UNIVERSIDADE TECNOLÓGICA FEDERAL DO PARANÁ
DAINF - DEPARTAMENTO ACADÊMICO DE INFORMÁTICA
CURSO DE BACHARELADO EM SISTEMAS DE INFORMAÇÃO

AMIR LEONARDO KESSLER ANNAHAS

MONITORAMENTO DE FONTES ABERTAS PARA APOIO DE ATIVIDADE DE INTELIGÊNCIA PÚBLICA

TRABALHO DE CONCLUSÃO DE CURSO

CURITIBA
2020

AMIR LEONARDO KESSLER ANNAHAS

MONITORAMENTO DE FONTES ABERTAS PARA APOIO DE ATIVIDADE DE INTELIGÊNCIA PÚBLICA

Trabalho de Conclusão de Curso apresentado ao Curso de Bacharelado em Sistemas de Informação da Universidade Tecnológica Federal do Paraná, como requisito parcial para a obtenção do título de Bacharel.

Orientador: Mauro Sergio Pereira Fonseca
DAINF - Departamento Acadêmico de Informática -UTFPR

CURITIBA
2020

AMIR LEONARDO KESSLER ANNAHAS

MONITORAMENTO DE FONTES ABERTAS PARA APOIO DE ATIVIDADE DE INTELIGÊNCIA PÚBLICA

Trabalho de Conclusão de Curso de Graduação apresentado como requisito para obtenção do título de Bacharel em Sistemas de Informação da Universidade Tecnológica Federal do Paraná (UTFPR).

Data de aprovação: 14/Setembro/2020

MARCELO MIKOSZ GONÇALVES
Doutorado
Universidade Tecnológica Federal do Paraná (UTFPR)

FABIANO SCRIPTORE DE CARVALHO
Doutorado
Universidade Tecnológica Federal do Paraná (UTFPR)

MAURO SERGIO PEREIRA FONSECA
Doutorado
Universidade Tecnológica Federal do Paraná (UTFPR)

CURITIBA

2020

RESUMO

Annahas, Amir. Monitoramento de fontes abertas para apoio de atividade de inteligência pública. 2020. 32 f. Trabalho de Conclusão de Curso – Curso de Bacharelado em Sistemas de Informação, Universidade Tecnológica Federal do Paraná. Curitiba, 2020.

A Secretaria de Estado de Segurança Pública e Administração Penitenciária do Paraná é responsável por planejar, dirigir, executar, coordenar, fiscalizar e controlar as atividades do setor de segurança pública do Estado do Paraná. Entretanto, a obtenção das informações sobre segurança no estado do Paraná, principalmente no meio digital, está precisando de melhoramentos. Visando aprimorar a eficiência de toda essa estrutura, o objetivo deste trabalho é facilitar a obtenção de dados pela SESP/PR, coletando, filtrando e classificando dados digitais públicos na Internet que contenham informações relevantes para uma análise manual por especialistas. A metodologia foi baseada no desenvolvimento de um sistema dividido em três módulos, um módulo para a coleta de dados, outro módulo para a aplicação de técnicas de recuperação de informação e um último módulo para a apresentação dos dados ao usuário. A solução proposta utilizou somente ferramentas de código aberto e está em funcionamento por mais de dois meses. Nesse tempo, já foram coletadas mais de três milhões de notícias, a partir de uma lista de portais de notícias de um cenário de teste, sendo possível identificar o texto, o autor, o título e a data de publicação delas. O sistema de monitoramento possibilitou facilitar a coleta de informações relevantes a SESP/PR, aumentando a agilidade para a coleta de informações críticas e contribuiu para a análise das informações com recursos gráficos que auxiliam na identificação de padrões, além de reunir todas essas informações em uma interface de fácil usabilidade. A ferramenta já está disponível para homologação e utilização pela SESP/PR, sendo uma contribuição à segurança pública no estado do Paraná.

Palavras-chave: Recuperação de Informação. Web Crawler. Segurança Pública.

ABSTRACT

Annahas, Amir. Monitoring system of open sources to support public intelligence activity. 2020. 32 f. Trabalho de Conclusão de Curso – Curso de Bacharelado em Sistemas de Informação, Universidade Tecnológica Federal do Paraná. Curitiba, 2020.

The Paraná's Department of Public Security and Penitentiary Administration (Secretaria de Estado de Segurança Pública e Administração Penitenciária do Paraná - SESP/PR) is responsible for planning, directing, executing, coordinating, supervising and controlling the activities of the public security sector in the State of Paraná. However, obtaining security information in the state of Paraná, especially in the digital environment, is limited. In order to improve the efficiency of this entire structure, the objective of this work is to facilitate data collection by SESP / PR, collecting, filtering and classifying public digital data on the Internet that contain relevant information for a manual analysis by specialists. The methodology was based on the development of a system divided into three modules, one module for data collection, another module for the application of information retrieval techniques and one last module for the presentation of data to the user. The proposed solution used only open source tools and has been in operation for more than two months. In that time, more than three million news items have already been collected, from a list of news portals in a test scenario, being possible to identify their text, author, title and publication date. The monitoring system made it possible to facilitate the collection of information relevant to SESP / PR, increasing the agility for the collection of critical information and contributed to the analysis of the information with graphic resources that help in the identification of patterns, in addition to gathering all this information in one user-friendly interface. The tool is already available for approval and use by SESP / PR, being a contribution to public security in the state of Paraná.

Keywords: Information Retrieval. Public Security. Web Crawler.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

GUI	*Graphical User Interface*, do inglês Interface gráfica do utilizador
HTML	*HyperText Markup Language*, do inglês Linguagem de Marcação de Hipertexto
HTTP	*Hypertext Transfer Protocol*, do inglês Protocolo de Transferência de Hipertexto
HTTPS	*Hyper Text Transfer Protocol Secure*, do inglês Protocolo de Transferência de Hipertexto Seguro
IR	*Information Retrieval*, do inglês Recuperação de Informação
PR	Paraná
SESP	Secretaria de Estado de Segurança Pública e Administração Penitenciária
SO	Sistema Operacional
URL	*Uniform Resource Locator*, do inglês Localizador Uniforme de Recursos

SUMÁRIO

1 INTRODUÇÃO

O estado do Paraná possui um vasto território, contendo 399 municípios (IPARDES, 2019) onde são distribuídos 14 Distritos Policiais na Capital e 375 Delegacias Policiais no Interior do Estado (PARANÁ, 1999). Toda essa estrutura está sob responsabilidade da Secretaria de Estado da Segurança Pública e Administração Penitenciária do Paraná (SESP/PR) (PARANÁ, 1977), cuja competência desta Secretaria é "planejar, dirigir, executar, coordenar, fiscalizar e controlar as atividades do setor de segurança pública do Estado".

Uma central única, localizada em Curitiba, fiscaliza a atuação dessas divisões distribuídas pelo estado, que tem aproximadamente 200 mil quilômetros quadrados de território (IBGE, 2019). Esta central é responsável pela segurança de toda a população paranaense, composta por pouco mais de 11 milhões de habitantes (IBGE, 2019), além de emitir documentos, administrar presídios e o corpo de bombeiros, dentre outras atuações.

Por decorrência dessa responsabilidade é comum haver reclamações, denúncias e outras informações relevantes, provindas da população, que devem ser levadas em consideração para que seja possível aprimorar a eficiência de toda essa estrutura. É cada vez mais frequente a transmissão dessa informação através de veículos de mídia digitais e redes sociais (AGNEZ, 2009; MITCHELSTEIN; BOCZKOWSKI, 2010), acarretando em um aumento expressivo de usuários que utilizam esses meios para se informar diariamente. Porém, a fonte dessas informações não necessariamente consideram a importância de fazer tais informações chegarem ao conhecimento das pessoas responsáveis.

Consequentemente, é uma dificuldade da SESP/PR ter conhecimento do que está sendo falado sobre segurança no estado do Paraná, principalmente no meio digital, que é o objeto de estudo deste trabalho. A relevância desse problema é notável, uma vez que atualmente a própria secretaria possui funcionários que utilizam ferramentas de busca digitais para reunir informações, sem uma ferramenta apropriada e específica. Tais funcionários dedicam parte de seu tempo, em vez de atuar em suas atribuições usuais, para esse fim.

Por se tratar de uma secretaria de segurança pública, é de extrema importância que informações contextuais relevantes cheguem ao conhecimento da SESP/PR o mais rápido possível. Desta forma, a ferramenta desenvolvida neste trabalho permite que informações relevantes para a SESP/PR, que antes demoravam muito tempo ou nem sequer chegavam ao conhecimento dos especialistas da secretaria, chegue agora no tempo certo e às pessoas certas, permitindo diminuir o tempo para que a secretaria tome as devidas providências.

Desta forma, como objetivo geral, este trabalho se propõe a facilitar a obtenção de dados pela SESP/PR, coletando, filtrando e classificando dados digitais que contenham informações relevantes para uma análise manual por especialistas. Como objetivos específi-

cos, este trabalho irá coletar e analisar dados provindos de veículos de mídia pré definidos que falem sobre a segurança pública do estado do Paraná, desenvolver uma ferramenta automática de visualização dos dados, auxiliar na identificação de padrões de relevância e possibilitar a análise de dados sobre a segurança pública no estado do Paraná.

Portanto, como contribuições este trabalho propõe-se a:

- facilitar a coleta de informações relevantes a SESP/PR;
- agilizar a coleta de informações críticas;
- identificar padrões que possam contribuir para a análise de especialistas; e
- reunir e apresentar as informações em uma plataforma de fácil usabilidade.

Os arquivos de configuração das ferramentas utilizadas no projeto estão disponíveis em Annahas (2020).

Este trabalho está estruturado da seguinte forma: o Capítulo 1 (Introdução) aborda o problema da pesquisa e os objetivos; o Capítulo 2 apresenta o Referencial Teórico, abordando os assuntos importantes para a execução deste trabalho; o Capítulo 3 detalha a metodologia utilizada, assim como o planejamento de execução; o Capítulo 4 apresenta os Resultados obtidos; e finalmente as Considerações Finais e os trabalhos futuros são apresentadas no Capítulo 5.

2 FUNDAMENTAÇÃO TEÓRICA

Neste capítulo serão discutidos os assuntos considerados importantes para o trabalho, como *Web Crawler*, Recuperação de Informação ,*Relevance Feedback* e ferramentas de busca. Conceitos teóricos serão abordados para embasar o assunto aqui discutido.

2.1 World Wide Web - WWW

Em 1990 foi proposta a forma de compartilhamento de documentos pela Internet na forma que conhecemos e utilizamos hoje (BERNERS-LEE; CAILLIAU, 1990), chamada de *World Wide Web*. Seguindo esse padrão, os documentos podem ser disponibilizados de diversas formas (imagens, vídeos, textos, entre outras), possíveis de serem acessados pelo usuário utilizando ferramentas que utilizem o protocolo de comunicação HTTP ou HTTPS (IETF, 2014a), como, por exemplo, navegadores WEB. Os navegadores WEB acessam URLs (Uniform Resource Locator) (IETF, 2014b) pelo protocolo HTTP ou HTTPs, que retornam o seu conteúdo, que normalmente é uma página HTML (W3C, 1999). Nesta página está todo o conteúdo que o autor quer disponibilizar para o usuário, que também pode conter textos, imagens, vídeos e outras formas de mídia.

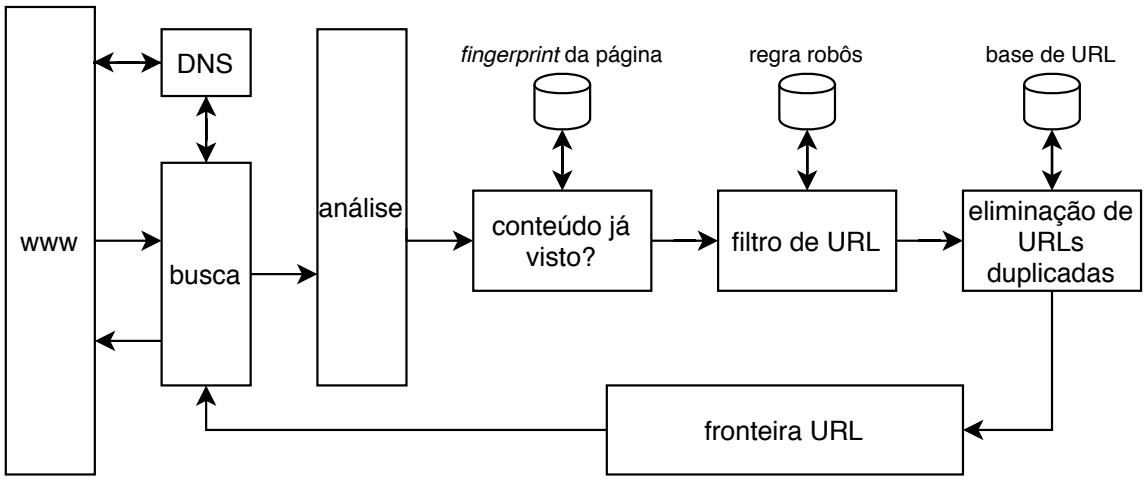
2.2 *Web Crawling* e *Indexing*

Web Crawling é o processo de procurar e indexar (armazenar URLs e seus respectivos conteúdos) páginas WEB para servir de base para um sistema de buscas (NAJORK; HEYDON, 2002; MANNING; RAGHAVAN; SCHÜTZE, 2010). O sistema que realiza essa função descrita se chama *Web Crawler* e tem o objetivo de recolher o máximo de páginas WEB úteis possíveis.

O fluxo demonstrado na Figura 1 inicia na fronteira URL. A fronteira URL é uma estrutura de dados que contém todas as URLs conhecidas, ainda não visitadas, que devem ser visitadas. É necessário ter ao menos uma URL conhecida para o *Web Crawler* iniciar o funcionamento. Na etapa de busca, o conteúdo de uma URL é obtido pelo protocolo HTTP ou HTTPS. Já na parte de análise, a resposta da requisição (geralmente uma página HTML) é processada, a fim de analisar o seu conteúdo (texto, imagens, vídeos, entre outros).

Após essas etapas iniciais é feita a parte teste do conteúdo, com a finalidade de avaliar se a URL deve ser adicionada na fronteira URL. O primeiro teste é se o conteúdo já foi visto anteriormente, utilizando uma verificação de *fingerprint*. Uma *fingerprint* é um código identificador gerado com o conteúdo de uma página, sem conter a URL, garantindo assim que o mesmo conteúdo não seja adicionado duas vezes à fronteira URL.

Figura 1 – Funcionamento de um *Web Crawler* simples.



Fonte: Adaptado de Manning, Raghavan e Schütze (2010)

O segundo teste é o filtro de URL, que verifica se a URL passa nas regras descritas no *robots.txt* da página. É um padrão os domínios conterem esse arquivo com regras de URLs que um *crawler* pode ou não visitar (SUN; ZHUANG; GILES, 2007). Passando deste teste, a URL visitada é comparada com outras já salvas, evitando assim uma duplicidade.

Um *framework* amplamente utilizado para auxiliar no processo de procurar e indexar as páginas WEB é o Scrapy. Esse *framework*, que é *open source* e foi desenvolvido em Python, é presente em muitas empresas e projetos acadêmicos como parte da solução na hora de realizar a coleta de dados na Web (KOUZIS-LOUKAS, 2016; MYERS; MCGUFFEE, 2015).

Outro *framework* que propõe resolver os problemas relacionados à coleta e indexação de dados da Web é o news-please. O news-please também é *open source* e foi desenvolvido em Python e é especializado em realizar o processo de coleta de indexação de dados em páginas de notícias. Além de funcionar bem com vários processos em paralelo, otimizando assim a busca em diversas páginas diferentes ao mesmo tempo, o news-please é utilizado pela comunidade cientifica para coletar artigos de notícias (HAMBORG; DONNAY; GIPP, 2019; HAMBORG; BREITINGER; GIPP, 2019; SARR; OUSMANE; DIALLO, 2018; LIU et al., 2019).

2.3 Recuperação de Informação e Sistemas de Busca

A Recuperação de Informação, ou *Information Retrieval* (IR), é a área da Ciência da Computação que abrange a busca de informações e sua representação. A informação pode estar estruturada, semiestruturada, como tabelas e colunas no banco de dados, ou até mesmo não estruturada, na forma de texto, imagens, sons, vídeos, grafos, entre outros (BAEZA-YATES; RIBEIRO-NETO, 2013).

Com relação a textos, existem diversas técnicas relacionadas à IR que visam resolver o problema de busca de palavras, ou termos, dentro de múltiplos documentos. Compreende-se por documento um texto ou um conjunto de textos, como por exemplo uma página WEB e uma notícia extraída de um portal de notícias. Alguns dos modelos mais clássicos da IR em texto são modelo Booleano, modelo vetorial e os modelos probabilísticos.

O modelo Booleano de busca, ou *Boolean Retrieval*, consiste em procurar no índice invertido qual o documento que contém todos os termos da busca, sendo o modelo mais básico (ANH; MOFFAT, 2006). O modelo vetorial, ou *Space Vector Model*, se baseia em representar um documento como um vetor no espaço, sendo detalhado na subseção 2.3.1. Já os modelos estatísticos levam em consideração conceitos probabilísticos, tendendo ser mais matematicamente e computacionalmente complexos (FUHR; PFEIFER, 1994).

Os modelos costumam utilizar a estrutura de dados chamada de índice invertido, que é uma lista de cada palavra (ou termo) da coleção (conjunto de todos os documentos conhecidos) junto com todos os ponteiros dos documentos em que ela aparece. Existem variações do índice invertido, como a adição de detalhes de ranqueamento ou posição da palavra no documento por exemplo.

Também fazem parte da área de IR modelos e técnicas que resolvem problemas de ranqueamento por relevância, que serão detalhados nas próximas subseções.

2.3.1 Modelo Vetorial

No modelo vetorial cada documento é um vetor no espaço de N dimensões, onde N é o número de palavras únicas em todos os documentos conhecidos (ANH; KRETSER; MOFFAT, 2001; WONG; ZIARKO; WONG, 1985). Um documento ($\vec{d}$) pode ser representado como:

$$\vec{d} = (\omega_1, \omega_2, ..., \omega_N) \tag{1}$$

em que ω_i é o peso da palavra i naquele documento.

Estipular o peso de um documento não é uma tarefa trivial, porém existe um conjunto de técnicas que abstraem essa dificuldade considerando algumas características facilmente mensuráveis. As técnicas clássicas consistem em estipular o peso considerando a frequência do termo entre os documentos específicos e a frequência do termo no documento sendo analisado. Nesse sentido, os termos com menor frequência global (em todos os documentos) possuem o peso aumentado, assim como os termos com maior frequência local (no documento sendo analisado).

Neste modelo, a busca é tratada como um documento, considerada como um vetor $\vec{q_k}$, onde o resultado é o conjunto dos vetores $\vec{d}$ próximo à $\vec{q_k}$. Para determinar a distância entre vetores, podem ser utilizadas diversas técnicas, como a distância euclidiana ou a semelhança entre cossenos, por exemplo.

2.3.2 Ranqueamento por Relevância

Entende-se por relevância o quão importante o documento encontrado na busca é para o usuário que a realizou. Nesse sentido, o ranqueamento por relevância pode ser definido pela quantificação e/ou ordenação pela relevância dos documentos encontrados. Existem diversos métodos para definir quais documentos são relevantes ou até mesmo quantificar a relevância (KEKÄLÄINEN, 2005).

A importância do Ranqueamento por Relevância torna-se notável na utilização de qualquer ferramenta de buscas. Quando o usuário pesquisa um conjunto de palavras, não é esperado somente a correspondência exata dos termos utilizados, mas também algumas variações dos termos, como por exemplo, diferentes tempos verbais. Outro tópico importante a ser considerado é a ambiguidade das palavras. Por exemplo, em um caso em que a palavra "manga" é pesquisada, o usuário pode estar procurando a fruta ou a parte da camisa e, dependendo da intenção do usuário, deve ser considerada relevante apenas uma dessas opções.

Várias técnicas são úteis para lidar com os problemas apresentados (MANNING; RAGHAVAN; SCHÜTZE, 2010; BAEZA-YATES; RIBEIRO-NETO, 2013), como a retirada de chamadas *stopwords*, que se resume em palavras extremamente comuns em todos os documentos, não representando assim uma real correspondência de busca. O tratamento de caracteres no texto também pode ser útil dependendo da situação, como deixar todas as letras em maiúsculo ou minúsculo e retirar dígitos, marcações, hifens, etc.

Para lidar com a sintaxe e importância relativa dos termos existem mais técnicas, como o *Stemming* (HULL, 1996; LOVINS, 1968; TOMLINSON, 2003), que substitui as variações sintáticas das palavras pela palavra base (plural por singular, tempos verbais diferentes para um único, etc). A seleção de termos ou palavras para serem definidos como principais na indexação também é uma técnica relevante, assim como atribuir um peso para cada uma das palavras da busca (ROBERTSON; JONES, 1976).

É importante ressaltar que cada técnica é desenvolvida tendo por objetivo uma categoria de problemas específica, e portanto, a aplicabilidade de uma técnica é dependente do problema, devendo ser avaliada na fase de análise.

2.3.3 *Relevance Feedback*

As técnicas discutidas até o momento não levam em consideração que o resultado esperado pelo usuário possa não corresponder aos termos da busca. Isso pode ocorrer pela existência de sinônimos na linguagem natural, ou ainda pela falta (ou excesso) de palavras na busca. Com esse problema em consideração, a técnica de realimentação de relevância, ou *Relevance Feedback* (ROCCHIO, 1971), visa adicionar ou eliminar termos da busca, ajustando os pesos de cada termo presente, a partir dos resultados da busca inicial.

Existem algumas formas de aplicar a técnica de *Relevante Feedback*, podendo ser

divida em realimentação explicita, com interação direta com o usuário, ou implícita, sem a interação direta com o usuário. A realimentação explicita pode ser feita com a execução assistida pelo usuário ao classificar a relevância de cada documento visualizado ou com os cliques na busca, considerando cada documento da busca acessado para a realimentação (ROCCHIO, 1971). Já a realimentação implícita considera os documentos no resultado da busca com maior relevância pela métrica de ranqueamento por relevância para a sua realimentação e/ou obtém informações de fontes externas (XU; CROFT, 2017).

O algoritmo clássico de ROCCHIO resolve o problema de realimentação explicita com execução assistida pelo usuário, utilizando o modelo vetorial de representação de documentos, discutido na subseção 2.3.1, para realizar a realimentação. Nesse algoritmo, o usuário determina quais são os documentos relevantes do resultado da busca inicial e essa informação é utilizada para formular uma busca expandida.

Uma busca inicial $\vec{q}_k$ pode ser expandida para uma busca $\vec{q}_{k+1}$ atualizando os pesos ω de $\vec{q}_k$ da seguinte maneira:

$$\vec{q}_{k+1} = \alpha\ \vec{q}_k + \frac{\beta}{N_r} \sum_{\forall \vec{d_j} \in D_r} \vec{d_j} - \frac{\gamma}{N_n} \sum_{\forall \vec{d_j} \in D_n} \vec{d_j} \tag{2}$$

em que D_r é o conjunto de documentos considerados relevantes, $\vec{d_j}$ é um documento pertencente a este conjunto, N_r é o número de documentos contidos em D_r, D_n é o conjunto de documentos considerados não relevantes, N_n é o número de documentos contidos em D_n e α, β e γ são constantes de ajuste.

Variações da expansão original de ROCCHIO são propostas em Ide (1971), como as representadas nas equações 3 e 4:

$$\vec{q}_{k+1} = \alpha\ \vec{q}_k + \beta \sum_{\forall \vec{d_j} \in D_r} \vec{d_j} - \gamma \sum_{\forall \vec{d_j} \in D_n} \vec{d_j} \tag{3}$$

$$\vec{q}_{k+1} = \alpha\ \vec{q}_k + \beta \sum_{\forall \vec{d_j} \in D_r} \vec{d_j} - \gamma\ max\_rank(D_n) \tag{4}$$

onde $max\_rank(D_n)$ representa o documento considerado não relevante melhor ranqueado.

Em relação às constantes de ajuste, ROCCHIO fixou $\alpha = 1$ e IDE fixou $\alpha = \beta = \gamma = 1$. Já Salton e McGill (1983) afirma que os documentos considerados relevantes devem influenciar mais do que os considerados não relevantes, logo β deve ser maior do que γ.

As expressões 2, 3 e 4 produzem resultados semelhantes e se destacam por sua simplicidade e bons resultados (BAEZA-YATES; RIBEIRO-NETO, 2013).

Foi discutida nesta seção sobre a Recuperação de Informação e Sistemas de Busca. Como o escopo deste trabalho se limita em texto, mais especificamente em artigos de notícias, foram citadas algumas técnicas de IR com esse foco.

2.4 Estado da Arte

Nesta seção serão discutidas e revisadas as ferramentas desenvolvidas na área deste trabalho, assim como outros trabalhos correlatos.

Em Hamborg et al. (2017) é criada uma ferramenta que realiza o *crawling* de notícias, o news-please. Essa ferramenta é capaz de explorar os *links* dentro de um domínio e extrair as informações de cada notícia, separando, com alta porcentagem de acerto, o título (82%), a descrição (76%), a data de publicação (70%) e a imagem principal (76%). Também foram atingidos bons índices ao separar o texto principal (62%) e houve uma maior dificuldade ao identificar o autor (34%).

No artigo Hamborg, Breitinger e Gipp (2019), que é uma continuação do estudo realizado no artigo citado anteriormente, é criado uma outra ferramenta, o Giveme5W1H, que faz a análise dos textos das notícias coletadas e são respondidas perguntas cruciais sobre o seu conteúdo: quem fez o que, quando, onde, por que e como aconteceu, atingindo um índice de acerto de 73% na extração das respostas de todas essas perguntas. Esse estudo possui um foco na interpretação da notícia (processamento de linguagem natural), divergindo do foco deste trabalho, que visa identificar quais são as notícias de interesse.

O funcionamento de um *Web Crawler* que leva em consideração palavras-chave para priorizar a coleta de URLs é estudada em Kumar et al. (2018). O artigo é parte de um estudo que visa encontrar páginas web de acadêmicos indianos trabalhando fora da Índia. No exemplo apresentado, o *crawler* visitou menos páginas para atingir as páginas de interesse do que o método de ordenação de URLs. Sendo assim, é notável a importância de um *crawler* com funcionamento específico quando o escopo é bem definido.

Um novo modelo de recuperação de informação é proposto em Agbele, Ayetiran e Babalola (2018), onde seu contexto é levado em consideração para determinar a relevância de um documento. O autor afirma que é possível detectar e classificar contextos utilizando sistemas de recuperação de informação adaptativos e utiliza essa informação no modelo proposto. Esse modelo leva em consideração o *feedback* do usuário, prometendo assim melhores resultados, com mais documentos relevantes e menos documentos não relevantes. O modelo é uma técnica nova e pouco utilizada, com resultados ainda não completamente validados pela comunidade cientifica.

A recuperação de informação é utilizada em Carnaz et al. (2018) a fim de auxiliar no trabalho policial. Segundo CARNAZ et al., dados não estruturados, semiestruturados e estruturados podem ser utilizados como fonte para um sistema de suporte à decisão forense. No sistema criado, a informação é obtida de relatórios policiais com diferentes tipos e formatos, automaticamente extraindo, limpando e formatando a informação com a finalidade de facilitar os investigadores policiais a identificar correlações. O trabalho se assimila a este, uma vez que também utiliza a recuperação de informação na área de segurança pública, porém possui uma aplicação com o foco de suporte à decisão, diferente deste.

Em Hanadi (2019) é proposto um *framework* que utiliza a realimentação de relevância para forense digital. O *framework* integra a mineração de dados (extração de textos) com técnicas de *relevance feedback* com o objetivo de diminuir o tempo dos investigadores para obter informação útil. O trabalho inclui uma prova de conceito na forma de aplicação do *framework* criado em um departamento de pesquisa forense, resultando em bons resultados. O *framework* foi desenvolvido com foco na forense digital, o que não será abordado neste trabalho.

O trabalho Hamborg et al. (2017) apresentou ser uma ótima ferramenta para a obtenção de dados de portais de notícias, e portanto possui utilidade em potencial para este trabalho. O escopo deste trabalho é restrito a portais de notícias abertos, justificando assim, como visto nesta seção, o desenvolvimento de uma aplicação específica. Modelos clássicos e de resultados já conhecidos serão utilizados no desenvolvimento, adaptando-os para o caso específico deste trabalho. A finalidade da aplicação das técnicas de recuperação de informação apresentadas é para reunir informações relevantes para o apoio de atividade de inteligência pública.

Neste capítulo foi abordada a *World Wide Web*, assim como o conceito de *crawling* e de indexação. Também foi discutido o funcionamento de sistemas de busca e modelos de representação de documentos da recuperação de informação. Foi apresentado o conceito de ranqueamento por relevância e técnicas de aperfeiçoamento de obtenção de documentos relevantes. No próximo capítulo será abordada a metodologia utilizada neste trabalho.

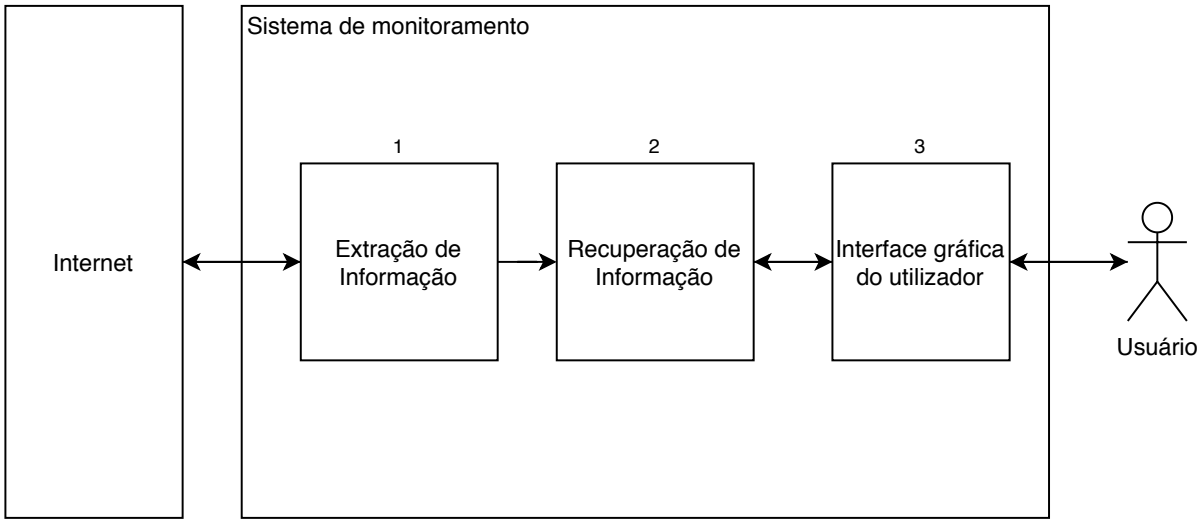
3 METODOLOGIA

Neste capítulo a metodologia do trabalho é apresentada. O capítulo está dividido em 3.1, onde a abordagem computacional é introduzida, 3.2, 3.3 e 3.4, onde os módulos da solução e a Interface gráfica do utilizador são apresentados, e 3.5, onde o escopo do trabalho é definido.

3.1 Estrutura da Solução

A solução adotada para o problema apresentado foi o desenvolvimento de um sistema cliente-servidor dividido em módulos. Foram identificados três módulos principais, demonstrados na Figura 2, um com a finalidade de extrair informações (um *Web Crawler*, abordado em 2.2), outro de aplicação de técnicas de recuperação de informação e o último responsável pela apresentação dos resultados relevantes ao usuário. Os módulos serão detalhados nas subseções 3.2, 3.3 e 3.4.

Figura 2 – Divisão do sistema em módulos.



Fonte: autoria própria

3.2 Extração de Informação

O módulo de Extração de Informação é responsável pelo monitoramento dos portais de notícias com a finalidade de encontrar novas notícias e pela coleta das notícias. A principal ferramenta utilizada neste módulo foi o *framework* news-please, uma vez que o *framework* possui funcionalidades de *crawling* e de indexação de documentos.

Conforme visto em 2.2, o *framework* news-please se sobressai ao escolher uma ferramenta para a implementação deste módulo.

Tabela 1 – Configuração do news-please

Configuração	Valor definido
Tipo de *crawling*	*RecursiveCrawler*
Número de *crawlers* em paralelo	10
Número de *daemons* em paralelo	30
Idioma do analizador do elasticsearch	*portuguese*
Data limite para baixar as notícias	1 semana antes do início do *crawling*

Fonte: autoria própria

A ferramenta news-please é um *crawler open source* focado em páginas de notícias (HAMBORG et al., 2017). O *crawler* foi desenvolvido em *Python*, compatível com as versões do *Python* 3.5 ou superiores, e está disponível no sistema de gerenciamento de pacotes padrão da linguagem de programação, o PIP, sendo de fácil instalação.

Essa ferramenta é compatível para a utilização dentro de um projeto, como uma biblioteca que fornece um conjunto de funcionalidades desenvolvidas para a obtenção de notícias e tratamento das informações obtidas, ou como um sistema *crawler* autônomo e independente. Como um *crawler* independente, news-please navega pelos portais de notícias a partir de um conjunto de endereços WEB utilizando uma heurística própria para identificar quais páginas são consideradas notícias para categorizá-las e realizar o armazenamento das mesmas.

Para realizar o armazenamento das notícias encontradas, cada notícia é convertida em um objeto JSON e é fornecido uma série de opções de armazenamento, como em arquivos ou em bancos de dados como o PostgreSQL ou o Elasticsearch.

O *framework* news-please foi configurado para acompanhar uma lista de portais de notícias. Foi configurado o *crawler* do framework na opção *RecursiveCrawler*, onde o *crawling* parte da URL informada e passa por cada URL que é referenciada, desde que não exista uma configuração dizendo para ignorá-la e que ela pertença ao mesmo domínio da URL que referenciou.

A ferramenta news-please suporta a configuração do número de *crawlers* que funcionam em paralelo ao iniciar a aplicação e, para conseguir monitorar o máximo de páginas possíveis na maior frequência possível, foi configurado um valor que se aproximasse ao máximo que os recursos computacionais disponíveis suportassem. Os valores encontrados para a configuração de *numberof_parallel_crawlers* e *number_of_parallel_daemons* foram respectivamente 10 e 30. As configurações utilizadas para a realização do projeto estão descritas na Tabela 1.

Também foi notado que apenas a configuração dos parâmetros do *crawler* não foi o suficiente para obter as notícias em tempo útil, uma vez que a função primária do *crawler* escolhido não é a de realizar o monitoramento em tempo real de portais de notícias e sim realizar a indexação de todas as notícias que esse portal possui. Para contornar esse

problema, foi configurado por meio da execução de um *script* cadastrado na tabela *crontab* do SO que todas as tarefas paralelas do news-please fossem re-executadas após um período de duas horas.

3.3 Recuperação de Informação

O módulo de Recuperação de Informação é responsável por aplicar técnicas de Recuperação de Informação ao gravar e apresentar os dados para o usuário de forma que seja possível a extração de informações úteis e relevantes.

Muitos bancos de dados estruturados e não estruturados utilizam técnicas de IR a fim de aumentar a eficiência de suas buscas. Como o escopo deste projeto é trabalhar com uma grande quantidade de texto, os bancos de dados não estruturados tendem a ser mais recomendados (STONEBRAKER, 2010). Dentre os bancos de dados não estruturados mais populares no mercado, se encontram o MongoDB (BANKER, 2011) e o Elasticsearch (ELASTIC, 2020a).

O Elasticsearch é uma ferramenta de busca *open source* que pode funcionar de forma distribuída e também funciona como um banco de dados não relacional mantida pela empresa elastic. A comunicação do Elasticsearch é realizada por meio de serviços WEB, facilitando assim a interoperabilidade e a compatibilidade com as linguagens de programação e com os sistemas operacionais. Um servidor Elasticsearch pode funcionar tanto no sistema operacional Windows quanto nos sistemas operacionais baseados no Linux.

O Apache Lucene é uma biblioteca Java que é utilizada pelo Elasticsearch nos processos de recuperação de informação, como indexação e busca de documentos (APACHE, 2020). Assim, o Elasticsearch se destaca por possuir funcionalidades essenciais para a execução deste trabalho, principalmente por ser um banco de dados especializado em armazenar grandes quantidades de textos e um mecanismo de buscas eficiente (THACKER; PANDEY; RAUTARAY, 2016).

As informações coletadas pelo módulo de Extração de Informação são armazenadas e indexadas pelo Elasticsearch, uma vez que o *framework* news-please possui uma compatibilidade nativa com a ferramenta Elasticsearch. Após aproximadamente dois meses de funcionamento do sistema, ele já tem mais de três milhões de documentos e mais de 12.0GB de dados coletados.

3.4 Interface gráfica do utilizador

O terceiro e último módulo da solução adotada age como a Interface gráfica do utilizador, ou *Graphical User Interface* (GUI). A GUI é a parte do sistema responsável pela interação do usuário com o sistema, por meio de elementos gráficos e outros indicadores visuais (ZHANG, 1996). Para obter esse resultado, foi utilizada a ferramenta Kibana.

Kibana é uma ferramenta *open source* mantida pela empresa elastic que fornece uma interface gráfica para o Elasticsearch, além de possuir diversas funcionalidades de análise de dados (ELASTIC, 2020b). Um servidor do Kibana pode ser instalado em uma máquina com o sistema operacional Windows ou com um sistema operacional baseado em Linux e a interface gráfica pode ser acessada por qualquer dispositivo com um navegador WEB.

Utilizando essa ferramenta é possível criar diversos tipos de gráficos e *dashboards* com os dados existentes no Elasticsearch, além de possuir integração com o Canvas. Também é possível acessar e alterar as configurações do Elasticsearch por meio de uma página WEB. Além disso, a ferramenta fornece funcionalidades para desenvolvedores, incluindo uma interface de requisições WEB para o Elasticsearch, auxiliando na comunicação com o mesmo.

A ferramenta Kibana é responsável pela visualização em tempo real dos dados armazenados no Elasticsearch, fornecendo uma interface de fácil utilização e que não demanda um conhecimento técnico apurado para a utilização e obtenção de informações cruciais à SESP/PR.

3.5 Restrições e Limitações

A solução implementada realiza a coleta de dados WEB a partir de um conjunto de URLs fornecida por membros da SESP/PR. São coletadas informações apenas de URLs que pertencem aos domínios das URLs do conjunto informado.

É objeto de estudo deste trabalho notícias sobre segurança pública no estado do Paraná veiculadas digitalmente em portais públicos de notícias.

Há uma limitação por parte dos usuários do sistema, uma vez que os integrantes da SESP/PR podem não possuir um conhecimento técnico elevado. Foi desenvolvido um manual de usuário com exemplos, para facilitar a utilização do mesmo.

Neste capítulo a solução adotada, implementada e em funcionamento foi detalhada. No próximo capítulo os resultados serão apresentados.

4 RESULTADOS

Neste capítulo serão apresentados os resultados obtidos com a solução adotada, como a solução atende os objetivos deste trabalho e suas limitações, assim como a avaliação da ferramenta implementada e disponibilizada.

Como resultado este trabalho apresenta uma solução implementada e em funcionamento para o problema apresentado, aplicado em um cenário de testes. A solução consiste em um sistema separado em três módulos, onde um módulo é responsável pela coleta de dados, outro módulo é responsável por aplicar técnicas de Recuperação de Informação nos dados coletados e o último módulo é responsável por apresentar os dados coletados de forma intuitiva ao usuário.

O cenário de testes foi aplicado em uma máquina virtual com o sistema operacional *Debian GNU/Linux 10 (buster)*. A máquina disponível estava configurada com 100 GB de espaço físico e 6 GB de memória RAM. Foi configurado o módulo da extração das informações e o módulo da recuperação de informação na mesma máquina, utilizando as tecnologias descritas nas seções 3.2 e 3.3. Para o acesso da interface gráfica, é necessário um dispositivo com um navegador WEB e acesso pela rede à uma máquina que esteja responsável pela GUI.

O sistema implantado consegue coletar novas notícias de uma lista de portais de notícias do cenário de testes e fornecer ferramentas importantes para análise e acompanhamento das notícias coletadas, como gráficos, *dashboards* e ferramentas avançadas de busca. As ferramentas utilizadas permitem que o usuário consiga analisar as notícias publicadas pelos portais presentes na lista do cenário de testes de forma facilitada, além de possibilitar a identificação de padrões com as diversas opções de formas de visualizações disponíveis dentro de uma interface de fácil usabilidade.

A Figura 3 mostra um *dashboard* configurado na ferramenta. Os *dashboards* são personalizáveis, sendo possível adicionar diversas formas de gráficos, tabelas e indicadores. No exemplo, foi filtrado o atributo *"text"* por incidências da palavra "drogas" e o *dashboard* atualizou em tempo real todos os seus componentes aplicando o filtro, ou seja, os gráficos e tabelas foram atualizados para mostrar o comportamento de portais e de publicação de notícias que envolvam "drogas". É importante ressaltar que o filtro por palavras não necessariamente busca uma correspondência literal, podendo até ser adicionado uma lista de sinônimos nas configurações para aumentar a eficiência da busca. No próprio exemplo, a busca realizada por "drogas" teve como resultado notícias que possuem a palavra "droga" em seu texto, como destacado.

A Figura 4 demonstra um exemplo de busca realizada na ferramenta. É possível notar a possibilidade de filtrar a partir de texto (palavras-chave) com operações lógicas (AND e OR) e, também, pela data, inclusive adicionando a data inicial e data final

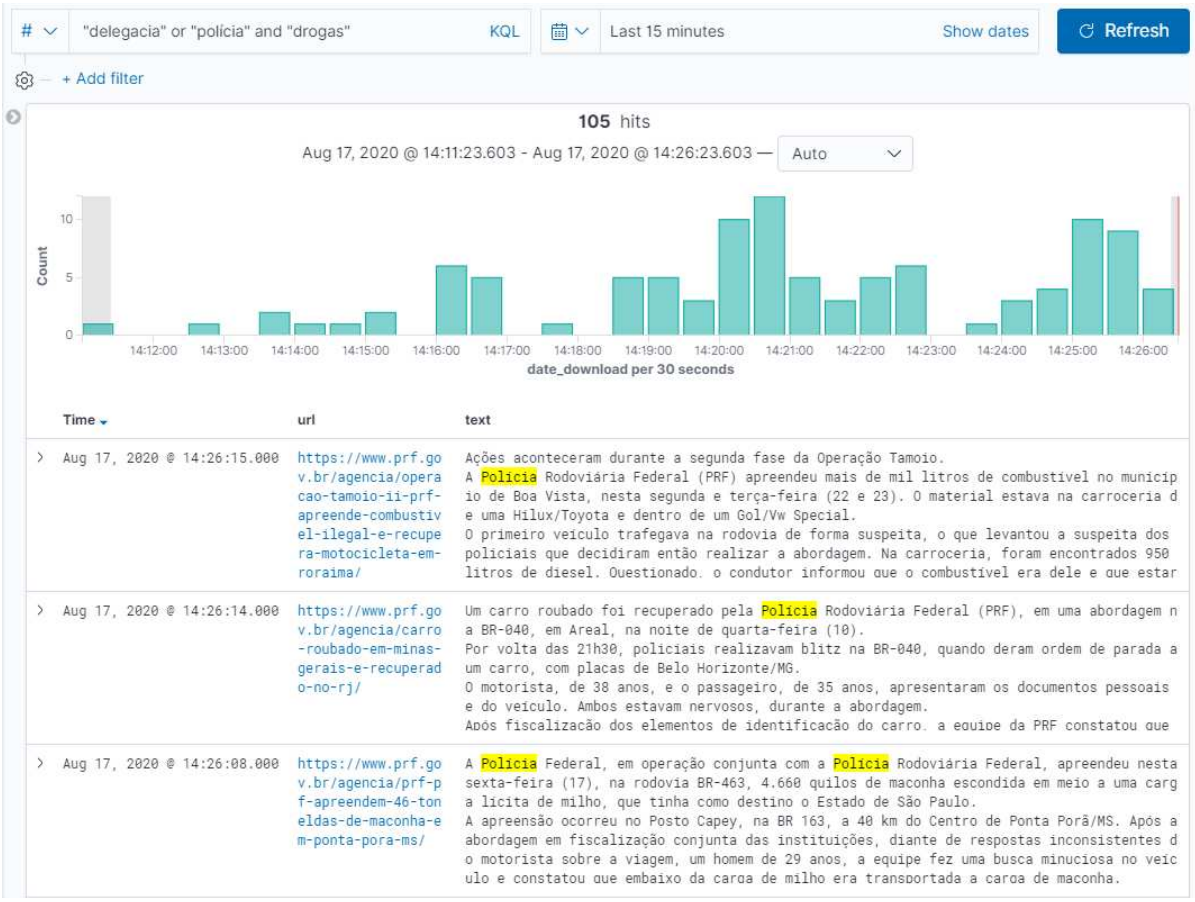
Figura 3 – *Dashboard* criado para a análise dos dados.

Fonte: autoria própria

de forma relativa ou absoluta. No resultado da busca, as partes do texto que possuem correspondência com o filtro são realçadas para agilizar a identificação da parte relevante do texto. As colunas do resultado são personalizáveis, sendo possível selecionar apenas os atributos relevantes para o usuário, diminuindo a poluição visual. Além disso, um histograma dos resultados é gerado, o que é útil para a identificação de padrões.

A ferramenta ficou em funcionamento por um período superior a dois meses, e neste período coletou mais de três milhões de documentos, totalizando mais de 12.0 GB de dados coletados. A Figura 5 demonstra a relação de notícias baixadas pelo tempo, dentro de um período de sete dias. O gráfico mostra que há períodos de tempo em que a ferramenta implantada consegue coletar mais notícias e há períodos de tempo em que a ferramenta implantada consegue coletar menos notícias num mesmo intervalo. Essa sazonalidade pode ser melhor estudada futuramente, gerando informações de grande relevância para a SESP/PR.

Figura 4 – Exemplo de busca na ferramenta.



Fonte: autoria própria

Figura 5 – Quantidade de notícias baixadas no intervalo de sete dias.

Fonte: autoria própria

A lista de portais do cenário de testes continha 51 portais de notícias e a aplicação desenvolvida e em funcionamento consegue realizar o monitoramento de forma satisfatória de 35 desses portais. Sugere-se que a dificuldade encontrada no acompanhamento dos portais exista pela falta de adequação aos padrões de codificação dos mesmos, uma vez

que muitos dos portais inclusos na lista são portais locais e/ou de nicho, que possuem poucos recursos para a construção e manutenção de seus *websites*.

Na Figura 6 é representada a quantidade de notícias que foram baixadas separada por portal de notícias. É visível que a maioria das notícias coletadas são de um grupo pequeno de portais, em comparação ao total. Esse comportamento se justifica pelo fato de existirem portais maiores e com mais notícias publicadas.

Figura 6 – Quantidade de notícias por portal.

Fonte: autoria própria

Foi encontrada uma dificuldade na identificação da data que as notícias são publicadas. Dos 35 portais que estão sendo monitorados, apenas 19 deles possuem a data de publicação confiável. Essa situação tende a ser minimizada quanto maior for o tempo em que o monitoramento do portal é realizado, uma vez que ao realizar o *crawling* de notícias de um portal é uma questão de tempo até que todas as notícias publicadas por este portal sejam indexadas, assim a data de *download* da notícia poderá ser considerada a data de publicação da mesma.

O *Portal 01*, que tem uma quantidade de notícias coletadas desproporcional do resto do conjunto analisado na Figura 6 possui notícias que a ferramenta consegue realizar a identificação da data de publicação corretamente e também notícias que a ferramenta falha nesse ponto, justificando a discrepância da quantidade de notícias coletadas provindas desta fonte. Na Figura 7 são analisadas apenas as notícias onde a ferramenta identificou a data de publicação. Nota-se um comportamento diferente na proporção de notícias pelo tempo, sem a ocorrência de portais de notícias com discrepância relevante. Também é possível observar que a ordem dos portais em relação a quantidade de notícias coletadas é alterada.

A solução foi avaliada por um especialista de alto escalão da SESP/PR, com mais de quinze anos de experiência na área da segurança pública, que constatou que é possível

Figura 7 – Quantidade de notícias com a data de publicação obtida por portal.

Fonte: autoria própria

obter informação útil de relevância do sistema implementado, sendo assim uma grande contribuição para a segurança pública do estado do Paraná. Também foi desenvolvido neste trabalho um manual de utilização da Interface gráfica do utilizador, que apresenta de modo intuitivo diversas funcionalidades da interface, com a finalidade de auxiliar o trabalho dos profissionais da SESP/PR na análise dos dados coletados.

Atualmente os profissionais da SESP/PR utilizam sistemas de busca comuns para realizar atividades de levantar informações sobre assuntos relevantes à própria SESP/PR. Com a solução desenvolvida neste trabalho, ao realizar uma busca, o usuário terá controle e conhecimento de como a busca funciona, assim como em quais portais de notícias estão sendo consultados e quais características estão sendo consideradas, o que não ocorre em um sistema de busca proprietário. Além disso, a plataforma oferece todas as informações relevantes ao usuário de forma única e simplificada, evitando assim que o usuário tenha que navegar por várias páginas até encontrar a informação que está buscando.

5 CONSIDERAÇÕES FINAIS

Foram apresentados estudos na área de Recuperação de Informação que buscam resolver problemas de segurança pública. Este trabalho abordou esse tema e contribui para diminuir o tempo de acesso a informações importantes, aumentando assim a produtividade da equipe pertencente à SESP/PR, conforme o problema apresentado no Capítulo 1.

A solução aplicada conta com a implementação de um software que realiza a interoperabilidade entre três módulos, sendo um módulo responsável pela coleta de dados, outro módulo pela seleção de documentos relevantes e o último módulo pela apresentação dos dados coletados para o usuário. A solução foi avaliada por especialistas da própria SESP/PR de forma positiva. Essa solução atende os objetivos deste trabalho e foi desenvolvida no intervalo de um ano, resultando em um sistema de fácil usabilidade que auxilia a SESP/PR a obter informações cruciais em tempo hábil.

Como trabalhos futuros, podem ser realizadas melhorias das funcionalidades desenvolvidas neste trabalho, ou estudos que envolvam os seguintes temas: uma aplicação de técnicas de mineração de dados para realizar a análise e identificações de padrões dos dados coletados de forma autônoma; uma aplicação de técnicas de classificação a fim de identificar a relevância de notícias e de suas características para a SESP/PR; e uma aplicação que colete informações de interesse da SESP/PR de redes sociais.

Referências

AGBELE, K.; AYETIRAN, E.; BABALOLA, O. A context-adaptive ranking model for effective information retrieval system. **International Journal of Information Science**, v. 8, n. 1, p. 1–12, 2018. Citado na página 17.

AGNEZ, L. F. Consumo da informação na sociedade contemporânea. In: **CONGRESSO BRASILEIRO DE CIÊNCIAS DA COMUNICAÇÃO, de**. [S.l.: s.n.], 2009. v. 4, p. 01–15. Citado na página 10.

ANH, V. N.; KRETSER, O. de; MOFFAT, A. Vector-space ranking with effective early termination. In: **Proc. SIGIR**. New York, NY: ACM Press, 2001. p. 35–42. ISBN 1-58113-331-6. Citado na página 14.

ANH, V. N.; MOFFAT, A. Improved word-aligned binary compression for text indexing. **IEEE Transactions on Knowledge and Data Engineering**, v. 18, n. 6, p. 857–861, 2006. Citado na página 14.

ANNAHAS, A. **Monitoramento de Fontes Abertas para Apoio de Atividade de Inteligência Pública v1.0.0**. [S.l.]: GitLab, 2020. <https://gitlab.com/alka1000/monitoramento-de-fontes-abertas-para-apoio-de-atividade-de-inteligencia-publica> v1.0.0. Acessado em 16/09/2020. Citado na página 11.

APACHE. **Lucene**. [S.l.]: Apache, 2020. <https://lucene.apache.org/>. Acessado em 04/05/2020. Citado na página 21.

BAEZA-YATES, R.; RIBEIRO-NETO, B. **Recuperação de Informação-: Conceitos e Tecnologia das Máquinas de Busca**. [S.l.]: Bookman Editora, 2013. 1–20,187–247,339–407 p. Citado 3 vezes nas páginas 13, 15 e 16.

BANKER, K. **MongoDB in action**. [S.l.]: Manning Publications Co., 2011. Citado na página 21.

BERNERS-LEE, T. J.; CAILLIAU, R. Worldwideweb: Proposal for a hypertext project. 1990. Citado na página 12.

CARNAZ, G. et al. An automated system for criminal police reports analysis. In: MADUREIRA, A. M. et al. (Ed.). **Proceedings of the Tenth International Conference on Soft Computing and Pattern Recognition (SoCPaR 2018)**. Cham: Springer International Publishing, 2018. p. 360–369. ISBN 978-3-030-17065-3. Citado na página 17.

ELASTIC. **Elasticsearch**. [S.l.]: elastic, 2020. <https://www.elastic.co/pt/elasticsearch>. Acessado em 04/05/2020. Citado na página 21.

ELASTIC. **Kibana**. [S.l.]: elastic, 2020. <https://www.elastic.co/pt/kibana>. Acessado em 04/05/2020. Citado na página 22.

FUHR, N.; PFEIFER, U. Probabilistic information retrieval as a combination of abstraction, inductive learning, and probabilistic assumptions. **TOIS**, ACM Press, New York, NY, USA, v. 12, n. 1, p. 92–115, 1994. ISSN 1046-8188. Citado na página 14.

HAMBORG, F.; BREITINGER, C.; GIPP, B. Giveme5w1h: A universal system for extracting main events from news articles. In: **Proceedings of the 13th ACM Conference on Recommender Systems, 7th International Workshop on News Recommendation and Analytics (INRA 2019)**. [S.l.: s.n.], 2019. Citado 2 vezes nas páginas 13 e 17.

HAMBORG, F.; DONNAY, K.; GIPP, B. Automated identification of media bias in news articles: an interdisciplinary literature review. **International Journal on Digital Libraries**, Springer, v. 20, n. 4, p. 391–415, 2019. Citado na página 13.

HAMBORG, F. et al. news-please: A generic news crawler and extractor. In: GAEDE, M.; TRKULJA, V.; PETRA, V. (Ed.). **Proceedings of the 15th International Symposium of Information Science**. [S.l.: s.n.], 2017. p. 218–223. Citado 3 vezes nas páginas 17, 18 e 20.

HANADI, A. S. **Relevance Feedback Optimization for Digital Forensic Investigations**. Tese (Doutorado) — The British University in Dubai (BUiD), 2019. Citado na página 18.

HULL, D. Stemming algorithms – A case study for detailed evaluation. **JASIS**, v. 47, n. 1, p. 70–84, 1996. Citado na página 15.

IBGE. **Panorama do estado do Paraná**. [S.l.]: Instituto Brasileiro de Geografia e Estatística, 2019. <https://cidades.ibge.gov.br/brasil/pr/panorama>. Acessado em 18/09/2019. Citado na página 10.

IDE, E. New experiments in relevance feedback. **The SMART retrieval system: Experiments in automatic document processing**, Prentice-Hall Inc., p. 337–354, 1971. Citado na página 16.

IETF. **Hypertext Transfer Protocol (HTTP/1.1): Message Syntax and Routing**. [S.l.]: Internet Engineering Task Force, 2014. <https://tools.ietf.org/html/rfc7230>. Acessado em 07/10/2019. Citado na página 12.

IETF. **URI Design and Ownership**. [S.l.]: Internet Engineering Task Force, 2014. <https://tools.ietf.org/html/rfc7320>. Acessado em 07/10/2019. Citado na página 12.

IPARDES. **Relação dos municípios do estado, ano de criação e respectivas mesorregiões e mocrorregiões geográficas e regiões geográficas - PARANÁ**. [S.l.]: Instituto Paranaense de Desenvolvimento Econômico e Social, 2019. <http://www.ipardes.gov.br/pdf/mapas/base_fisica/relacao_mun_regiao_geografica_parana.pdf>. Acessado em 18/09/2019. Citado na página 10.

KEKÄLÄINEN, J. Binary and graded relevance in IR evaluations – Comparison of the effects on ranking of IR systems. **IP&M**, v. 41, p. 1019–1033, 2005. Citado na página 15.

KOUZIS-LOUKAS, D. **Learning scrapy**. [S.l.]: Packt Publishing Ltd, 2016. Citado na página 13.

KUMAR, M. et al. Keyword query based focused web crawler. **Procedia Computer Science**, v. 125, p. 584 – 590, 2018. ISSN 1877-0509. The 6th International Conference on Smart Computing and Communications. Disponível em: <http://www.sciencedirect.com/science/article/pii/S1877050917328399>. Citado na página 17.

LIU, Y. et al. Roberta: A robustly optimized bert pretraining approach. **arXiv preprint arXiv:1907.11692**, 2019. Citado na página 13.

LOVINS, J. B. Development of a stemming algorithm. **Translation and Computational Linguistics**, v. 11, n. 1, p. 22–31, 1968. Citado na página 15.

MANNING, C.; RAGHAVAN, P.; SCHÜTZE, H. Introduction to information retrieval. **Natural Language Engineering**, Cambridge university press, v. 16, n. 1, p. 177–194,443–460, 2010. Citado 3 vezes nas páginas 12, 13 e 15.

MITCHELSTEIN, E.; BOCZKOWSKI, P. Online news consumption research: An assessment of past work and an agenda for the future. **New Media and Society**, SAGE Publications Ltd, v. 12, n. 7, p. 1085–1102, 11 2010. ISSN 1461-4448. Citado na página 10.

MYERS, D.; MCGUFFEE, J. W. Choosing scrapy. **Journal of Computing Sciences in Colleges**, Consortium for Computing Sciences in Colleges, v. 31, n. 1, p. 83–89, 2015. Citado na página 13.

NAJORK, M.; HEYDON, A. High-performance web crawling. In: **Handbook of massive data sets**. [S.l.]: Springer, 2002. p. 25–45. Citado na página 12.

PARANÁ. **Decreto Estadual nº 3.700/77**. 1977. Citado na página 10.

PARANÁ. **Decreto Estadual nº 1.045/99**. 1999. Citado na página 10.

ROBERTSON, S. E.; JONES, K. S. Relevance weighting of search terms. **JASIS**, v. 27, p. 129–146, 1976. Citado na página 15.

ROCCHIO, J. J. Relevance feedback in information retrieval. In: . [S.l.: s.n.], 1971. p. 313–323. Citado 2 vezes nas páginas 15 e 16.

SALTON, G.; MCGILL, M. J. **Introduction to modern information retrieval**. [S.l.]: mcgraw-hill, 1983. Citado na página 16.

SARR, E. N.; OUSMANE, S.; DIALLO, A. Factextract: automatic collection and aggregation of articles and journalistic factual claims from online newspaper. In: IEEE. **2018 Fifth International Conference on Social Networks Analysis, Management and Security (SNAMS)**. [S.l.], 2018. p. 336–341. Citado na página 13.

STONEBRAKER, M. Sql databases v. nosql databases. **Communications of the ACM**, ACM New York, NY, USA, v. 53, n. 4, p. 10–11, 2010. Citado na página 21.

SUN, Y.; ZHUANG, Z.; GILES, C. L. A large-scale study of robots. txt. In: ACM. **Proceedings of the 16th international conference on World Wide Web**. [S.l.], 2007. p. 1123–1124. Citado na página 13.

THACKER, U.; PANDEY, M.; RAUTARAY, S. S. Performance of elasticsearch in cloud environment with ngram and non-ngram indexing. In: IEEE. **2016 International Conference on Electrical, Electronics, and Optimization Techniques (ICEEOT)**. [S.l.], 2016. p. 3624–3628. Citado na página 21.

TOMLINSON, S. Lexical and algorithmic stemming compared for 9 European languages with Hummingbird Searchserver at CLEF 2003. In: **Proc. Cross-Language Evaluation Forum**. [S.l.: s.n.], 2003. p. 286–300. Citado na página 15.

W3C. **HTML 4.01 Specification**. [S.l.]: World Wide Web Consortium, 1999. <https://www.w3.org/TR/html401/>. Acessado em 07/10/2019. Citado na página 12.

WONG, S. M.; ZIARKO, W.; WONG, P. C. Generalized vector spaces model in information retrieval. In: ACM. **Proceedings of the 8th annual international ACM SIGIR conference on Research and development in information retrieval**. [S.l.], 1985. p. 18–25. Citado na página 14.

XU, J.; CROFT, W. B. Quary expansion using local and global document analysis. In: ACM. **Acm sigir forum**. [S.l.], 2017. v. 51, n. 2, p. 168–175. Citado na página 16.

ZHANG, W. **Formal description and development of graphical user interfaces**. [S.l.]: Herbert Utz Verlag, 1996. Citado na página 21.