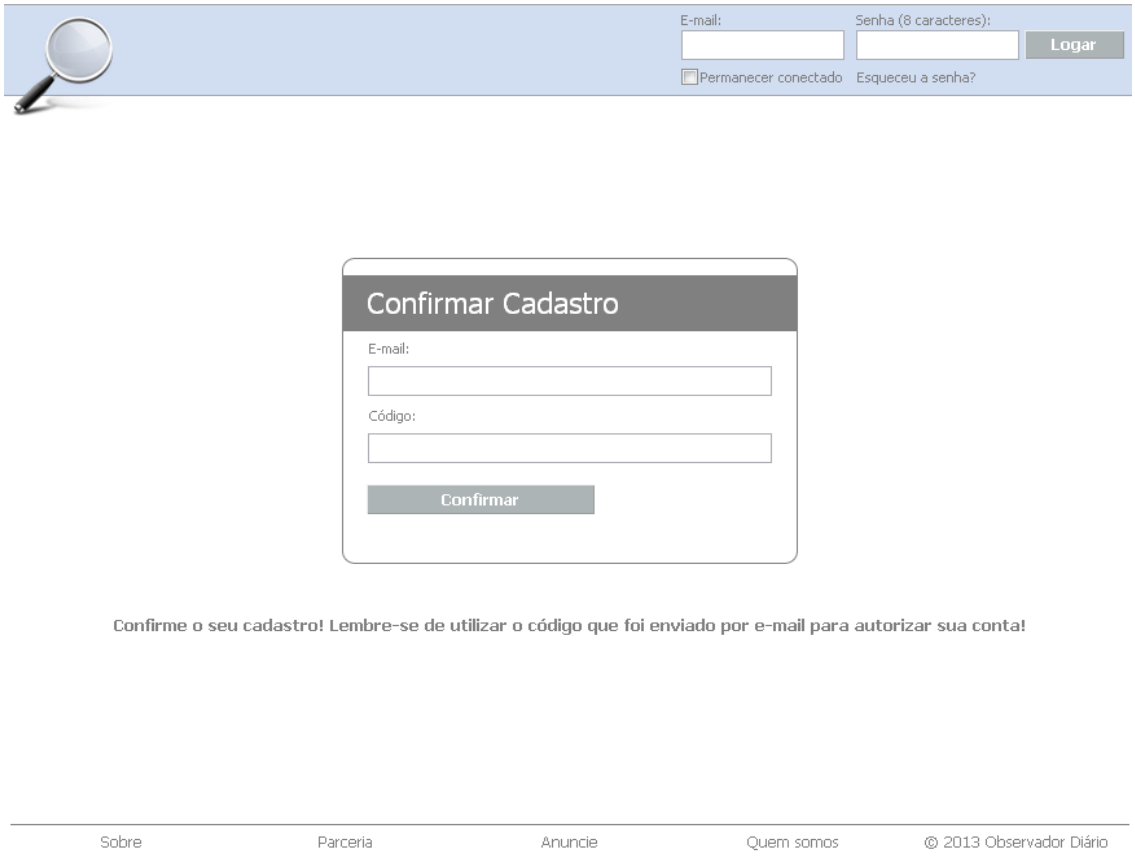


Figura 12 - Tela de Confirmação de Cadastro da Aplicação Web

Após realizar a confirmação de cadastro, uma tela confirmando o cadastro é exibida.

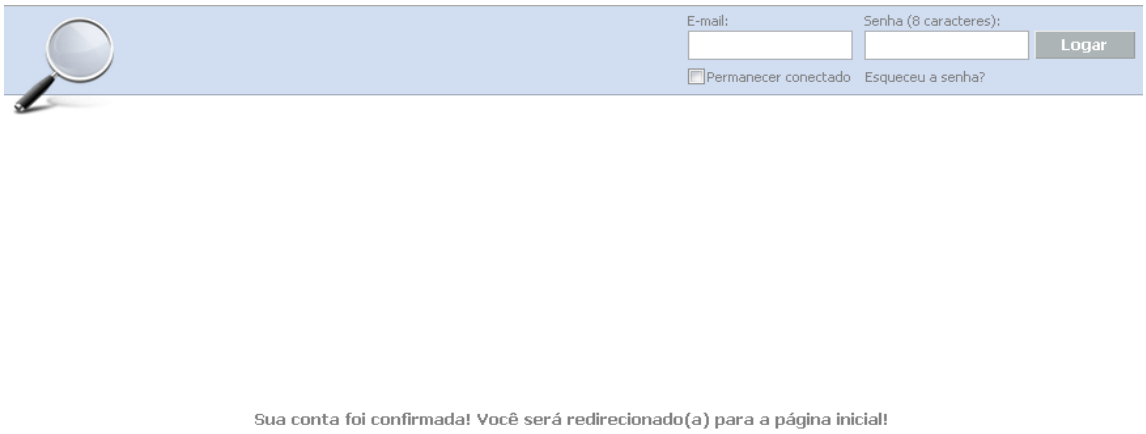


Figura 13 – Tela Informando a Confirmação do Cadastro da Aplicação Web

Após a conclusão do cadastro, o usuário poderá realizar o login utilizando o formulário na barra superior. A imagem abaixo exibe uma tentativa de login utilizando os dados cadastrados errados. Esta tela é exibida também caso o usuário tente realizar o login sem realizar sua confirmação de cadastro.

Figura 14 - Tela de Falha de Login na Aplicação Web

Ao realizar o login corretamente, o usuário é redirecionado para sua página principal. A página principal do usuário pode ser visualizada na imagem abaixo.

Figura 15 - Tela Principal do Usuário Logado da Aplicação Web

Na tela principal do usuário é possível cadastrar ou excluir uma consulta, acessar a página de edição de perfil e se deslogar. A imagem abaixo exibe a tela de edição de perfil.

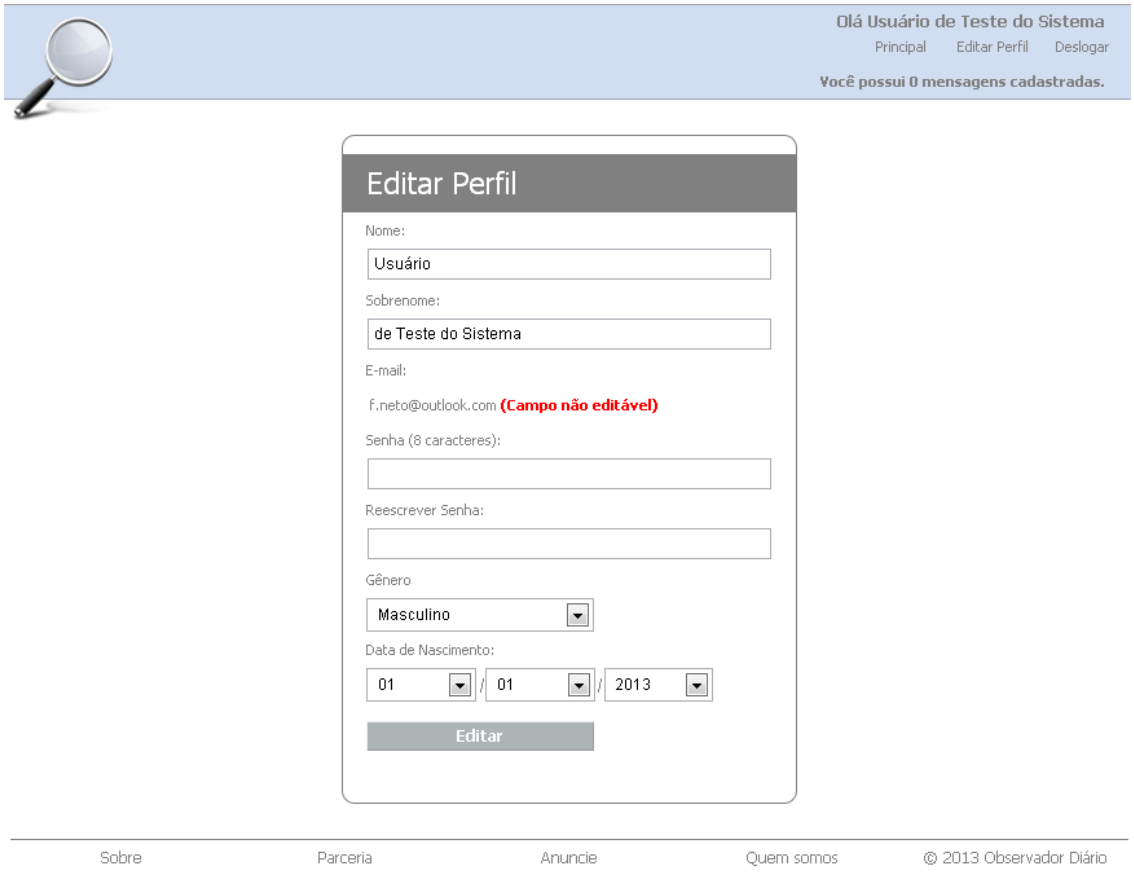


Figura 16 - Tela de Edição de Perfil do Usuário da Aplicação Web

As imagens a seguir exibe uma consulta realizada no portal da Imprensa Nacional utilizando como parâmetro o nome da presidente “Dilma Rousseff” realizada no dia 29 jan. 2013, informando que tal nome foi citado no jornal na Sessão 1, página 26, e uma consulta cadastrada na tela principal do usuário na Aplicação Web utilizando o nome da presidente para que posteriormente sirva como teste do sistema.

Figura 17 - Pesquisa Realizada no Portal da Imprensa Nacional

Figura 18 - Cadastro de Consulta Na Aplicação Web

Após o cadastro da consulta, o usuário precisa esperar a execução do Gerenciador de Atividades que ocorrerá diariamente ás 08h30min da manhã, devido ao fato das publicações serem normalmente disponibilizadas no Portal da Imprensa Nacional às 8 horas da manhã, dando assim uma margem de segurança de 30 minutos. Isso significa que, caso o usuário cadastre sua consulta após as 8 horas da manhã e o Gerenciador de Atividades já tenha iniciado o processo de pesquisa e notificação, a consulta do usuário só será realizada a partir do dia seguinte.

Durante a execução o sistema criará quatro pastas, sendo uma para os arquivos pdfs, outra para os arquivos txts, outra para os arquivos txts que passaram pelo tratamento do Adaptador TXT e outra para o arquivo de índice.

A pasta pdfs é criada no momento em que o Crawler realiza o download dos pdfs, e é dividida em subpastas que classificam o tipo da publicação, como jornais e suplementos, e dentro destas pastas, para cada publicação, serão criadas subpastas nomeadas de acordo com o nome da publicação, contendo os documentos pdfs destas.

A pasta txts2 é criada no momento da extração do conteúdo dos pdfs. Para cada pdf é criado um arquivo correspondente, seguindo a configuração de pastas e subpastas contida na pasta pdfs.

A pasta txts é criada no momento em que os arquivos da pasta txts2 são submetidos ao tratamento do Adaptador TXT, mantendo a mesma nomenclatura e Hierarquia de pastas contida na pasta txts2.

A pasta de índice é criada no momento em que o indexador inicia o processo de indexação dos arquivos contidos na pasta txts, os índices criados serão utilizados para a realização das buscas das consultas cadastradas pelo usuário.

Após o processo de extração e tratamento dos textos e indexação dos documentos txts, o Gerenciador de Atividades verifica se as consultas cadastradas pelos usuários foram encontradas, e caso tenham sido, um e-mail de notificação é enviado ao usuário. Levando de consideração a consulta cadastrada anteriormente e a pesquisa realizada no portal da Imprensa Nacional, um e-mail de notificação deve ter sido enviado informando que a consulta foi encontrada no Diário Oficial da União.

A imagem abaixo exibe o e-mail de notificação enviado pelo sistema informando ao usuário cadastrado que sua consulta foi encontrada em uma publicação do Diário Oficial da União.

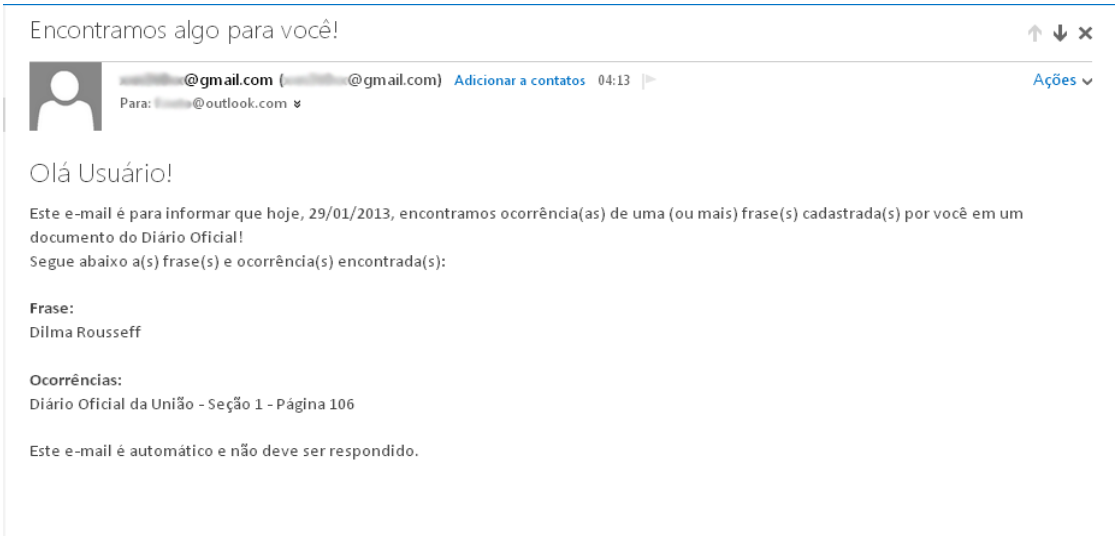
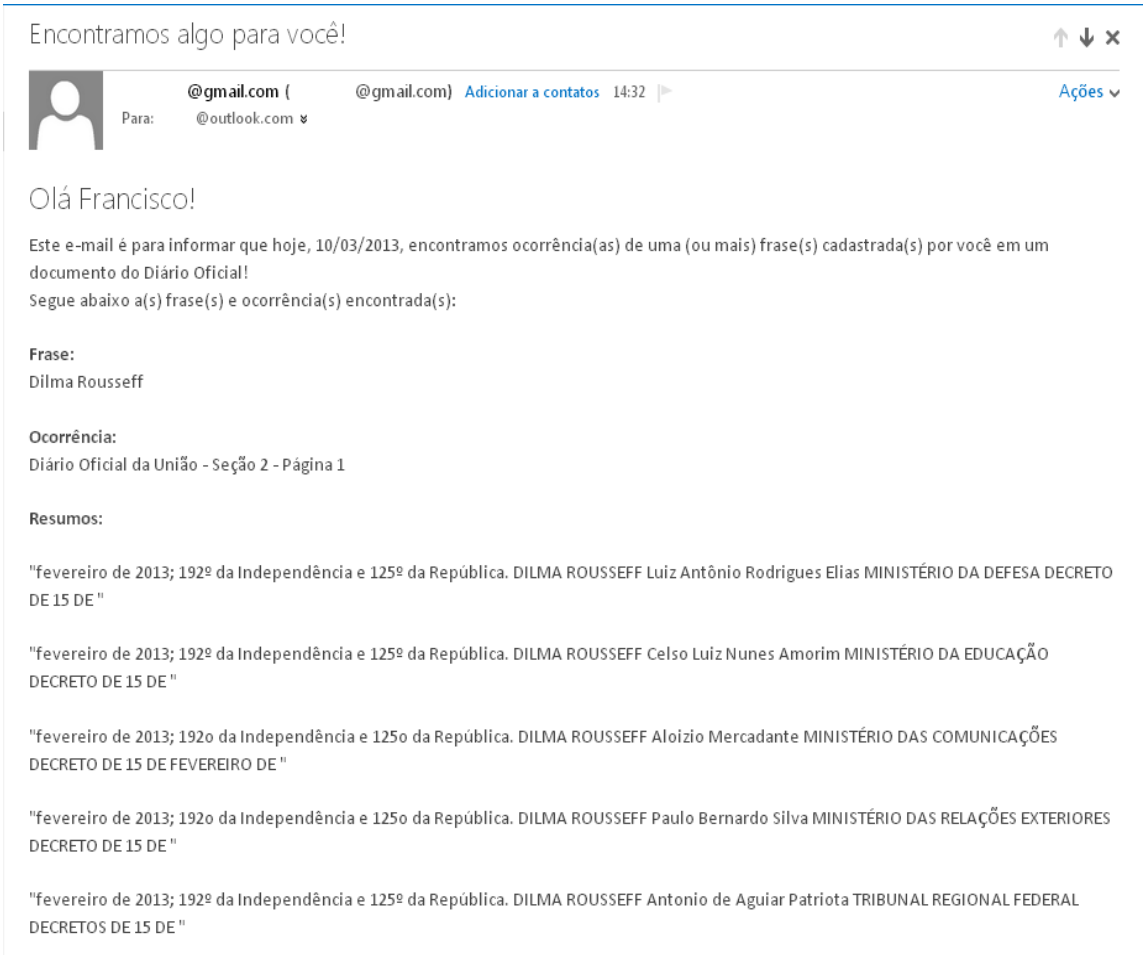


Figura 19 - Notificação por E-mail de Pesquisa Encontrada pelo Gerenciador de Atividades

A Figura 19 exibe uma comparação de resultados de busca, a figura 20 exibe o resultado de uma consulta após a adição de uma nova funcionalidade ao sistema, a obtenção de resumos, onde, após uma busca o sistema além de retornar a o nome do jornal e a página, o sistema retorna resumos do texto onde foi encontrado o texto pesquisado.

Para o teste dessa funcionalidade foi utilizado "Dilma Rousseff" como termo de pesquisa, e a pesquisa foi realizada no dia 18 de fevereiro de 2013.

A imagem a seguir exibe um exemplo do e-mail de notificação enviado ao usuário informando a pesquisa que retornou resultado, o arquivo, a página, e os resumos das ocorrências encontradas na página.


Encontramos algo para você!

@gmail.com (@gmail.com) Adicionar a contatos 14:32 Ações

Para: @outlook.com

Olá Francisco!

Este e-mail é para informar que hoje, 10/03/2013, encontramos ocorrência(as) de uma (ou mais) frase(s) cadastrada(s) por você em um documento do Diário Oficial!
Segue abaixo a(s) frase(s) e ocorrência(s) encontrada(s):

Frase:
Dilma Rousseff

Ocorrência:
Diário Oficial da União - Seção 2 - Página 1

Resumos:

"fevereiro de 2013; 192º da Independência e 125º da República. DILMA ROUSSEFF Luiz Antônio Rodrigues Elias MINISTÉRIO DA DEFESA DECRETO DE 15 DE "

"fevereiro de 2013; 192º da Independência e 125º da República. DILMA ROUSSEFF Celso Luiz Nunes Amorim MINISTÉRIO DA EDUCAÇÃO DECRETO DE 15 DE "

"fevereiro de 2013; 192o da Independência e 125o da República. DILMA ROUSSEFF Aloizio Mercadante MINISTÉRIO DAS COMUNICAÇÕES DECRETO DE 15 DE FEVEREIRO DE "

"fevereiro de 2013; 192o da Independência e 125o da República. DILMA ROUSSEFF Paulo Bernardo Silva MINISTÉRIO DAS RELAÇÕES EXTERIORES DECRETO DE 15 DE "

"fevereiro de 2013; 192º da Independência e 125º da República. DILMA ROUSSEFF Antonio de Aguiar Patriota TRIBUNAL REGIONAL FEDERAL DECRETOS DE 15 DE "

Figura 20 - E-mail de notificação com resumos de ocorrências

De posse desse e-mail percebe-se que a funcionalidade proposta pelo sistema está funcionando como esperado. O sistema foi capaz de baixar os pdfs das publicações do Diário Oficial, indexá-los e realizar pesquisas, notificando, por fim, o usuário de que sua consulta foi encontrada em uma publicação.

7 CONCLUSÃO

Este capítulo apresenta as considerações finais sobre o trabalho, analisando o problema e a solução proposta, se os objetivos foram alcançados, problemas enfrentados durante a elaboração da solução e possíveis melhorias na solução do problema, as limitações encontradas no decorrer da elaboração do trabalho e sugestões para trabalhos futuros.

7.1 CONCLUSÃO

O objetivo deste trabalho foi apresentar uma solução para um problema levantado observando-se a forma de recuperação da informação das publicações do Diário Oficial da União. Notou-se que existe uma grande quantidade de informação nas publicações do Diário, e que a forma de recuperação disponibilizada no portal da Imprensa Nacional não é, de fato, satisfatória para toda a população.

Percebe-se que o meio de pesquisa atual impõe a interação do usuário com o sistema, o que nem sempre é possível, trazendo assim a possibilidade de que alguma informação relevante possa ser perdida devido ao fato do usuário não estar disponível, por algum motivo, para realizar uma busca diária no sistema do portal da Imprensa Nacional.

Como solução para este problema foi proposto um sistema capaz de recuperar estas informações automaticamente, e notificar os usuários caso alguma informação relevante tenha sido encontrada.

Durante o projeto, foram identificadas necessidades de utilização de algumas bibliotecas e frameworks que auxiliaram as atividades e facilitaram o desenvolvimento do sistema. Algumas destas bibliotecas e frameworks agiram como um risco para a conclusão do sistema, pois houve a necessidade do estudo de seu funcionamento impactando assim no tempo previsto para o desenvolvimento, porém suas utilizações foram satisfatórias, pois, supriram as necessidades e viabilizaram o desenvolvimento do projeto.

Também foi necessário o estudo do funcionamento de sistemas de recuperação da informação e de sistemas Crawler, componentes essenciais para que o sistema alcançasse seu objetivo.

Durante o desenvolvimento do sistema, os maiores problemas encontrados foram na utilização novas tecnologias, como a biblioteca Lucene e o framework Spring Security, que demandaram tempo e estudo.

Foram realizados testes para garantir o funcionamento correto do sistema. Durante estes testes foram comparados resultados gerados entre a ferramenta disponível no portal da Imprensa Nacional e o resultado gerado pelo sistema proposto. O sistema proposto cumpriu com o seu objetivo e retornou resultados similares a ferramenta existente no portal, porém com a vantagem da solução proposta, removendo a necessidade de interação do usuário com o sistema, notificando-o por e-mail quando sua pesquisa cadastrada apresenta resultados relevantes.

Este sistema será de grande valia para todos os usuários que possuem interesse nas publicações do Diário Oficial da União, pois, através da sua utilização, elimina-se a cansativa necessidade de interação diária com o sistema de busca hoje disponibilizado.

7.2 LIMITAÇÕES

As maiores limitações encontradas na elaboração deste projeto foram a não conclusão das telas administrativas do sistema e a falta de realização de alguns testes como o de unidade e o de carga.

7.3 SUGESTÕES PARA TRABALHOS FUTUROS

A sugestão para trabalhos futuros seria um aprimoramento deste projeto através da adição de outros tipos de funcionalidade, como integração com redes sociais como o Twitter e o Facebook, que podem ser utilizados como ferramentas para facilitar o cadastro e as notificações, e também adicionar novas formas de notificações ao usuário como, por exemplo, o uso de SMS.

REFERÊNCIAS BIBLIOGRÁFICAS

AIRES, Rachel V.X. **Uso de marcadores estilísticos para a busca na Web em português.** 2005. 202 p. Tese (Doutorado em Ciências da Computação e Matemática Computacional) – USP, São Carlos.

CARDOSO, O.N.P. Artigo Cientifico: **Recuperação da Informação:** Anais da III SECICON, vol. 2, n. 1, nov. 2000, Disponível em: < http://www.dcc.ufla.br/infocomp/index.php?option=com_content&view=article&id=47:number-1&catid=39:volume-2&Itemid=73>. Acesso em 22 jan. 2013.

FERNEDA, E. **Recuperação da Informação: Análise sobre a contribuição da Ciência da Computação para a Ciência da Informação.** 2003. 137 p. Tese (Doutorado em Ciência da Comunicação) – USP, São Paulo. Disponível em: < www.teses.usp.br/teses/disponiveis/27/27143/tde-15032004-130230/publico/Tese.pdf>. Acesso em 26 jan. 2013.

GAMMA E., et al. **Padrões de Projeto: Soluções reutilizáveis de software orientado a objetos.** Porto Alegre: Bookman, 2000.

HENRIQUE, W.A. **Verificação de unicidade de urls em coletores de páginas web.** 2011. 45 p. Dissertação (Mestrado em Ciência da Computação) – UFMG, Belo Horizonte. Disponível em < http://www.bibliotecadigital.ufmg.br/dspace/bitstream/handle/1843/SLSS-8GQJNA/wallacefavoreto.pdf?sequence=1>. Acesso em 24 jan. 2013.

SILVA, Altigran; MOURA, Edleno. **Web Crawling: Coleta Automática na Web.** 2002. 77 p. Disponível em: <http://www.eicstes.org/EICSTES_PDF/PRESENTATIONS/Web crawling - Coleta automática na web (Silva-Moura).pdf>. Acesso em: 25 jan. 2013.