UNIVERSIDADE TIRADENTES

PRÓ-REITORIA DE PÓS-GRADUAÇÃO, PESQUISA E EXTENSÃO - PPgPE

PROGRAMA DE PÓS-GRADUAÇÃO EM EDUCAÇÃO - PPED

DOUTORADO EM EDUCAÇÃO

FABIO GOMES ROCHA

O USO DA INTELIGÊNCIA ARTIFICAL PARA A PESQUISA EM HISTÓRIA DA EDUCAÇÃO: UMA PROPOSTA DE INSTRUMENTO TECNÓLOGICO

ARACAJU (SE)

2020

FABIO GOMES ROCHA

O USO DA INTELIGÊNCIA ARTIFICAL PARA A PESQUISA EM HISTÓRIA DA EDUCAÇÃO: UMA PROPOSTA DE INSTRUMENTO TECNÓLOGICO

Tese apresentada como pré-requisito parcial para obtenção do título de Doutor Programa de Pós-Graduação em Educação na linha Educação e Formação Docente – Universidade Tiradentes.

Orientadora: Profa. Dra. Ester Fraga Vilas-Bôas Carvalho do Nascimento

ARACAJU (SE)
2020

Rocha, Fábio Gomes
R672u O uso da inteligência artificial para a pesquisa em história da educação: uma proposta de instrumento tecnológico/ Fabio Gomes Rocha; orientação [de] Prof.ª Dr.ª Ester Fraga Vilas- Bôas Carvalho do Nascimento– Aracaju: UNIT, 2020.

128 f. il ; 30 cm
Tese (Doutorado em Educação) - Universidade Tiradentes, 2020
Inclui bibliografia.

1. História da educação 2 Inteligência artificial 3. Recuperação de informação 4. Processamento de imagem 5. Tecnologia para pesquisa I. Rocha, Fábio Gomes II. Nascimento, Ester Fraga Vilas-Bôas Carvalho do (orient.). III. Universidade Tiradentes. IV. Título.

CDU: 37:004.8

SIB- Sistema Integrado de Bibliotecas

FABIO GOMES ROCHA

O USO DA INTELIGÊNCIA ARTIFICAL PARA A PESQUISA EM HISTÓRIA DA EDUCAÇÃO: UMA PROPOSTA DE INSTRUMENTO TECNÓLOGICO

Tese apresentada como pré-requisito parcial para obtenção do título de Doutor Programa de Pós-Graduação em Educação na linha Educação e Formação Docente – Universidade Tiradentes.

APROVADO EM:

BANCA EXAMINADORA

Profa. Dra. Ester Fraga Vilas-Bôas Carvalho do Nascimento (Orientadora)

Prof. Dr. Alejandro C. Frery (Universidade Federal de Alagoas)

Prof. Dr. Paulo Sérgio Marchelli (Universidade Federal de Sergipe)

Profa. Dra. Andréa Karla Ferreira Nunes (Universidade Tiradentes)

Prof. Dr. Ronaldo Nunes Linhares (Universidade Tiradentes)

ARACAJU (SE)

2020

DEDICATÓRIA

A minha esposa, Rosimeri, pelo apoio incondicional e pela dedicação a mim.

A minha filhinha de quatro patas, Melissa, por me animar e pelo carinho em todos os momentos.

Aos meus pais, Ilson e Rosa, por me preparar para a vida por meio de seus exemplos e servindo-me de guia para a minha jornada.

Ao Prof. Alejandro C. Frery por ter sido um mentor e uma permanente fonte de inspiração.

AGRADECIMENTOS

Este trabalho não seria possível sem a ajuda e colaboração de inúmeras pessoas, para as quais seguem os meus sinceros agradecimentos. Mas, antes de tudo e todos, agradeço a Deus, pelas bênçãos cotidianas que me fortaleceram e ampararam no caminho desta pesquisa.

Aos alunos do projeto DigitalSE e do GPITIC por atuarem ativamente no grupo de pesquisa e no projeto.

Ao colega e parceiro de pesquisa Guillermo Rodriguez.

Ao Wallace e ao Ricardo Roriz por acreditarem no projeto Hemeroteca Digital, dando início ao que, posteriormente, viria a ser o DigitalSE.

Aos colegas de doutorado Elias, Verônica, Arleide, Janilce, Fabio, Maria, Michelline, Salete e João que compartilharam diversos momentos e debates em sala de aula, que auxiliaram para o amadurecimento das pesquisas.

Aos professores do Doutorado em Educação da Unit e, em especial, à Professora Dra. Ada Augusta (*in memoriam*) por inspirar a todos durante suas aulas, pelos seus ensinamentos, pelas oportunidades de compartilhar ideias e reflexões, essenciais para a construção de conhecimento.

Um agradecimento especial à Professora Dra. Andrea Karla, que me deu apoio em diversos momentos.

Ao meu coordenador nos cursos de computação da UNIT, Professor Me. Fabio Batista Santos, pela compreensão e apoio.

À Professora Dra. Ester, por acreditar neste projeto e me aceitar como seu orientando.

Impossível não destacar o meu mentor nesta jornada, Professor Dr. Alejandro C. Frery, que acreditou e incentivou para que a minha pesquisa avançasse. O seu exemplo como professor, pesquisador, orientador e amigo vai muito além da sala de aula. O senhor, Professor Alejandro, é uma fonte contínua de inspiração pessoal e profissional.

Aos mestres que, tão atenciosamente, participaram dos momentos de qualificação e defesa desta tese, contribuindo com relevantes sugestões para a pesquisa: Alejandro C. Frery, Paulo Sérgio Marchelli, Andrea Karla Nunes e Ronaldo Linhares.

À UNIT, pela oportunidade de bolsa para o meu aperfeiçoamento como pesquisador, viabilizando a execução desta pesquisa.

E a vocês, Rosimeri Ferraz Sabino e Melissa, minha companheira de todas as horas e minha filha de quatro patas. Rosi, não há como expressar minha gratidão, por todo o apoio,

amor, carinho, por dividir comigo as descobertas, pela compreensão sobre os momentos de ausência, pelo apoio nas horas difíceis, pelo estímulo, auxílios e conselhos neste percurso. Melissa, minha eterna “bebê, que Deus lhe dê muitos anos conosco e que, quando chegada a hora de você virar uma estrelinha, nós tenhamos a certeza do reencontro.

EPÍGRAFE

“Por meio da tecnologia, obstáculos ancestrais à interação humana, como geografia, linguagem e informação limitada, vão cedendo, e uma nova onda de criatividade e potencial humanos vai se elevando. A adesão em massa à internet está promovendo uma das mais empolgantes transformações sociais, culturais e políticas da história, e, ao contrário do que ocorreu nos períodos de mudança anteriores, desta vez os efeitos são globais”. (SCHMIDT; COHEN, 2013, p. 6).

RESUMO

Esta tese teve como objetivo o desenvolvimento de um instrumento digital, voltado à descoberta de padrões em imagens de fontes documentais histórico-educacionais, com o emprego da inteligência artificial (IA). A pesquisa classifica-se como aplicada, utilizando a metodologia *design science research*, que se volta a soluções de problemas ou a produção de um artefato. As análises do estudo foram realizadas sob abordagem quantitativa, no tocante ao desempenho técnico-operacional do instrumento, e qualitativa em relação à contribuição dele para as buscas documentais em História da Educação. Inicialmente, foi realizada a identificação de instrumentos que empregam tecnologias contemporâneas para a pesquisa em História da Educação e caracterizados os métodos de extração de dados das fontes históricas. Em prosseguimento, foram caracterizadas as técnicas de indexação de informações e aplicadas em classificação de documentos da História da Educação, bem como foram analisados os modelos existentes de busca inteligente de informação e a sua eficiência em relação a documentos histórico-educacionais. Por fim, implementou-se o instrumento denominado SPEdu, que permitiu a integração dos itens de extração, indexação e busca de fontes da História da Educação. Essa ação foi validada por meio de um estudo de caso, aplicado em números do Diário Oficial do Estado de Sergipe, para a verificação do desempenho técnico-operacional do instrumento e seu auxílio para as buscas documentais em História da Educação. Como parte da metodologia *design science research*, o estudo de caso permitiu a avaliação observacional sobre o instrumento elaborado. Em conclusão, comprovou-se que a inclusão de IA nas TICs adotadas pelos pesquisadores da História da Educação auxilia na ampliação de fontes de análise e de resultados de buscas, reduzindo o trabalho manual e permitindo a gestão dos dados.

Palavras-chave: História da Educação. Inteligência artificial. Recuperação de informação. Processamento de imagem. Tecnologia para pesquisa.

ABSTRACT

This thesis aimed at the development of a digital instrument, focused on the discovery of patterns in images from historical and educational documental sources, with the use of artificial intelligence (AI). The research is classified as applied, using the design science research methodology, which turns to problem solutions or the production of an artifact. The analysis of the study was carried out under a quantitative approach, regarding the technical-operational performance of the instrument, and qualitative in relation to its contribution to the documental searches in History of Education. Initially, the identification of instruments that employ contemporary technologies for research in the History of Education and characterized the methods of data extraction from historical sources were carried out. In continuation, the techniques of information indexing were characterized and applied to the classification of documents in the History of Education, as well as the existing models of intelligent information search and their efficiency in relation to historical-educational documents were analyzed. Finally, the instrument called SPEdu was implemented, which allowed the integration of extraction, indexation and search items from sources of the History of Education. This action was validated by means of a case study, applied in numbers of the Official Gazette of the State of Sergipe, to verify the technical-operational performance of the instrument and its assistance for document searches in History of Education. As part of the design science research methodology, the case study allowed the observational evaluation of the instrument. In conclusion, it was proved that the inclusion of AI in the ICT's adopted by researchers in the History of Education helps to expand sources of analysis and search results, reducing manual work and allowing data management.

Keywords: Artificial Intelligence. History of Education. Image processing. Information retrieval.

RESUMEN

Esta tesis tuvo como objetivo el desarrollo de un instrumento digital, centrado en el descubrimiento de patrones en imágenes de fuentes documentales históricas y educativas, con el uso de la inteligencia artificial (IA). La investigación se clasifica como aplicada, utilizando la metodología de investigación de la ciencia del diseño, que se convierte en soluciones de problemas o en la producción de un artefacto. El análisis del estudio se realizó bajo un enfoque cuantitativo, en lo que respecta a las prestaciones técnico-operativas del instrumento, y cualitativo en lo que respecta a su contribución a las búsquedas documentales en Historia de la Educación. Inicialmente, se llevó a cabo la identificación de los instrumentos que emplean tecnologías contemporáneas para la investigación en la Historia de la Educación y se caracterizaron los métodos de extracción de datos de fuentes históricas. A continuación, se caracterizaron y aplicaron las técnicas de indización de la información para la clasificación de documentos en la Historia de la Educación, y se analizaron los modelos existentes de búsqueda inteligente de información y su eficacia en relación con los documentos histórico-educativos. Por último, se puso en marcha el instrumento denominado SPEdu, que permitió la integración de elementos de extracción, indexación y búsqueda de fuentes de la Historia de la Educación. Esta acción fue validada mediante un estudio de caso, aplicado en números del Boletín Oficial del Estado de Sergipe, para verificar el rendimiento técnico-operativo del instrumento y su ayuda para la búsqueda de documentos en Historia de la Educación. Como parte de la metodología de investigación de la ciencia del diseño, el estudio de caso permitió la evaluación observacional del instrumento. En conclusión, se demostró que la inclusión de la IA en las TIC adoptadas por los investigadores de Historia de la Educación ayuda a ampliar las fuentes de análisis y los resultados de las búsquedas, reduciendo el trabajo manual y permitiendo la gestión de los datos.

Palabras clave: Historia de la Educación. Inteligencia Artificial. Procesamiento de imágenes. Recuperación de información. Tecnología para la investigación.

LISTA DE QUADROS

LISTA DE GRÁFICOS

LISTA DE FIGURAS

LISTA DE FÓRMULAS

LISTA DE CÓDIGOS

LISTA DE TABELAS

LISTA DE SIGLAS E ABREVIATURAS

CAPES	Coordenação de Aperfeiçoamento de Pessoal de Nível Superior
CBSE	Corpo de Bombeiros do Estado de Sergipe
CD	Compact Disk
CSV	Comma-Separated Vallues
DDD	Domain Driven Design
DER	Diagrama de Entidades Relacionais
DVD	Digital Versatile Disk
FAPESP	Fundação de Amparo à Pesquisa do Estado de São Paulo
GB	Gigabyte
GQM	Goal Question Metric
IA	Inteligência Artificial
JSF	Java Server Faces
KDD	Knowledge Discovery in Databases
MLP	Perpecptron Multicamadas
NLP	Natual Language Process
OCR	Reconhecimento Ótico de Caracteres
OS	Open Source
PDF	Portable Document Format
RGB	Vermelho Verde e Azul
RI	Recuperação de Informação
SEGRASE	Serviços Gráficos do Estado de Sergipe
SEFAZ	Secretaria da Fazenda do Estado de Sergipe
SPEdu	Sistema de Pesquisas Educacionais
SVM	Máquina Vetor de Suporte
TIC	Tecnologia da Informação e Comunicação

SUMÁRIO

1 INTRODUÇÃO

O objetivo desta tese foi o desenvolvimento de um instrumento digital, voltado à recuperação de informações de fontes documentais histórico-educacionais, com o emprego da inteligência artificial (IA). Para isso, foi realizada a identificação de instrumentos que empregam tecnologias contemporâneas para a pesquisa em História da Educação e caracterizados os métodos de extração de dados das fontes históricas. Em prosseguimento, foram caracterizadas as técnicas de indexação de informações a serem aplicadas em classificação e documentos da História da Educação, bem como se analisou os modelos existentes de busca de informação e a sua eficiência em relação a documentos histórico-educacionais. Por fim, implementou-se o instrumento que permite a integração dos itens de extração, indexação e busca de fontes da História da Educação. Essa ação foi validada por meio de um estudo de caso para a verificação do desempenho técnico-operacional do instrumento e seu auxílio para as buscas documentais em História da Educação.

O instrumento desenvolvido vai ao encontro da necessidade de criação de planos de gestão de dados pelos pesquisadores. A Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP) já exige formas de gestão e compartilhamento de dados de pesquisas, cujos responsáveis recebam financiamentos. Conforme a Fapesp (2017), essa demanda já é obrigatória nos Estados Unidos, Europa e Austrália. Nesse sentido, o objetivo é que os pesquisadores expliquem como os dados produzidos pela pesquisa serão tratados e disponibilizados de forma aberta, quando advindas de pesquisas com financiamento público.

Como exemplo de volume de dados de pesquisa que demanda organização, indexação e formas de busca, tem-se o projeto de produtividade em educação "Brasil, Portugal e Inglaterra: circulação de impressos protestantes e outros impressos durante a segunda metade dos oitocentos", de autoria da Profa. Dra. Ester Fraga Vilas-Bôas Carvalho do Nascimento, para o período de 2016-2019[1]. Com o objetivo de efetuar o levantamento da documentação da Sociedade Bíblica Britânica (BFBS), bem como de analisar outros documentos, a pesquisa pretende sistematizar a documentação levantada, criando um banco para catalogar e armazenar os dados. Já em 2017, tais dados ultrapassavam 45 gigabytes (GB) de documentos, os quais, após digitalizados e sistematizados, foram publicados em

[1] Edital MCT/CNPq 02/2009; Edital Universal CNPq 14/2011 Faixa B; Edital Universal, 2015; Bolsa de Produtividade em Educação/CNPq 2012, 2015, 2018

compact disk (CD) ou *digital versatile disk* (DVD) para a disponibilização à comunidade acadêmica.

Na referida pesquisa, foi necessário fotografar e digitalizar documentos, identificados em diversos países. Isso exigiu um tempo considerável, além do trabalho manual na catalogação de uma das fontes, com o objetivo de criar um banco com os diversos metadados do documento. Após, foi necessário gravar todos os documentos em mídia. O tamanho da base documental levantada ultrapassou os limites de armazenamento de dez DVDs. Além dos custos desse processo, os dados não contam com um modelo de indexação automática, o que onera o tempo de busca. A observação sobre as ações que envolviam o projeto de produtividade da Profa. Dra. Ester permitiu identificar as dificuldades dos pesquisadores para a gestão e disponibilização dos dados de pesquisas no campo da História da Educação. Isso motivou o desenvolvimento da tese, com foco na criação de um artefato para auxiliar os pesquisadores na gestão documental, por meio da recuperação de informação, com a adoção da IA.

Considerando-se a relevância dos documentos como fonte para investigações históricas (LE GOFF, 2003) e o cenário atual de informações eletrônicas e em rede, a transformação do documento físico para o digital se coloca como imperativo na contribuição à preservação, ao acesso, à disseminação e às análises de dados documentais (CHARTIER, 1999, 1999a; DARTON, 2010; MELLO, SANTOS, OLIVEIRA, 2011). Nesse contexto, explora-se o uso da IA na classificação e recuperação de informações de fontes documentais (ROCHA, CORTEZ, NEVES, 2008; BAEZA-YATES, RIBEIRO-NETO, 2013; RUSSEL, NORVIG, 2013).

As fontes, matéria-prima do historiador, surgem como meio enriquecedor para o conhecimento do contexto que as envolvem. Em sua busca de elementos que façam emergir evidências sobre o objeto investigado, o pesquisador deve, então, considerar o maior número de fonte, questionando-as e relacionando-as de forma a obter a maior proximidade possível do cenário investigado, diante das comprovações encontradas.

No âmbito das fontes documentais, Le Goff (2003) alerta que o documento não é um retrato da realidade, mas o resultado de escolhas conscientes ou inconscientes, tanto na sua etapa de elaboração como na de preservação. Toma-se conhecimento do passado com base nos documentos preservados, os quais foram escritos por pessoas que ali aplicaram os seus pontos de vista. A busca pelas condições de produção de um documento (LE GOFF, 2003) impele o historiador a considerações de fontes adicionais que elucidem uma situação, inicialmente, dada como fato documental. O exame apurado sobre fontes diversas impõe o

uso de recursos que possibilitem a identificação do maior número de evidências que se relacionem ao cenário e objeto investigado. Para isso, é necessário que o pesquisador adote tecnologias que permitam armazenar, catalogar e consultar fontes de investigação.

Tais tecnologias tiveram a sua utilização multiplicada pelos historiadores partir da década de 1980 (LOMBARDI, 2000). Alguns dos fatores que contribuíram para isso foram o barateamento dos equipamentos e a ampliação do uso dos microcomputadores, surgidos na década de 1970, nos Estados Unidos, e no Brasil, em 1974. Na atualidade, no entanto, a aplicação dessas tecnologias ainda não se mostra consolidada na produção intelectual brasileira. Segundo Amorim (2000), referindo-se às investigações no campo da História da Educação, a desorganização e baixa utilização de tecnologias contemporâneas por parte dos órgãos de documentos resultam em obstáculos aos pesquisadores.

Além disso, bases dispersas inviabilizam ou, no mínimo, dificultam a descoberta de informações relevantes ao trabalho do pesquisador. Assim, o emprego de tecnologias digitais contemporâneas para o trato documental pode subsidiar pesquisas históricas no tocante à conservação dos dados, consulta integrada e exploração de novos conhecimentos em base de dados. Essas ações se relacionam com a área da inteligência artificial (IA), a qual visa criar entidades computacionais inteligentes capazes de perceber e resolver problemas (RUSSELL, NORVIG, 2013).

No âmbito da IA, o processo denominado "*Knowledge Discovery in Databases*" (KDD) está relacionado ao tratamento de dados, os quais constam, inclusive, em documentos. O termo KDD foi cunhado em 1989, no âmbito das discussões do *International Joint Conferencce on Artificial Intelligence*, em Detroit (Michigan/EUA). A ocasião reuniu os principais pesquisadores em aprendizagem de máquina, banco de dados, lógica difusa, aquisição de conhecimentos, entre outras áreas (PIATETSKY-SHAPIRO, 1990). O processo KDD visa transformar dados brutos em conhecimento de alto-nível (ROCHA, CORTEZ, NEVES, 2008), o que possibilita aplicá-lo em contextos específicos no âmbito educacional. A extração, organização e descoberta de indícios, de forma sistêmica, pode apoiar análises históricas (GOLDSCHMIDT, PASSOS, BEZERRA, 2015).

O emprego de tecnologia contemporânea para suporte à pesquisa histórica, segundo Gondra (2000), é uma imposição da atualidade aos investigadores da área da História da Educação, não devendo se limitar à aplicação técnica, mas se estender ao exame de processos tecnológicos que venham a potencializar a exploração de fontes. Disso depreende-se a necessidade de uma relação com outros campos disciplinares, focalizando eventuais

contribuições para o acesso, análise e disponibilização de materiais ou acervos documentais. Neste sentido, Bonato (2004) afirma que o campo da História da Educação amplia-se nas suas possibilidades de pesquisa por meio das novas tecnologias, visto que elas podem fornecer novos suportes a registros, armazenamento e recuperação de informações. Além disso, Tavares (2015) afirma que padronizações e sistematizações de dados por meio de *softwares* de gerenciamento de banco de dados servem de apoio na consolidação de pesquisas para mais de um pesquisador.

Dessa forma, tornam-se possíveis investigações sobre materiais tratados com a aplicação de tecnologias contemporâneas para a conversão de impressos para o meio digital, constituindo uma "biblioteca sem muros", segundo a noção de Chartier (1999a). A etimologia do termo "biblioteca", na busca de Chartier (1999a), indica tanto um espaço organizado para livros, como livros que contenham catálogos de livros de bibliotecas ou, ainda, "[...] uma compilação de várias obras da mesma natureza ou de autores que compilaram tudo que se pode dizer sobre um mesmo tema" (CHARTIER, 1999a, p. 70). Nesse último sentido, é possível identificar que a reunião de materiais documentais sobre um tema investigado contribui para a mobilidade do conhecimento sobre ele. Situando tal conhecimento diante da atualidade dos meios eletrônicos, Chartier (2017, p. 20) menciona, também, que "[...] a conversão das coleções existentes promete a construção de uma biblioteca sem muros"

Uma vez transposto para o meio digital, determinado volume de documentos pode ser objeto de análise por meio da IA. Considerando-se que os documentos, nas suas mais variadas formas, não podem ser tomados como um mero dado em uma investigação, a IA viabiliza um exame minucioso sobre fatos e relações que envolvem o contexto do teor documental. Isso se torna possível pela ampliação de materiais da análise trazida pelo pluralismo de documentos de onde venham emergir indícios sobre um tema investigado. Além de constituir-se em uma via para uma interpretação interdisciplinar por parte do pesquisador, a IA também permite a redução da escala de observação, conforme defendida por Ginzburg (1989, p. 150), quando menciona que "[...] pistas talvez infinitesimais permitem captar uma realidade mais profunda, de outra forma inatingível"

Observa-se, assim, o necessário diálogo do campo da História da Educação com a Arquivologia e a Tecnologia da Informação e Comunicação (TIC). Assim, para a abordagem sobre um "instrumento digital de pesquisa" considerou-se o pressuposto de que o uso de IA na História da Educação resulta em eficiência na busca de informação. Isso se confirma no contexto das ações necessárias para recuperação da informação (RI), campo da IA, em um

universo de grande volume. Essa é uma área que evoluiu para além da mera indexação de textos e documentos. Na atualidade, a RI envolve desde a arquitetura de sistemas a interfaces com o usuário, considerando as formas mais eficientes para a visualização e filtragem de dados. Tais elementos voltam-se à "[...] construção de índices eficientes, [ao] processamento de consultas com alto desempenho e [ao] desenvolvimento de algoritmos de ranqueamento, a fim de melhorar os resultados" (BAEZA-YATES, RIBEIRO NETO, 2013, p. 1).

A aplicação da IA sobre materiais históricos também encontra respaldo no trabalho Piotrowski (2012), que explora o processamento de linguagem natural, campo da IA, na análise documental. Para aquele autor, além da IA permitir maior grau de preservação de acervos, ela repercute diretamente na ampliação da acessibilidade a documentos históricos, bem como na eficiência da busca em larga escala. Isso ocorre pela oportunidade que a IA apresenta para a elaboração de estratégias de buscas (RUSSELL; NORVIG, 2013), permitindo a construção de caminhos entrelaçados, sob intersecções temáticas, que levam à maior assertividade nos resultados.

O campo da História da Educação, embora venha adotando TICs como forma de gerenciar acervos documentais, emprega, majoritariamente, CD, DVD, banco de dados, microfilmes, etc. Tais recursos não auxiliam na busca de informações. Diante desse contexto, a tese que se defende é que a inclusão de IA nas TICs adotadas pelos pesquisadores da História da Educação permite a automação de extração, indexação e busca documental, auxiliando na ampliação de fontes de análise e de resultados de buscas, reduzindo o trabalho manual e permitindo a gestão dos dados.

A pesquisa classifica-se, assim, como aplicada (GIL, 2017) utilizando a metodologia *design science*, que se volta a "[prescrever] soluções e métodos para resolver determinado problema ou projetar um novo artefato" (DRESH, LACERDA, ANTUNES JÚNIOR, 2015, p. 52). A adoção dessa metodologia é defendida por Johannesson e Perjons (2014) não apenas como meio de criação de artefatos, mas, também, pela possibilidade trazida por ela à geração de conhecimentos sobre esses artefatos e de seu ambiente de aplicação. Obteve-se como resultados os modelos mais adequados para as tarefas aqui explanadas e um sistema de código aberto, o qual visa permitir que outros pesquisadores utilizem e colaborem futuramente com o instrumento.

Raymond (1999) apresentou um ensaio denominado "A catedral e o bazar", sendo este considerado o manifesto do movimento de software aberto (Open Source). Aquele autor faz alusão ao código fechado como uma catedral e o modelo Open Source (OS) como um bazar, desenvolvido de forma aberta e pública. Raymund (1999) enfatiza que o modelo OS

traz diversos benefícios, como a colaboração, o aprendizado, a velocidade da criação de novas versões e a integração do desenvolvimento com a comunidade de usuários. Outro ponto que a destacar é que um sistema OS permite que outros utilizem, aprendam, melhorem a ferramenta, criem versões, ou seja, além da liberdade, tem-se um aprendizado contínuo com as colaborações. Há, ainda, a redução de custo com licenças. Nesse sentido, os artefatos, resultado desta pesquisa, serão liberados de forma aberta, sendo possível que pesquisadores utilizem, compartilhem, melhorem e colaborem com a evolução do sistema.

A validação do instrumento de pesquisa foi realizada por meio de estudo caso, empregando o jornal Diário Oficial do Estado de Sergipe. A empresa estatal Imprensa Oficial, surgida da necessidade de publicação de atos oficiais do Governo de Sergipe, foi criada em 1895, pelo, então, presidente da província Manoel Prisciliano de Oliveira Valadão. Atualmente denominada Serviços Gráficos do Estado de Sergipe (SEGRASE), a empresa ainda configura-se como a gráfica oficial do estado, sendo responsável por publicar muitas das obras dos escritores regionais, como cordéis, livros, revistas e, também, pela produção e impressão de outros documentos oficiais, como o Diário da Justiça. A SEGRASE está localizada na Rua Propriá, número 227, centro de Aracaju/Sergipe. A gráfica tem como principal função produzir o Diário Oficial do Estado, documento onde são registrados as licitações, decretos, portarias, nomeações, exonerações e outras informações que o estado deve disponibilizar de maneira pública. Além disso, esse jornal tem grande representatividade histórica, já que muitas de suas páginas foram dedicadas a reportar notícias memoráveis, como fatos e acontecimentos da II Guerra Mundial. A partir de 2012, o Diário Oficial passou a ser digital, havendo, portanto, números impressos no período de 1895 a 2011 disponíveis para uso na presente tese. Como resultados da etapa da validação do instrumento aqui proposto, foram digitalizados números do mencionado jornal, referentes a quatorze anos de publicação em papel, relativos aos anos de 1997 a 2011, totalizando 32.004 páginas de jornal, liberados por meio do Termo de Convênio SEGRASE n° 001/2017 para o desenvolvimento da Hemeroteca Digital da SEGRASE.

Como parte da metodologia *design science research*, o estudo de caso permitiu a avaliação observacional sobre o instrumento elaborado. As análises do estudo foram realizadas sob abordagem quantitativa, no tocante ao desempenho técnico-operacional do instrumento, e qualitativa em relação à contribuição dele para as buscas documentais em História da Educação.

Como etapa inicial da pesquisa, foram realizadas buscas no Portal de Periódicos e no Banco de Teses da Coordenação de Aperfeiçoamento de Pessoal do Nível Superior

(CAPES), identificando-se trabalhos, publicados até o primeiro semestre de 2018, que empregam tecnologias contemporâneas no contexto da História da Educação, e verificando-se o objetivo do uso desses recursos nas respectivas pesquisas. Os termos empregados para as buscas foram: tecnologia digital, novas tecnologias, história e história da educação. Essas palavras foram organizadas da seguinte forma: (("tecnologia digital" OR "novas tecnologias") AND ("história da educação" OR "história")). Após a busca, foi realizada a leitura do título e resumo para uma seleção inicial. Considerou-se nessa ação os trabalhos em que os pesquisadores adotaram TICs aplicadas à pesquisa em História ou História da Educação. Como resultado, obteve-se dezesseis produções. Constatou-se que a tecnologia mais empregada pelos pesquisadores é a digitalização, encontrada em quinze trabalhos. O foco dos autores é o uso da digitalização para a preservação documental.

Considerando-se que um trabalho pode abordar a aplicação de mais de uma tecnologia, identificou-se dez produções em que seus autores também implementaram bases de dados. Ressalta-se que essa é uma ação relevante para a consolidação de repositórios de armazenamento e catalogação de informações, facilitando o desenvolvimento de novas investigações ou ampliando as já existentes. Para o armazenamento, preservação e distribuição de conteúdo identificou-se nove trabalhos que utilizaram as mídias CD ou DVD. Com menor número de trabalhos, o site, como tecnologia que também permite a distribuição de conteúdo, foi abordado em sete produções. E, por fim, constatou-se apenas uma produção com proposta desenvolvida para um museu virtual.

Em relação ao contexto em que a tecnologia esteve empregada nas produções, constatou-se que há maior tendência de uso para a preservação documental e redução de espaço físico para o armazenamento. No trabalho de Vieira (2011), que trata sobre Arquivo Público do Estado do Paraná, são apresentados os avanços que a microfilmagem e a digitalização de documentos públicos podem trazer aos pesquisadores da História da Educação. Já no trabalho de Bonato (2004) tem-se a abordagem sobre a efetiva contribuição das novas tecnologias para a ampliação e diversificação de fontes nas pesquisas de História da Educação. No entanto, essa autora expõe a necessidade de atenção sobre o acompanhamento dos meios utilizados para o armazenamento. Segundo Bonato (2004), é importante que seja realizada a transferência da informação para novos suportes, sob o risco de, em um futuro próximo, não ser possível obter o acesso a ela devido a um formato tecnológico obsoleto.

Em se tratando de preservação, o trabalho de Pena e Silva (2008) apresenta a digitalização documental como forma de proteger documentos históricos. Para uma gestão

eficaz dessa documentação, as autoras propõem uma metodologia para implementação de um sistema de gestão eletrônica de documentos (GED), o que promove a democratização no acesso das informações on-line. O foco sobre digitalização documental também foi identificado na produção de Lopes et al. (2016). Esses autores descrevem o seu trabalho na criação de um DVD com fontes para a História da Educação de Ouro Preto do Oeste, Estado de Rondônia, contendo vídeos de sessenta e quatro entrevistas, além de outras fontes documentais e iconográficas. O DVD visou permitir uma preservação da História da Educação, muitas vezes desconhecida, sobre aquela cidade. Ainda voltado para a preservação histórica, o trabalho de Cabral (2002) apresenta o processo de digitalização como forma de conservar e preservar materiais informativos em bibliotecas e arquivos, salientando a importância do planejamento para a adoção das novas tecnologias e apresentando como vantagens a economia de espaço físico e reunião de documentos, ora dispersos em um mesmo local.

No trabalho de Siqueira (2005) é apresentado o resultado do Grupo Educação e Memória (GEM) da Universidade Federal de Mato Grosso (UFMT), com a organização de fontes escritas e entrevistas por meio da criação de um banco de dados, banco de vozes e banco de fontes, e gerando CDs que integraram um conjunto de fontes privadas, públicas e familiares. Essas ações visaram facilitar o acesso aos pesquisadores sobre a História da Educação daquele estado. Na mesma linha da preservação de fontes, identificou-se o trabalho de Souza (2013), que aponta como imprescindível o debate sobre a conservação e proteção documental como condição para o desenvolvimento do patrimônio educativo do país.

Em consonância às reflexões promovidas por Souza (2013), o trabalho de Silva (2011) também aborda a importância da preservação de documentos de arquivos escolares, apontando como urgente a intervenção de áreas como Arquivologia e Ciências da Informação junto aos pesquisadores da História e História da Educação. O diálogo e o trabalho interdisciplinar entre esses campos se fazem necessário à busca de meios atualizados para a salvaguarda e disponibilização documental eficiente. Essa perspectiva sobre os documentos escolares também foi identificada no trabalho de Fernandes (2010), o qual aponta a fragilidade de recursos tecnológicos e humanos para as atividades de arquivamento e preservação documental nas escolas, comprometendo a recuperação de memórias históricas. Esse autor destaca a urgente necessidade de digitalização do acervo dessas instituições. Por óbvio, essa ação envolve a questão da gestão escolar, aspecto presente na produção de Toschi e Rodrigues (2003), onde é exposto o trabalho de elaboração

de um CD com materiais audiovisuais para as disciplinas de História da Educação e Prática de Ensino. O mesmo conteúdo foi utilizado para a criação, pelos autores, de um museu virtual sobre educação. As fontes para o estudo de Toschi e Rodrigues (2003) foram coletadas a partir do ambiente de escolas nas cidades de Goiânia, Anápolis, Jataí e Catalão. Em suas reflexões, os autores afirmam a necessidade de articulação entre as dimensões acadêmica, técnica e de gestão.

Voltado também para a realidade das escolas, Pereira (2011) trabalhou as fontes documentais para a História da Educação, empregando fontes diversas, entre elas as de gênero denominado "informático", as quais são constituídas de CDs, DVDs e disquetes, buscando classificar e hierarquizar o conteúdo encontrado, a fim de constituir uma memória da educação básica pública no Distrito Federal. Essa autora também criou um modelo de processo de digitalização para a elaboração de acervo documental. Esse recurso também foi empregado no trabalho de Soares, Braga e Lima (2015), que apresentam a digitalização de documentos referentes à educação durante o regime militar. O processo foi desenvolvido pela classificação, digitalização, armazenamento e disponibilização, por meio de um repositório digital. Esse mesmo meio tecnológico foi utilizado por Louveira e Ferro (2013). Essas autoras elaboraram um banco de dados fotográfico de crianças e infância do sul do Mato Grosso, criando um catálogo denominado "FotoMemo", o qual foi disponibilizado para uso por meio de CDs e DVDs. A mídia CD também foi adotada no trabalho de Werle (2007) como forma de apoio à pesquisa de identidade e história institucional, discutindo a digitalização e a organização em CDs de documentos que permitam a reconstrução da história de instituições escolares.

No trabalho de Góes (2008) é apresentado um banco de dados sobre a História da Educação, relativo a produções em nível *stricto sensu* de instituições de ensino soteropolitanas, com foco a reduzir o tempo de busca pelos pesquisadores, além de garantir a integridade e a preservação de acervo, e o acesso ao público. Também sobre a região nordeste, Andrade (2017) apresenta a experiência da criação de um repositório digital com fontes de História da Educação sobre a cidade de Bananeiras, no Estado da Paraíba, com o objetivo de permitir o compartilhamento de documentos para fins de pesquisa.

Esse levantamento permite constatar que a IA ainda não é foco de investigação sobre a sua aplicação e eventual contribuição ao campo da História da Educação. Em tempos em que a digitalização de acervos bibliográficos mostra-se uma realidade irreversível (DARTON, 2010), repercutindo "[...] nas estruturas do suporte material do escrito assim como nas maneiras de ler" (CHARTIER, 1999, p. 13), o uso de tecnologias contemporâneas

no campo educacional surge como oportunidade para a construção de uma "[...] república digital do saber" (DARTON, 2010 p. 13). Como exemplo da aplicação dessas tecnologias tem-se o processamento de imagens, podendo ser desenvolvido em jornais, componentes do estudo de caso desta tese.

Isso se desenvolve pela manipulação da imagem após a sua captura, atuando na melhoria de informação visual para interpretação do computador, por meio da redução de ruídos, realce e recuperação de imagens. Tal processo envolve procedimentos que são expressos, inicialmente, em forma algorítmica, prosseguindo para o tratamento de imagens despadronizadas, com vistas à criação de um modelo de leitura de imagens (GONZALEZ, WOODS, 2000). Embora melhore a visualização, esse processo deve estar associado à análise nos documentos em busca das informações desejadas. Para isso, é utilizada a visão computacional, a qual integra a IA e trata da extração de informações, da identificação e classificação de objetos presentes nas imagens, neste caso, os textos, possibilitando a criação de uma base de dados textuais (CONCI, AZEVEDO, LETA, 2008).

Essas bases, segundo Bousbia e Belamri (2014), constituem importantes fontes que, uma vez tratadas sob técnicas específicas, poderão gerar informações e conhecimentos. Tem-se, assim, as técnicas de classificação de textos, que permitem gerenciar, classificar e identificar informações, auxiliando a geração de conhecimento para o pesquisador, subsidiando análises e implicando novas descobertas. Dessa forma, as bases de textos podem ser analisadas por meio da IA, permitindo explorar e examinar, de modo automático ou semiautomático, grandes quantidades de dados, viabilizando a identificação de padrões e regras significativas (ROMERO et al, 2010) e, ainda, sob análise de conteúdo. Essa última, como um conjunto de técnicas para o exame de comunicações, permite a interpretação sobre os seus significados (BARDIN, 2016; HENRY, MOSCOVICI, 1968). Observa-se, portanto, que as informações necessárias para a geração de conhecimento sobre qualquer cenário demandam a adoção e aplicação de tecnologias de forma estratégica, considerando desde métodos a ferramentas de captura, tratamento e organização para análises.

Assim, para a implementação do instrumento desenvolvido desta tese, foram utilizadas as linguagens de programação Python v. 3.7.4[2] e Java 8[3]. Para os testes estatísticos

[2] PYTHON. Programming language. Disponível em: https://www.python.org/. Acesso em: 02 dez 2019.
[3] JAVA. Programming language. Disponível em: https://www.java.com/pt_BR/. Acesso em: 02 dez 2019.

foi adotada a linguagem R[4]. As bibliotecas Matplotlib[5], Scikit-image[6], Numpy[7] foram utilizadas no processo de extração e processamento de imagem e, por fim, para validar os modelos de IA, foi utilizada a biblioteca Weka 3[8], que permitiu testar os dados extraídos das imagens em modelos de IA.

A partir dessas considerações, prossegue-se para a apresentação dos procedimentos metodológicos a serem adotados para a elaboração do instrumento, mencionando os principais conceitos que o envolvem.

1.1 PROCEDIMENTOS METODOLÓGICOS

Inicialmente, cabe ressaltar que o instrumento de pesquisa elaborado, denominado Sistema de Pesquisa para Educaçao (SPEdu), se volta à análise de documentos digitais. Segundo o Conselho Nacional de Arquivos (CONARQ) um documento digital é a informação "[...] registrada, codificada em dígitos binários, acessível e interpretável por meio de sistema computacional" (2016, p. 21). Esse termo está relacionado a fontes textuais, iconográficas e audiovisuais, podendo ser nativas ou migradas. As primeiras são aquelas elaboradas originalmente no meio eletrônico. Já as fontes migradas são as que sofreram um processo de digitalização, ou seja, são encontradas de forma primária em meio físico. Nesta tese, os documentos considerados são os textuais e iconográficos na forma migrada.

Para as fontes migradas são necessários vários processos. O primeiro é a própria migração da forma física para a digital. A seguir, é necessário efetuar o processamento digital da imagem obtida. Tal processamento, segundo Gonzalez e Woods (2010), visa à melhoria da imagem do documento, permitindo a extração de informações e percepções automatizadas por meio do computador (Apêndices A e B). Aqueles autores indicam que a compreensão de imagens se relaciona ao seu processamento e à visão computacional, envolvendo "[...] operações primitivas, como o pré-processamento de imagens para reduzir

[4] R. The R Project for Statistical Computing. Disponível em: https://www.r-project.org/. Acesso em: 02 dez 2019.
[5] MATPLOTLIB. Visualization with Python. Disponível em: https://matplotlib.org/. Acesso em: 02 dez 2019.
[6] SCIKIT-IMAGE. Image processing in Python. Disponível em: https://scikit-image.org/. Acesso em: 02 dez 2019.
[7] NUMPY. Fundamental package for scientific computing. Disponível em: https://numpy.org/. Acesso em: 02 dez 2019.
[8] WEKA. The workbench for machine learning. Disponível em: https://www.cs.waikato.ac.nz/ml/weka//. Acesso em: 02 dez 2019.

o ruído, o realce de contraste e o aguçamento de imagens" (GONZALEZ, WOODS, 2010, p. 2). Isso ocorre por meio de três níveis de processos, conforme o Quadro 1, a seguir:

Quadro 1 – Níveis de processamento de imagens

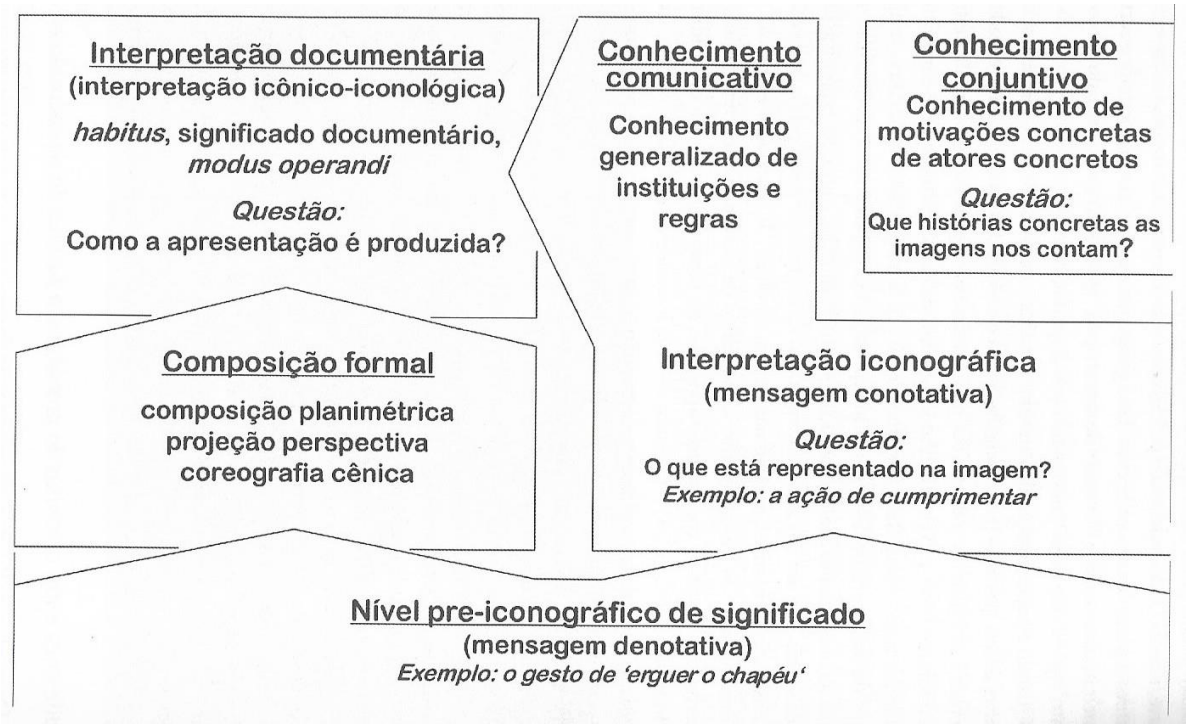
Nível do processo	Operações
Baixo	Pré-processamento de imagens para reduzir o ruído, o realce de contraste e o aguçamento de imagens. É caracterizado pela entrada e saída serem imagens
Médio	Tarefas como a segmentação (separação de uma imagem em regiões ou objetos), a descrição desses objetos para reduzi-los a uma forma adequada para o processamento computacional e a classificação (reconhecimento) de objetos individuais. É caracterizado pelas entradas serem imagens, mas as saídas são atributos extraídos dessas imagens
Alto	Atribui sentido a um conjunto de objetos reconhecidos, como na análise de imagens e possibilita realizar as funções cognitivas normalmente associadas à visão.

Fonte: Elaborado pelo autor, com base em Gonzalez e Woods (2010).

Como exemplo de percepção, tem-se a detecção facial nas imagens. Em um conjunto de fotografias, poderia ser identificada a frequência da presença de determinada figura na totalidade dos materiais. A possibilidade do uso de imagens em investigações no campo da educação é confirmada por Banks (2009, p. 17), ao mencionar que elas "[...] são onipresentes na sociedade e por isso, algum exame de representação visual pode ser potencialmente incluído em todos os estudos de sociedade". Aquele autor afirma, ainda, que a incorporação de análise de imagens "[...] pode revelar algum conhecimento sociológico que não é acessível por nenhum outro meio" (BANKS, 2009, p. 18).

A relevância das imagens também é corroborada por Bohnsack (2010), que indica a necessidade de incorporação delas para a evolução de métodos qualitativos em pesquisas. As análises sobre as imagens devem considerar não apenas o que foi produzido, mas como e os motivos pelo qual foi produzido. Nesse sentido, a busca do pesquisador deve ampliar-se para materiais que permitam identificar aspectos anteriores relacionados à própria imagem analisada. O instrumento de pesquisa proposto na presente tese poderá, assim, contribuir para a identificação de fontes iconográficas e textuais relacionadas. A seguir, são apresentadas as dimensões do significado e interpretações de imagens indicadas por Bohnsack (2010).

Figura 1 – Dimensões do significado e interpretação de imagens



Fonte: Bohnsack (2010, p. 117).

A possibilidade de relacionar fontes iconográficas e textuais ocorre por meio da identificação de expressões ou termos presentes no conjunto de material iconográfico. Dessa forma, materiais como jornais, diários de classe, relatórios e outras espécies documentais, existentes primariamente na forma física e migradas para a forma eletrônica por meio da digitalização, podem ser analisados a partir de termos determinados pelo pesquisador e automaticamente encontrados pelo processamento de imagem.

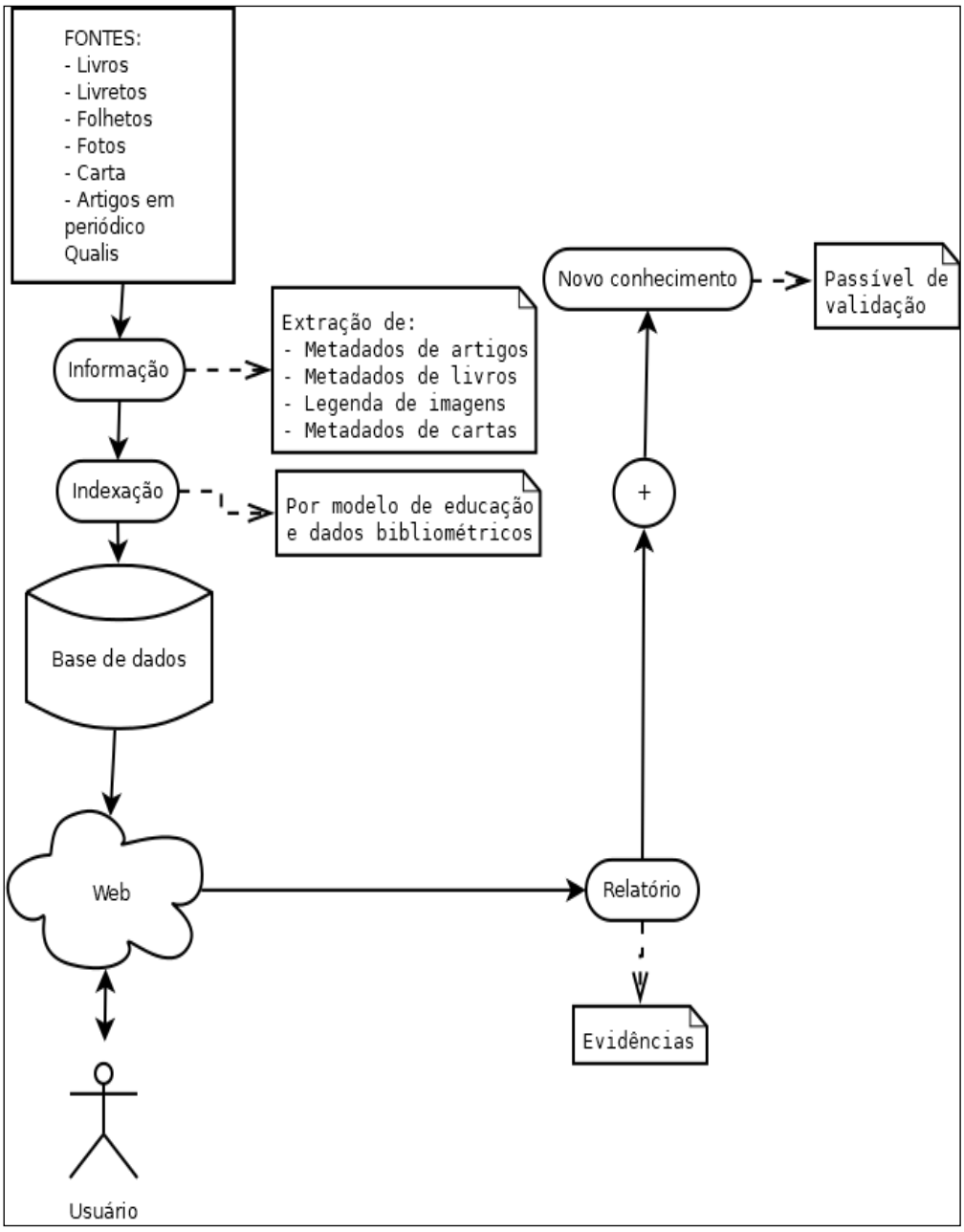
Após a seleção dos documentos é realizada a extração de informações sobre materiais textuais para compor um banco de dados. Dessa forma, os textos que compõem os documentos digitais são extraídos e vinculados à imagem original. Esse processo, denominado extração de características, emprega o reconhecimento ótico de caracteres (OCR) que é "[...] um processo de leitura de caracteres manuscritos e impressos. Ele é amplamente utilizado como forma de entrada de dados, a partir de registros de dados impressos" (CHUDHURI, BADELIA, GHOSH, 2017, p. 1) (tradução nossa)[9].

Em etapa seguinte, é realizada a indexação e classificação de documentos histórico-educacionais. A indexação por termos é a extração de dados que indiquem do que se trata o

[9] *Texto original: "[...] a process of reading handwritten and printed characters. It is widely used as a form of data entry from printed data records.*

documento (LANCASTER, 2004), gerando um índice. Segundo Baeza-Yates e Ribeiro-Neto (2013) um índice é composto por um vocabulário e suas ocorrências. Uma vez criado o índice, é necessário realizar a indexação dos termos encontrados por conceitos. Essa é uma forma de classificação que "[...] provê um meio para organizar a informação, [permitindo] a melhor compreensão e interpretação dos dados" (BAEZA-YATES E RIBEIRO-NETO, 2013, p. 278). Essa e as demais etapas até aqui descritas, podem ser identificadas no instrumento desenvolvido, SPEdu, na Figura 2, a seguir.

Figura 2 - Ciclo de etapas do SPEdu



Fonte: Elaborado pelo autor (2019).

A IA surge, assim, como técnica de aprendizagem de máquina para classificação por meio de treinamento. Na prática, o pesquisador deverá submeter um conjunto de documentos como modelo de treinamento e, em próprias inserções, o processo será automatizado. Segundo Russell e Norvig (2013, p. 605), "[...] um agente estará aprendendo a melhorar o seu desempenho nas tarefas futuras de aprendizagem após fazer observações sobre o mundo". O aprendizado do computador, como agente, e suas observações sobre as fontes documentais serão realizadas no instrumento de pesquisa aqui apresentado, sob o paradigma indiciário (GINZBURG, 1989). Isso resultará na identificação de indícios durante a realização de buscas pelo pesquisador, uma vez que permitirá a microanálise sobre pontos específicos investigados. Esse é um dos princípios do método indiciário, o qual pauta-se pela exploração minuciosa e interdisciplinar sobre os mais variados aspectos que envolvem o objeto estudado.

Dessa forma, o instrumento SPEdu integra as tecnologias contemporâneas para o tratamento de documentos digitais por meio da IA na recuperação de informações. A partir dessas considerações sobre cada etapa dos procedimentos metodológicos, apresenta-se a estrutura da tese.

1.2 A ESTRUTURA DA TESE

A partir da explanação sobre as etapas da pesquisa e dos principais conceitos adotados para as análises, apresenta-se a estrutura da tese. Na primeira seção, denominada Introdução, apontou-se o objetivo central do estudo e as fases que o envolvem, culminando na validação do instrumento tecnológico desenvolvido.

Em prosseguimento, a segunda seção trata sobre a aquisição e o pré-processamento de imagens, abordando análise de *layout* e detecção automatizada de imagens. É, também, apresentado um primeiro experimento de extração de dados históricos, utilizando o jornal Diário Oficial do Estado de Sergipe.

Na terceira seção, é abordada a indexação, classificação e busca de informações histórico-educacionais. A partir dos resultados sobre a extração de dados realizada no experimento, essa seção também traz a avaliação referente ao modelo de indexação adotado para os documentos históricos. Os resultados obtidos compõem a base de dados utilizada para a busca e geração de evidências.

A quarta seção da tese apresenta o instrumento tecnológico proposto, consolidado como Sistema de Pesquisas Educacionais (SPEdu), denominação atribuída pelo autor. Tal instrumento é exposto em relação as suas funcionalidades, sendo aplicado em um estudo de caso com o jornal Diário Oficial do Estado de Sergipe. As análises e discussões sobre essa aplicação compõem a quinta seção.

Por fim, nas considerações finais, apresenta-se a síntese sobre a aplicação e possível contribuição da IA para a extração, catalogação e padronização de fontes documentais para a História da Educação.

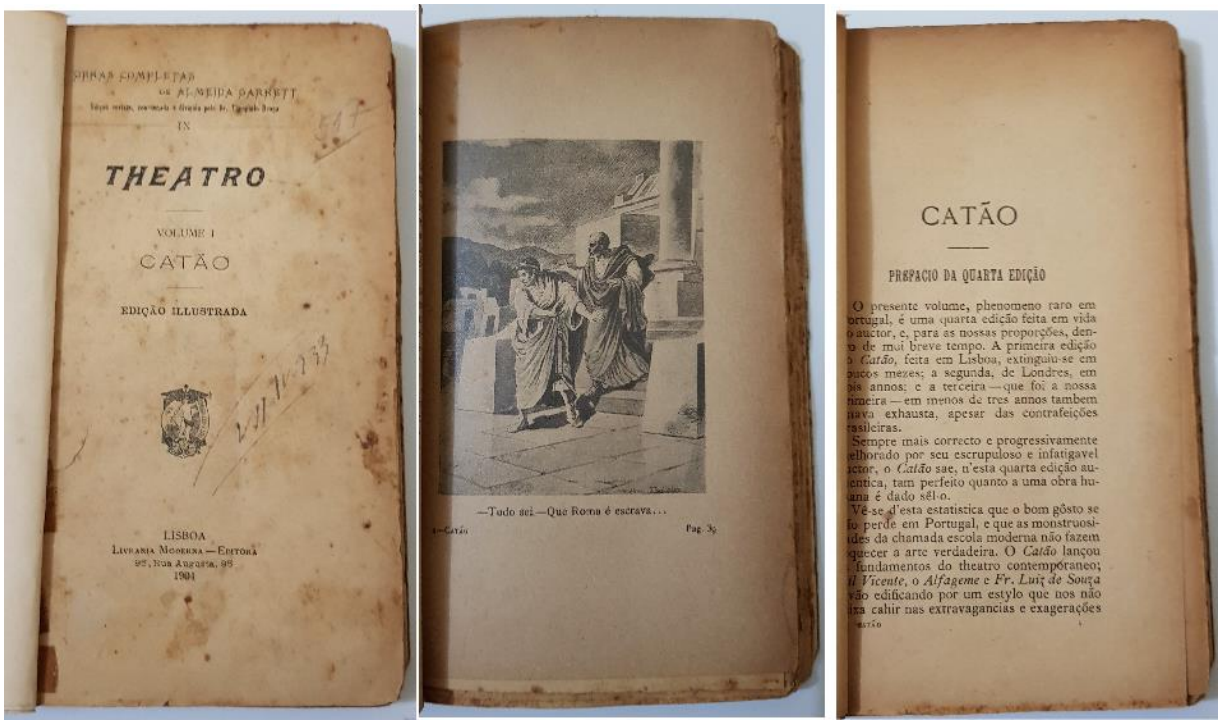
2 EXTRAÇÃO DE DADOS HISTÓRICOS EDUCACIONAIS

Entre as variadas possibilidades de fontes para a História da Educação, os documentos assumem um papel central para os pesquisadores (BACELLAR, 2018). Um dos meios de acesso a tal material são as bibliotecas físicas ou digitais. Sobre essas últimas, Reffle e Ringlstter (2013) apontam o crescimento, em todo o mundo, de unidades que disponibilizam documentação histórica digital. Tal fato representa grande contribuição ao trabalho de pesquisadores dos diversos campos, mas, principalmente, para os da História e História da Educação (JANOTTI, 2018; BACELLAR, 2018). O processo de disponibilização digital é fundamental no cenário de convergência tecnológica contemporâneo, pois, conforme afirma Darton (2009, p. 8): "O futuro, seja ele qual for, será digital".

Os pesquisadores já vêm disponibilizando corpus documental digitalizado, utilizando mídias como CD, DVD, entre outras. Como exemplo, tem-se o trabalho de Nascimento (2008), com a disponibilização de CD de documentos sobre a Missão Central do Brasil, organização da Igreja Presbiteriana do Norte dos Estados Unidos. Porém, esse tipo de ação ainda é escassa por parte dos arquivos públicos brasileiros. Segundo Bacellar (2018), os arquivos possuem, em geral, sérios problemas como infraestrutura inadequada, má preservação documental e falta de pessoal qualificado. Soma-se a isso, o fato de grande parte dos documentos não estarem digitalizados, dificultando pesquisas. Assim, para que o pesquisador utilize suas fontes impressas, quando estas não estão disponíveis digitalmente, há necessidade de transcrição, digitalização e catalogação, implicando dificuldades para análises de grandes volumes de dados.

Constata-se, então, a relevância que assume a automatização do processo de extração de dados para o trabalho que envolva documentos históricos. Tal processo, no entanto, não é nada trivial, visto que documentos históricos possuem ruídos causados pelo desgaste, amarelamento das páginas, entre outros problemas. Esses aspectos prejudicam o trabalho de extração de dados de fontes documentais. Um exemplo pode ser observado na Figura 3, a seguir, do livro Theatro, de 1904, onde se constata a deteriorização do material.

Figura 3 - Páginas do livro Theatro, publicado em 1904, em Lisboa



Fonte: Acervo do autor.

Entre as fontes de pesquisas históricas têm-se livros, diários de classe, leis, jornais etc. Dessas, a que representa um desafio mais complexo são os jornais impressos (ZENI; WELDERNARIAM, 2017). Eles contêm vários *layouts* de páginas com múltiplos artigos, em que estes são projetados para permitir que o interessado defina sua própria ordem de leitura. Os parágrafos e as imagens são distribuídos em diversas páginas de forma imprevisível, tornando a extração de dados de jornais um grande desafio. Segundo Jana et al. (2018), o processamento de dados de jornais é uma tarefa com alto grau de dificuldade, já que eles são impressos em papéis de baixa qualidade, os quais têm como tendência a mudança de cor com o decorrer do tempo. Tais mudanças geram ruídos que aumentam com o tempo de existência do documento. Assim, é alta a probabilidade de um jornal, com aproximadamente vinte anos, ter suas páginas amareladas, conforme demonstrado na Figura 4, a seguir.

Figura 4 - Parte da capa do Diário Oficial do Estado de Sergipe, de 28 de agosto de 1997

Diário O

ESTADO DE SERGIPE

Nº 22.866 ARACAJU, Quinta - feira, 28 de agosto de 199

PODER EXECUTIVO

GOVERNADOR DO ESTADO
ALBANO DO PRADO PIMENTEL FRANCO

VICE GOVERNADOR DO ESTADO
JOSÉ CARLOS MACHADO

SECRETÁRIOS DE ESTADO

LUCIANO AUGUSTO BARRETO CARVALHO
Secretário-Chefe do Gabinete do Governador

FRANCISCO GUIMARÃES ROLLEMBERG
Secretário-Chefe da Casa Civil

JOSÉ AUGUSTO CHAVES REZENDE
Secretário-Chefe da Controladoria Geral do Estado

ROBERTO EUGÊNIO DA FONSECA PORTO
Procurador Geral do Estado

CLÁUDIO MAYNART RABELO
Defensor Público Geral do Estado

MARCOS ANTONIO DE MELO
Secretário de Estado do Planejamento e da Ciência e Tecnologia

MARIA ISABEL CARVALHO NABUCO D'ÁVILA
Secretária de Estado da Administração

O GOVERNADOR DO ESTADO DE SERGIPE, no uso das atribuições que lhe são conferidas nos termos do Art. 84, incisos V, VII e VIII, da Constituição Estadual; de acordo com o disposto no art. 74, inciso I, da Lei nº 2.148, de 21 de dezembro de 1977 (Estatuto dos Funcionários Públicos Civis do Estado de Sergipe), resolve

EXONERAR

A pedido, SÉRGIO RICARDO LEITE BARBOSA, R.G. nº 468.358/SE, do cargo em comissão de Delegado Especial de Polícia, Símbolo CCS-12, da Polícia Civil, a partir de 30 de junho de 1997.

Aracaju, 27 de agosto de 1997, 176º da Independência e 109º da República.

ALBANO FRANCO
GOVERNADOR DO ESTADO

Fonte: Figura capturada pelo autor do acervo SEGRASE (1997).

Nesse sentido, Rajeswari e Magapu (2018) afirmam que na era digital, a conversão de documentos impressos para a forma eletrônica tornou-se uma necessidade para a disponibilização de informações. Mas os autores afirmam, também que, para permitir que os documentos sejam encontrados, é necessário que os seus metadados sejam extraídos por meio de ferramentas de reconhecimento ótico de caracteres (OCR). Um OCR é uma tecnologia que permite converter diferentes tipos de documentos como arquivos *Portable Document Format* (PDF), imagens capturadas em câmera digital e documentos digitalizados em um formato editável e pesquisável (RAJESWARI; MAGAPU, 2018).

Porém, para que seja possível extrair dados por meio do OCR, Vasilopoulos e Kavallieratou (2017) afirmam que é necessário empregar métodos que combinem a análise de layout de documentos com a detecção de texto. A primeira etapa do processo ocorre com a digitalização do documento. Isso pode ser realizado por meio de fotografias ou escaneamento das fontes. No entanto, o processo de digitalização visa transformar o documento impresso ou manuscrito em um documento digital, mas isso não permite que o

documento seja pesquisado automaticamente. Dessa forma, o pesquisador permanecerá dependente de transcrição, catalogação e indexação dos documentos.

Kaur e Jindal (2018) indicam que, para que seja possível reconhecer e detectar os textos, é necessário realizar um pré-processamento das imagens, removendo ruídos indesejados. Considerando que documentos históricos são fontes complexas, após o pré-processamento é necessária a análise de layout da página. Isso permitirá o reconhecimento adequado dos textos. A análise de layout visa, então, reconhecer a distinção entre as regiões que são textuais e das que não são, viabilizando, a seguir, a extração dos textos (KAUR; JINDAL, 2018).

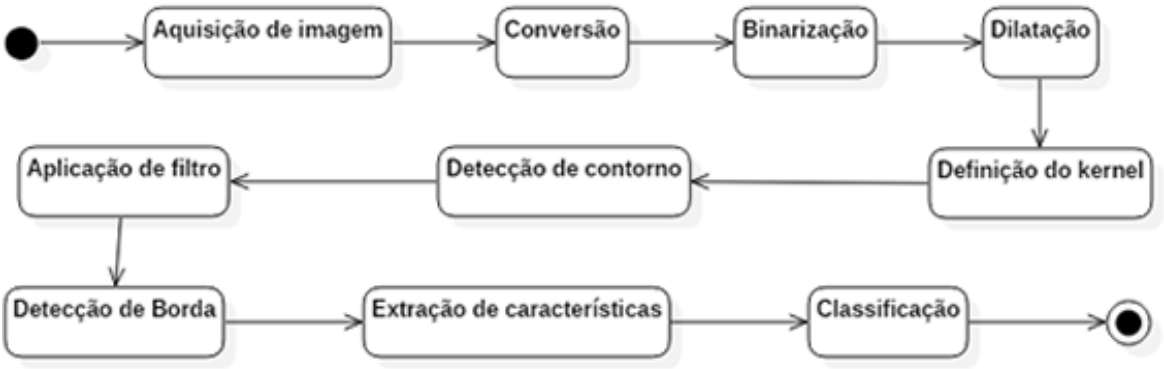
Nesse contexto, o objetivo da presente seção é discutir o processo de detecção e extração de textos de documentos históricos. Inicialmente, foram analisados os trabalhos de Zeni e Weldernarian (2017), Bukhari et al (2010), Palfray et al (2012), Hebert et al (2014), Pramanik e Bag (2018), Chathuranga e Ranathunga (2017) e Vasilopoulos, Wasfi, Kadallieratou (2018)que abordam tais operações em materiais históricos e jornais. Esses últimos foram selecionados por configurarem uma fonte histórica com layout complexo e, ainda, suscetível a grande desgaste devido ao papel de baixa qualidade.

Para a realização de testes dos algoritmos desenvolvidos optou-se pelo jornal Diário Oficial do Estado de Sergipe, o qual possui mais de 100 anos de publicação em papel. Esse jornal contempla o histórico político da região, já que todos os atos governamentais devem ser publicados nele. Como exemplos têm-se os atos de abertura de escolas e posses de secretários do governo, os quais podem subsidiar pesquisadores do campo da educação. Para validar os modelos propostos nesta tese, os testes foram realizados a partir de uma seleção aleatória de dez números do Diário Oficial do Estado de Sergipe.

2.1 AQUISIÇÃO E PRÉ-PROCESSAMENTO DE IMAGENS

Para o trabalho com documentos históricos, adotou-se um ciclo que envolve da aquisição, o pré-processamento e a classificação da imagem, conforme demonstrado na Figura 5, a seguir.

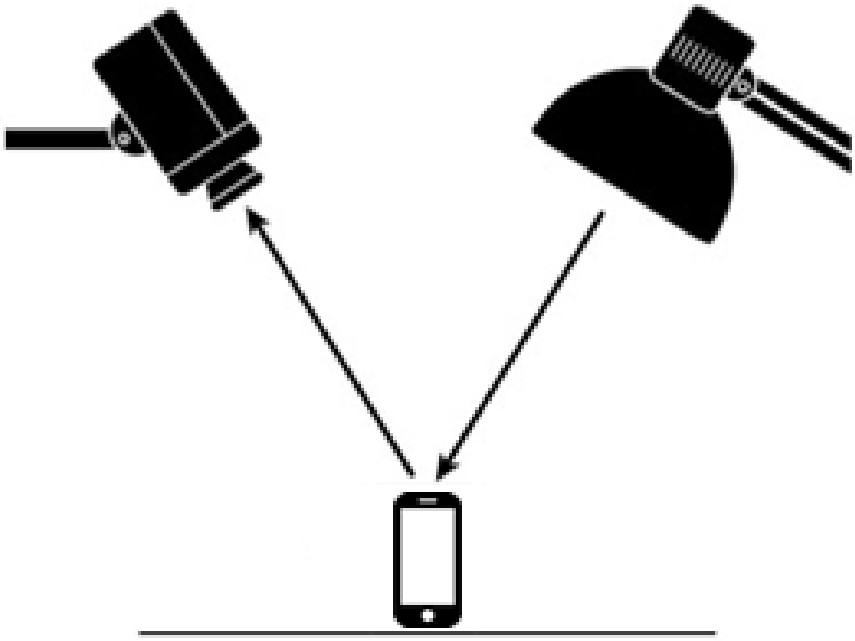
Figura 5 - Ciclo de aquisição, pré-processamento e classificação



Fonte: Elaborada pelo autor (2019).

A primeira etapa do processo é a aquisição de imagem. Isso ocorre com a utilização de um equipamento como escâner, câmera digital ou sensores para gerar imagens digitais, a partir da digitalização dos documentos físicos. Estes equipamentos capturam a luz refletida dos objetos (Figura 6).

Figura 6 - Captura de imagem



Fonte: Elaborada pelo autor (2019).

Cabe ressaltar que objetos diferentes exigem condições de iluminação diferentes, principalmente, quando se faz uso de câmeras ou celulares. Para documentos, esse aspecto pode ser minimizado com a utilização de um escâner para a digitalização.

As imagens, ao serem digitalizadas, possuem resolução, que é a quantidade de pontos (*pixels*) que representados na imagem, organizados em uma matriz “M x N” Por exemplo, na Figura 6, a seguir, com resolução de 2560 x 1390 (matriz MxN), o “M” é de 2560 *pixels* de largura e o “N” é de 1390 *pixels* de altura. Isso significa que a imagem é representada como uma grade de *pixels*, sendo 2560 colunas e 1390 linhas, resultando: 2560 x1390 = 3.558.400 *pixels* na imagem, Figura 7.

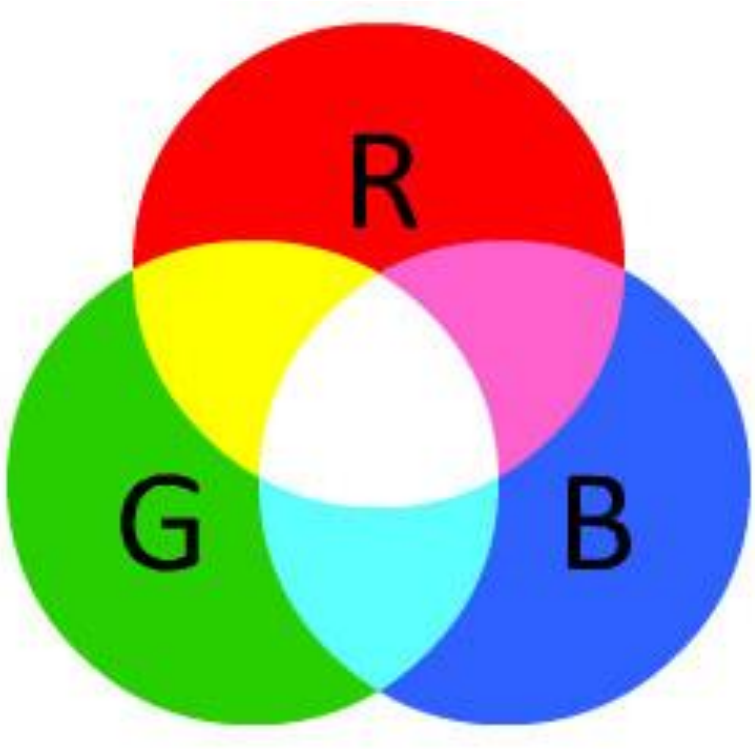
Figura 7 - Capa do Diário Oficial do Estado de Sergipe, com resolução 2560 x 1390

Fonte: Elaborada pelo autor, adaptada do acervo SEGRASE (1997).

Os documentos, ao serem digitalizados utilizando-se câmeras ou escâner, podem ser coloridos em tons de cinza ou em preto e branco. As imagens coloridas são formadas por uma combinação de cores. No caso das imagens digitais, geradas por escâner ou câmeras, o sistema de cores mais adotado é o RGB (vermelho, verde e azul como cores primárias), o qual é, também, o sistema de cor dos monitores dos computadores. No sistema RGB cada cor é codificada em uma sequência de 24 *bits*, sendo 8 *bits* para cada componente de cor (vermelho, verde e azul). Cada componente, por sua vez, é definido por um valor inteiro de

0 a 255, e o valor triplo (R, G, B) define a cor. Na Figura 8, a seguir, é apresentada a paleta de construção de cores RGB.

Figura 8 - Construção de cores por meio do padrão RGB

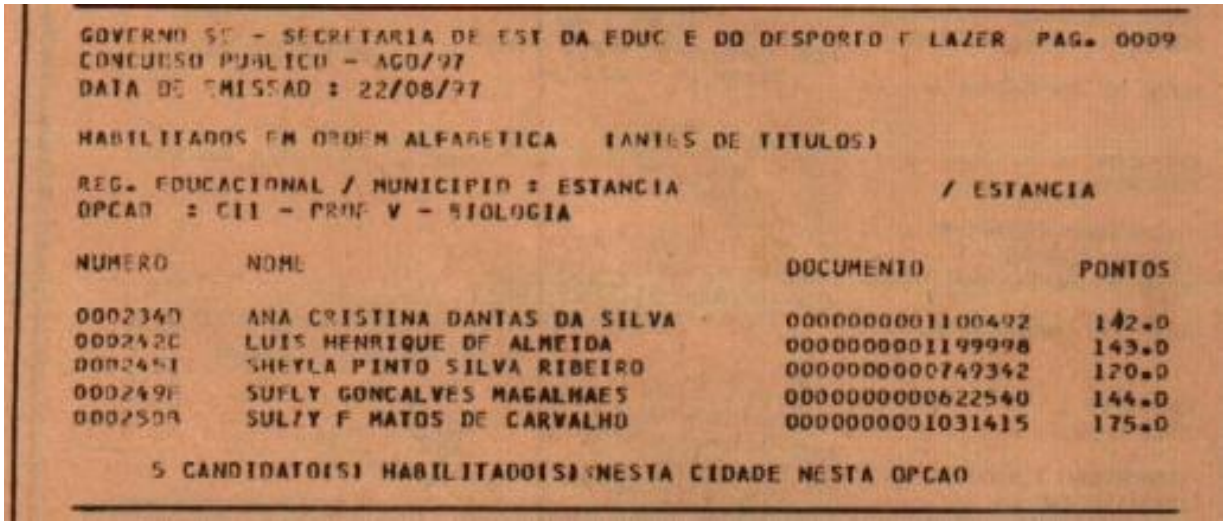


Fonte: Elaborada pelo autor (2019).

Em análise à Figura 6, por exemplo, o *pixel* 600 X 1000 (apenas um *pixel*), encontra-se uma cor próxima a preta, com um tom de cinza escuro. Esse *pixel* tem um valor matemático de 34, 27 e 21, representando a quantidade de cores vermelho, verde e azul, respectivamente. Cabe lembrar que na ausência de cor (representada pelo valor matemático 0 0 0) tem-se o preto e o branco representado pelo valor máximo em todos os componentes (neste caso, pelos valores 255 255 255). Observa-se, assim, que para o computador, a figura é um conjunto numérico, representando pontos para a construção de uma imagem. Dessa forma, não é possível efetuar a busca de um texto diretamente na figura. Para isso é necessário um padrão numérico que represente um texto.

Após digitalizar a imagem, um pesquisador já poderá fazer uso desse material. Porém, dependendo da qualidade da imagem e o desgaste do documento, a leitura direta poderá se tornar difícil, conforme demonstrado na Figura 9, a seguir.

Figura 9 - Parte do Diário Oficial do Estado de Sergipe, publicado em 1997



GOVERNO SE - SECRETARIA DE EST DA EDUC E DO DESPORTO E LAZER PAG. 0009
CONCURSO PUBLICO - AGO/97
DATA DE EMISSAO : 22/08/97

HABILITADOS EM ORDEM ALFABETICA (ANTES DE TITULOS)

REG. EDUCACIONAL / MUNICIPIO : ESTANCIA / ESTANCIA
OPCAO : C11 - PROF V - BIOLOGIA

NUMERO	NOME	DOCUMENTO	PONTOS
0002340	ANA CRISTINA DANTAS DA SILVA	0000000001100492	142.0
0002420	LUIS HENRIQUE DE ALMEIDA	0000000001199998	143.0
0002451	SHEYLA PINTO SILVA RIBEIRO	0000000000749342	120.0
0002490	SUELY GONCALVES MAGALHAES	0000000000622540	144.0
0002509	SULZY F MATOS DE CARVALHO	0000000001031415	175.0

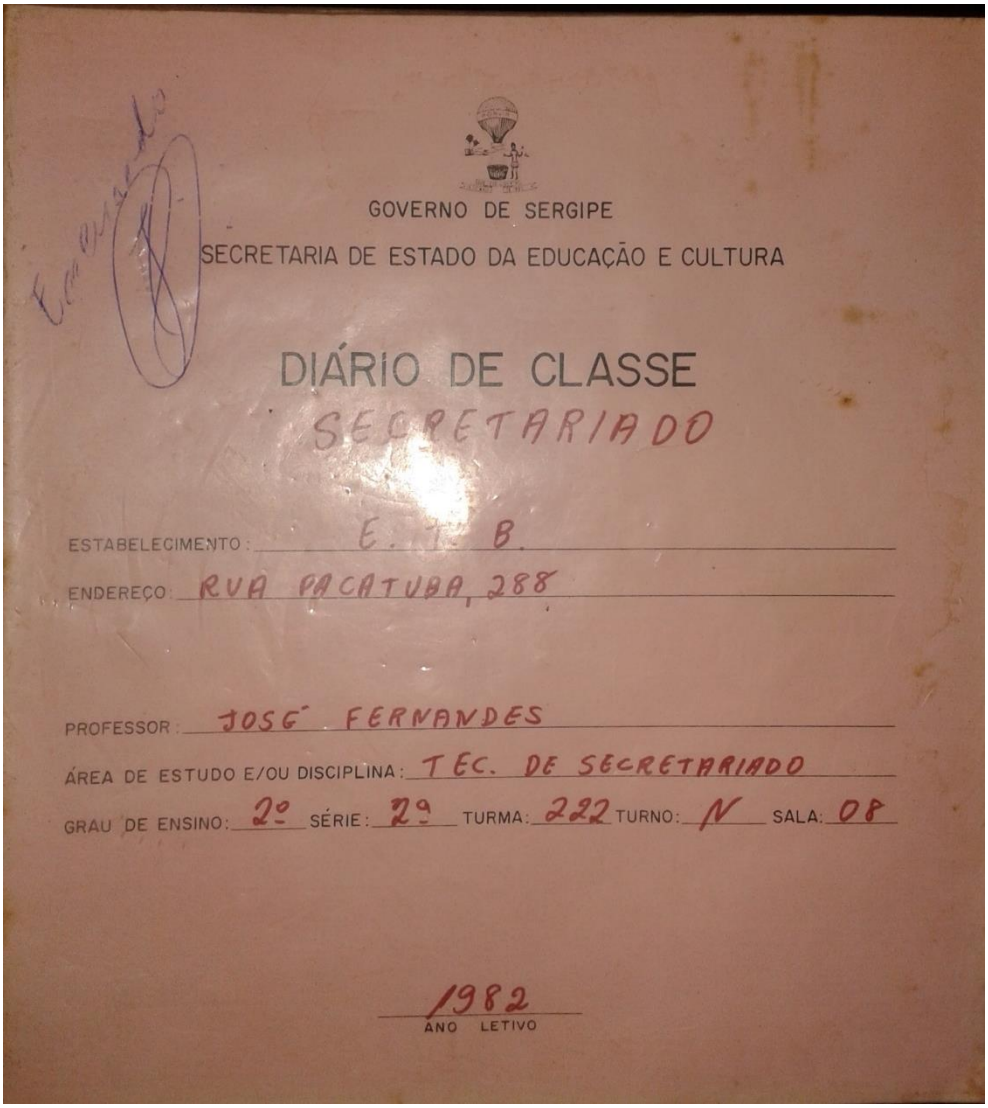
5 CANDIDATO(S) HABILITADO(S) NESTA CIDADE NESTA OPCAO

Fonte: Capturada pelo autor do acervo SEGRASE (1997).

Constata-se, portanto, a necessidade do pré-processamento, o qual visa melhorar a qualidade da imagem. Porém, para um contexto histórico, neste trabalho mantemos a imagem original e processamos os dados em uma cópia. Isso permite que um pesquisador possa analisar os ruídos e degradações do documento na imagem original, como marcas de manuseio, anotações marginais, desgaste do papel, etc, sem comprometer a extração dos dados.

Para documentos históricos, é necessário adotar alguns cuidados na digitalização, certificando-se que nenhuma informação foi excluída da imagem. No caso de fotos, a iluminação deve ser levada em consideração. Como exemplo desse aspecto tem-se o trabalho de Sabino (2017), onde constam fotos de documentos histórico-escolares, como diários de classe. Nesse caso específico, os documentos estavam plastificados, produzindo reflexos do *flash* da câmera. Isso acaba por dificultar o trabalho do pesquisador e da leitura pelos interessados. O exemplo é exibido na Figura 10, a seguir.

Figura 10 - Diário de classe

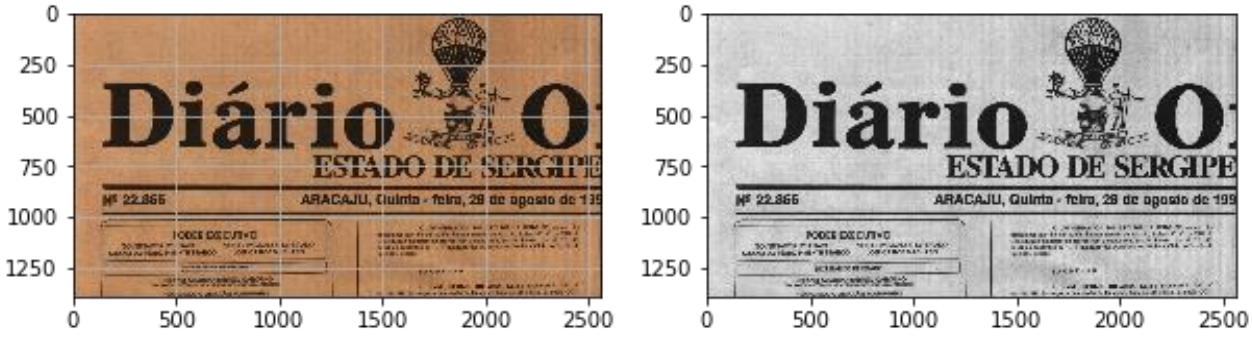


Fonte: Sabino (2017).

Se por um lado a fotografia evita o manuseio excessivo do documento, a iluminação é um fator essencial para a qualidade da imagem. Já no uso de escâners o cuidado recai no manuseio do documento, evitando que ele seja danificado. As imagens empregadas neste trabalho foram capturadas em um escâner de mão, modelo NIP-A4 portátil, da Marc Nipponic. Esse aparelho foi escolhido devido ao fato de que jornais antigos são altamente frágeis, rasgando-se com facilidade. Dessa forma, não seria possível utilizar um escâner de mesa para a tarefa. Também foram realizados testes com celulares e câmaras para as fotografias dos jornais. Porém, no processo de detecção, constatou-se um resultado de menor qualidade do que o obtido com o uso do escâner.

Uma imagem pode apresentar diversos problemas, como: ruídos, baixo contraste, inclinação inadequada, elementos que dificultam o reconhecimento de características, entre outros. Assim, para a realização da detecção e do reconhecimento de textos, procedeu-se a segunda etapa do modelo proposto: a conversão da imagem em tons de cinza. Com isso, a imagem colorida, que possui três canais de cores, é convertida em uma imagem com um único canal de tons de cinza, reduzindo a quantidade de informações relativas à imagem. O resultado desse procedimento é apresentado na Figura 11, a seguir.

Figura 11 - Conversão de imagem em tons de cinza



Fonte: Elaborada pelo autor, adaptada do acervo SEGRASE (1997).

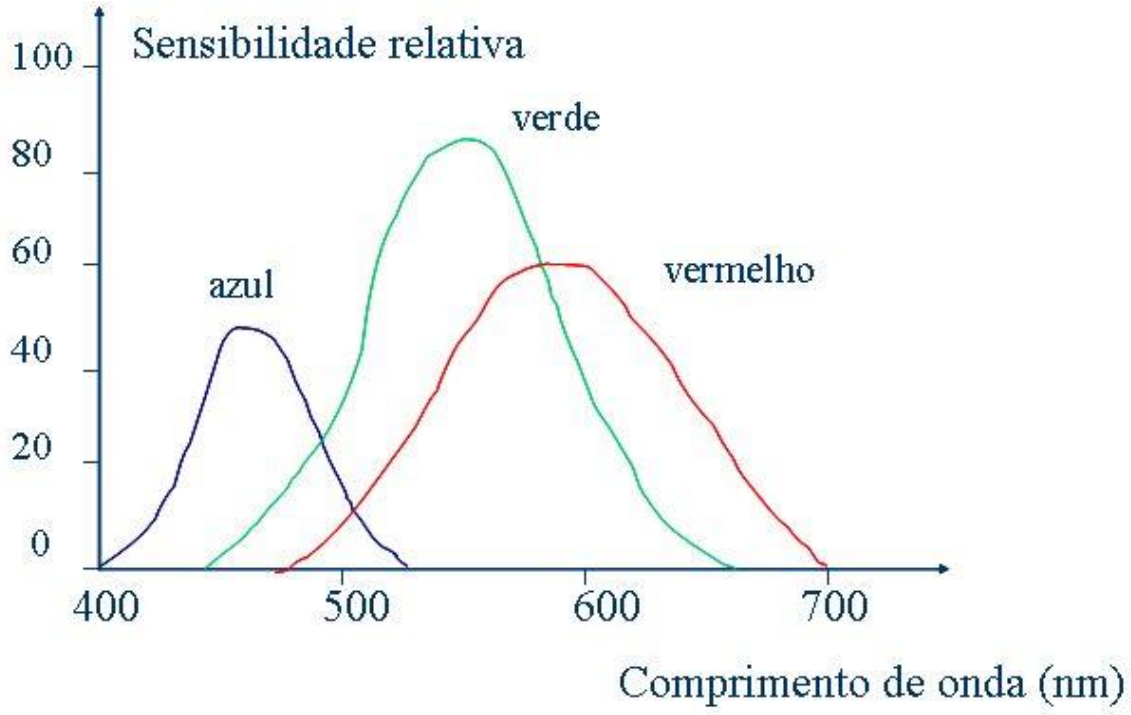
A imagem em tons de cinza é gerada após uma conversão realizada a partir da soma ponderada dos canais de cores, considerando a capacidade do olho humano em absorver a luz emitida por cada uma das cores, conforme exposto na Figura 12, a seguir. Assim, a representação da imagem por meio da luminância produz uma imagem em escala de cinza, onde cada cor (RGB) é associada a um valor de luminância, dado pela fórmula1 (com base em MELLO; SANTOS; OLIVEIRA, 2011):

Fórmula 1 - Conversão de imagem colorida em tons de cinza

Imagem_cinza=(imagemvermelho*0,299)+(imagemverde*0,587)+(imagemazul*0,114)

Fonte: Fórmula elaborada pelo autor com base em Mello, Santos e Oliveira (2011).

Figura 12 - Curvas de sensibilidade relativa do olho humano para cada uma das componentes R, G e B



Fonte: Felgueiras (2019).

A imagem cinza é o resultado do valor de luminância, o qual foi encontrado por meio da fórmula 1, baseado nos componentes de cor R (vermelho), G (verde) e B (azul). Como resultado tem-se que a imagem gerada emprega níveis de cinza que variam do 0 (preto absoluto) ao 255 (branco absoluto) (MELLO; SANTOS; OLIVEIRA, 2011), conforme exibido na Figura 13, a seguir. Como a imagem possuía três canais de cores e foi convertida para um canal com apenas tons de cinza, o tamanho final da imagem é reduzida, implicando na quantidade de informações disponíveis para o computador processar.

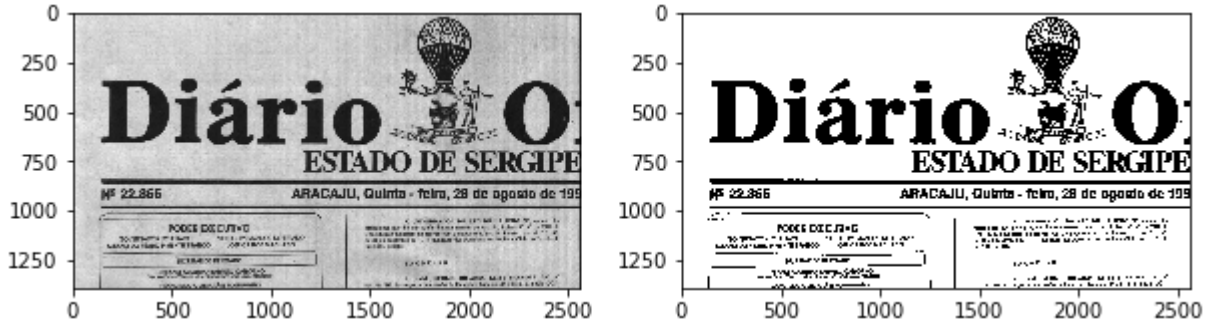
Figura 13 - Tabela de tons de cinza

Fonte: Elaborada pelo autor (2019).

O próximo passo do processamento da imagem é a binarização (limiar) (RAJESWARI; MAGAPU, 2018), a qual pode ser categorizada como global (GONZALEZ;

WOODS, 2010) e local (GLLAVATA et al, 2003). Esse processo transforma a imagem que está em tons de cinza em uma imagem binária, que tenha apenas duas cores - o preto (representado pelo 0) e o branco (representado pelo 1), removendo as demais cores. Há, ainda, uma forma de limiar com base em um único limite, denominado “binarização de Otsu” (OTSU 1979). Uma alternativa é o limiar adaptativo proposto por Sauvola et al. (1997). A binarização é uma técnica que permite segmentar os objetos do fundo. No caso deste trabalho, tal técnica foi adotada para remover o fundo amarelado, o que dificulta a análise e classificação dos objetos da imagem. Esse processo é apresentado na Figura 14, a seguir.

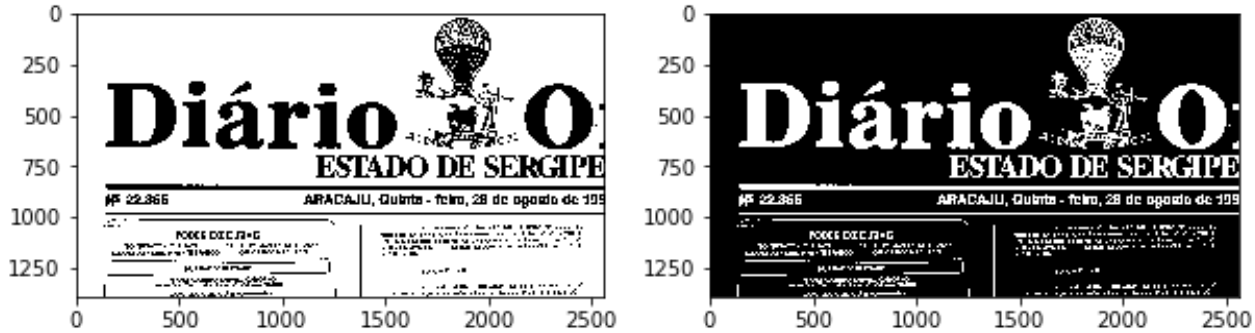
Figura 14 - Conversão de imagem de tons de cinza para preto e branco



Fonte: Elaborada pelo autor, adaptada do acervo SEGRASE (1997).

Porém, para que se possa marcar o texto é necessário aplicar a inversão da binarização, onde o branco ficará preto e o preto ficará banco, conforme exposto na Figura 15, a seguir. Isso possibilitará a marcação de cada quadro branco no fundo preto. Cabe lembrar, no entanto, que a figura original foi mantida, sendo processada uma cópia da imagem, gerada automaticamente pelo sistema.

Figura 15 - Binarização invertida de imagem



Fonte: Elaborada pelo autor, adaptada do acervo SEGRASE (1997).

A seguir, foi realizada uma operação morfológica, que é uma transformação aplicada a uma imagem binária ou em escala de cinza. As operações podem ser feitas para expandir ou reduzir o tamanho de objetos em imagem, fechar lacunas, entre outros aspectos. Para a detecção de texto, a operação morfológica é utilizada para reunir os textos em um único grupo. Dessa forma, os textos podem ser detectados em blocos. Há diversas operações morfológicas. Neste trabalho foi adotada a operação dilatação. Para isso, foi definido um elemento estruturante, ou seja, uma matriz a ser utilizada para realizar a morfologia.

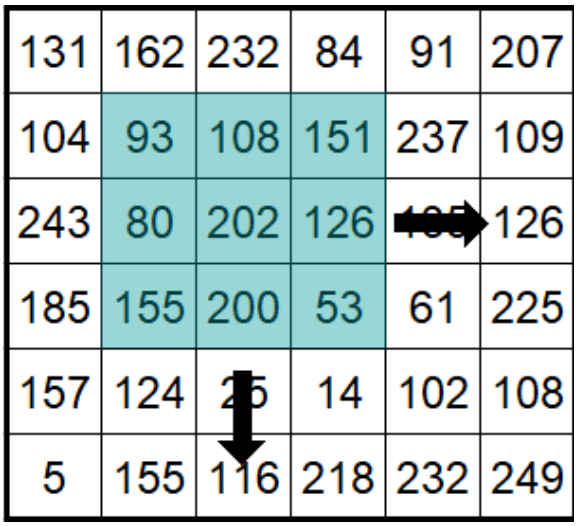
Conforme Gonzales e Woods (2010), a operação de dilatação permite que os *pixels* se expandam, tendo como resultado a união de lacunas. Na operação realizada neste trabalho, o elemento estruturante que apresentou melhor resultado foi o de tamanho 51 x 51. No caso dos textos, a tendência é a unificação em um bloco, conforme exposto na Figura 16, a seguir.

Figura 16 - Imagem dilatada

Fonte: Elaborada pelo autor, adaptada do acervo SEGRASE (1997).

Buscando-se reduzir ruídos gerados pela imagem devido ao desgaste, após aplicar a dilatação, foi empregado o filtro mediana. Um filtro é uma forma de processar os *pixels* de uma imagem, gerando uma nova imagem de saída por meio de um *kernel*. Já o *kernel* é uma pequena matriz, a qual pode ser usada para desfocar a imagem, melhorar a nitidez, detectar bordas, entre outras funcionalidades. Basicamente, o *kernel* é deslizado sobre a imagem, da esquerda para a direita e do canto superior para o inferior, permitindo a realização de operações matemáticas na imagem original e criando uma imagem nova como resultado. Esse processo é demonstrado na Figura 17, a seguir.

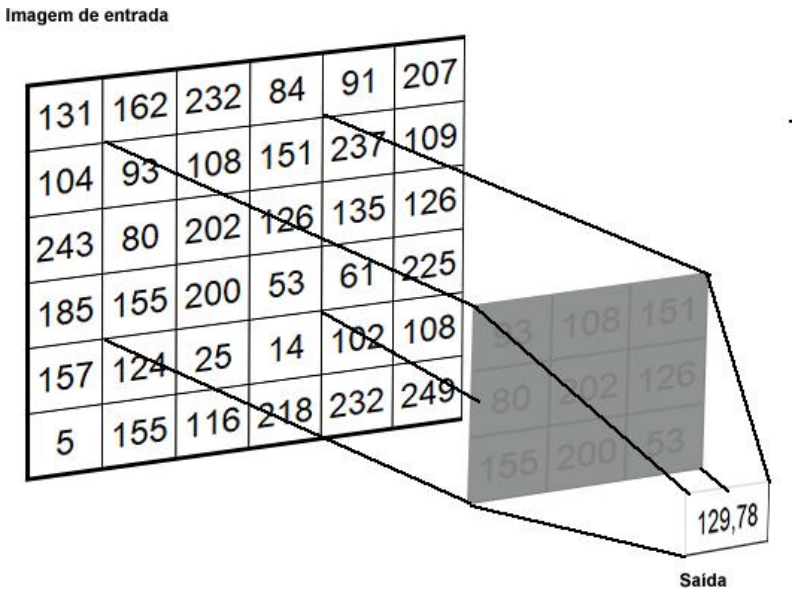
Figura 17 - Kernel para processamento de imagens



Fonte: Elaborada pelo autor (2019).

O *kernel* é uma matriz quadrada, ou seja, tem o mesmo tamanho de largura e altura, porém, deve possuir tamanho ímpar, para ter um ponto central (FRERY, 2011). Esse será o ponto a ser modificado na imagem. Tem-se, então, um *kernel* com um determinado tamanho e algumas informações, dependendo da operação e, por fim, a imagem (matriz) de saída. Esse processo é exposto na Figura 18, a seguir.

Figura 18 - Exemplo de *kernel*



Fonte: Elaborada pelo autor (2019).

Segundo Frery (2011), o filtro mediano tem como saída um *pixel*, que por sua vez tem como valor a mediana das observações. No caso da Figura 16, o *pixel* retornado é a média dos *pixels* analisados no *kernel*. Embora tenha um efeito similar ao do filtro média, o filtro mediana tem como ponto positivo a preservação das bordas e dos contornos de um objeto. Considerando tal aspecto, adotou-se esse filtro para reduzir alguns pontos da imagem, mantendo as bordas que se pretendia detectar. Nos testes realizados para a obtenção de um resultado com bom retorno de objetos detectados, tanto para jornais como para imagens de livros e documentos, o tamanho do *kernel* definido foi de 25. O resultado obtido é apresentado na Figura 19, a seguir.

Figura 19 - Imagem binária após aplicação do filtro mediana sobre a imagem dilatada

Fonte: Elaborada pelo autor, adaptada do acervo SEGRASE (1997).

Para identificar os objetos, foi utilizada a detecção de contornos do OpenCV, que tem como base o algoritmo proposto por Suzuki e Abe (1983). Como resultado, a detecção localizou contornos dos objetos que são diferentes do plano de fundo. Assim, um contorno é uma representação do limite de uma forma. Os contornos foram utilizados com base no Código 1, a seguir.

Código 1 - Detecção de contorno de objetos

```
cnts = cv2.findContours(imagem, cv2.RETR_EXTERNAL, cv2.CHAIN_APPROX_SIMPLE)
```

Fonte: Código elaborado pelo autor (2019).

A função *findCountours*, da biblioteca OpenCV, detecta o contorno em uma imagem. Para que seja selecionado apenas o contorno externo em uma figura, e não o contorno dentro de contorno, utilizou-se o cv2.RETR_EXTERNAL. Já o parâmetro cv2.CHAIN_APPROX_SIMPLE foi utilizado por reduzir o número de pontos redundantes, deixando apenas os pontos chaves da detecção. Assim, o resultado é a detecção de contorno sobre os objetos que podem ser marcados em uma imagem, conforme demonstrado na Figura 20, a seguir, sendo um dos passos para a busca de dados em fontes históricas.

Figura 20 - Imagem com detecção de objetos

Fonte: Elaborada pelo autor, adaptada do acervo SEGRASE (1997).

Uma fonte histórica deve ser devidamente identificada e organizada (BACELLAR, 2018). Para isso, é necessário extrair algumas características relativas aos contornos detectados, permitindo a classificação sobre um item detectado, identificando se ele é um texto ou uma figura no documento digitalizado. No presente trabalho, foram extraídas informações de 100 números do jornal Diário Oficial do Estado de Sergipe, totalizando 1.100

páginas, visando caracterizar o objeto detectado, testar e treinar o modelo de inteligência artificial. Assim, foram utilizadas as seguintes informações de cada um dos itens detectados: a) Proporção, que é o resultado da largura da imagem dividido pela altura da imagem. Dessa forma, se o valor obtido como resultado for menor do que 1, o objeto ou figura tem uma altura maior que a largura. Se a proporção for maior que 1, o objeto tem a largura maior que a altura e, se for igual a 1, tem-se um objeto quadrado. No trabalho de Quddus, Cheikh e Gabbouj (2016) a proporção é utilizada para reduzir o espaço de pesquisa. Neste trabalho, a proporção foi uma das características adotadas na detecção. Segundo Kumar, Sailaja e Begum (2019), a proporção é a relação entre a largura e a altura de uma imagem, conforme Fórmula 2, a seguir.

Fórmula 2 - Proporção de uma imagem

$$proporção = \frac{largura\ da\ imagem}{altura\ da\ imagem}$$

Fonte: Fórmula adaptada de Kumar, Sailaja e Begum (2019).

b) Extensão, que é a área da forma dividida pela área da caixa delimitadora do objeto. Como resultado, toda extensão é menor do que 1, visto que o objeto deve sempre ser menor que a figura como um todo, ou seja, a extensão é a área coberta por algo (KUMAR; SAILAJA; BEGUM, 2019), conforme a fórmula3, a seguir.

Fórmula 3 - Extensão de um objeto

$$Extensão = \frac{área\ do\ objeto}{área\ do\ retangulo}$$

Fonte: Fórmula adaptada de Kumar, Sailaja e Begum (2019).

c) Casco convexo, que é usado para verificar a curva quanto a defeitos de convexidade e corrigi-la (KUMAR; SAILAJA; BEGUM, 2019). Assim, dado um conjunto de pontos no espaço euclidiano, o casco convexo é o menor conjunto possível que contém esses pontos, conforme apresentado na Figura 21, a seguir.

Figura 21 - Aplicação do casco convexo

Fonte: Elaborada pelo autor (2019).

Dessa forma, o espaço da imagem é menor do que o espaço utilizado pelo casco convexo para delimitar o objeto.

d) Solidez, que é a qualidade ou estado de ser firme ou forte em relação a sua estrutura (KUMAR; SAILAJA; BEGUM, 2019), sendo o resultado da divisão da área do contorno pela área do casco convexo, conforme a fórmula 4, a seguir.

Fórmula 4 - Solidez de um objeto

$$solidez = \frac{área\ do\ contorno}{área\ do\ casco\ convexo}$$

Fonte: Fórmula elaborada pelo autor (2019).

e) Área do objeto, que representa o número de pixels dentro do contorno, ou seja, o total de pontos em um objeto, dado pela Fórmula 5, a seguir.

Fórmula 5 - Área de um objeto

$$area = largura\ do\ objeto\ X\ altura\ do\ objeto$$

Fonte: Fórmula elaborada pelo autor (2019).

Por fim, foram utilizados ainda, isoladamente, a largura e altura dos objetos detectados.

2.2 CLASSIFICAÇÃO: ANÁLISE DE *LAYOUT* E DETECÇÃO AUTOMATIZADA DE TEXTO E IMAGENS

Para classificar os itens de uma imagem é necessário efetuar a análise do layout. Assim, com base nos dados extraídos, foi criada uma base de informação para preparar a Inteligência Artificial (IA). Isso permitiria que, nas próximas imagens, o sistema detectasse, automaticamente, se os itens de um documento eram textos, figuras ou, ainda, apenas ruídos. Nesse processo foram utilizadas 5.000 informações relativas a dez números do jornal pesquisado, com as características listadas anteriormente. Entre as técnicas de IA, optou-se para essa tarefa a adoção do "aprendizado supervisionado".

O aprendizado supervisionado é uma técnica onde, com base em exemplos de pares de entradas e saídas, o sistema aprende uma função que realiza o mapeamento das entradas e saídas (RUSSEAL; NORVIG, 2013). Nesse contexto, o aprendizado supervisionado é um modelo de aprendizado indutivo. Segundo Goldschimidt, Passos e Bezerra (2015), o aprendizado supervisionado faz com que o algoritmo compreenda e abstraia os relacionamentos e as informações, criando um modelo de conhecimento a partir de um conjunto de dados, os quais são apresentados em um formato de pares ordenados – entradas e saídas desejadas.

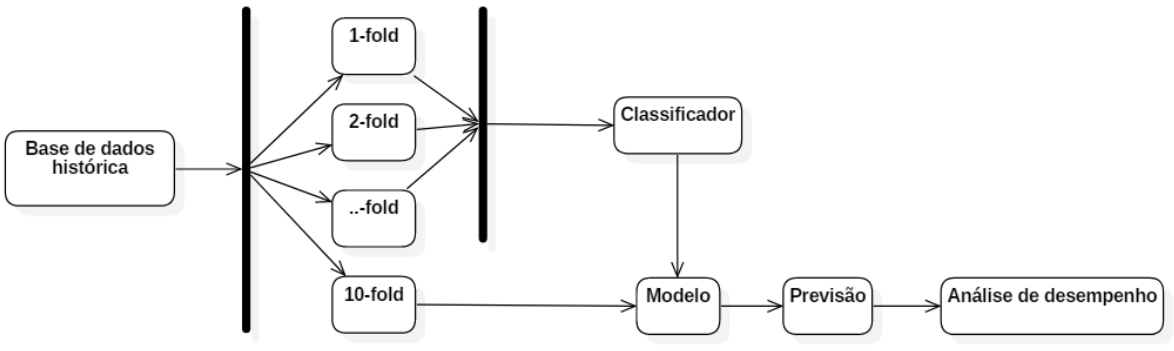
O x representa um conjunto de atributos que são utilizados para previsão. Já o f(x) é o alvo, ou seja, corresponde a uma função f, desconhecida, sendo que cada um dos pares é denominado um exemplo de função f (GOLDSCHIMIDT; PASSOS; BEZERRA,2015). A função f atua sobre um conjunto de hipóteses, que são os alvos. Assim, o modelo é criado para analisar os atributos previsores, com o objetivo de aceitar uma das hipóteses e derrubar as demais. No caso do presente trabalho, as hipóteses são: o objeto é uma imagem, objeto é um texto, o objeto é um ruído. Cada um desses itens pode aceitar apenas uma hipótese e negar as demais. Esta é uma técnica útil na análise de layout, visando validar se os itens detectados são texto, imagens etc.

Diversos autores vêm estudando técnicas que permitem a análise de layout, utilizando aprendizado de máquina em fontes impressas, principalmente, jornais. Zeni e Weldermarian (2017) realizaram análise de layout em um experimento com cem jornais, empregando o algoritmo de árvore de decisão e obtendo uma precisão média de 84%. Já Bukhari et al. (2010) utilizaram o classificador Perceptron Multicamadas (MLP) na separação de texto e não texto. O trabalho, no entanto, não traz a precisão relacionada ao algoritmo. Palfray et al. (2012) utilizaram para um experimento quarenta e duas imagens do

Jornal de Rouen, em francês antigo, obtendo uma precisão de 85,84%. Outro trabalho que utiliza fontes antigas é o de Hebertet al (2014), com o sistema PlaIR, realizando um experimento em dois livros,em língua francesa, obtendo como resultado a acurácia de 77,07 e 87,61%. Já Pramanik e Bag (2018) obtiveram precisão global de 88,74%, sendo que o MLP obteve precisão de 88,74%, o classificador Máquina de Vetor de Suporte (SVM)obteve precisão de 86,45% e o Random Forest com precisão de 86,17% para classificação de documentos em Bengla (um dos idiomas indianos). Os pesquisadores Chathuranga e Ranathunga (2017) propuseram um algoritmo para extração de conteúdo de jornais antigos. No teste com quarenta e quatro imagens, eles obtiveram uma precisão de 69.79%. No trabalho de Vasilopoulos (2018), utilizando setenta e quatro páginas digitalizadas de jornais, em Árabe, com um algoritmo de análise de layout elaborado pelo pesquisador, foi obtida uma precisão de 90,4%

Assim, para identificar se uma determinada região segmentada de uma imagem é um texto ou não, foram testadas as seguintes técnicas de aprendizado de máquina: NaiveBayes, MLP, Regressão Logística, Árvore, Árvore Randômica e Floresta Randômica. O objetivo foi determinar qual dessas técnicas teria um melhor resultado para a detecção se o objeto era um texto, uma figura ou um ruído. Para avaliar as técnicas, foram utilizadas as características extraídas dos jornais, buscando-se treinar e validar os algoritmos de aprendizado de máquina, conforme exposto na Figura 22, a seguir.

Figura 22 - Ciclo do aprendizado de máquina



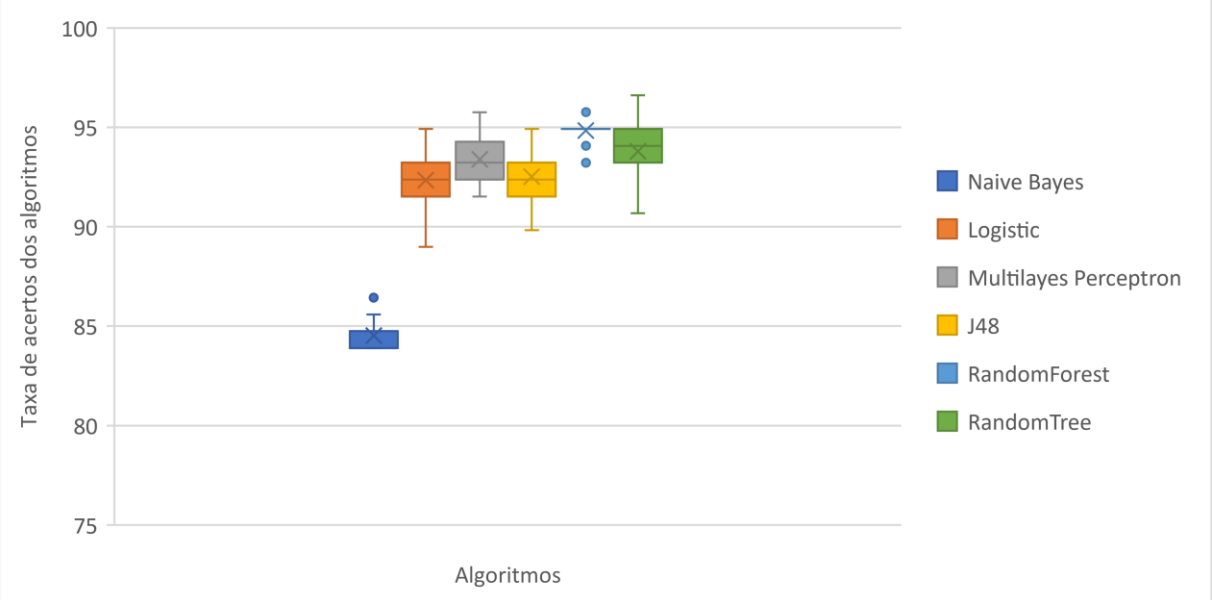
Fonte: Elaborada pelo autor (2019).

Para que fosse possível testar os algoritmos, foi utilizada a ferramenta "*Waikato Environment for Knowledge Analysis*" (WEKA) na base de dados de características. Para a validação dos algoritmos, foi adotada a técnica cruzada "k-fold", a qual é aplicada de forma

que toda a base de dados seja utilizada em treinamento e teste (DUCHESNE; RÉMILLARD, 2005). Assim, dado uma base de dados com 2000 registros e sendo definido o k = 10, a base será dividida em 10 subconjuntos, onde cada um terá 200 registros cada, sendo utilizados 9 subconjuntos para o treinamento e um para o teste, sendo rotacionado, até que todos os subconjuntos tenham sido utilizados (DUCHESNE; RÉMILLARD, 2005).

Dessa forma, a base de dados de características extraídas dos jornais foi dividida em *folds*. A seguir, foram realizados trinta testes com cada algoritmo, empregando-se como variação os valores de 1 a 30. Isso permitiu a análise da média de acerto das seguintes técnicas de aprendizado de máquina: J48 (Árvore), *Logistic* (Regressão Logistica), Perceptron Multicamadas (MLP), *Naive Bayes*, *Random Forest e Random Tree*, totalizando 180 testes. Como resultado, o algoritmo *Random Forest* apresentou a melhor média, além de ter uma baixa variância dos dados, demonstrando que este algoritmo possui maior precisão e acurácia, seguido do Random Tree, conforme pode ser visto no Gráfico 1 que apresenta os *boxplots* das porcentagens de acerto, a seguir.

Gráfico 1 - Comparação das técnicas de aprendizado de máquina



Fonte: Elaborado pelo autor (2019).

O Quadro 2, a seguir, apresenta a média de acertos dos algoritmos testados. Contata-se que o algoritmo *Random Forest* obteve um melhor desempenho, com uma taxa média de acerto de 95% na classificação de itens como texto, imagens e ruídos.

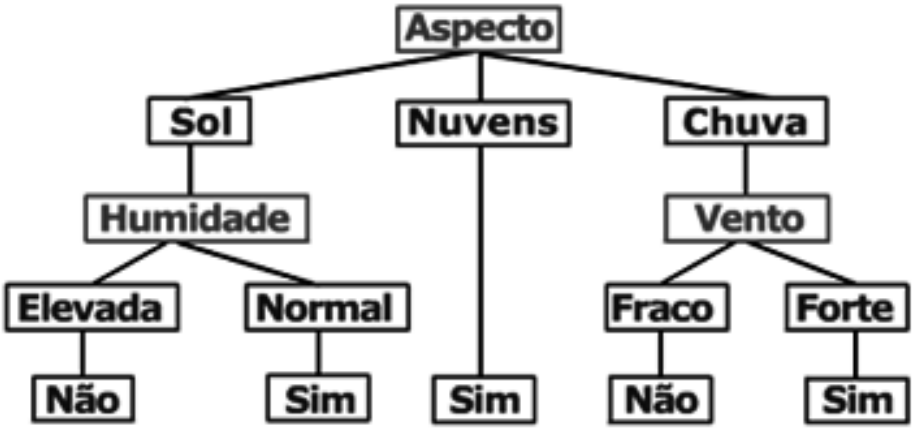
Quadro 2 - Média de acerto dos algoritmos de aprendizado de máquina

Algoritmos	*NaiveBayes*	*Logistic*	*MLP*	*J48*	*Random Forest*	*RandomTree*
Média	85	92	93	93	95	94

Fonte: Elaborado pelo autor (2019).

Com base no desempenho médio, o *Random Forest* foi selecionado como algoritmo de aprendizado de máquina supervisionado para a etapa de classificação na análise de layouts. O método *Random Forest* surgiu em 1995, no trabalho de Kan, apresentado na "*International Conferenceon Document Analysis and Regonition*". Esse método é uma combinação de classificadores do tipo árvore, a qual é, conforme Goldschimidt, Passos e Bezerra (2015), um modelo de representação de conhecimento que utiliza os nós para representar decisões. Segundo Baeza-Yates e Ribeiro-Neto (2013), a árvore é um método preditivo que, por meio de regras, utiliza um caminho em um formato de árvore, a qual é utilizada para classificação, conforme demonstrado na Figura 23, a seguir.

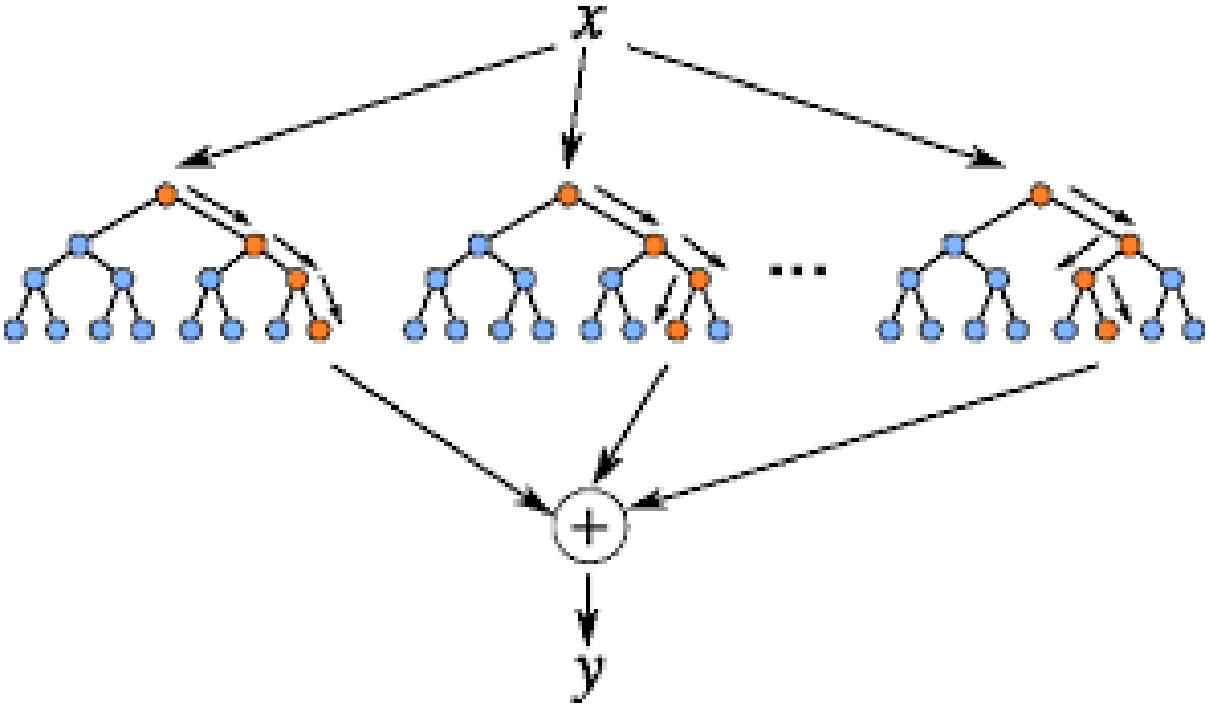
Figura 23 - Exemplo de árvore de decisão



Fonte: Elaborada pelo autor (2019).

Dessa forma, uma *Random Forest* combina diversos classificadores do tipo árvores, sendo um algoritmo mais robusto em relação a ruídos, porém, ainda considerado um algoritmo leve (GISLASON; BENEDIKTSSON; SVEINSSO, 2006; BREIMAN, 2001). A *Random Forest* é, assim, um conjunto de árvores atuando para obter uma classificação, conforme exposto na Figura 24, a seguir.

Figura 24 - Floresta randômica



Fonte: Elaborada pelo autor (2019).

Após a classificação, os itens textuais devem ser então extraídos, para que seja possível inserir as informações em um banco de dados.

2.3 EXTRAÇÃO DE DADOS HISTÓRICOS

A transcrição de documentos impressos históricos é uma tarefa essencial para o pesquisador (BACELLAR, 2018). Como exemplo tem-se o trabalho de Sabino (2017), onde houve a necessidade de transcrição de vários trechos de jornais antigos, a fim de facilitar a leitura e visualização das informações em sua tese. Porém, transcrever um grande volume de documentos demanda esforço e tempo por parte do pesquisador. Segundo Bacellar (2018) essa é uma tarefa que pode se estender entre dias a meses. Dessa forma, um sistema automático reduz o trabalho manual do pesquisador, fazendo a transcrição do documento e extraindo textos de imagens. Esse processo é denominado reconhecimento ótico de caracteres (*Optical Character Recognition* - OCR), que tem como benefício a possibilidade de realização de buscas nos textos, auxiliando o pesquisador na tarefa de catalogar informações.

Os primeiros OCRs surgiram na década de 1930. Tauscheck (1935) propôs um equipamento mecânico que permitia o reconhecimento de caracteres e números. Porém, foi apenas na década de 1950, com o surgimento dos computadores, que os OCRs se tornaram produtos de mercado. Em 1953, David H. Shepard obteve uma patente para o primeiro dispositivo de OCR, denominado Gismo (SHERPARD, 1953). Posteriormente, em 1962, Hannan, do Grupo RCA, elaborou um OCR que combinava técnicas eletrônicas e óticas, permitindo reconhecer os idiomas inglês e russo (MORI et al., 1992).

Em 1984, a empresa HP patrocinou um projeto de doutorado que deu origem ao Tesseract, sendo desenvolvido entre 1984 e 1994, no laboratório HP Labs Bristol em conjunto com a divisão de Scanners da HP do Colorado. Em 1995, o Tesseract foi apresentado durante o Teste anual da UNLV, sendo lançado para Windows em 1996. No ano de 2005 a HP liberou o Tesseract como código aberto, permitindo assim, a sua melhoria e atualização do sistema. A partir de 2006, a empresa Google assumiu a gestão do Tesseract, auxiliando no seu desenvolvimento. Desde então, houve a melhora do suporte a múltiplas linguagens, alcançando, em 2010, mais de 60 idiomas suportados, entre eles o português.

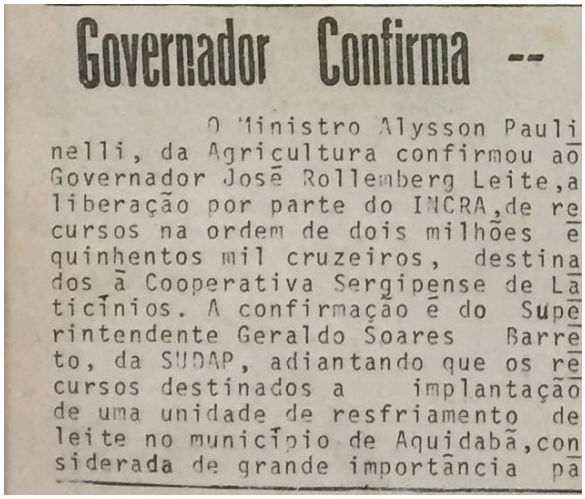
Para o presente estudo, foi utilizado o motor de OCR Tesseract, versão 4.0.0.20181030. A sua escolha deve-se ao fato de ser um *software open-source*, multiplataforma (disponível para Windows, Linux e MacOS), tendo base de treinamento para o idioma português. O Tesseract assume que a entrada é uma imagem binária, com regiões de texto, e segue o ciclo básico de extração do texto (SMITH, 2007). Assim, a arquitetura proposta nesta tese para a extração de texto segue o fluxo apresentado na Figura 25, a seguir. A caixa em azul (Figura 25) representa o Tesseract que recebe uma imagem já preparada e tem como saída o texto referente à imagem.

Figura 25 - Fluxo de trabalho para extração de dados

Fonte: Elaborada pelo autor (2019).

Como resultado, ao ser realizada a entrada de uma imagem que contenha texto (Figura 26), ele é extraído automaticamente, possibilitando que seja inserido em um banco de dados para, posteriormente, efetuar a indexação e busca.

Figura 26 - Transcrição automática de textos de imagem



(*) Governador confirma –
O ministro Alyson Paulinelli, da Agricultura confirmou ao Governador José Rollemberg Leite, a liberação por parte do INCRA, de re cursos na ordem de dois milhões e quinhentos mil cruzeiros, destina dos a Cooperativa Sergipense de La ticinios. A confirmação é do Superintendente Geraldo Soares Barre to, da SUDAP, adiantando que os re cursos destinados a implantação de uma unidade de resfriamento de leite no município de Aquidabã, com siderada de grande importância pa

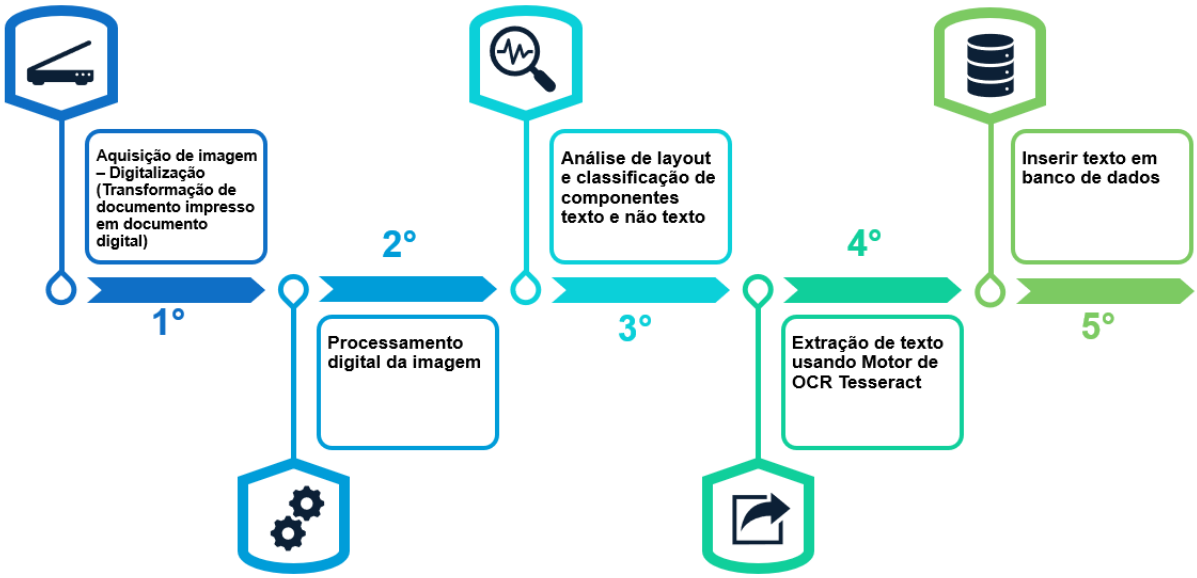
Nota (*): O texto apresentado está, em sua integridade, de acordo com o obtido pela extração automatizada, incluindo eventuais falhas em espaços e letras.
Fonte: Figura elaborada pelo autor, adaptado do acervo SEGRASE (1997).

Silva (2014) aponta que a transcrição automática de documentos, quando em bom estado de conservação, possui uma alta taxa de transcrição (maior que 90%) em ferramentas comerciais, porém em documentos históricos, que possuem degradação e ruídos, o desempenho é reduzido. Assim, a etapa de análise de layout, apresentada na Figura 23, efetua ainda um processamento digital da imagem, já que, conforme Mello (2002) os OCRs perdem informações quando há alguma degradação ou ruído. Aquele autor aponta, ainda, que o reconhecimento ótico, ao analisar uma imagem, separa as linhas de texto e, posteriormente, analisa cada caractere de forma individual. A partir disso, o reconhecimento ótico classifica o objeto como um dos caracteres do alfabeto. Assim, caso tenha algum ruído, é possível que o OCR classifique um caractere de forma incorreta.

Ao utilizar o Tesseract diretamente no jornal do Diário Oficial do Estado de Sergipe, de 28 de agosto de 1997, a velocidade de extração de texto de metade de uma página de jornal, sem pré-processamento prévio, foi de 9,833 segundos com variações de 0,8 segundos para mais ou para menos, e com uma taxa de reconhecimento de 83%. Já com o pré-processamento, que leva 1,132 segundos, a velocidade de extração foi de 5,733 segundos e a taxa de reconhecimento foi de 96%. Embora se tenha obtido uma baixa taxa de erro, considerada tecnicamente aceitável, na próxima etapa de indexação será realizada uma análise do texto, buscando-se corrigir eventuais erros. Assim, como resultado do processo de análise de layout e extração de dados, obteve-se o processo tecnológico, conforme a Figura 26, a seguir. Tal processo pode ser utilizado tanto por pesquisadores, em pequena

escala, como por bibliotecas que pretendam digitalizar um acervo. As ferramentas desenvolvidas neste trabalho possibilitam que os passos 2 ao 5 sejam realizados de forma automatizada (Figura 27).

Figura 27 - Fluxo técnico desenvolvido para extração de dados de fontes históricas



Fonte: Elaborada pelo autor (2019).

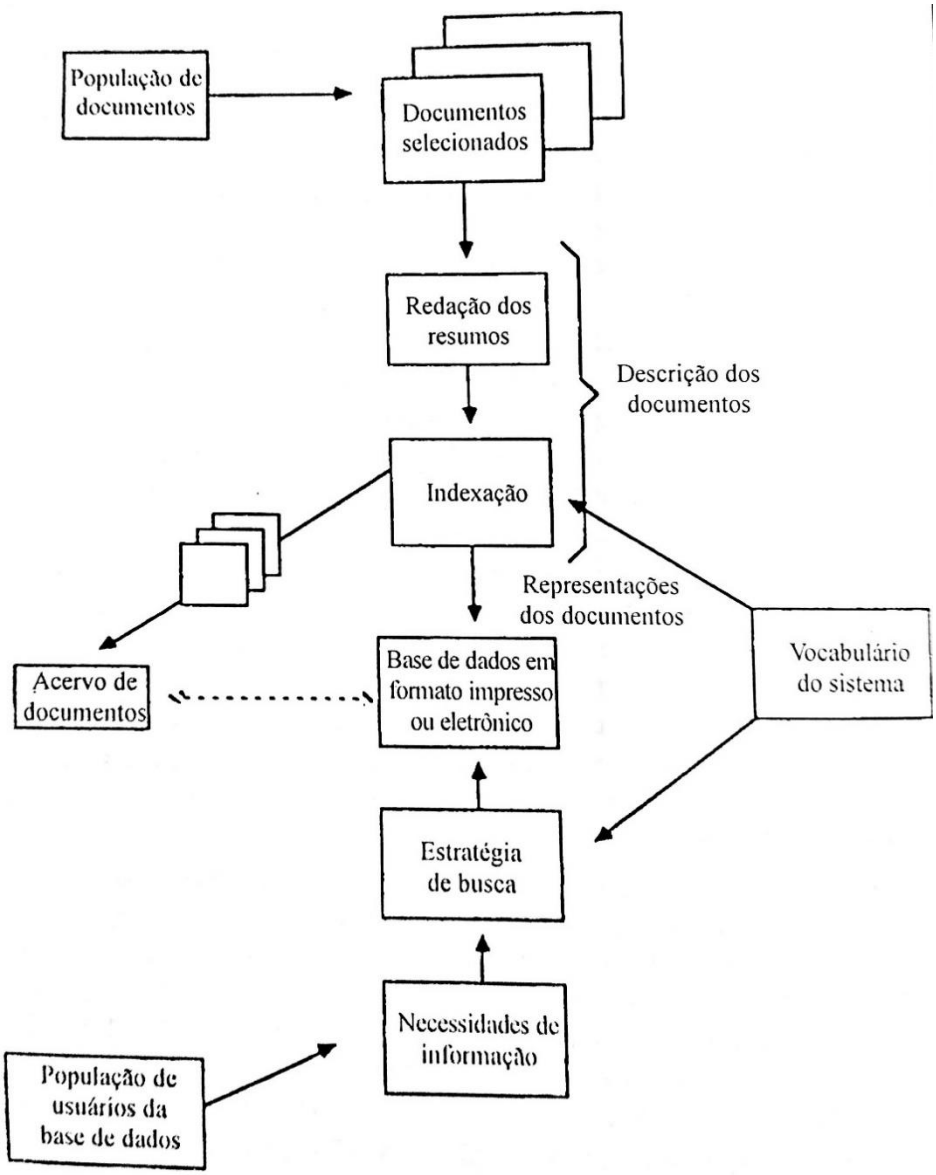
O fluxo da Figura 27, permite que os pesquisadores possam utilizar a tecnologia no processo de transcrição do documento, reduzindo o tempo necessário para essa tarefa essencial ao pesquisador da educação. O procedimento manual de transcrição exige cuidados que são, por vezes, ignorados pelo pesquisador e que torna o processo demorado ou comprometido em sua qualidade (BACELLAR, 2018). Outro ponto a ser destacado como benefício da adoção da transcrição automatizada é que ela mantém a grafia original, um cuidado recomendado aos pesquisadores durante o processo manual (BACELLAR, 2018). Após a extração de dados, o passo seguinte será a análise dos textos e a indexação, porém, sem perder a imagem original. A reserva do material original é útil ao pesquisador que intente analisar o estado da imagem, as marcar do papel, etc. Em prosseguimento, na seção 3, é tratado sobre a indexação dos dados e a busca de informações.

3 INDEXAÇÃO E BUSCA DE INFORMAÇÕES HISTÓRICO-EDUCACIONAIS

O pesquisador que realiza análises históricas e documentais, ao obter um conjunto de fontes e extrair, de forma manual ou automatizada, tem agora uma nova tarefa. Partindo-se das recomendações de Bacelar (2018), os investigadores devem se questionar sobre como organizar o conjunto de documentos obtidos durante as visitas a escolas, órgãos de documentação, editoras, etc. A massa documental obtida nas pesquisas deve ser classificada e ordenada de forma a facilitar a busca de dados. É necessário que os pesquisadores analisem e adotem meios para evitar a perda de informações e obter, com maior assertividade e em menor tempo possível, os dados de seu interesse no universo de documentos levantados.

Bacelar (2018) aponta que, para a tarefa de organização e busca de informação, os pesquisadores podem fazer uso de cartões ou outros instrumentos como forma de indexar as fontes com base nos dados transcritos. Isso permitirá ao consulente remeter-se às fontes disponíveis quando necessário. De Luca (2018) menciona que ao lidar com fontes como jornais, periódicos ou revistas para o tratamento histórico, é necessária uma peregrinação em um conjunto grande de materiais, exigindo uma busca minuciosa pelas informações. Em posse dessas informações, é necessária a ordenação do material e a organização e caracterização do conteúdo para que seja possível proceder uma análise sobre as fontes encontradas. Neste contexto de organização e busca de informação, Lancaster (2004) apresenta um modelo amplo, demonstrando o local da indexação e dos resumos em um sistema de recuperação de informação, conforme exposto na Figura 28, a seguir.

Figura 28 - Função da elaboração de índices e resumos no quadro mais amplo da recuperação de informação



Fonte: Lancaster (2004, p. 2)

No contexto da História da Educação, Werle (2001) apresenta o fluxo adotado em seu trabalho doutoral, partindo da coleta de dados, seguido de uma organização deles por meio de fichas. Aquele autor obteve um volume de documentos que precisou ser estruturado para o adequado armazenamento. Isso viabilizou um sistema de busca, a partir da ficha criada, a qual conduziu à definição de um banco de dados. Esse processo, feito manualmente, poderia ter sido automatizado, porém exigiria ferramentas tecnológicas diversas. O foco daquele autor foi a criação de um banco de dados, sendo utilizado o sistema Dialog para

montar uma base de informações. O processo adotado por Werle (2001) é comum a diversos pesquisadores do campo da História da Educação. Como exemplo tem-se o projeto desenvolvido em meados da década de 1990, com foco na construção de um banco de dados temático, adotando o *software FolioViews* (WERLE, 2001). Esse programa foi utilizado devido à capacidade para suportar um volume grande de informações em formatos livres e semiestruturadas. O *FolioViews* utiliza o Infobases para articular as informações, as quais podem ser vinculadas como hipertexto. No entanto, independente da ferramenta adotada pelos pesquisadores, o processo é manual e, em geral, não apresenta integridade à indexação de informações, além de não possibilitar a criação de um vocabulário documental. Esse aspecto é apontado por Lancaster (2004) como sendo essencial para a busca de informação. Dessa forma, o SPEdu automatiza o processo de indexação e geração de vocabulário do documento, facilitando a busca de informações.

Para a automatização no SPEdu, foi adotado o processamento de linguagem natural (NLP), a qual é uma subárea da IA que, segundo Eisensten (2019), possui um conjunto de métodos para tornar a linguagem humana acessível ao computador.

3.1 INDEXAÇÃO DE DOCUMENTOS HISTÓRICO-EDUCACIONAIS

A indexação e o resumo de fontes são essenciais para a catalogação e organização de acervos (LANCASTER, 2004), compondo um índice estruturado de dados que mapeia termos para representar documentos. Dessa forma, indexar é representar um documento por meio da uma descrição abreviada (BORGES, 2016), sendo que esta deve ser ajustada às necessidades do usuário (HJORLAND, 2001). No caso desta tese, tais necessidades estão relacionadas ao campo da História da Educação.

Antes de descrever um documento, visando a sua representação, é necessário modelar como a indexação funcionará. Independentemente do tipo de documento, a indexação deve permitir que os pesquisadores possam realizar buscas e encontrar o maior número de dados relacionados. Esse aspecto é abordado por Bonato (2004), referindo-se ao uso da tecnologia para permitir a recuperação de informações histórico-educacionais. O pesquisador, ao criar uma coleção de documentos, forma um acervo, sendo representado pela Fórmula 6, a seguir, onde "A" indica o Acervo como sendo a união dos documentos (d).

Fórmula 6 - Representação do acervo documental

$$A = d1 \cup dn$$

Fonte: Fórmula elaborado pelo autor (2020)

Dada a diversidade de documentos, eles devem ser, então, classificados e organizados manualmente segundo a sua tipologia. Cabe considerar que cada pesquisa, do campo da História da Educação pode possuir um conjunto próprio de documentos. Assim, o modelo de indexação desenvolvido nesta tese prevê uma etapa manual, em que o pesquisador, em posse dos documentos, cadastra os tipos de documentos que sua pesquisa utilizará e, a seguir, insere os documentos no sistema. A extração da informação e a inserção no banco de dados é automatizada. Cada documento (D) do acervo, possui um conjunto de palavras (p) que formam o seu vocabulário, conforme a Fórmula 7, a seguir.

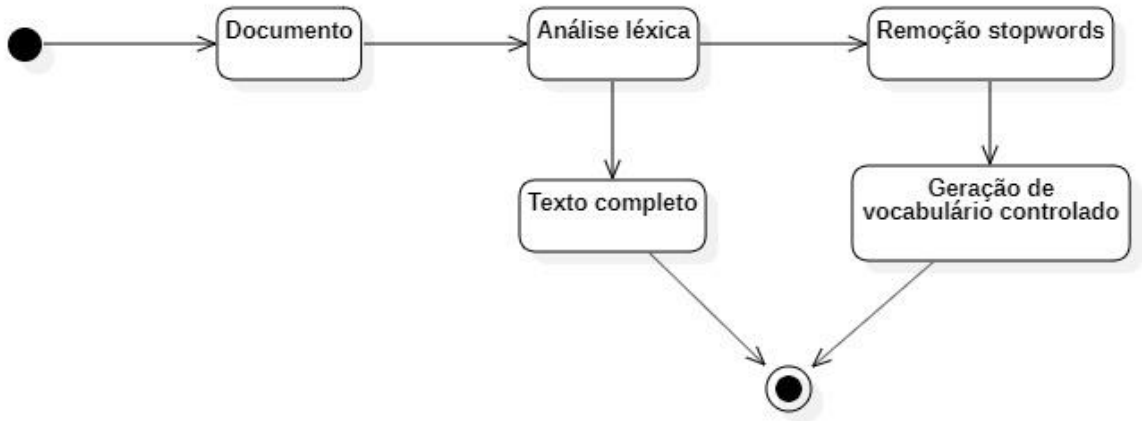
Fórmula 7 - Representação do vocabulário do documento

$$D = p1 \cup pn$$

Fonte: Fórmula elaborada pelo autor (2020).

Assim, para que se extraia o vocabulário, os documentos passam por uma etapa de pré-processamento de informação. Conforme Baeza-Yates e Ribeiro-Neto (2013) esse procedimento é denominado *stopwords*, envolvendo a análise léxica e eliminando palavras que não tenham relevância para a recuperação de informação, (a, as, e, o, que), conforme a Figura 29, a seguir.

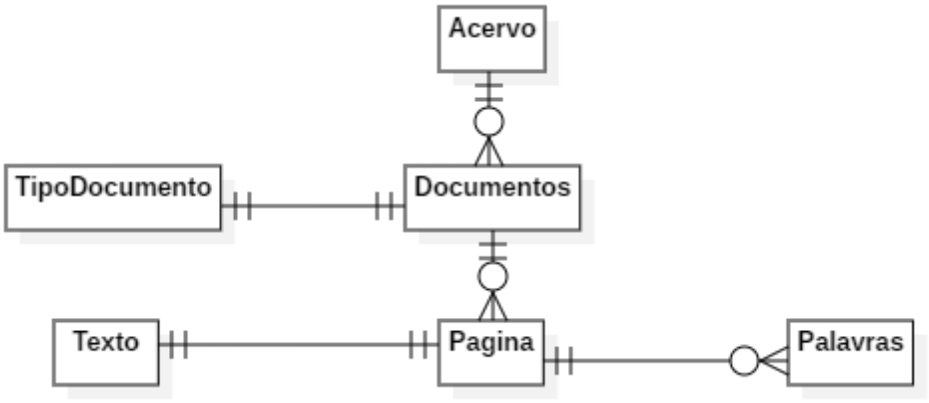
Figura 29 - Visão lógica do documento por meio do pré-processamento de texto



Fonte: Elaborada pelo autor, adaptada de Baeza-Yates e Ribeiro-Neto (2013).

Essas tarefas estão relacionadas com a NLP, que é parte da IA. Embora as palavras sejam extraídas, os textos são mantidos na sua integralidade, como as imagens originais. Assim, para que se possa fazer um bom relacionamento dos dados, é necessário modelar uma organização documental definindo, em primeira instância, os tipos de documentos a serem armazenados. Nesse contexto, como o campo da História da Educação possui um vasto conjunto de documentos, em vez de se criar um modelo fixo, definiu-se para o SPEdu apenas uma entrada de tipo de documento. Dessa forma, cada pesquisador, ao utilizar o instrumento, definirá os tipos para a organização de documentos, conforme Figura 30, a seguir.

Figura 30 - Diagrama de entidade relacional do SPEdu, representando a organização dos documentos, palavras e texto dentro do acervo



Fonte: Elaborado pelo autor (2020).

O modelo adotado no SPEdu (Figura 29), permite a criação de uma taxonomia: acervo – tipo de documento – documento, sendo que acervo e tipo é definido manualmente pelo usuário. Tais dados são utilizados durante as buscas, permitindo uma catalogação do conteúdo por assunto. Esse processo, conforme Lancaster (2004), representa o conteúdo baseado no total de itens por um tipo, como livros, relatórios, etc. Em complemento, a análise léxica, realizada na etapa de pré-processamento, tem como principal função segmentar o texto em palavras, as quais serão utilizadas na etapa de busca. Nesse sentido Khalfallah, Aloulou e Belguith (2016) desenvolveram uma plataforma que permite a criação de um dicionário histórico padrão para documentos em Árabe, empregando NLP para buscas de documentos por meio do corpus dos textos naquele idioma. Já o trabalho de Fagan (2017) apresenta um método para indexar os documentos por meio de frases, demonstrando a recuperação de cinco coleções de documentos.

Colavizza, Ehrmann e Bortoluzzi (2019) afirmam que o trabalho de indexação por conteúdo, normalmente realizado por palavras e frases, se feito manualmente não dimensiona o processo, dificultando o trabalho do pesquisador. Assim, os autores apresentam um método para inicializar a implantação de um sistema de informação baseado em conteúdo de arquivos históricos. Nesse sentido, o SPEdu atua na indexação com base no conteúdo completo para a recuperação de informação. Segundo Colavizza, Ehrmann e Bortoluzzi (2019), tal técnica requer o conteúdo total do documento, adquirido na etapa de extração. Como exemplo de uma abordagem de pesquisa de conteúdo total têm-se a pesquisa do Google Books, com uma coleção de livros digitalizados. Essa forma de indexação de conteúdo se faz necessária para o campo da História da Educação, pois possibilita uma análise completa das informações, ampliando as possibilidades de busca por parte do pesquisador do campo (BONATO, 2004).

Zhang e El-Gohary (2016) apontam que a extração de informação para a criação de um vocabulário documental é uma tarefa desafiadora, necessitando um processamento de texto eficiente. Conforme Croft, Metzler e Strohman (2009), é necessário que um sistema de indexação possua um componente de análise de dados, o qual deve ser responsável por processar as sequências de texto de forma a identificar palavras, efetuar limpeza nos dados, etc. O primeiro passo para a indexação, denominado *tokenização* do texto, tem como principal função a quebra dos textos em palavras (GREFENSTETTE, TAPANAINEN, 1994; EISENSTEIN, 2019). Ao processo o texto é importante definir, ainda, se ele sofrerá alguma modificação. Como exemplo, pode ser definido como critério se os textos deverão ser convertidos em letras maiúsculas ou minúsculas, uma vez que isso afetará a busca posteriormente. No SPEdu foi adotado que o texto completo seria mantido com a grafia original. Já os *tokens*, seriam convertidos para letras minúsculas. Tal ação é necessária para que a busca seja mais simples de executar.

Após a *tokenização*, é necessário a eliminação de termos, pontuações e símbolos, os quais são irrelevantes em buscas por não agregarem valor as pesquisas (REESE, 2015). Essa tarefa de eliminação é denominada remoção de *stopwords*. Palavras consideradas irrelevantes à pesquisa, como artigos, preposições, numerais e pronomes, ocorrem em mais de 80% dos documentos, devendo ser removidas dos termos de indexação (BAEZA-YATES; RIBEIRO-NETO, 2013). Isso viabilizará que o mecanismo de busca seja mais eficiente (CROFT; METZLER; STROHMAN, 2009). O Quadro 3, a seguir, apresenta o processo de *tokenização* e remoção de *stopwords*.

Quadro 3 - Lista de *tokens* (palavras) obtidas após a remoção de *stopwords*

Texto original	O ministro Alyson Paulinelli, da Agricultura confirmou ao Governador José Rollemberg Leite, a liberação por parte do INCRA,
Tokenização	O | ministro | Alyson | Paulinelli | , | da | Agricultura | confirmou | ao | Governador | José | Rollemberg | Leite | , | a | liberação | por | parte | do | INCRA | ,
Remoção de *stopwords* e sinais	ministro | Alyson | Paulinelli | Agricultura | confirmou | Governador | José | Rollemberg | Leite | liberação | parte | INCRA

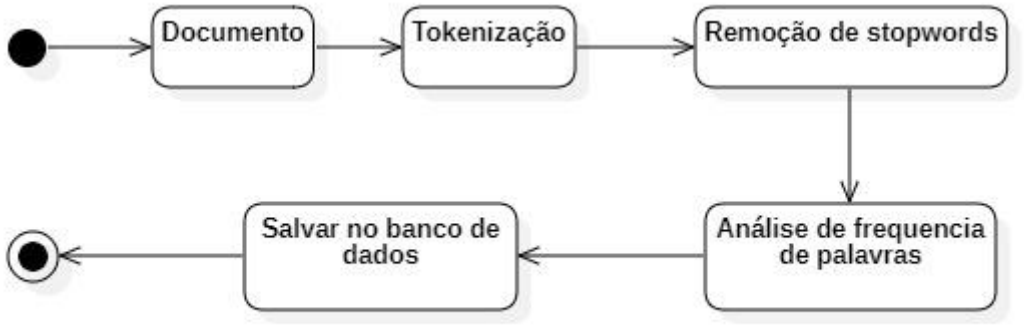
Fonte: Elaborado pelo autor (2020).

Em seguida, os termos restantes são passados para análise das palavras e, posteriormente, para o armazenamento no banco de dados. Essa análise tem como objetivo verificar a frequência de ocorrência das palavras no documento. Por exemplo, um documento que mencione cinco vezes a palavra “escola”, comparado a outro que mencione a mesma palavra apenas uma vez, deve ser mais importante. Ou seja, o instrumento SPEdu deve identificar em primeiro lugar o documento com cinco ocorrências do termo “escola”. Se um pesquisador, em posse de diários escolares, livros e jornais, desejar encontrar nesse acervo o nome de determinada pessoa, é comum que o documento que apresente mais ocorrências do nome buscado apareça no topo da lista, e o documento com menos ocorrências apareça ao final. Nesse contexto, Luhn (1958) menciona que a média de frequência são bons indicadores. Assim, conforme Rijsbergen (1979), com base nos estudos de Luhn (1958), é possível estabelecer limites de cortes relacionados a palavras que não contribuem significantemente para indexar um documento. Essa afirmação é fundamentada pela Lei de Zipf, a qual aponta que 20% das palavras mais frequentes representam 80% das buscas (WALKINSHW; MINKU, 2018). Jones (2004) sugere que os termos com menor frequência pode ser utilizados como forma de *ranking* de raridade, o que, dependendo do tipo de pesquisa, pode ser útil. Evidencia-se, portanto, a necessidade de medir a frequência das palavras (LUHN 1958; JONES, 2004). Nesta tese utilizam-se ambos os modelos, mantendo a contagem, sem remoção de palavras com baixa frequência, porém, permitindo que o usuário, ao efetuar a busca, possa selecionar se deseja os documentos que tenham maior ou menor incidência da palavra.

Além disso, a frequência das palavras pode ser utilizada, posteriormente pelo pesquisador no exame do *corpus* documental, sob o prisma da análise de conteúdo. Segundo

Bardin (2016), essa análise é uma das medidas mais utilizadas na enumeração dos termos, sendo que a frequência da palavra em um determinado documento indica a importância daquela unidade de requisitos. Assim, é também necessária uma análise estatística do vocabulário. Essa etapa tem como função reunir e registrar informações estatísticas sobre as palavras e o documento, senda utilizada posteriormente na fase de classificação dos resultados de busca (CROFT, METZLER, STROHMAN; 2009). Esse processo, no SPEdu, ficou dividido sequencialmente, conforme demonstrado na Figura 31, a seguir.

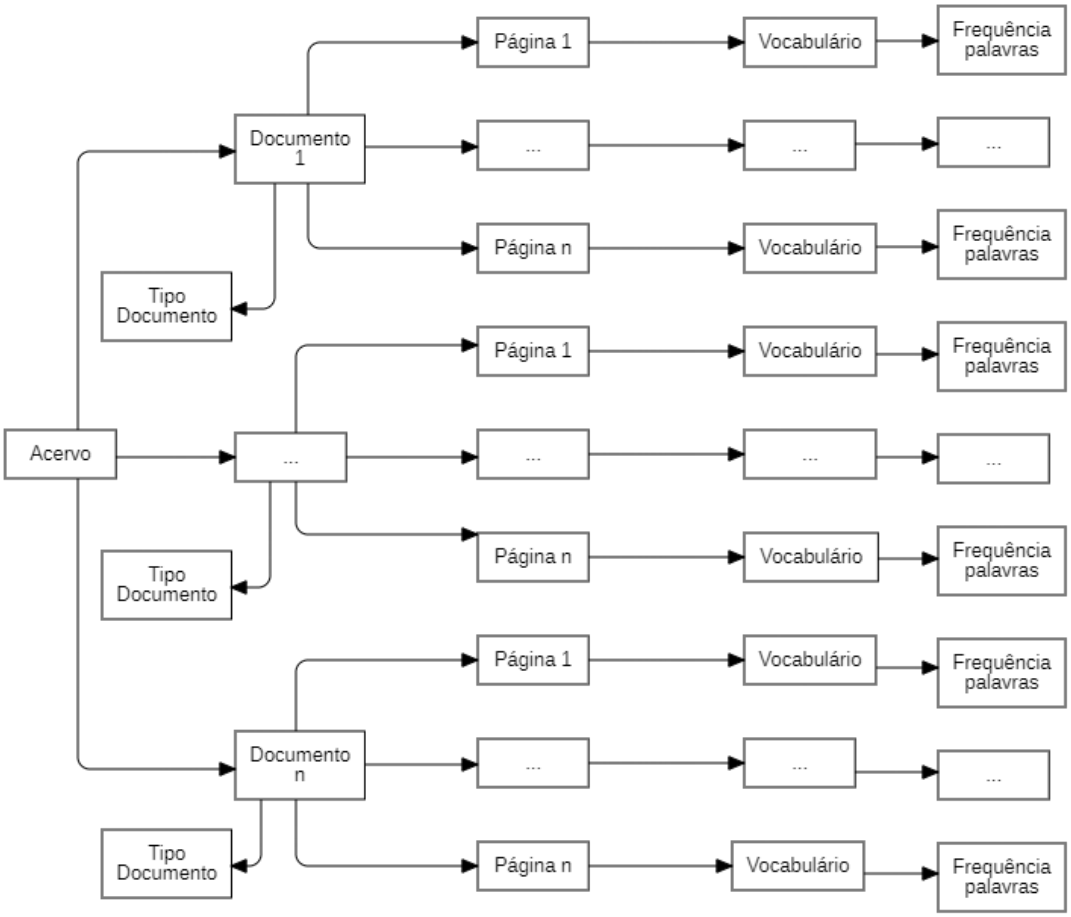
Figura 31 - Fluxo de sequência do processo de indexação por conteúdo



Fonte: Elaborado pelo autor (2020).

O fluxo gera um conjunto de informações organizadas em acervo: documentos, tipo de documento, páginas dos documentos, vocabulário das páginas e, por fim, frequência, conforme exposto na Figura 32, a seguir.

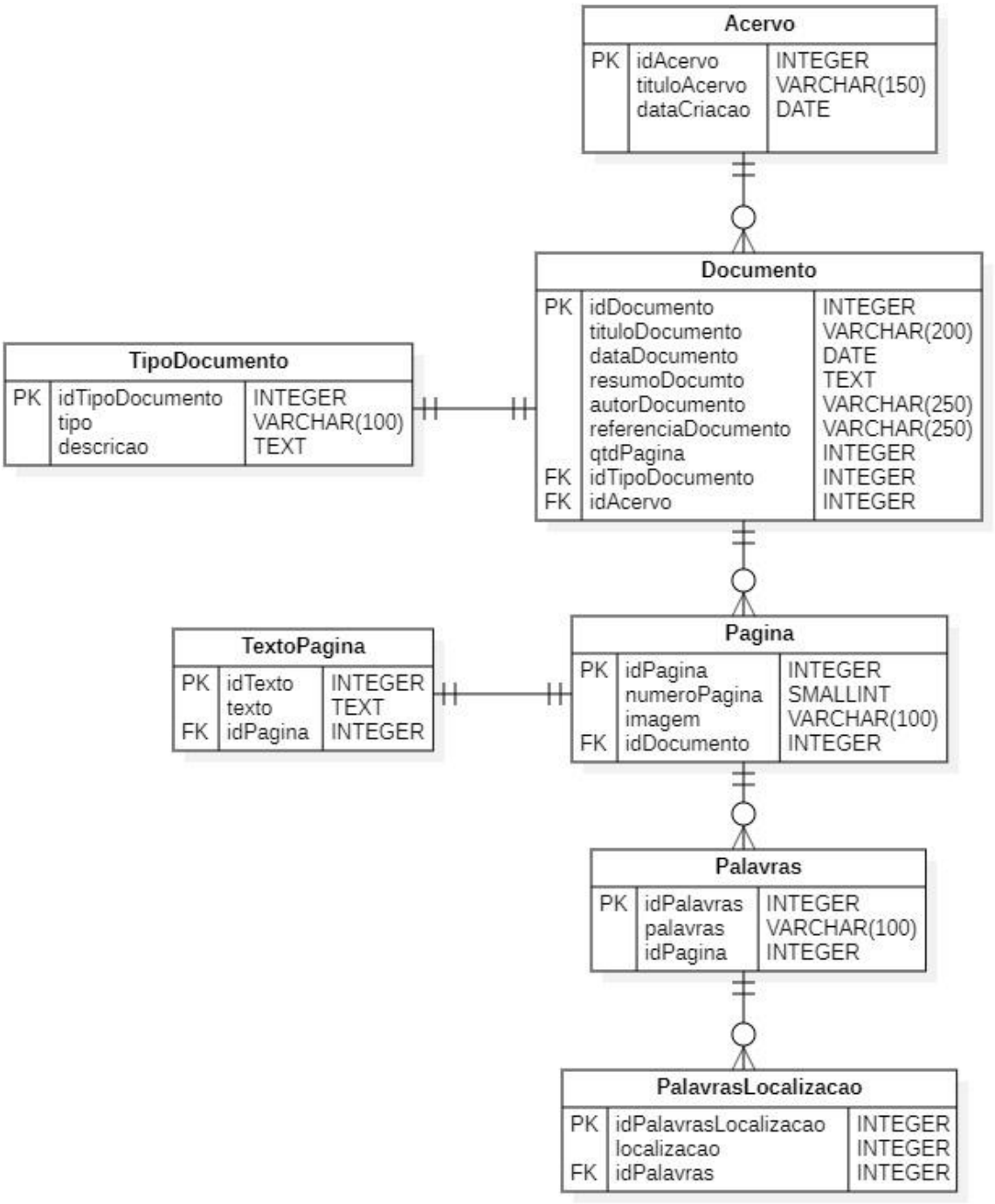
Figura 32 - Fluxo de informações para indexação documental



Fonte: Elaborado pelo autor (2020).

Os dados são, então, salvos no banco de dados, na ordem do fluxo. O vocabulário é salvo na tabela de palavras, e a frequência é contabilizada com base na análise da tabela palavrasLocalizacao. Isso permite, ainda, identificar se a palavra está no início ou no final de uma página, sendo a que a tabela palavrasLocalizacao. relacionada à página. Dessa forma, é obtido o resultado do processamento dos dados para a geração do vocabulário documental. Para ilustrar essa relação, foi criado o Diagrama de Entidades Relacional (DER) do banco de dados, conforme Figura 33, a seguir.

Figura 33 - Diagrama de Entidade Relacional (DER) do banco de dados do SPEdu



Fonte: Elaborado pelo autor (2020).

As relações apresentadas no DER (Figura 33), permitem que seja realizada consulta das palavras pertencente ao vocabulário documental, ligado às páginas de um documento de todo o acervo. A tabela PalavrasLocalizacao possui a relação entre palavras, representando

a importância das palavras em uma página de documento. Isso significa que quanto maior frequência de uma palavra em um documento, maior a sua importância (LUHN, 1958; BAEZA-YATES; RIBEIRO-NETO 2013). Com base nesses dados será possível a realização de buscas mais eficientes.

3.2 BUSCA E ACESSO DE INFORMAÇÕES HISTÓRICO-EDUCACIONAIS

A criação de motores de busca para que sejam realizadas pesquisas e recuperação de informação por meio de computadores iniciou na década de 1950 (LANCASTER, 2004). Esses motores constituem um sistema de recuperação de informação, permitindo que os usuários interajam, efetivamente, com os dados disponíveis, e oferecendo suporte a pesquisas e exibição dos itens da coleção de forma organizada (SCHATZ, 1997). Assim, um motor de busca é uma ferramenta que permite encontrar documentos em uma coleção. Há diversos tipos de coleção de documentos. No caso desta tese, as coleções estão relacionadas com tipos documentais históricos-educacionais, como diários de aulas, livros antigos, cartas, jornais, etc. As coleções variam em tamanho, dependendo do tipo do acervo, podendo se expandir lenta ou rapidamente, criando dificuldades na busca manual.

O modelo tradicional de busca adotado para os sistemas está relacionado à pesquisa por palavras-chaves, extraídas manual ou automaticamente, e inseridas em um catálogo. Outro formato é a realização de pesquisas estruturadas, como pesquisa por autor, título ou palavras-chaves. Porém, em bases textuais, o modo dominante de pesquisa é pelo conteúdo (ZOBEL, MOFFAF; 2006). Tais bases, geralmente, satisfazem às necessidades de informações do pesquisador. A grande maioria dos sistemas de busca adotam consultas no modelo “saco de palavras” (ZOBEL; MOFFAF, 2006). Algumas adotam operadores booleanos como “E” e “OU”, permitindo restringir respostas a uma linguagem específica.

Croft, Metzler e Strohman (2009) apontam que um sistema de busca fornece uma interface com o usuário, permitindo que seja realizada a busca, e um analisador de conteúdo para a consulta, que viabiliza a utilização dos operadores booleanos. Dessa forma, com base nas informações extraídas e registradas no banco de dados, é possível realizar buscas ou navegação nos registros disponíveis no banco de dados por meio de um motor de busca.

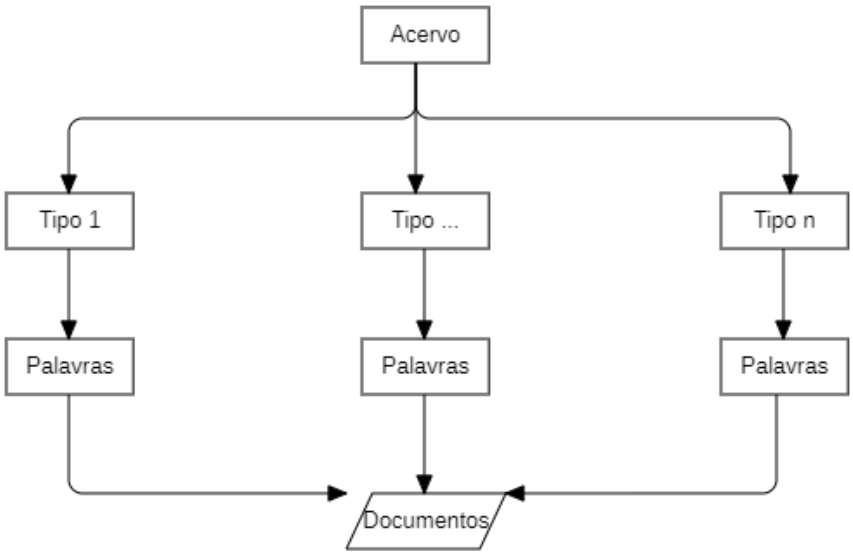
O processo de recuperação de informação permite uma abordagem baseada em tipos de documentos ou em dados relacionados aos arquivos disponíveis. Esse procedimento depende das etapas de extração e indexação para que se tenha um conjunto de informações.

Esse último, por sua vez, permitirá a busca e recuperação dos dados que tenham um formato de texto eletrônico. Conforme Lima (2016) isso possibilita novas maneiras de armazenamento e acesso de forma interativa. Segundo Zobel e Moffaf (2006), os motores de busca, em um sistema de recuperação de informação, devem possuir os seguintes requisitos técnicos:

a) Resolução eficaz de consultas;
b) Utilização de características do texto convencional, tais como proximidade do termo de consulta, que melhoram a eficácia;
c) Resolução rápida de consultas;
d) Uso mínimo de outros recursos (disco, memória, largura de banda);
e) Escala para grandes volumes de dados;
f) Provisão de recursos avançados, como restrição booleana e consulta de frases.

Para uma resolução eficaz da consulta, pode-se, com base nas informações armazenadas, criar automaticamente uma navegação hierárquica. Isso permite que o usuário acesse os documentos disponíveis, ou seja, que ele navegue por meio do catálogo de assuntos. Essa navegação partirá do nó raiz (acervo), prosseguindo para os tipos de documentos, palavras relacionadas aos documentos e, por fim, ao próprio documento. O formato mais comum de navegação hierárquica é o *outline* (SILVA, 2016), o qual foi adotado no SPEdu como forma de recuperar informações. Isso visa permitir o acesso às informações, seguindo a hierarquia, como apresentado na Figura 34, a seguir.

Figura 34 - Navegação hierárquica SPEdu



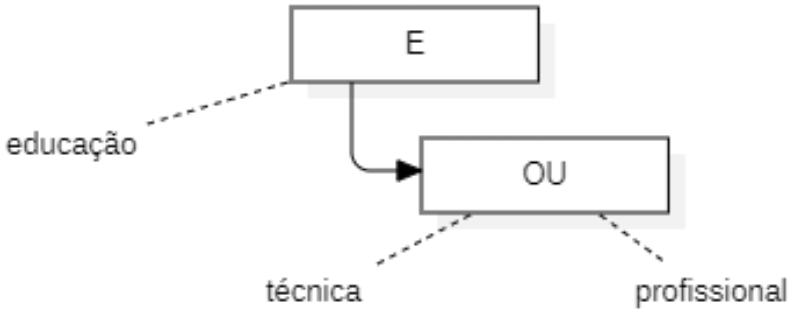
Fonte: Elaborado pelo autor (2020).

Porém, como o propósito de um sistema de recuperação de informação é permitir que os usuários encontrem diversas informações disponíveis no banco de dados (BAEZA-YATES; RIBEIRO-NETO, 2013), há momentos em que será necessária uma consulta de palavras nos documentos. Nesta tese, optou-se por manter todas as palavras do documento, criando uma indexação que utiliza a localização e a frequência de palavras por documentos. Isso facilita a localização e ordenação dos resultados da busca, visando maximizar a satisfação do usuário em relação aos resultados e reduzir os esforços para encontrar informações. Assim, há duas formas de acesso aos dados: a primeira por navegação hierárquica e a segunda por consulta, baseada em palavras nos documentos, tendo como unidades de recuperação os documentos disponíveis no acervo.

A consulta baseada em palavras, conforme Baeza-Yates e Ribeiro-Neto (2013), é a mais empregada em sistemas de recuperação de informação quando estes são baseados em consultas textuais. Como resultado da consulta, deve-se obter um conjunto de documentos que possuam, ao menos, uma das palavras consultadas. Para definir qual documento deve ser exibido em primeiro lugar e qual em segundo, foi implementado no SPEdu o módulo de análise de frequência de palavras e a sua localização no documento. Dessa forma, os documentos que contém as palavras, são organizados de forma que sejam exibidos em primeiro lugar os documentos com a palavra de maior frequência, em diversas localizações do documento e, por último, os documentos com menor incidência da palavra.

Há ainda a busca conjuntiva, ou seja, dado um conjunto de palavras, o sistema deve trazer todos os documentos que possuam todas as palavras (BAEZA-YATES; RIBEIRO-NETO, 2013). Outro modelo adotado é a consulta booleana, a qual emprega combinações entre palavras chaves e os operadores “E” e “OU” para recuperar informações. Nesse contexto, caso o usuário deseje pesquisar dois itens, na condição de obrigatórios, emprega-se o “E”. Já para o caso de uma busca por itens que podem ter proximidade, pode-se utilizar o “OU”. Conforme Baeza-Yates e Ribeiro-Neto (2013) a consulta que utiliza o “OU” traz todos os documentos que têm a primeira ou a segunda palavra buscada. Já no caso do “E”, deve-se selecionar todos os documentos que satisfaçam as duas palavras buscadas – primeira e segunda. Como o exemplo apresentado na Figura 35, a seguir, a consulta vai recuperar todos os documentos que possuam a palavra “educação”, assim como as palavras “técnica” ou “profissional”, ou ambas.

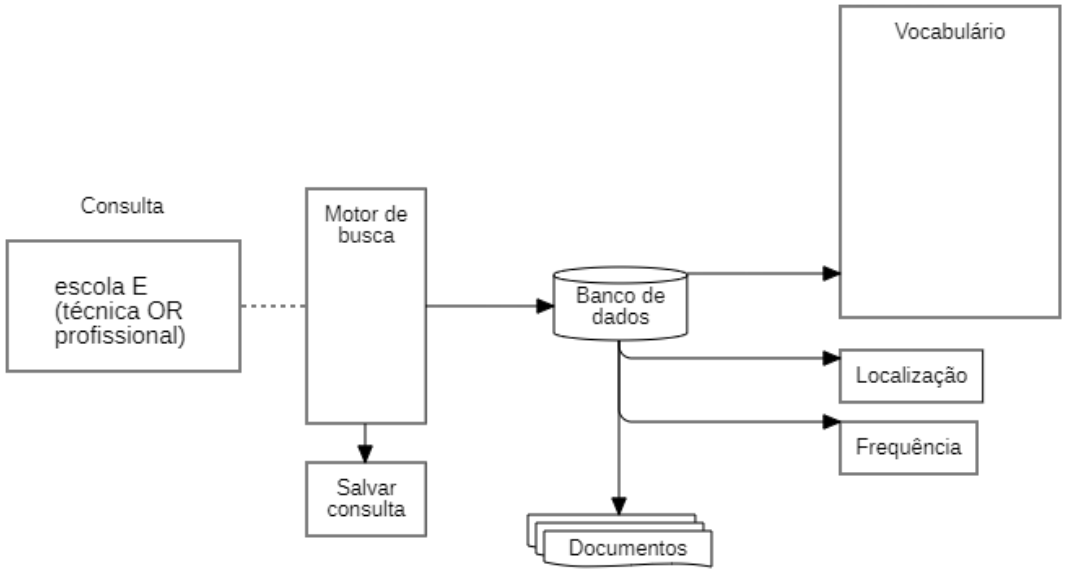
Figura 35 - Árvore de sintaxe de consulta



Fonte: Elaborado pelo autor, adaptado em Baeza-Yates e Ribeiro-Neto (2013).

O motor de busca é responsável por processar a consulta, enviá-la ao banco de dados para analisar os vocabulários disponíveis do acervo e, por fim, localizar os documentos que satisfaçam a consulta. Esse processo é exposto no exemplo da Figura 36, a seguir.

Figura 36 - Lógica do motor de busca



Fonte: Elaborada pelo autor (2020).

Com base na Figura 36, o motor de busca do SPEdu é o responsável por transformar a consulta em uma busca funcional, analisando o vocabulário, a sua localização e frequência nos documentos, retornando como resultado os documentos que atendem a consulta que é “escola” mais a palavra “técnica” ou “profissional”. Os resultados, em tempo de execução e

salvos em *cache* do sistema, são, então, analisados pelo pesquisador, que pode selecionar os documentos para gerar uma evidência de pesquisa. Além disso, o motor de busca salva a consulta, permitindo a utilização dela em buscas posteriores, ou ainda, em casos de evidências, a comparação com novos resultados. Isso contribui com o pesquisador, pois ele poderá consultar um número maior de documentos e salvar a busca como evidência para que possa ser validada ou revisitada por outros pesquisadores. Isso atende à necessidade de ampliação e diversificação de fontes na pesquisa por meio da tecnologia, como apontado por Bonato (2004).

Outro ponto a se destacar é que, além de preservar os documentos, o SPEdu permite que o pesquisador encontre informações que, em busca manual, dificilmente seriam obtidas. Como exemplo tem-se casos em que o pesquisador não possui a data correta em que uma escola sergipana foi criada, buscando informações no Diário Oficial do Estado. Se essa busca ocorrer de forma manual, demandará várias visitas à editora do respectivo jornal para encontrar a informação desejada. Cabe ressaltar que o jornal tem publicação diária, com exceção a domingos e feriados. Se considerado apenas o ano de 1997, o investigador precisará manipular um total de 1.636 páginas para leitura. Se a escola ter sido criada no início do ano, o pesquisador tem um problema menor. Mas, no caso da escola procurada ter sido criada no mês de dezembro o trabalho de busca se tornará árduo. Já com o uso do SPEdu, a busca será facilitada, uma vez que o pesquisador encontrará apenas o que precisa, podendo ainda, salvar a informação, para que seja uma evidência dos resultados.

Em outro exemplo, o pesquisador pode desejar encontrar, por meio dos atos oficiais do Governo, publicados no Diário Oficial, todas as escolas públicas estaduais abertas no período de 1997 a 2007. O número de documentos relativos a tal período totaliza 20.157 páginas, dificultando a busca ou, ainda, favorecendo o esquecimento de alguns documentos nos resultados. Outra limitação nesse exemplo é que, tratando-se apenas de documentos físicos, dificilmente o pesquisador terá como disponibilizar a evidência de suas pesquisas. Uma possibilidade para isso, ainda que limitada, é que ele necessite disponibilizar um CD/DVD ou um site com as fotos, como já vem ocorrendo no campo da História da Educação. Porém, dependendo do volume de dados, para que o leitor encontre determinada informação precisará acessar muitas imagens.

A disponibilização dos dados da pesquisa é necessária, como aponta a Fapesp (2017), principalmente, quando há investimento público. Isso permitirá que outros pesquisadores acessem os materiais e possam ampliar a pesquisa. Investigações como as de Nascimento (2008), em que os documentos foram disponibilizados em CD, as fontes empregadas

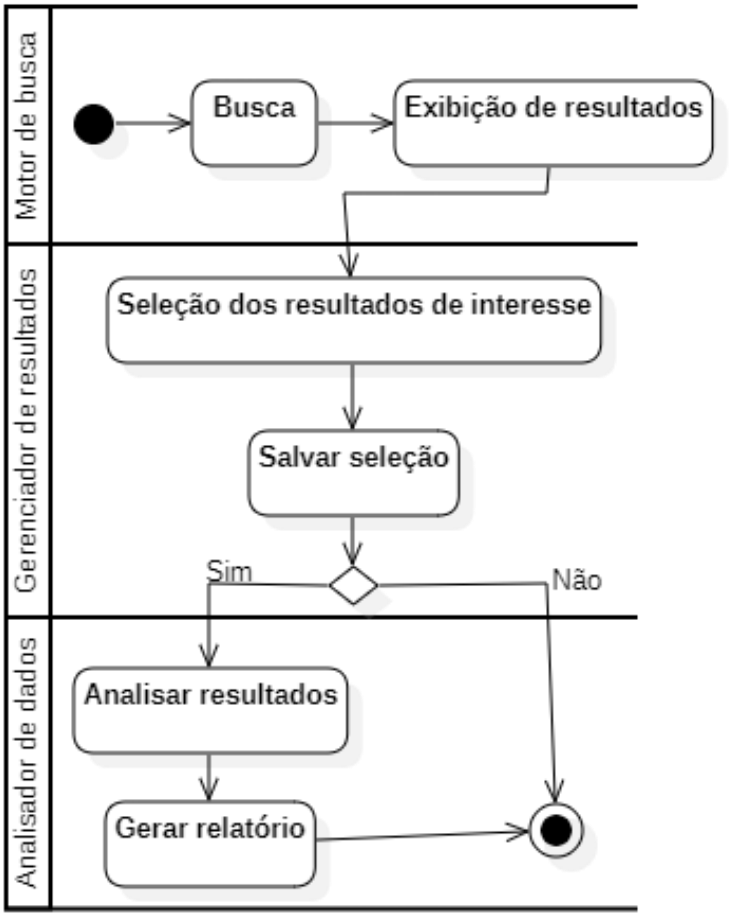
oferecem um *corpus* documental importante. Porém os pesquisadores precisam, ainda, acessar cada arquivo para efetuar a leitura e verificar o que cada um contém. No modelo proposto, com o SPEdu, além de ser possível compartilhar a evidência, é possível dar andamento a pesquisa, inserindo novos documentos e criando, assim, novas evidências. Outro ponto importante é que o SPEdu automatiza o processo de análise. Não basta a um pesquisador encontrar determinado documento. É necessário, ainda, analisar o material, como por exemplo, examinar um diário de classe identificando-se os alunos de uma turma, quem foi o professor, etc. Ou seja, é necessário explorar os dados encontrados na pesquisa, ação que pode ser otimizada com a adoção do SPEdu.

3.3 EXPLORANDO RESULTADOS E GERANDO EVIDÊNCIAS HISTÓRICO-EDUCACIONAIS

A exploração dos resultados implica, em primeira instância, a seleção dos documentos a serem analisados, filtrados diante do conjunto de materiais encontrados pelo pesquisador (BACELAR, 2018). Essa seleção de ser, então, salva para subsidiar tanto a investigação em curso, como, também, servir de evidências da pesquisa para outros investigadores. Nesta tese, foi adotado um modelo de armazenamento de resultados para replicação da pesquisa no SPEdu.

Parte da análise deve ser realizada pelo pesquisador, por ações como entender o contexto e o significado das palavras e expressões do documento (BACELAR, 2018). Porém, outros tipos de análise podem ser realizadas de forma automatizada com o uso da análise de conteúdo. Segundo Bardim (2016), essa última é um conjunto de recursos metodológicos que se aplicam a conteúdos diversificados, empregando técnicas como, por exemplo, análise de frequência de palavras. Assim, o ciclo de busca, seleção de resultados e análise do SPEdu é apresentado na Figura 37, a seguir.

Figura 37 - Ciclo de busca, geração de evidências e análise de resultados



Fonte: Elaborada pelo autor (2020).

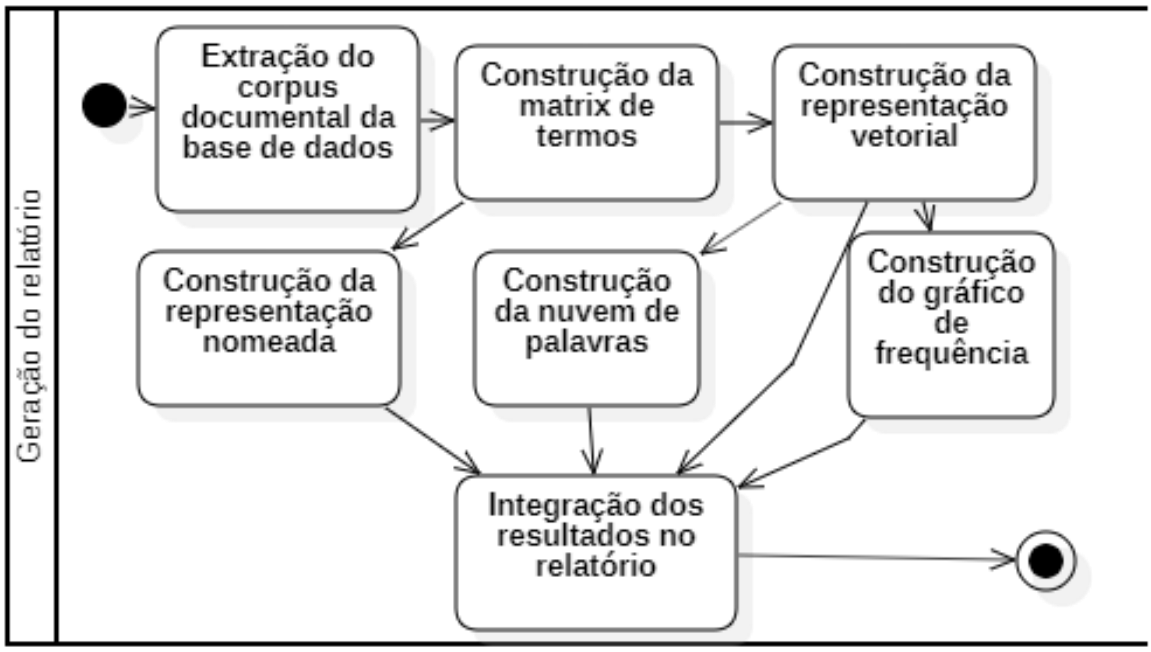
Com base na Figura 37, tem-se, então, duas formas de gerar evidências: uma durante a etapa de gerenciamento de resultados, que seria uma evidência bruta, ou seja, apenas o conjunto de documentos selecionados para compor o corpo documental da pesquisa, e outra após a análise dos documentos, que é o relatório com os dados de pesquisa somados às análises de conteúdos geradas pelo processamento de linguagem natural.

Dessa forma, para a análise de conteúdo, após a seleção dos documentos, deve realizada a análise de frequência, gerando uma tabela Documento – Palavra – Frequência. Com base na análise de frequência, é gerada uma nuvem de palavras por documento e pelo conjunto documental, permitindo uma análise sobre o conteúdo das mensagens disponíveis nos diversos documentos selecionados. Com base nos resultados, o pesquisador pode, conforme aponta Bardin (2016), comparar mensagens dos documentos. Por exemplo, o investigador pode observar a quantidade de escolas nos documentos de um determinado

período em relação a outro período. O pesquisador realizará a análise, porém o SPEdu apoiará esse processo com os resultados da análise textual dos documentos. Assim, o sistema permite extrair dados de forma a gerar diversas pistas que o pesquisador pode, então, analisar de forma a obter uma realidade sobre os as fontes estudadas, indo ao encontro do método indiciário (GINSBURG, 1989).

Além disso, o SPEdu traz anotações nas palavras de documentos, permitindo que o pesquisador identifique ocorrências no vocabulário de adjetivos, verbos, substantivos etc. Isso é feito por meio da análise léxica e sintática do documento, empregando processamento de linguagem natural. Dessa forma, é possível realizar uma análise não apenas da frequência das palavras, mas da frequência de substantivos, adjetivos, verbos, etc. Como resultado, obtém-se um relatório que possui um sumário de frequência, seguido da frequência das palavras e os seus tipos (adjetivos, verbos, substantivos etc.). Por fim, são gerados dois gráficos: a nuvem de palavras e o gráfico de barras. A figura 38, a seguir, apresenta a ordem da geração do relatório de análise.

Figura 38 - Ciclo para geração de relatório de análise de dados documentais



Fonte: Elaborada pelo autor (2020).

Com a tendência de dados abertos - Open Data, o relatório é uma evidência de pesquisa que pode ser distribuída, revisada e replicada, indo ao encontro da necessidade apontada por diversos órgãos de fomento à pesquisa. A FAPESP, por exemplo, solicita aos pesquisadores bolsistas que tenham um plano de gestão dos dados (FAPESP, 2020). Um dos objetivos daquela instituição é o compartilhamento dos dados como forma de racionalizar recursos. Isso se deve ao fato de que diversas pesquisas no mesmo campo investem tempo e recursos financeiros para o levantamento dos mesmos dados. Nesse sentido, o SPEdu, auxilia no compartilhamento não só da massa de dados, como, também, dos resultados de uma determinada pesquisa e das evidências de análises geradas automaticamente.

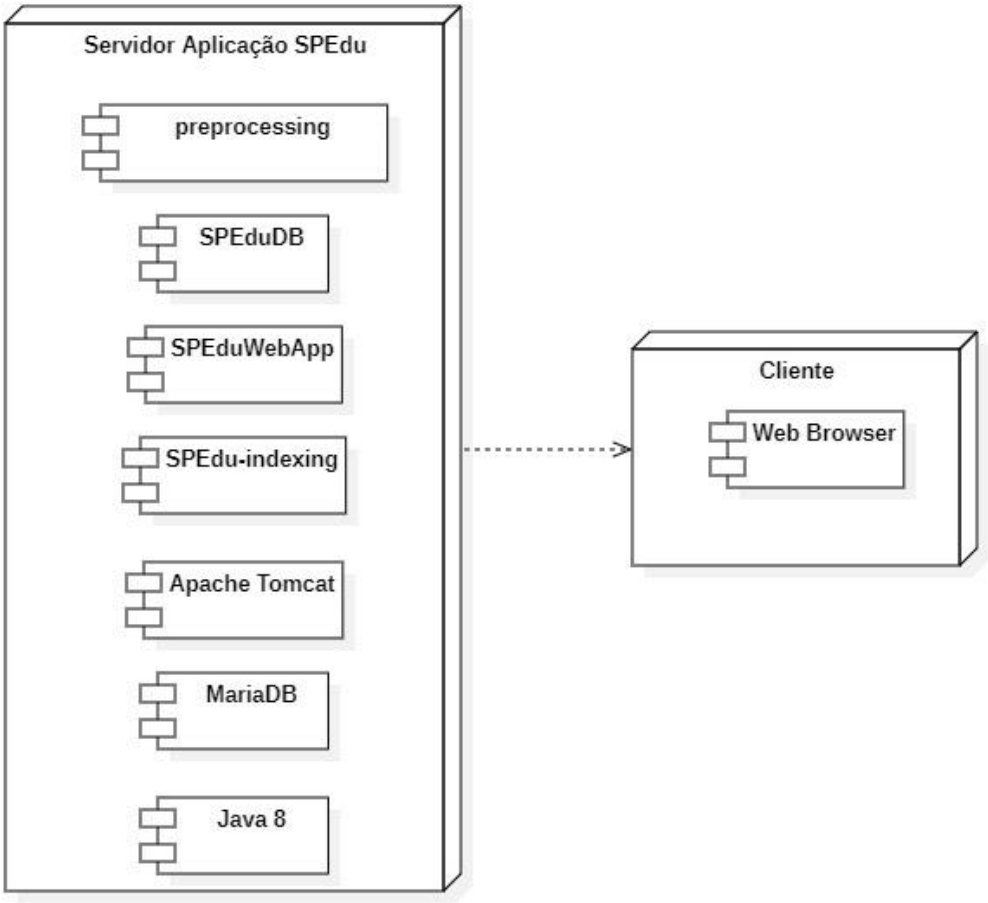
Na prática, o pesquisador, após colher diversas fontes documentais e incluir no sistema, realizar uma busca, selecionar os documentos que comporão sua evidência e solicitar ao sistema a geração da análise dos documentos, receberá como resultado um relatório com documentos, análises e gráficos. Ao salvar tal relatório, o pesquisador poderá compartilhar os seus resultados com a comunidade científica, garantindo, assim, a replicação e validação do estudo. Nos trabalhos encontrados, o comum é a preocupação do pesquisador com a preservação documental. Já com a utilização do SPEdu, os investigadores poderão avançar na gestão de seus dados, não se restringindo apenas a manter os materiais encontrados em formato digital.

Outra contribuição do instrumento SPEdu é que o seu usuário poderá, ainda, realizar uma análise exploratória dos documentos selecionados, já que estes podem ser vistos sob outros vieses como, marcas, manchas e anotações. Na próxima seção, apresenta-se detalhadamente o SPEdu, com o estudo de caso nos jornais do Diário Oficial de Sergipe.

4 SPEDU: INSTRUMENTO E ESTUDO DE CASO

Como resultado da adoção da metodologia *design science research*, desenvolveu-se o instrumento de software SPEdu, que integra os recursos apresentados nas seções 2 e 3 desta tese. O SPEdu é um sistema que pode ser utilizado por pesquisadores, bem como órgãos de documentação. Os pilares do SPEdu são: o sistema de extração automatizada de dados – módulo *preprocessing*, o sistema de indexação–SPEdu-indexing, e o SPEduWebApp, responsável pelo o motor de busca e pelo sistema de geração de evidências. Tais módulos são independentes, permitindo um reaproveitamento de parte do sistema. Um ponto importante é que o sistema foi desenvolvido sob o paradigma OS, permitindo que outros pesquisadores possam colaborar com o sistema, melhorando seus recursos, além de ser gratuito para uso por parte dos pesquisadores e instituições que necessitem gerir suas fontes. A visão da implementação da arquitetura do SPEdu é exposta na Figura 39, a seguir.

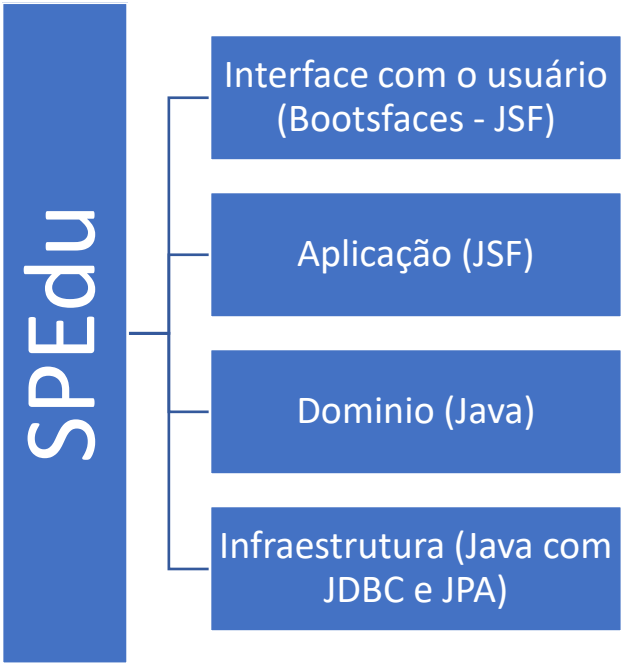
Figura 39 - Visão da implementação da arquitetura do SPEdu



Fonte: Elaborada pelo autor (2020).

O SPEdu foi desenvolvido segundo o padrão de camadas Domain Driven Design (DDD) (EVANS, 2017), dividido em quatro camadas lógicas. Tal separação possibilita que as partes do sistema sejam intercambiáveis, permitindo o reuso das camadas individualmente. A Figura 40, a seguir, demonstra a arquitetura das camadas da aplicação.

Figura 40 - Camadas do SPEdu, segundo o DDD



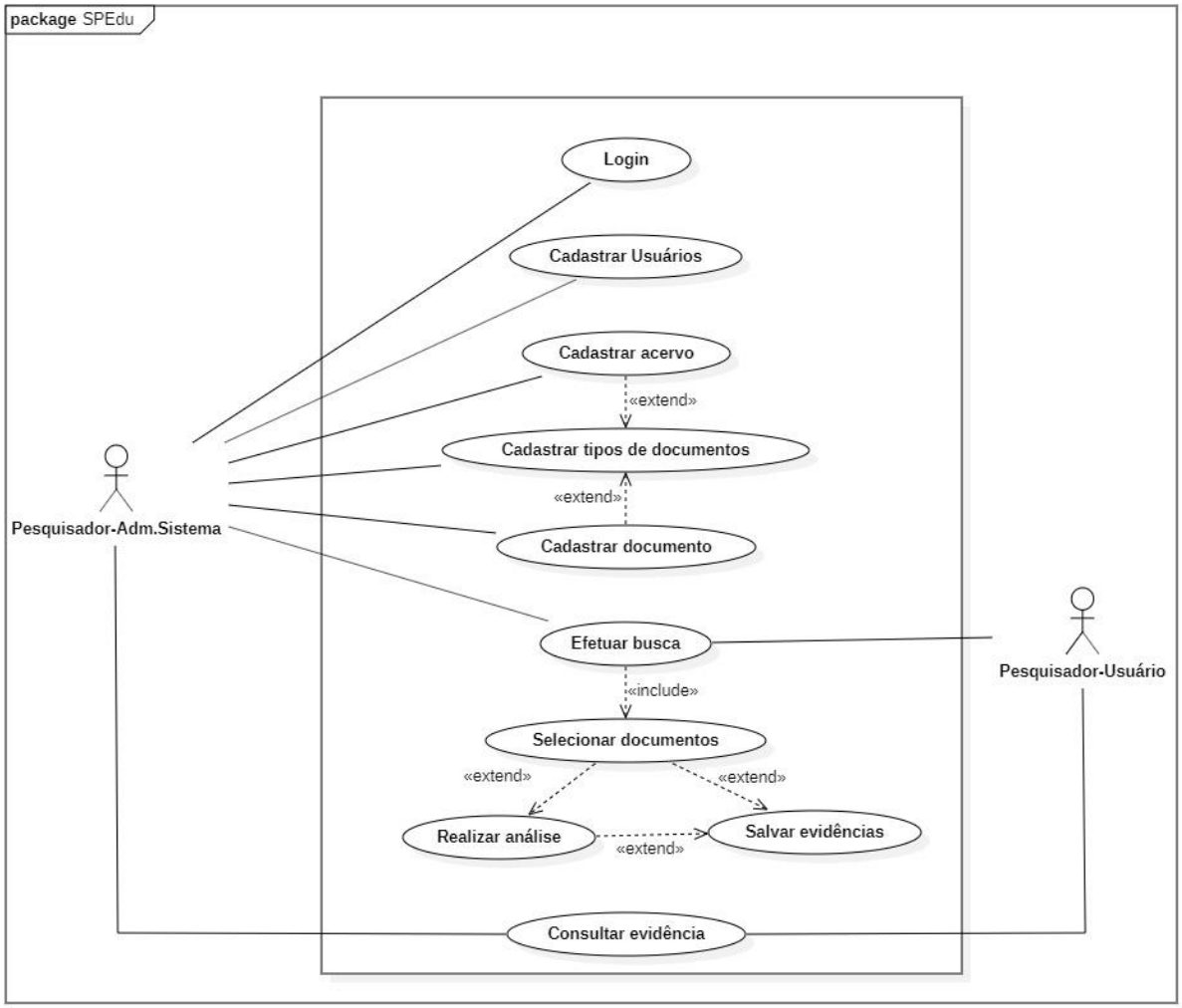
Fonte: Elaborada pelo autor (2020).

Para o desenvolvimento da interface com o usuário, que funciona em plataforma Web, foi utilizado o Java Server Faces (JSF), selecionado por ser um *framework* baseado em componentes (GEARY; HORSTMANN, 2012) e por possibilitar o uso de bibliotecas gráficas. Para esse fim, foi selecionada a biblioteca Bootsfaces que, segundo os criadores, é uma estrutura leve e poderosa, baseada na Bootstrap3 e na jQuery UI. Essas bibliotecas de componentes Java Script permitem o desenvolvimento de interfaces com o usuário, de forma fácil e rápida, além de ser OS (BOOTSFACES, 2020). Esses recursos possibilitaram um desenvolvimento ágil, multiplataforma, de fácil manutenção, de forma aberta (OS) e com uma interface responsiva, ou seja, permitindo que a aplicação seja acessada por computador ou celulares, se adaptando a cada um dos ambientes. Além disso, para melhorar o desempenho das consultas e manter os resultados de busca para novas consultas, foi adotado

sistema de cache EhCache. Essa é uma biblioteca de cache OS, com capacidade de escalabilidade e uma boa performance. Um sistema de cache salva os dados comuns em memória, proporcionando um melhor desempenho para acesso aos dados.

As funcionalidades do sistema são apresentadas no diagrama de caso de uso, na Figura 41, onde é possível identificar os tipos de usuários que o sistema possui, bem como as funcionalidades disponíveis para cada tipo de usuário.

Figura 41 - Diagrama de caso de uso do SPEdu



Fonte: Elaborada pelo autor (2020).

A camada de negócio define nove classes, sendo elas:

a) Acervo: responsável por gerenciar o acervo de documentos;

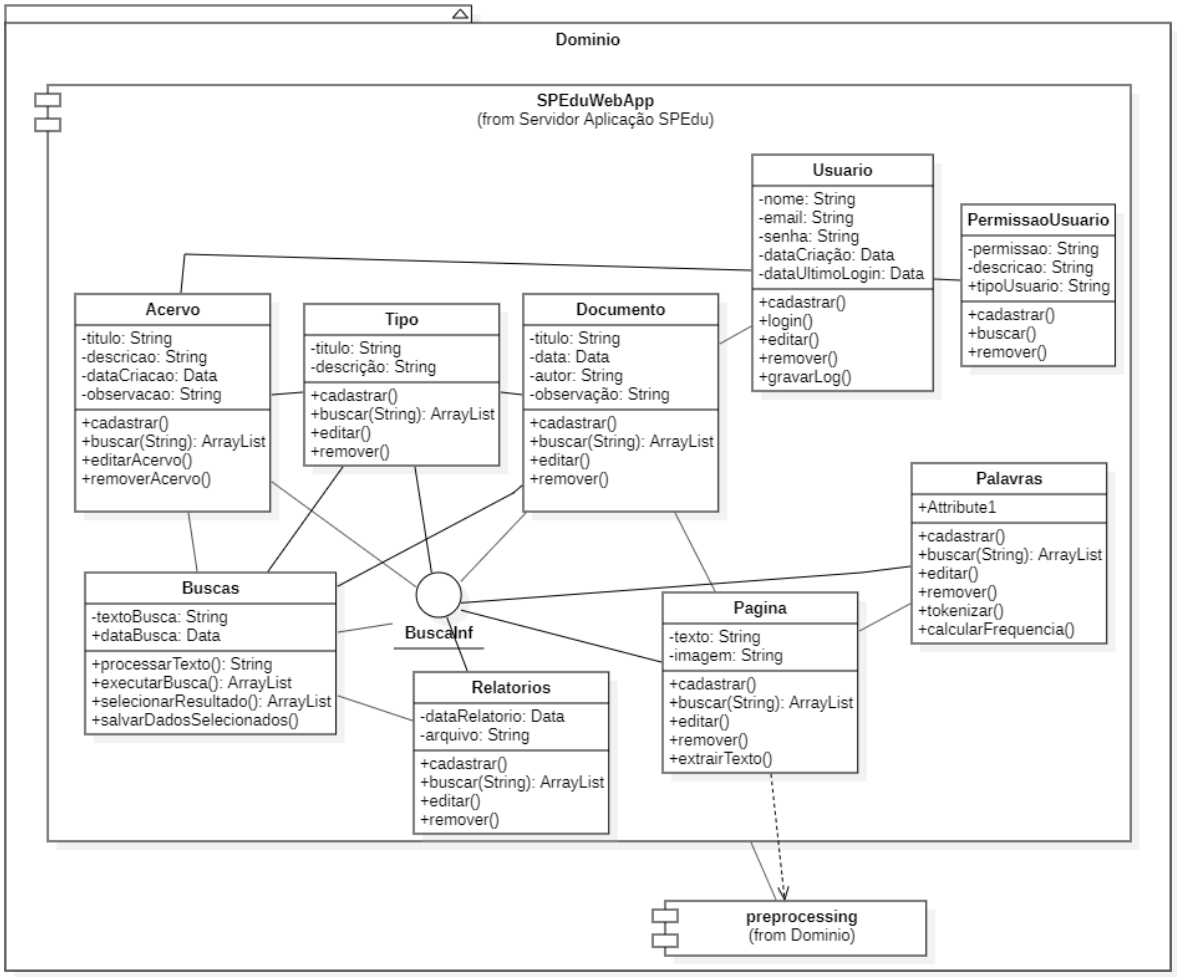
b) Tipo: responsável pela taxonomia de tipos de documentos do acervo;

c) Documentos: responsável pelo próprio documento, com informações relacionadas a autor, data de publicação, etc.;

d) Página: um documento pode possuir diversas páginas. Essa classe é responsável por cadastrar a página, extraindo o texto automaticamente e salvando no banco de dados;
e) Palavras: uma página possui diversas palavras. Essa classe é responsável por processar as páginas de um documento, extraindo as palavras (*tokens*), efetuando a limpeza e contabilizando a frequência das palavras, para, então, salvar no banco de dados;
f) Busca: responsável por processar o texto de busca do usuário, realizando as pesquisas no banco de dado. É também responsável por salvar as pesquisas para buscas futuras e pela seleção de documentos para o relatório;
g) Relatórios: responsável por gerar e salvar o relatório;
h) Usuário: responsável por gerenciar os usuários e permitir o acesso ao sistema;
i) Permissão: responsável por gerenciar as permissões de usuários e controle de acesso às informações do sistema.

Além disso, há uma interface comum, denominada BuscaInf, que obriga que as classes tenham funcionalidades similares, possibilitando a adoção de polimorfismo na busca. Isso que permite buscar acervos, tipos, documentos, páginas e palavras. Há, ainda, uma ligação com a biblioteca *preprocessing*, criada separadamente para permitir seu uso em outros aplicativos. As classes são representadas na Figura 42, a seguir.

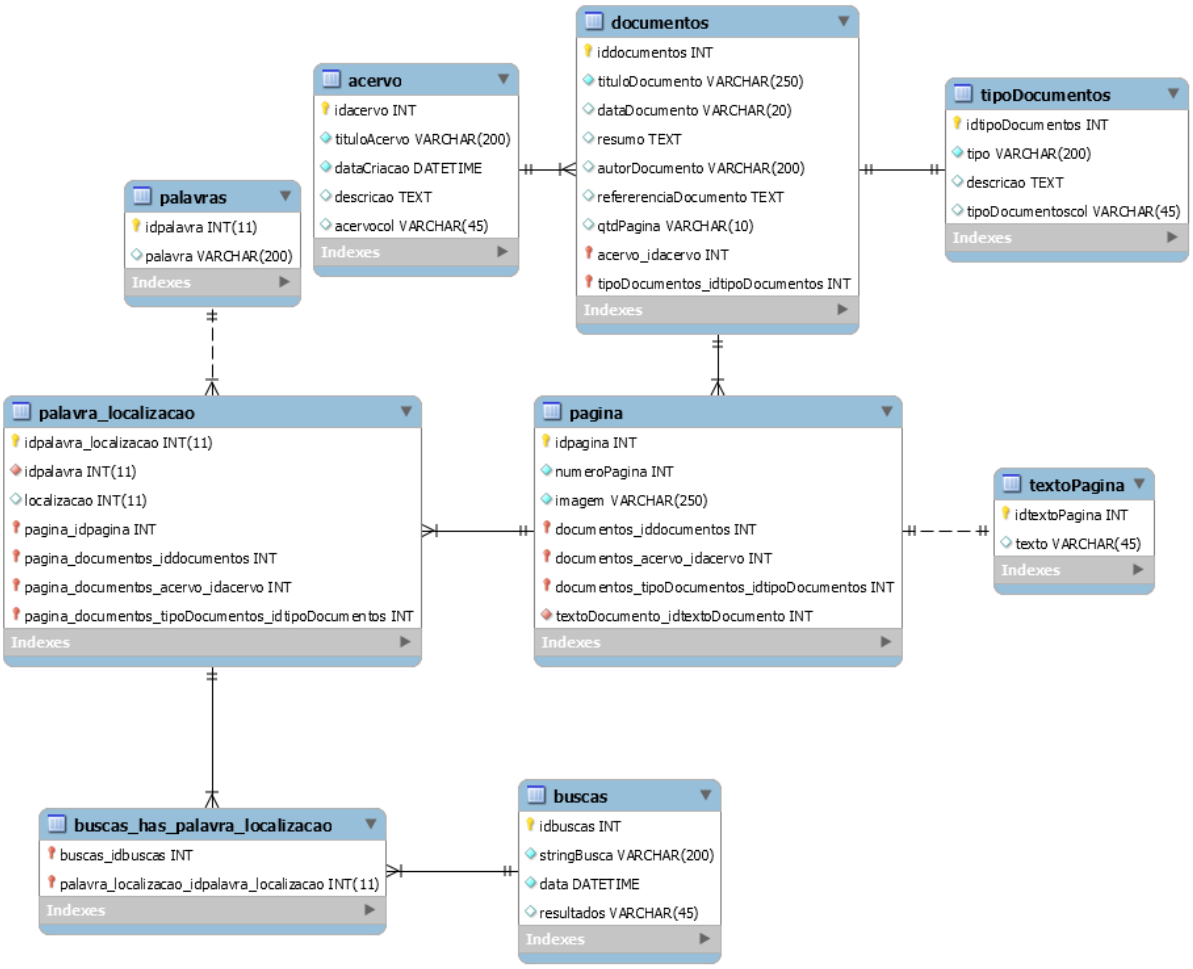
Figura 42 - Diagrama de classes do SPEdu



Fonte: Elaborada pelo autor (2020).

Para que os dados fossem salvos (persistência de dados), foi utilizado o sistema de gerenciamento de banco de dados MariaDB, selecionado por ser OS e possuir os recursos necessário para o SPEdu. O modelo do banco de dados foi criado empregando a ferramenta MySQL Workbench, conforme Figura 43, permitindo o armazenamento dos dados.

Figura 43 - Diagrama de banco de dados do SPEdu



Fonte: Elaborada pelo autor (2020)

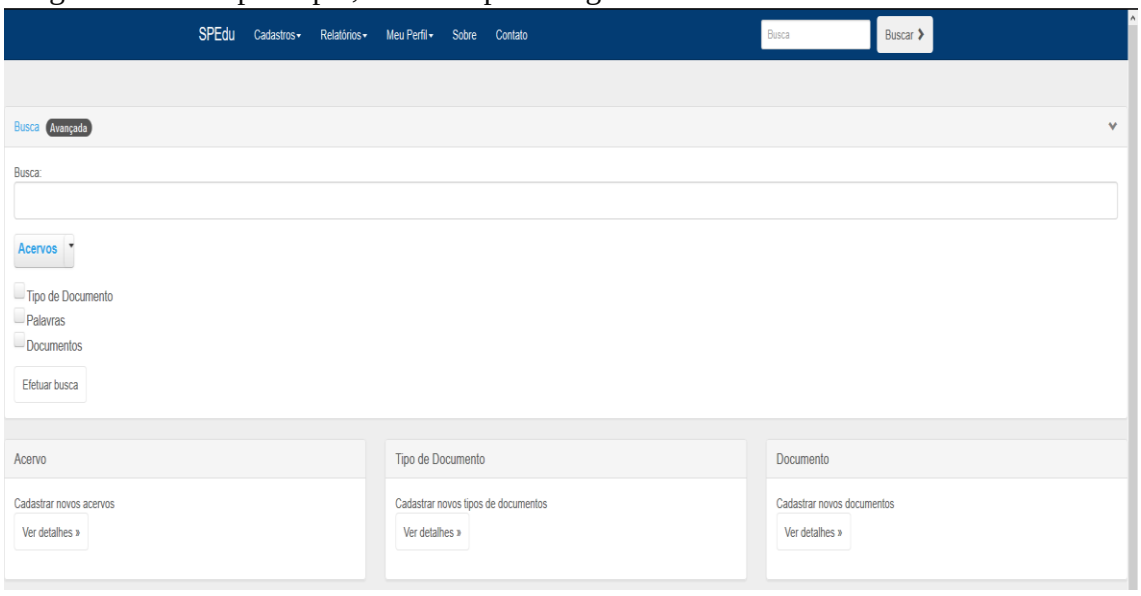
O núcleo do sistema, comportando a camada de negócios e a interface, foi desenvolvida na linguagem de programação Java, versão 8, que é multiplataforma e OS. Ela foi criada em 1995, pela SUN Microsystem, sendo a mais utilizada do mundo (TIOBE, 2020). Já a interface com o usuário, foi desenvolvida com JSF, empregando a biblioteca Bootsfaces, com uso de *template* de padronização de interface desenvolvida pelo projeto. A tela principal tem apenas informações do *login* e sistema (Figura 44).

Figura 44 - Tela inicial do sistema SPEdu

Fonte: Elaborada pelo autor (2020).

Ao acessar o sistema, o usuário pode realizar as ações descritas no diagrama de caso de uso (Figura 40). O sistema permite que sejam inseridos documentos apenas após o usuário criar um acervo, tendo no mínimo, um tipo de documento. A tela principal é apresentada na Figura 45, a seguir.

Figura 45 - Tela principal, exibida após o login do usuário

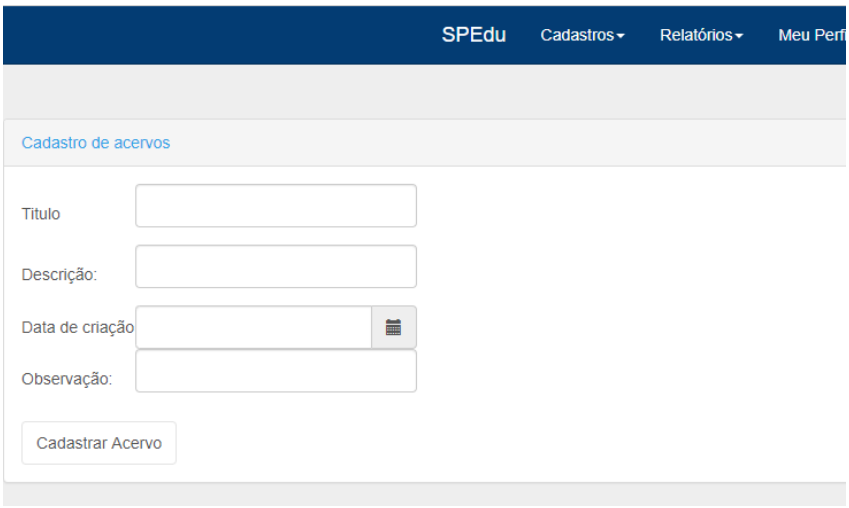


Fonte: Elaborada pelo autor (2020).

Após o usuário efetuar o *login* no sistema, ele poderá executar as diversas tarefas apresentadas no diagrama de caso de uso (Figura 41). Porém, se for o seu primeiro acesso, o usuário deverá criar acervos cujos documentos estejam agrupados com alguma relação entre eles. Há diversas pesquisas no campo de História da Educação que exigem análises

sobre acervos distintos. Como exemplo tem-se o trabalho de Sabino (2017) que examinou documentos de seis escolas e duas universidades. Nesse caso, seria necessária criar um acervo para cada instituição. A Figura 46, a seguir, apresenta a tela de cadastro de acervos.

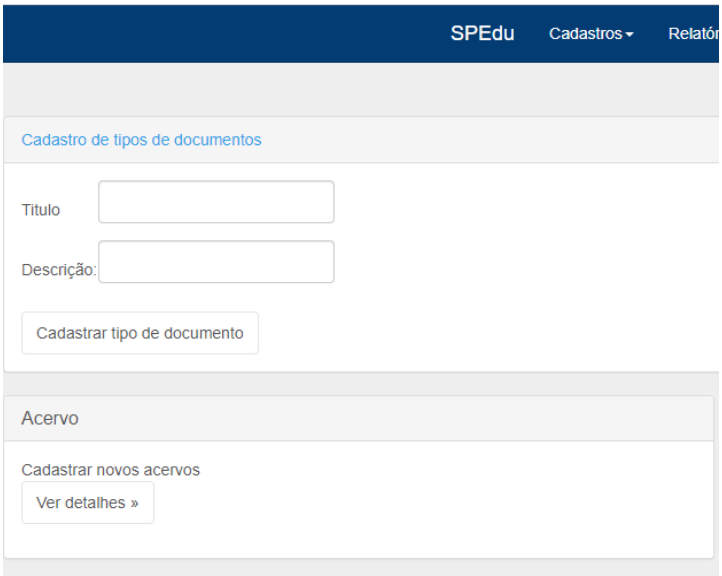
Figura 46 - Cadastro de acervos de pesquisa



Fonte: Elaborada pelo autor (2020).

Outro ponto importante é que cada pesquisa contém tipos distintos de documentos, podendo ter cartas, livros, diários, jornais, etc. Assim, se faz necessário que o pesquisador cadastre, manualmente, os tipos de documentos que sua pesquisa conterá. A Figura 47, a seguir, apresenta a tela de cadastros de tipo documental, que servirá de taxonomia durante o cadastro de documentos.

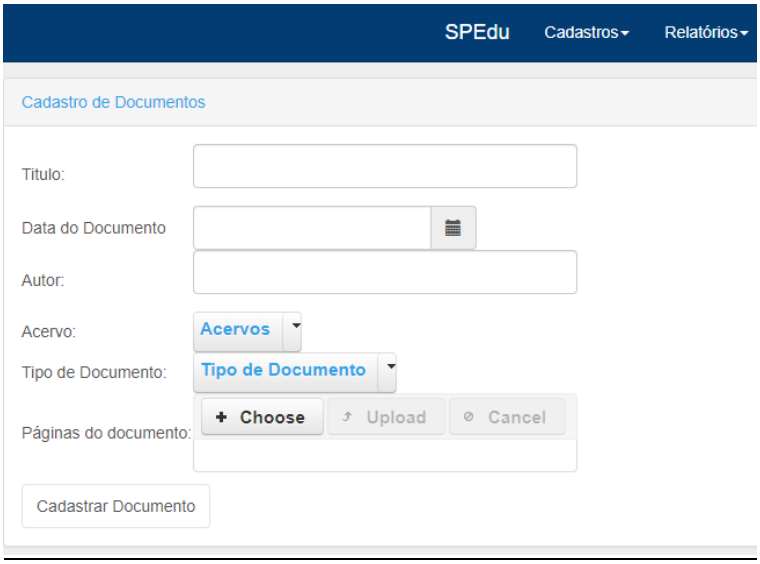
Figura 47 - Cadastro de tipos de documentos



Fonte: Elaborada pelo autor (2020).

Com o acervo e os tipos de documentos cadastrados, o usuário do SPEdu pode, então, iniciar o cadastro dos documentos. O sistema fará, automaticamente, a extração do texto e o processamento da informação para o usuário. Nesse contexto, para o processo de extração automatizada, conforme apresentado na seção 2 desta tese, foi adotado o Tesseract como motor de OCR. A tela de cadastro de documentos permite que o pesquisador insira o título do documento, data de criação do documento, autor (es) do documento, selecione o acervo o qual o documento faz parte, o tipo de documento, permitindo, ainda, a inserção das páginas do documento (uma ou mais páginas) (Figura 48). Tais dados vão permitir que o pesquisador tenha uma referência sobre o documento.

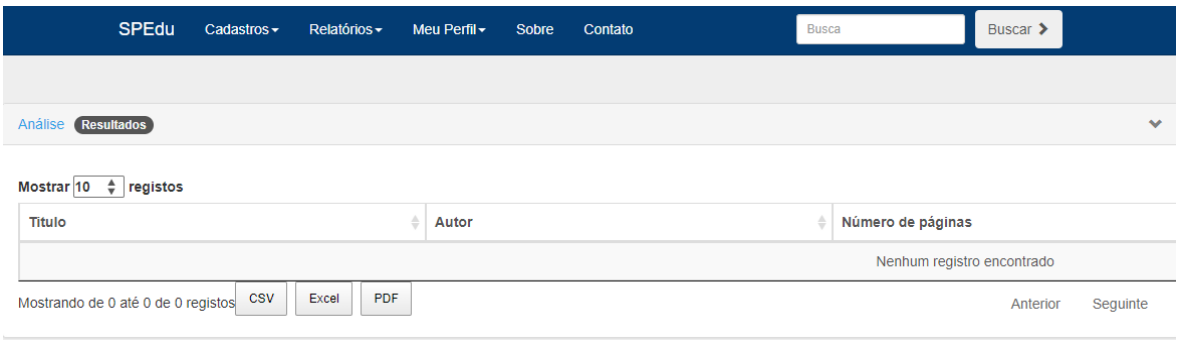
Figura 48 - Cadastro de documentos



Fonte: Elaborada pelo autor (2020)

O usuário, ao realizar a busca, receberá os resultados exibidos em um quadro, o que permitirá a seleção dos documentos que irão compor a evidência de sua pesquisa. A tela de resultados é apresentada na Figura 49, a seguir. Os resultados podem, ainda, ser exportados diretamente em formato PDF, CSV e Excel, e explorados em novas buscas.

Figura 49 - Resultados de pesquisa do SPEdu



Fonte: Elaborada pelo autor (2020).

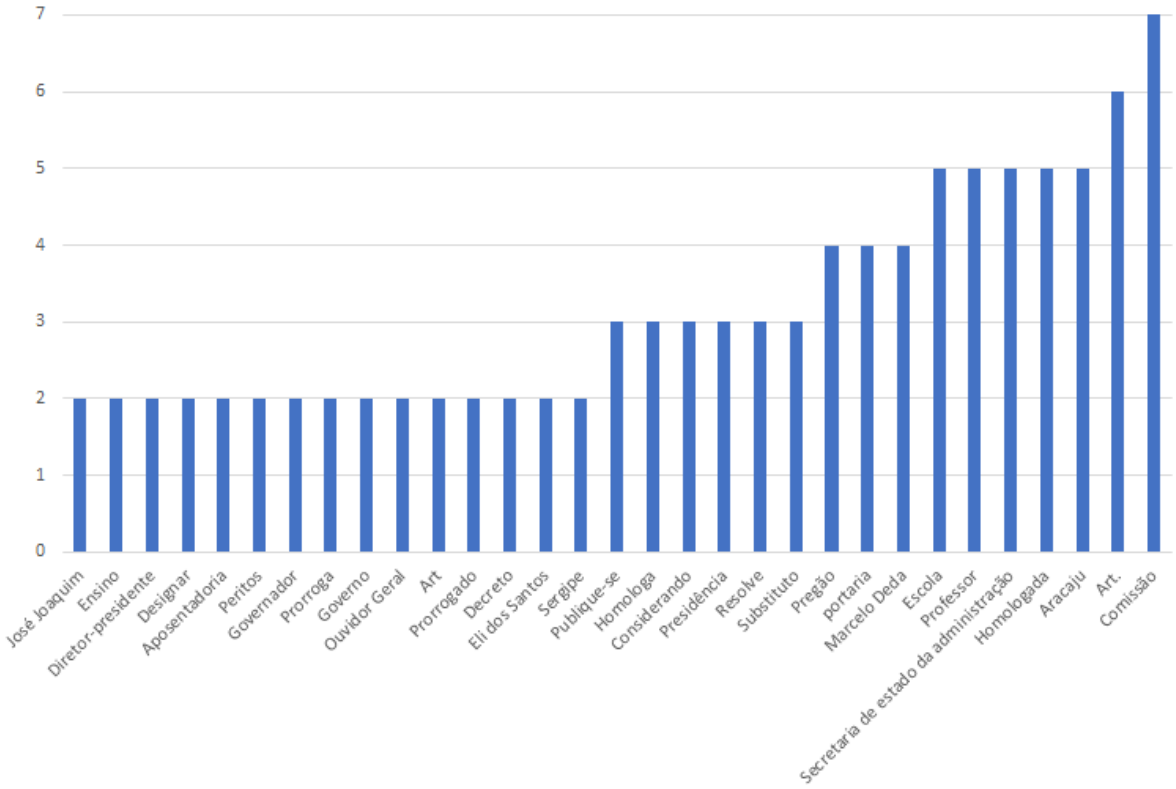
Todas as telas do SPEdu seguem um padrão, buscando facilitar o entendimento e o acesso às informações. Isso se deve à capacidade do JSF em suportar *templates*, que é um modelo que todas as telas seguem. Por fim, com os dados salvos, é possível que o pesquisador solicite análise dos resultados. Entre os resultados de análise de dados, tem-se o gráfico de nuvem de palavras. Esse tipo de gráfico vem sendo adotado em pesquisas para analisar a frequência de palavras. A Figura 50, a seguir, representa o gráfico do SPEdu gerado com base na análise do Diário Oficial de Sergipe do dia 1º de fevereiro de 2011.

Figura 50 - Nuvem de palavras geradas pelo SPEdu

Fonte: Elaborada pelo autor (2020).

O sistema gera, ainda, o gráfico de barras para palavras que tenham frequência maior ou igual a dois. A frequência a ser exibida pode ser configurada para mostrar valores mínimos ou máximos, conforme exposto no Gráfico 2, a seguir.

Gráfico 2 - Gráfico de barras de frequência de palavras



Fonte: Elaborado pelo autor (2020).

Por fim, na análise, o sistema extrai as informações para que o pesquisador possa realizar suas análises de forma sistemática. Isso ocorre pela geração de agrupamentos possíveis, bem como uma tabela de frequência de palavras e uma tabela de palavras e seus tipos (adjetivo, substantivo, verbo etc.) com foco na análise de conteúdo. O Quadro 4, a seguir, apresenta as palavras com frequência maior ou igual a quatro, totalizando nove palavras etiquetadas para análise de conteúdo.

Quadro 4 - Etiquetamento de palavras para análise de conteúdo

Palavra	Etiqueta
pregão	Nome Própio Pessoa
portaria	Nome Comun Feminino Singular
marcelo deda	Nome Própio Pessoa
escola	Nome Própio Organização
professor	Nome Própio Pessoa
secretaria	Nome Própio Organização
homologada	Nome Própio Pessoa
aracaju	Nome Própio Lugar
comissão	Nome Própio Organização

Fonte: Elaborado pelo autor (2020)

Tais resultados podem ser salvos como relatórios de pesquisa. O sistema permite um número ilimitado de acervos, tipos de documentos e documentos, ficando limitado apenas aos recursos físicos do usuário no momento da implementação do sistema, como espaço em disco e memória disponível em seu equipamento. Na próxima subseção apresenta-se o estudo de caso realizado nesta tese, abordando o seu planejamento e a sua execução.

4.1 ESTUDO DE CASO

Esta subseção apresenta o estudo de caso nos Diários Oficiais do Estado de Sergipe, abordando-se o seu planejamento, descrição do local de realização, preparação do estudo e, por fim, a sua execução. O estudo foi realizado em duas etapas: a primeira com o SPEdu original e, a segunda, com uma versão modificada desse instrumento, visando servir de sistema de gestão documental da SEGRASE, responsável pelo citado jornal.

4.1.1 PLANEJAMENTO

Esta subseção apresenta o protocolo empregado para o estudo de caso, adotado como parte da metodologia *design science research* para validar o instrumento construído durante a pesquisa. O objetivo do estudo de caso foi formalizado usando parte do modelo GQM (WHOLIN et al. 2012), a saber: analisar o SPEdu, com a finalidade de avaliação, em relação à funcionalidade, do ponto de vista da pesquisa, no contexto da Segrase.

As questões adotadas para o estudo de caso foram:

Q1. O sistema permite extrair automaticamente os dados, reduzindo os problemas relacionados à transcrição manual de documentos pelo pesquisador?

Q2. O sistema permite a indexação dos documentos?

Q3. Dado um conjunto de dados de busca, o sistema consegue obter os resultados de forma satisfatória?

Q4. O sistema possibilita a replicação da busca?

Q5. O sistema realiza análise documental, gerando dados relacionados à frequência de palavras, gráficos de frequência e de nuvem de palavras?

Q6. O sistema permite salvar o relatório, gerando evidências de pesquisa?

4.1.2 DESCRIÇÃO DO LOCAL

O Diário Oficial do Estado de Sergipe, teve origem na aquisição de equipamento de tipografia, autorizada pelo, então, Presidente do Estado, Manuel Prisciliano de Oliveira Valladão. Essa compra foi autorizada pela Lei n° 106, de 05 de dezembro de 1894. Em 24 de agosto de 1985, por meio do Decreto de Lei n° 141, foi criada a Imprensa Oficial do Estado de Sergipe (SILVA, 2018), responsável pela publicação dos atos oficiais do governo estadual. Atualmente, denominada Serviços Gráficos do Estado de Sergipe (SEGRASE), a empresa ainda se configura como a gráfica oficial do estado. A SEGRASE está localizada na Rua Propriá, número 227, centro de Aracaju/Sergipe. A gráfica tem como principal função produzir o Diário Oficial do Estado, documento de publicação de licitações, decretos, portarias, nomeações, exonerações e outras informações que o estado deve disponibilizar de maneira pública. A partir de 2012, o Diário Oficial passou a ser digital, havendo, portanto, números impressos, no período de 1895 a 2011, disponíveis para uso no presente estudo.

A Hemeroteca da SEGRASE, local responsável por armazenar os diários físicos, está localizada no segundo andar da SEGRASE, contando com dois funcionários, responsáveis por manter e realizar as pesquisas de números específicos no acervo de 116 anos de jornais impressos. O acervo conta com armários deslizantes para o acondicionamento dos jornais (Figura 51), sendo que cada grupo de jornais está organizado em caixas ou envelopado em papel especial para proteger do desgaste (Figura 52). Para os interessados, há uma sala de apoio para leitura e pesquisa no acervo (Figura 53).

Figura 51 - Armário deslizante da hemeroteca da SEGRASE

Fonte: Elaborada pelo autor (2020).

Figura 52 - Acervo de Diários Oficiais, envelopados para proteção ao desgaste

Fonte: Elaborada pelo autor (2020).

Figura 53 - Sala de pesquisa da hemeroteca da SEGRASE

Fonte: Elaborada pelo autor (2020)

Como é possível observar, a instituição conta com estrutura física e recursos humanos limitados para a manutenção de um grande acervo, possibilitando apenas pesquisas locais e de forma manual.

4.1.3 PREPARAÇÃO

A primeira parte do estudo de caso na SEGRASE iniciou por meio de um projeto, com foco em adaptar o SPEdu para ser o sistema da nova Hemeroteca Digital. O projeto envolve ainda alunos que realizam o trabalho de digitalização e inclusão no SPEdu adaptado para a instituição (SEGRASE, 2017). Para o desenvolvimento do projeto foram efetuadas reuniões presenciais com o presidente da SEGRASE, Sr. Ricardo José Roriz Silva Cruz, e com o responsável pela Hemeroteca, Sr. Wallace Douglas Nascimento dos Santos. Nesses encontros, identificou-se que, para o atendimento ao processo de gestão, o módulo de extração de informações deveria estar separado do SPEadu. Já para a parte pública, haveria apenas o sistema de busca e recuperação de informações.

A segunda parte do estudo de caso foi realizada com os mesmos dados, porém no sistema sem nenhuma adaptação, simulando uma busca por escolas, professores que se aposentaram, etc. No sistema adaptado, ao final da busca, foi solicitado que o usuário respondesse algumas perguntas, tendo como objetivo analisar se atendeu as necessidades.

Em agosto de 2017, para a validação dos modelos e do instrumento desta tese, foi estabelecida uma parceria com a SEGRASE (Figura 54), com foco na digitalização de todo o acervo documental. Esse material deveria estar disponibilizado, inicialmente, para uso interno e, posteriormente, para o público externo por meio da Internet. Assim, uma adaptação da interface do SPEdu foi desenvolvida, sendo separada a extração em um módulo externo. Isso viabiliza que os usuários externos apenas realizem buscas de informação, tendo sido removida a interface de geração de evidências.

Figura 54 - Reunião de parceria realizada com a SEGRASE, em agosto de 2017

Fonte: SEGRASE (2017).

Na subseção a seguir, apresenta-se a execução do estudo de caso junto à SEGRASE, detalhando as modificações de interfaces de usuários do SPEdu.

4.1.4 EXECUÇÃO DO ESTUDO DE CASO

Para atender as necessidades internas e externas da SEGRASE, uma adaptação do SPEdu foi realizada, com uma nova interface gráfica. Porém, as funcionalidades referentes a busca foram mantidas. A Figura 55 apresenta a tela principal de busca do SPEdu modificado. A demanda básica de busca era a pesquisa por data (período inicial e final), busca por palavras chaves ou CPF, e busca avançada que integra data e palavras chaves, disponíveis na tela de busca.

Figura 55 - SPEdu adaptado para a SEGRASE, na criação da Hemeroteca Digital

Fonte: Elaborada pelo autor (2020).

Como exemplo, em uma consulta pela palavra escola, apenas de 2011, obteve-se 715 resultados, nos quais são informadas a data e a página do jornal, como pode ser visto na Figura 56, a seguir.

Figura 56 - Tela com os resultados da busca pela palavra “escola” no ano de 2011

Fonte: Elaborada pelo autor (2020).

Ao clicar em visualizar, o pesquisador terá acesso ao documento que atendem aos critérios de busca, com opções de ampliação da imagem, conforme exposto Figura 57, a seguir.

Figura 57 - Tela de exibição do Diário Oficial, com opções de ampliação da imagem

Fonte: Elaborado pelo autor (2020).

Os dados utilizados nos testes, ou seja, na extração de dados e na indexação, são referentes aos Diários Oficiais digitalizados durante o projeto, totalizando 32.004 páginas, distribuídas entre os anos de 1997 a 2011, conforme Tabela 1, a seguir. Esses dados foram utilizados para medir o tempo de extração de dados, indexação e busca do sistema.

Tabela 1 - Total de páginas digitalizadas e processadas por ano

Ano	Página
2011	2.916
2010	3.150
2009	2.938
2008	2.843
2007	2.586
2006	2.320
2005	2.076
2004	1.836
2003	1.802
2002	1.808
2001	1.588
2000	1.395
1999	1.429
1998	1.681
1997	1.636
Total	32.004

Fonte: Elaborada pelo autor (2020).

Ao final da busca e do cadastro, o sistema, durante as primeiras 20 consultas, solicitou ao usuário (funcionários da Hemeroteca) uma validação sobre a facilidade de uso, sobre os resultados, buscando avaliar a qualidade sentida pelo usuário.

Já para a segunda parte do estudo de caso, com foco no teste da busca do SPEdu e na geração dos gráficos de análise, não disponíveis no sistema adaptado, optou-se por selecionar um grupo pequeno de jornais do ano de 2011. A intenção foi buscar informações relacionadas à educação, como concursos e aposentadoria de professores, validando se as informações obtidas. As análises sobre os resultados são apresentadas e discutidas na próxima seção.

5 RESULTADOS

Durante a análise da aplicação do SPEdu junto à SEGRASE, constatou-se que o instrumento permite a extração de dados em grandes volumes de documentos. Para órgãos de documentação, o sistema possibilita a criação de base de informação, podendo ser utilizado como ferramenta interna ou externa, para disponibilizar acervos por meio da Internet. Esse aspecto vai ao encontro da concepção de uma "biblioteca sem muros" (CHARTIER, 1999a). Já, para os pesquisadores, o SPEdu permite que a criação de acervos de pesquisa, auxiliando no contexto da gestão de dados, bem como no apoio à busca e geração de evidência de investigações. Dessa forma, foi possível responder as questões de estudo:

Q1. O sistema permite extrair automaticamente os dados, reduzindo os problemas relacionados a transcrição manual de documentos pelo pesquisador?

R.Q1. Sim, o sistema permite, tanto para usuários (pesquisadores) que tenham um volume reduzido de informações, como para centros de documentação, que possuam acervos maiores, requerendo um processo sólido e funcional. Assim, para a etapa de processamento automatizado da imagem e de extração de texto, foi observado como resultado que o processamento em escala possui tempo reduzido se comparado à extração de um único documento por vez. A extração de uma página de jornal atingiu 2 minutos e 53 segundos, e para indexar os dados 0,50 segundo. Assim, nas 32.004 páginas a extração do texto dos jornais levaram sessenta e quatro dias. Esse processo envolve extrair os documentos e inserir no banco de dados.

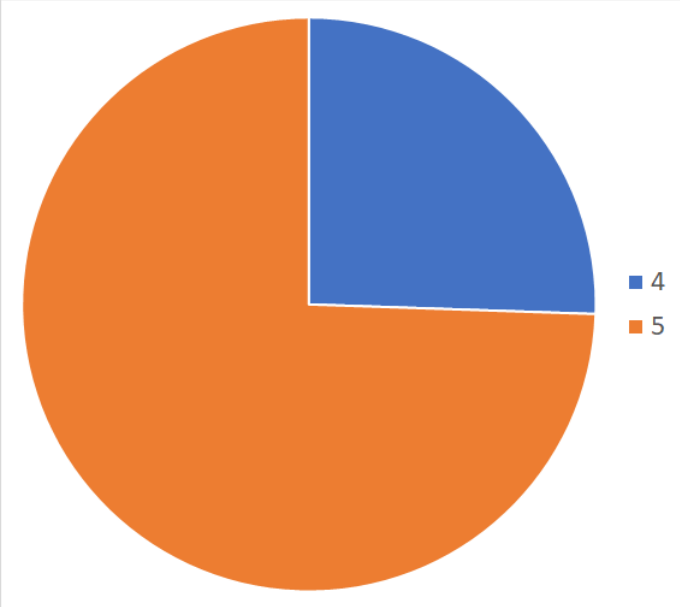
a) Q2. O sistema permite a indexação dos documentos?

RQ2. O sistema realiza a indexação de um ou vários documentos de forma automatizada, ou seja, sem a necessidade de intervenção do usuário. Nos testes realizados junto à SEGRASE o tempo aproximado de indexação de apenas uma página de jornal foi de 0,50 segundos. Mas o sistema permite que seja, ainda, indexado volume de documentos para acervo. Esse processo, para o conjunto de 32.004 páginas, demandou quatro dias e meio ou cento e seis horas de processamento ininterrupto. Nesse sentido, pode-se afirmar que o sistema permite a indexação de forma a atender pesquisadores e acervos documentais, com um desempenho satisfatório, principalmente, em grandes volumes de dados.

Q3. Dado um conjunto de dados de busca, o sistema consegue obter os resultados de forma satisfatória?

b) RQ3. Para a primeira parte do estudo de caso, o sistema aplicou um questionário, buscando validar se os resultados foram satisfatórios entre as primeiras vinte buscas, realizado pelos dois usuários da hemeroteca física da Segrase. A análise do SPEdu, no ponto de vista qualitativo, foi realizada com base na definição da ISO/IEC 25010 (2011). Essa norma menciona que qualidade é a capacidade de um sistema de ser compreendido, aprendido e utilizado quando em condições específicas. Nesse caso, as condições são de busca de informação. Assim, quando questionado se o sistema foi fácil de usar, com atribuição de notas entre 0 a 5, sendo 0 para difícil e 5 para muito fácil, obteve-se 74,47% das respostas com a nota 5 - muito fácil e 25,53% com nota 4 - fácil, conforme pode ser visto no Gráfico 3, a seguir.

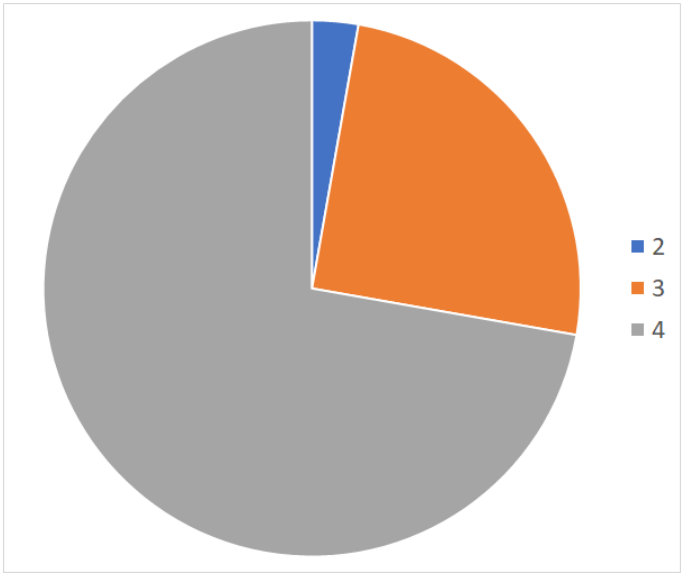
Gráfico 3 - O sistema foi fácil de utilizar?



Fonte: Elaborado pelo autor (2020).

Já quando questionado se o sistema atendeu às necessidades de pesquisa do usuário, uma vez que as pesquisas dentro da Hemeroteca podem ser realizadas por data, CPF (em caso de termos de posse, aposentadoria etc.) e por palavras, sob atribuição de notas entre 0 a 5, sendo 0 para “não atendeu” e 5 para “completamente satisfeito”, os resultados obtidos foram: 2,78% das buscas foram atendidas “parcialmente”, 25% delas foram atendidas como “parcialmente positivo” e 72,22%, sendo a maioria, foram atendidas como “completamente satisfeito”, conforme pode ser visto no Gráfico 4, a seguir

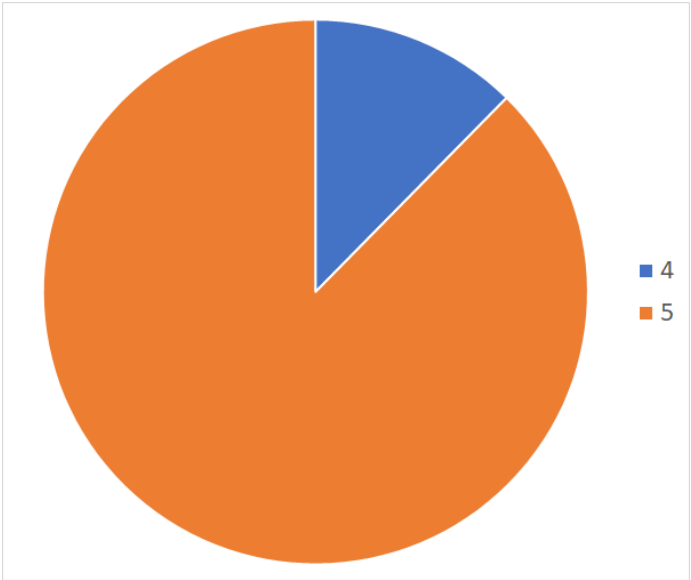
Gráfico 4 - O sistema atendeu às necessidades de pesquisa?



Fonte: Elaborado pelo autor (2020).

Quando questionados se o sistema atendeu às necessidades relacionadas à exibição dos resultados, com atribuição de notas entre 0 a 5, sendo que 0 para “não atendeu” e 5 para “atendeu completamente”, obteve-se como resultados: 12,37% com nota 4 e 87,63% com nota 5, conforme exposto no Gráfico 5, a seguir.

Gráfico 5 - O sistema atendeu às necessidades relacionadas à exibição dos resultados?



Fonte: Elaborado pelo autor (2020).

Quando questionado sobre se o usuário considera que o sistema auxilia e simplifica o processo de recuperação de informações da Hemeroteca, bem como se o sistema atendeu as expectativas, com atribuição de nota entre 0 e 5, sendo 0 para "não atendeu" e 5 "atendeu completamente", todas as 20 respostas das buscas, para ambas as perguntas, foram 5.

Já para os testes realizados como pesquisas no SPEdu, na segunda parte do estudo de caso, o sistema se mostrou relativamente rápido, demandando sete segundos para obter um volume de dados relativamente grande e ordenados. A ordenação por data e por importância facilita na análise de resultados, atendendo às necessidades do pesquisador do campo de História da Educação. Além disso, o sistema pode trazer todos os documentos de uma determinada data, acervo ou, ainda, todos os documentos de um determinado tipo, dando liberdade ao usuário para fazer suas pesquisas.

c) Q4. O sistema possibilita a replicação da busca?

RQ4. Há duas formas de replicar buscas: por meio da consulta salva e por meio do EhCache. No primeiro caso, as consultas salvas podem ser utilizadas com o uso do histórico de consultas. Além disso, um código para consulta é gerado, de forma que o usuário possa replicar as consultas. O usuário tem acesso a todas as suas buscas, permitindo que refaça ou compartilhe a busca com outros usuários. Já o EhCache permite uma busca mais rápida. Isso se dá a partir da segunda consulta semelhante, visto que tanto as consultas como os resultados são salvos em um *cache* de memória, que tem como função agilizar as buscas. Isso melhora a eficiência do sistema em um ambiente de consultas contínuas, tendo resultados mais rápidos. Já em caso da consulta ser em um ambiente fechado, em que apenas um pesquisador esteja utilizando o sistema, os resultados salvos facilitam sua replicação. Assim, nos testes de busca pela palavra "educação" em 2.916 páginas de documentos, referentes ao Diário Oficial do ano de 2011, obteve-se 1.235 resultados. Esse processo demandou 7 segundos. Na segunda busca, para o qual o sistema utilizou automaticamente o EhCache, com as mesmas informações, o tempo foi reduzido para 3 segundos. Constata-se, portanto, a ampliação de desempenho em pesquisas com resultados consultados anteriormente.

d) Q5. O sistema realiza análise documental, gerando dados relacionados à frequência de palavras, gráficos de frequência e de nuvem de palavras?

RQ5. Após a seleção dos dados que vão compor a análise, o sistema, por meio de NLP, realiza uma análise dos dados. Nesse contexto, o SPEdu traz, tanto em pequena escala como em grande escala, a possibilidade de analisar os resultados, gerando o gráfico denominado "nuvem de palavras". Isso permite que o pesquisador identifique as palavras com maior frequência e as com menor frequência. Além disso, um gráfico de barras é gerado com as

mesmas informações, porém, com um número reduzido de palavras. Com isso, evita-se um gráfico indecifrável pelo excesso de informação. Esse gráfico é gerado com as dez palavras com maior frequência, podendo ser configurado para exibir mais palavras de interesse do usuário. Por fim, é gerada uma tabela com os dados nomeados. A descrição dos dados é feita por meio de IA treinada, a qual é constantemente atualizada. O conjunto de informações deve subsidiar os pesquisadores nas análises documentais.

e) Q6. O sistema permite salvar o relatório, gerando evidências de pesquisa?

RQ6. No contexto das pesquisas, diversas vezes é necessário ter a evidência dos resultados. Assim, o SPEdu permite salvar os dados, replicar as buscas e, ainda, gerar o relatório com a análise dos dados. Esse relatório pode ser disponibilizado no formato PDF para o caso do sistema esteja implementado em um computador e não na Web. Com isso, o sistema gera o relatório que traz os resultados da consulta, bem como as análises realizadas automaticamente.

Para do estudo de caso, esta tese como resultados adicionais a criação do projeto de extensão contínua denominado “DigitalSE”, na Universidade Tiradentes (UNIT), campus Farolândia. Esse projeto tem como foco a modificação ou implementação do SPEdu na gestão de conhecimento e pesquisa, tendo sido iniciado em setembro de 2019, com a adoção do sistema em mais dois órgãos públicos do estado de Sergipe: Secretaria da Fazendo (SEFAZ) e Corpo de Bombeiros do Estado de Sergipe (CBSE). Já está em desenvolvimento nesses dois órgãos as fases de implementação do SPEdu e da digitalização do acervo documental. Na SEFAZ já há 100 mil páginas de documentos digitalizados e extraídos. Para esse órgão foram realizadas adaptações ao SPEdu, denominadas SGED, visando atender as especificidades daquela secretaria, conforme exposto na Figura 58.

Figura 58 - Tela principal do SGED (adaptação do SPEdu) para a SEFAZ

Fonte: Elaborada pelo autor (2020)

No CBSE já há 1.355 páginas de documentos digitalizados, sendo que SPEDu está em análise para eventuais adaptações. No caso desse órgão, o SPEdu será implementado em dois contextos: um para uso interno na gestão de conhecimento e outro para a disponibilização, por meio da Web, da história dos 100 anos do CBSE, devendo estar disponível em outubro de 2020.

No âmbito do projeto "DigitalSE", o SPEdu envolve, além do autor desta tese, alunos de graduação em Computação, os quais atuam na digitalização documental e na melhoria e adaptação do instrumento. Atualmente, há treze alunos participantes, sendo três alocados na SEGRASE, seis na SEFAZ e quatro no CBSE. O SPEdu vem recebendo o interesse da sociedade, tendo sido apresentado, em palestra, como resultado do projeto "Hemeroteca Virtual" na Quinta Bienal do Livro, na cidade de Itabaiana/SE, e abordado em diversas entrevistas em mídias televisivas (APÊNDICE C).

Entende-se, assim, que o SPEdu apresenta resultados sociais ao permitir que órgãos, sem recursos financeiros e tecnológicos, possam ter um ambiente para gestão da informação. Segundo aponta Baum (2020), os programas de pós-graduação strictu sensu serão avaliados por sua contribuição social e outros produtos. No caso desta tese, além da cooperação junto à sociedade, no âmbito de órgãos públicos, o trabalho produziu dois artefatos tecnológicos: o SPEdu e a biblioteca de extração de dados *preprocessing*. Ainda, como resultado da tese, houve a publicação ou aceite de cinco trabalhos sobre o tema diretamente ou, ainda, relacionado às metodologias adotadas (APÊNDICE D).

6 CONSIDERAÇÕES FINAIS

O tema do estudo surgiu das necessidades dos pesquisadores dos mais variados campos, mas, neste caso, aplicado à História da Educação. Um investigador deve apropriar-se do maior número possível de recursos tecnológicos para a realização de suas pesquisas. Diante disso, foi proposto o desenvolvimento de um artefato de *software*, aqui denominado instrumento SPEdu, o qual tem como foco o apoio aos pesquisadores na tarefa de compor acervos, efetuar buscar, compartilhar resultados e gerir informações. Além disso, o SPEdu se mostrou promissor para ser adotado na gestão de acervos de órgãos de documentação, como no caso da SEGRASE, bem como para disponibilizar documentos históricos, como caso do CBSE.

Assim, com o objetivo de desenvolver um modelo de instrumento digital, voltado à recuperação de informações de fontes documentais histórico-educacional, esta tese iniciou por identificar pesquisas no campo da História da Educação que adotaram instrumentos tecnológicos. Nesse contexto, foi possível identificar que, de forma geral, são adotados os CD´s e DVD´s para disponibilização de acervos de pesquisas. Porém, como apresentado, tem como problema a limitação da mídia, na qual estão dispostos os dados. Isso não permite uma busca eficiente. Esse aspecto foi tratado no artigo produzido a partir desta tese, intitulado "Novas tecnologias aplicadas à pesquisa em História da Educação" (APÊNDICE C), com publicação prevista na revista "Cadernos de História da Educação da UFU". No trabalho é apresentado o estado do conhecimento sobre a adoção de tecnologia nas pesquisas em História da Educação.

A adoção da metodologia *design science research* possibilitou a criação do SPEdu que, ao ser analisado, obteve êxito nas suas diversas partes. O processo de extração de dados resultou em um modelo que, com base em informações sobre a imagem do documento e com o uso do algoritmo RandonForest, possibilitou a análise de itens com foco na classificação de texto e imagem, melhorando o desempenho do OCR. Para a utilização dessa metodologia, foi realizado um estudo sobre ela na inovação de sistemas de informação. Isso gerou o trabalho intitulado "Design science in digital innovation: a literature review", aceito para apresentação o "XVI Simpósio Brasileiro de Sistemas de Informação" (APÊNDICE D).

No contexto da extração de dados, a redução do tempo de análise de imagem, possibilitada pela IA, integrada ao Tesseract, tornou o processo mais eficiente. Isso permitiu que o módulo de análise e extração de dados fosse utilizado tanto em pequena escala (quando

um pesquisador insere os dados no sistema de forma individual), como em grande escala (no caso de centros de documentação que desejam extrair dados de diversos documentos de forma simultânea).

Já sobre a indexação, adotou-se nesta tese o modelo utilizado pela maioria das pesquisas, que é a frequência de palavras, porém sem estabelecimento do mínimo de ocorrências. Também foi realizado um pré-processamento, com foco a limpar os dados que não auxiliam na recuperação da informação, criando, assim, uma base integrada das fontes documentais com eficiência para a etapa de recuperação de informação. Dessa forma, o pré-processamento realizado, em conjunto com a análise de frequência, possibilitou a criação de um vocabulário documental integrado de forma automatizada.

Como cada pesquisa tem características próprias em relação aos tipos de documentos, a criação desses tipos possibilita a geração de taxonomias documentais. Isso pode ser aplicado a qualquer volume de documentos. Outro ponto a se destacar é que o SPEdu, como foi pensado em relação ao usuário que tem conjuntos distintos de informações, permite a criação de diversos acervos, facilitando a organização dos dados e a sua recuperação.

Na recuperação de informação, com base nos modelos existentes, foi adotada a análise de frequência de palavras para a classificação dos documentos, em conjunto com os operadores booleanos. O SPEdu se mostrou eficiente tanto em conjunto amplo de documentos, como restrito a fontes documentais. É possível, ainda, ordenar os resultados por data do documento. Outro ponto a se destacar é a adoção de *cache* no sistema de busca, melhorando o desempenho durante a segunda busca aos mesmos dados. Todas as buscas são salvas, permitindo que sejam replicadas pelo pesquisador e, também, pelo sistema, o qual identifica se a busca tem resultados anteriores em *cachê*. Isso resulta em melhoria no tempo de recuperação de informação.

Tais resultados corroboram a tese de que a adoção da IA nas TICs utilizadas pelos pesquisadores permite a automação dos processos relacionados à extração, indexação e busca documental. O trabalho demonstrou, ainda, que a aplicação da IA auxilia o pesquisador durante a análise de um conjunto ampliado de fontes e dados, permitindo o exame minucioso, bem como a gestão e o compartilhamento das informações de pesquisa.

Dessa forma, como resultado tecnológico utilizando a IA tem-se a implementação do SPEdu, instrumento de gestão e recuperação de informação, que permite a extração e indexação automatizada de fontes documentais, além do motor de busca, que é responsável por recuperar as informações armazenadas. Todo o instrumento foi criado como *Open Source* e estará disponível, após o registro em: https://github.com/gomesrocha/spedu. Já a

biblioteca de extração de dados estará disponível em: https://github.com/gomesrocha/DigitalSE_InformationExtract-.

Como um produto *Open Source*, o SPEdu poderá ser modificado, utilizado por pesquisadores e órgãos de documentação sem custo, além de permitir a colaboração de diversos pesquisadores. Os sistemas livres são fundamentais para o desenvolvimento sustentável do país emergente, permitindo a inclusão digital. Assim, os pesquisadores e órgão de documentação não precisarão investir em ferramentas para a gestão da informação, podendo fazer uso do SPEdu sem custos e, ainda, podendo colaborar com o aperfeiçoamento do instrumento.

Destaca-se, aqui, o potencial da aplicabilidade dos resultados desta tese, que já vêm sendo adotados em diversos órgãos. Ressalta-se que o SPEdu também pode ser adotado em módulos isolados do sistema, de acordo com as necessidades do usuário. O instrumento demonstra, assim, o seu caráter social, permitindo a redução de custos a órgãos públicos, bem como a pesquisadores. Nesse sentido, a tese contribui para o avanço no campo da tecnologia aplicada à pesquisa em Educação e História, bem como para o campo de recuperação de informação, com um modelo de análise documental, busca estruturada, *cache* de informação e geração de evidências. Os resultados se mostram promissores e apontam para a viabilidade da adoção do SPEdu. Ressalta-se, no entanto, que como a validação do instrumento SPEdu foi realizada por meio de um estudo de caso, os resultados devem ser considerados diante dessa limitação, não podendo ser generalizados em relação a outras aplicabilidades similares. Além disso, todo o material desta tese, como artigos publicados, a tese em MS Word e Latex, estarão disponíveis em http://github.com/gomesrocha/doutorado.

Durante o desenvolvimento da tese, foi criado o projeto “DigitalSE” como parte dos estudos, o qual visa ampliar, melhorar e manter o SPEdu. Esse projeto contribuiu para:

a) Criação de base documental do Diário Oficial do Estado de Sergipe;
b) Criação de base histórica do Corpo de Bombeiros do Estado de Sergipe;
c) Criação de algoritmos para extração de dados, indexação e busca de informação e disponibilizado de forma OS;
d) Qualificação de vinte e dois alunos, que atuam ou atuaram no projeto dentro dos órgãos públicos na adaptação e implementação do SPEdu.

Para trabalhos futuros, intenta-se utilizar a biblioteca *tensorflow* na tarefa de análise de *layout* e de resultados, buscando melhorar a p*erformance* do sistema, bem como na sumarização textual dos documentos, com o objetivo de criar resumos automáticos. Outro ponto é o aprendizado por reforço, que pode ser adotado no sistema de busca, bem como a

inclusão de análise de dados para a criação de um módulo de indicação para os investigadores, auxiliando-os nas pesquisas.

REFERÊNCIAS

AMORIM, Eliane Dutra. Arquivos, pesquisa e as novas tecnologias. In: FARIA FILHO, Luciano Mendes. (Org.). **Arquivos, fontes e nova tecnologia:** questões para a história da educação. Campinas: Autores Associados/ Bragança Paulista: Universidade São Francisco, 2000, p. 89-99.

ANDRADE, Vivian Galdino. A experiência de criação de um repositório digital como fonte de pesquisa para a história da educação de Bananeiras. **Revista de História e Historiografia da Educação**, v. 1, n. 2, p. 266-284, 2017.

BACELLAR, Carlos. Fontes documentais: uso e mau uso dos arquivos. In: PINSKY, Carla Bassanezi. **Fontes históricas**. 3. ed. São Paulo: Contexto, 2018, p. 23-79.

BAEZA-YATES, Ricardo; RIBEIRO-NETO, Berthier. **Recuperação de informação**: conceitos e tecnologia das máquinas de busca. Tradução: Leandro Krug Wives. Porto Alegre: Bookman, 2013.

BANKS, Marcus. **Dados visuais para pesquisa qualitativa**. Tradução de José Fonseca. Porto Alegre: Artmed, 2009.

BARDIN, Laurence. **Análise de conteúdo**. São Paulo: Edições 70, 2016.

BAUM, Sonia Nair. O novo modelo de avaliação dos Programas de Pós-Graduação do Brasil. **Youtube**. 14 fev. 2020. Disponível em https://pt.wikihow.com/Citar-um-V%C3%ADdeo-do-YouTube. Acesso em: 20 fev. 2020.

BOHNSACK, Ralf. A interpretação de imagens segundo o método documentário. In: WELLER, Wivian; PFAFF, Nicolle. (Orgs). **Metodologia da pesquisa qualitativa em educação**: teoria e prática. Petrópolis: Vozes, 2010.

BONATO, Nailda Marinho da costa. O uso das fontes documentais na pesquisa em história da educação e as novas tecnologias. **Acervo**, Rio de Janeiro, v. 17, n. 2 jul-dez, p. 85-110, 2004.

BOOTSFACES. 2020. Conteúdo sobre o Bootsfaces. Disponível em https://bootsfaces.net/index.jsf. Acesso em 04 de janeiro de 2020.

BORGES, Graciane S. B. Indexação Automática de Documentos Textuais. In LIMA, Gercina Ângela. **Biblioteca Digital Hispertextual**: Caminhos para a Navegação em Contexto. Rio de Janeiro: Interciência, 2016.

BOUSBIA, Nabila; BELAMRI, Idriss. Which contribuition does EDM provide to computer-based learning environments? PEÑA-AYALA, Alejandro. (Org.). **Educational data mining**: applications and trends. Warsaw (Polônia); Springer, 201, p. 3-28, 2014.

BREIMAN, Leo. **Randomforests**. Machine learning, New York, v. 45, n. 1, p. 5-32, 2001.

BUKHARI, Syed Saqib. et al. Document image segmentation using discriminative learning over connected components. In: **Proceedings of the 9th IAPR International Workshop on Document Analysis Systems**, Boston, International Association for Pattern Recognition, 2010, p. 183-190.

CABRAL, Ana Maria Rezende. Tecnologia digital em bibliotecas e arquivos. **Transinformação**, Campinas, v. 14, n 2, p. 167-177, 2002.

CHARTIER, Roger. **A aventura do livro**: do leitor ao navegador. Tradução Reginaldo Carmello Corrêa Moraes. São Paulo: UNESP, 1999.

_____. **A ordem dos livros**: leitores, autores e bibliotecas na Europa entre os séculos XIV e XVIII. Tradução de Mary Del Priori. Brasília: Editora Universidade de Brasília, 1999a.

_____. **Novas tecnologias e a história da cultura escrita**. Obra, leitura, memória e apagamento. Leitura: Teoria e Prática, Campinas, São Paulo, v. 35, n.71, p.17-29, 2017.

CHATHURANGA, R.M. Samitha; RANATHUNGA, Lochandaka. Procedural approach for content segmentation of old newspaper pages. In. **Proceedings of IEEE International Conference on Industrial and Information Systems**. Peradeniya, University of Peradeniya, 2017, p. 1-6.

CHUDHURI, Arindan; BADELIA, Krupa Mandaviya Pratixa; GHOSH, Soumya K. **Optical character recognition systems for different languages with soft computing**. Gewerbestrasse: Springer, 2017.

COLAVIZZA, Giovanni; EHRMANN, Maud; BORTOLUZZI, Fabio. Index-driven digitization and indexation of historical archives. **Frontiers in Digital Humanities**, v. 6, p. 1-16, 2019.

CONCI, Aura; AZEVEDO, Eduardo; LETA, Fabiana R. **Computação gráfica V2**. Rio de Janeiro: Elsevier, 2008.

CONSELHO NACIONAL DE ARQUIVOS - CONARQ. Câmara Técnica de Documentos Eletrônicos - CTDE. **Glossário Documentos arquivísticos digitais**, 7. versão. Rio de Janeiro: 2016.

CROFT, Bruce; METZLER, Donald; STROHMAN, Trevor. **Search Engines: Information Retrieval in Practice**. Londres: Person, 2009.

DARTON, Robert. **A questão dos livros**: passado, presente e futuro. Tradução: Daniel. Pellizari. São Paulo: Companhia das Letras, 2010.

DE LUCA, Tania Regina. Fontes impressas: história dos, nos e por meio dos periódicos. In: PINSKY, Carla Bassanezi. **Fontes históricas**. 3. ed. São Paulo: Contexto, 2018, p. 23-79.

DRESH, Aline; LACERDA, Daniel Pacheco; ANTUNES JÚNIOR, José Antonio Valle. **Design science research**: método de pesquisa para avanço da ciência e tecnologia. Porto Alegre: Bookman, 2015.

DUCHESNE, Pierre; RÉMILLARD, Bruno (Ed.). **Statistical modeling and analysis for complex data problems**. Springer Science & Business Media, 2005.

EISENSTEIN, Jacob. **Introduction to natural language processing**. Boston: Mit Press, 2019.

Fagan, Joel L.. **Automatic Phrase Indexing for Document Retrieval**. ACM SIGIR Forum, New York: ACM, v. 51, n. 2, p. 51–61. 2017. doi:10.1145/3130348.3130355

FAPESP. **Planos de gestão de dados se incorporam a projetos de pesquisa no Brasil**. São Paulo: Pesquisa Fapesp, 2017. Disponível em https://revistapesquisa.fapesp.br/2017/10/25/planos-de-gestao-de-dados-se-incorporam-a-projetos-de-pesquisa-no-brasil/

FAPESP. **Plano de Gestão de Dados**. Disponível em http://www.fapesp.br/gestaodedados/#gestao. Acesso em 20 de fevereiro de 2020.

FARIA FILHO, Luciano Mendes. (Org.). **Arquivos, fontes e nova tecnologia:** questões para a história da educação. Campinas: Autores Associados/ Bragança Paulista: Universidade São Francisco, 2000.

FELGUEIRAS, Carlos Alberto. **Processamento digital de imagens - Processamento de Cores**. [s.d]. Disponível em: http://www.dpi.inpe.br/~carlos/Academicos/Cursos/Pdi/pdi_cores.html. Acesso em: 08 jan. 2019.

FERNANDES, Lincoln Christian. Arquivos escolares e memória: novas perspectivas da pesquisa histórica a partir das novas tecnologias da informação. In: **X Encontro de História de Mato Grosso do Sul.** Três Lagoas, UFMS, 2010, p. 1058-1069.

FRERY, Alejandro C. Image Filtering. In: MELLO, Carlos Alexandre Barros de; SANTOS, Wellington Pinheiro dos; OLIVEIRA, Adriano Lorena Inácio de. **Digital document:** analys and processing. New York: Nova Science Publishers, 2011, p.55-71

GEARY, David; HORSTMANN, Cay. **Core JavaServer Faces**. Tradução de Lúcia Helena. 3. ed. Rio de Janeiro: Altabooks, 2012.

GIL, Antonio Carlos. Como elaborar projetos de pesquisa. 6. ed. São Paulo: Atlas, 2017.

GINZBURG. Carlo. **Mitos, emblemas, sinais**: morfologia e história. Tradução: Frederico Carotti. São Paulo: Companhia das Letras, 1989.

GISLASON, PallOskar; BENEDIKTSSON, Jon Atli; SVEINSSON, Johannes R. **Random forests for land cover classification**. Pattern Recognition Letters, Heslington, v. 27, n. 4, p. 294-300, 2006.

GLLAVATA, Julinda; EWERTH, Ralph; FREISLEBEN, Bernd. A robust algorithm for text detection in images. In: **3rd International Symposium on Image and Signal Processing and Analysis**, ROME, University of Zagreb, 2003, p. 611-616.

GÓES, Camila Magalhães. A produção historiográfica educacional das ies soteropolitanas: arquivos e dados. In: **Anais V Congresso Brasileiro de História da Educação**. Aracaju, UFS/UNIT, 2008, p. 1-11.

GOLDSCHMIDT, Ronaldo; PASSOS, Emmanuel; BEZERRA, Eduardo. **Data mining**: conceitos, técnicas, algoritmos, orientações e aplicações. 2. ed. Rio de Janeiro: Elsevier, 2015.

GONDRA, José G. A leveza dos bits. In: FARIAS FILHO, Luciano Mendes. **Arquivos, fontes e novas tecnologias**: questões para a História da Educação. Campinas: Editores Associados, 2000, p. 3-17.

GONZALEZ, Rafael C. E WOODS, Richard E. **Processamento de imagens digitais**. Tradução Roberto Marcondes Cesar Junior e Luciano da Fontoura Costa São Paulo: Edgard. Blücher Ltda, 2000.

GREFENSTETTE, G., TAPANAINEN, P.. What is a word, what is a sentence? problems of tokenization. In: **Procedings 3rd Conf. Computational Lexicography and Text Research (COMPLEX'94)**, Research Institute for Linguistics Hungarian Academy of Sciences, Budapest, 1994, p. 79–87.

HEBERT, David et al. Pivaj: displaying and augmenting digitized newspapers on the web experimental feedback from the Journal de Rouen collection. In: **Proceedings of the First International Conference on Digital Access to Textual Cultural Heritage**. Universidade de Salford, Madri, 2014. p. 173-178.

HENRY, Paul; MOSCOVI, Serge. Probléms de l'analyse de contenu. **Langages**. Paris, n. 11, p. 36-60, set, 1968.

HJORLAND, B.; NIELSEN, Lykke K. **Subject Access Point in Electronic Retrieval.** Annual Review of Information Science and Technology (AFISI), Maryland, v. 35, 2001, p. 249-298.

ISO. International Organization for Standardization. **ISO/IEC 25010**: Systems and software engineering — Systems and software Quality Requirements and Evaluation (SQuaRE) — System and software quality models. Geneva, 2011.

JANA, Shantanu et al. Recognition System to Separate Text Graphics from Indian Newspaper. In: **Proceedings of International Conference on Frontiers in Optimization: Theory and Applications**, Kolkatta, Jadavpur University, 2018, p. 185-194.

JANOTTI, Maria de Lources. O livro "Fontes históricas" como fonte. In: PINSKY, Carla Bassanezi. **Fontes Históricas**. 3. ed. São Paulo: Contexto, 2018, p. 9-22.

JOHANNESSON, Paul; PERJONS, Erik. An introduction to design science. Kista: Springer, 2014.

JONES, Karen Sparck. A statistical interpretation of term specificity and its application in retrieval. **Journal of documentation**, London, v. 60, n. 5, p. 493-502 2004.

KAUR, Rupinder Pal; JINDAL, Manish Kumar. Head line and Column Segmentation in Printed Gurumukhi Script Newspapers. In: **Proceedings of International Conference Smart Innovations in Communication and Computational Sciences**. Punjab, North West Group of Institutions, 2017, p. 59-67.

KHALFALLAH, Faten; ALOULOU, Chafik; BELGUITH, Lamia. HAD, a platform to create a historical dictionary. (AICCSA) IEEE/ACS **13th International Conference of Computer Systems and Applications**. Agarir, 2016.

KUMAR, P. Ramesh; SAILAJA, K. L.; BEGUM, Shaik Mehatab. Human Identification Based on Ear Image Contour and Its Properties. In: **International Conference on ISMAC in Computational Vision and Bio-Engineering**. Vivekanandha College of Technology for Women, Elayampalayam, 2019. p. 1527-1536.

LANCASTER, Frederick Wilfrid. **Indexação e resumos**: teoria e prática. Tradução Antônio Agenor Briquet de Lemos. 2. ed. rev. Brasília: Briquet de Lemos, 2004.

LE GOFF, Jacques. **História e memória.** Tradução: Irene Ferreira, Bernardo Leitão, Suzana Ferreira Borges. 5. ed. Campinas: Editora UNICAMP, 2003.

LIMA, Gercina Ângela. Navegação Hipertextual em Contexto. In: LIMA, Gercina Ângela. **Biblioteca Digital Hipertextual:** Caminhos para a navegação em contexto. Rio de Janeiro: Interciencia, 2016.

LOMBARDI, José Claudinei. As novas tecnologias e a pesquisa em História da Educação. In: FARIAS FILHO, Luciano Mendes. **Arquivos, fontes e novas tecnologias**: questões para a História da Educação. Campinas: Editores Associados, 2000, p. 123-150.

LOPES, Ivone Goulart et al. O fio da história–nas trilhas de Ouro Preto do Oeste-RO. Vitrais da memória de professores e escolas. In: **Anais do XI Seminário Internacional de La Red Estrado**. Mexico: UPN, 2016, p. 1-17.

LOUVEIRA, Andreína de Melo; FERRO, Maria Eduarda. Constituição de um catálogo fotográfico digital como ferramenta de pesquisa em História da Educação. In**: Anais VII Congresso Brasileiro de História da Educação.** Cuiabá, UFMT, 2013, p. 1-14.

LUHN, Hans Peter. The automatic creation of literature abstracts. **IBM Journal of research and development**, v. 2, n. 2, p. 159-165, 1958.

MELLO, Carlos Alexandre Barros de. **Filtragem, Compressão e Síntese de Imagens de Documentos Históricos**. 2002. 119p, Tese (Doutorado em Ciências da Computação) - Universidade Federal de Pernambuco, Recife, 2002.

MELLO, Carlos Alexandre Barros de; SANTOS, Wellington Pinheiro dos; OLIVEIRA, Adriano Lorena Inácio de. **Digital document**: analys and processing. New York: Nova Science Publishers, 2011.

MORI, Shunji; SUEN, Ching Y.; YAMAMOTO, Kazuhiko. **Historical review of OCR research and development**. Proceedings of the IEEE, Torino, v. 80, n. 7, p. 1029-1058, 1992.

NASCIMENTO. Ester Fraga Vilas-Bôas Carvalho do. **Fontes para a história da educação**: Documentos da missão presbiteriana dos Estados Unidos no Brasil. Maceió: EDUfal, 2008.

OTSU, Nobuyuki. A Thresholds election method from gray-level histograms. **IEEE Transactions on Systems, Man, and Cybernetics**. Piscataway, v. 9, n.1, p.62-66, 1979.

PALFRAY, Thomas; et al. Logical segmentation for article extraction in digitized old Newspapers. In: **Proceedings of the ACM Symposium on Document Engineering,** Paris, Telecom ParisTech,2012, p.129-132.

PENA, M. G.; SILVA, A. C. A digitalização de documentos históricos e a gestão eletrônica de documentos para disponibilização online. **Saber Digital**, Valença, v. 1, n. 1, p. 85-102, 2008.

PEREIRA, Aracy Roza Sampaio. **Fontes documentais da história da educação escolar no Distrito Federal (1956-1960)**. Monográfia (Licenciatura em Pedagogia) - Faculdade de Educação, Universidade de Brasilia, Distrito Federal, p. 77, 2011.

PIATETSKY-SHAPIRO, Gregory. Knowledge Discovery in Real Databases: a report on the IJCAI-89 Workshop. **Artifical Intelligence Magazine**, Palo Alto, v. 11, n. 5, p. 68-70, 1990.

PRAMANIK, Rahul; BAG, Soumen. Shape decomposition-based handwritten compound character recognition for Bangla **OCR. Journal of Visual Communication and Image Representation**. Redmond, v. 50, p. 123-134, 2018.

QUDDUS, Azhar; CHEIKH, Faouzi Alaya; GABBOUJ, Moncef. Wavelet-based multi-level object retrieval in contour images. In: **Proceedings of the International Workshop on Very Low Bit Rate Video Coding**, Urbana, 2016. p. 1-5.

RAYMOND, Erik Steven. **The Cathedral & the Baazar**. Sebastopol: O´Reilly Media, 1999.

RAJESWARI, S.; MAGAPU, Sai Baba. Desenvolvimento e customização de OCR desenvolvido internamente e sua avaliação. **The Electronic Library**, Howard House, n. 5, vol. 36, p. 766-781, 2018.

REESE, Richard M. **Natural language processing with Java**. Birmingham: Packt Publishing Ltd, 2015.

REFFLE, Ulrich; RINGLSTETTER, Christoph. Unsupervised profiling of OCR ed historical documents. **Pattern Recognition**, Heslington, v. 46, n.5, p. 1346-1357, 2013.

RIJSBERGEN, C. J. Van. Information retrieval: theory and practice. In: **Proceedings of the Joint IBM/University of Newcastle upon Tyne Seminar on Data Base Systems**. New Castle, 1979. p. 1-14.

ROCHA, Miguel; CORTEZ, Paulo; NEVES, José Maria. **Análise inteligente de dados**: algoritmo e implementação em Java. Lisbo: FCA, 2008.

ROMERO, Cristóbal; BENTURA, Sebastian; PECHENIZKIY, Mykola; BAKER. Ryan. **Handbook of Educational Data Mining**. Minneapolis: CRC Press, 2010.

RUSSEL, Stuart; NORVIG, Peter. **Inteligência artificial**. Tradução Regina Célia Simille de Macedo. Rio de Janeiro: Elsevier, 2013.

SABINO, Rosimeri Ferraz. **A configuração da profissão de secretário em Sergipe**: educação, atuação e organização da área (1975-2010). 2017. 387p. Tese (Doutorado em Educação) – Universidade Federal de Sergipe, São Cristóvão, 2017.

SAUVOLA, Jaakko et al. Adaptive document binarization. In: **Proceedings of the Fourth International Conference on Document Analysis and Recognition**. International Association for Pattern Recognition, Ulm, 1997. p. 147-152.

SCHATZ, Bruce R. Information retrieval in digital libraries: Bringing search to the net. **Science**, v. 275, n. 5298, p. 327-334, 1997.

SCHMIDT, Eric; COHEN Jared. **A nova era digital**: como será o futuro das pessoas, das nações e dos negócios. Tradução de Ana Beatriz Rodrigues e Rogério Durst. Rio de Janeiro: Intrínseca, 2013.

SEGRASE. Serviços Gráficos de Sergipe. **SEGRASE firma convênio com a Universidade Tiradentes**, 2017. Disponível em https://www.segrase.se.gov.br/site/noticias/visualizar/563, acesso em 10 de dezembro de 2019.

SEGRASE. Serviços Gráficos de Sergipe. **Acervo do Diário Oficial do Estado de Sergipe**, período de 1990 a 2012.

SHEPARD, David H. **Apparatus for reading**. US Patent 2663758, 1953.

SILVA, Ana Lícia de Melo. **Imprensa Oficial do Estado de Sergipe**: 123 anos. Aracaju: Edise, 2018.

SILVA, Eva Cristina Leite da. Mapeamento dos arquivos escolares: história, memória e preservação de documentos. **Ágora**, Florianópolis, v. 21, n. 42, 2011, p. 111-125.

SILVA, Gabriel de França Pereira e. **Algoritmos para classificação, filtragem e transcrição de imagens de documentos**. 2014. 256p, Tese (Doutorado em Engenharia Elétrica) - Universidade Federal de Pernambuco, Recife, 2014.

SILVA, Marcel Ferrante. Interface para navegação em bibliotecas digitais. In: LIMA, Gercina Ângela. **Biblioteca Digital Hipertextual:** Caminhos para a navegação em contexto. Rio de Janeiro: Interciencia, 2016.

SILVEIRA, Sérgio Amadeu da. Inclusão Digital, Software Livre e Globalização Contra-Hegemônica. In: SILVEIRA, Sérgio Amadeu da; CASSINO, João. **Software Livre e inclusão digital.** São Paulo: Conrad, 2003.

SIQUEIRA, Elizabeth Madureira. Reconstruindo arquivos escolares: a experiência do GEM/MT. **Revista Brasileira de História da Educação.** Maringa, v. 5, n.2, 2005, p. 123-152.

SMITH, Ray. An overview of the Tesseract OCR engine. In: **Ninth International Conference on Document Analysis and Recognition (ICDAR 2007)**. IEEE Computer Society, Curitiba, 2007. p. 629-633.

SOARES, Irvin Bezerra; BRAGA, Mirella de Almeida; LIMA, Izabel França de. Digitalização de Documentos: disponibilizando as fichas do DOPS do período da ditadura militar na Internet. **InterScientia**. João Pessoa. v. 3, n. 2, 2015, p. 17-28.

SOUZA, Rosa Fátima de. Preservação do patrimônio histórico escolar no Brasil: notas para um debate. **Revista Linhas**. Florianópolis, v. 14, n. 26, 2013. p. 199 – 221.
SUZUKI, Satoshi; ABE, Keiichi. Topological structural analysis of digitized binary images by border following. **Computer vision, graphics, and image processing**. Orlando, v. 30, n. 1, p. 32-46, 1983.

TAUSCHECK, Gustav. Reading machine. US Patent 2026329, 1935.

TAVARES, Andrezza M. B. Pedagogia social e juventude em exclusão: compreensões necessárias à formação de professores. **Holos**, ano 31, v. 4, p. 18-32, 2015.

TIOBE. **TIOBE Index for February 2020**. Disponível em https://www.tiobe.com/tiobe-index/. Acesso em 04 de janeiro de 2020.

TOSCHI, Mirza Seabra; RODRIGUES, Maria Emília de Castro. Infovias e educação. **Educação e Pesquisa**. São Paulo, v. 29, n. 2, 2003, p. 313-326.

VASILOPOULOS, Nikos; KAVALLIERATOU, Ergina. Complex layout analysis based on contour classification and morphological operations. **Engineering Applications of Artificial Intelligence**, Laxenburg, v. 65. p. 220-229, 2017.

VASILOPOULOS, Nikos; WASFI, Yazan; KAVALLIERATOU, Ergina. Automatic text extraction from Arabic newspapers. In: **Proceedings of XV International Conference on Image Analysis and Recognition**. Póvoa de Varzim, Association for Image and Machine Intelligence, 2018, p. 505-510.

VIEIRA, Alboni Marisa Dudeque Pianovski. Os documentos microfilmados e/ou digitalizados como fonte para o estudo da história da educação: avanços e possibilidades. In: **Anais do VI Congresso Brasileiro de História da Educação.** Vila Velha: UFES, 2011, p. 1-11.

WALKINSHAW, Neil; MINKU, Leandro. Are 20% of files responsible for 80% of defects?. In: **Proceedings of the 12th ACM/IEEE International Symposium on Empirical Software Engineering and Measurement**. Onlu, University of Oulu 2018. p. 1-10.

WERLE, Flavia Obino Corrêa. Cd-rom como apoio na pesquisa sobre a identidade e a história institucional. **Educação Unisinos**. São Leopoldo, v.11, v. 2, 2007, p. 111-120.

WHOLIN, Claes et al. **Experimentation in software engineering.** Dordrecht: Springer Science & Business Media, 2012.

YIN, Robert K. **Estudo de caso**: planejamento e métodos. Tradução Cristhian Matheus Herrera. 5. ed. Porto Alegre: Bookman, 2015.

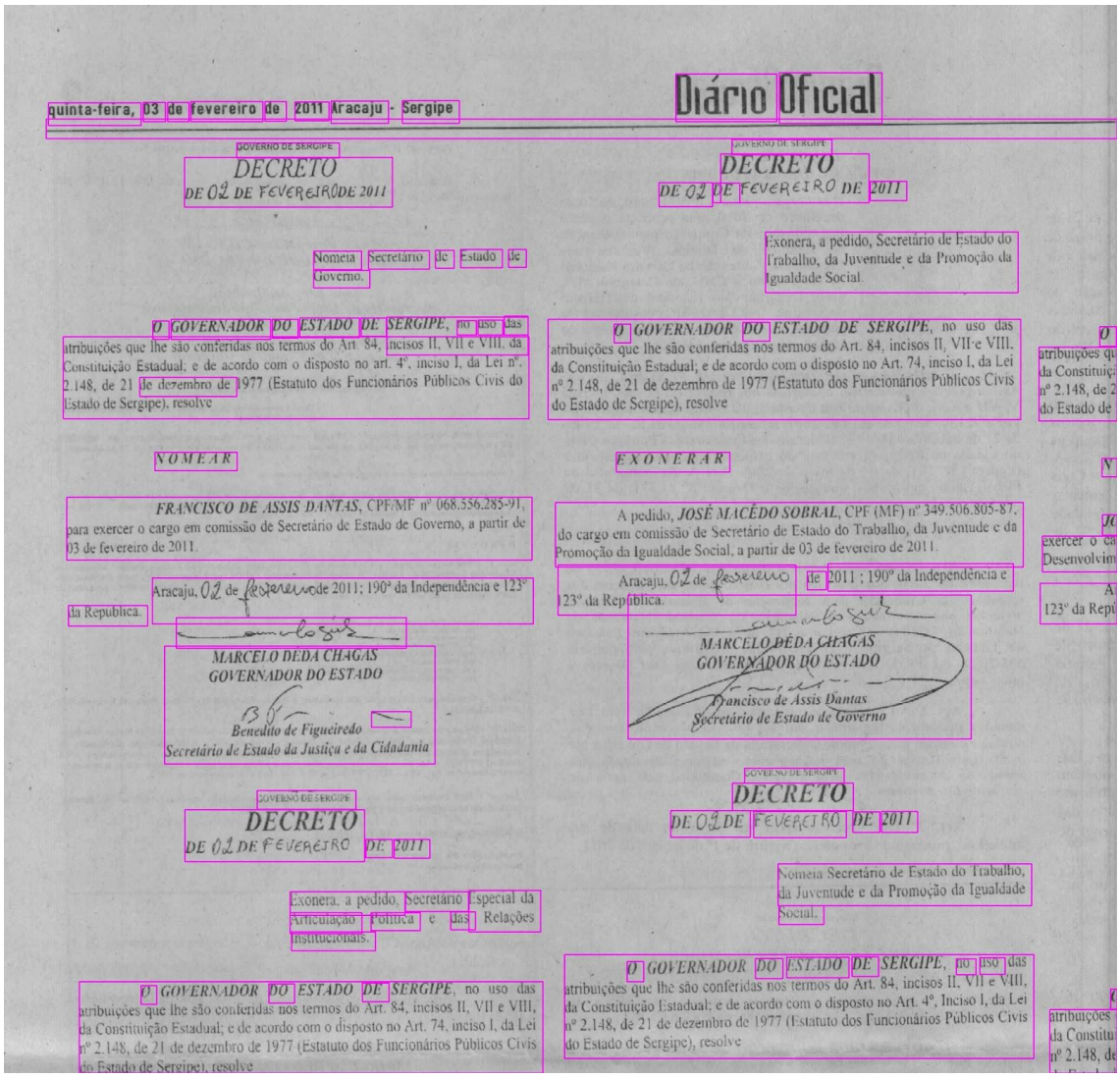
ZHANG, Jiansong; EL-GOHARY, Nora M. Semantic NLP-based information extraction from construction regulatory documents for automated compliance checking. **Journal of Computing in Civil Engineering**, v. 30, n. 2, p. 04015014/1-14, 2016, Disponível em: www.doi.org/10.1061/(asce)cp.1943-5487.0000346

ZENI, M.; WELDEMARIAM, K. Extracting informati on from newspaper archives in Africa. **IBM Journal of Research and Development**, Ossining, v. 61, n. 6, p. 1-12, 2017.

ZOBEL, Justin; MOFFAT, Alistair. Inverted files for text search engines. **ACM computing surveys (CSUR)**, v. 38, n. 2, p. 1-56, 2006.

APÊNDICES

APÊNDICE A – EXEMPLO DE DETECÇÃO DE TEXTO EM IMAGEM ORIGINAL DO DIÁRIO OFICIAL DO ESTADO DE SERGIPE



Fonte: Elaborada pelo autor, adaptado do acervo SEGRASE (1997).

APÊNDICE B – EXEMPLO DE EXTRAÇÃO DE TEXTO EM IMAGEM ORIGINAL DO DIÁRIO OFICIAL DO ESTADO DE SERGIPE

Ouvidor Geral do Estado
LUIZ EDUARDO COSTA

Procurador-Geral do Estado
MÁRCIO LEITE DE REZENDE

Secretário de Estado da Comunicação Social
FRANCISCO FERREIRA PEREIRA

Fonte: Figura elaborada pelo autor, adaptado do acervo SEGRASE (1997).

APÊNDICE C – ENTREVISTAS E PALESTRAS DO AUTOR SOBRE TEMA DA TESE

Local	Ano	Link
TV Aperipe	2019	https://www.youtube.com/watch?v=AziN_kU9uWc
TV Sergipe	2019	https://www.youtube.com/watch?v=cFvrAzjQyB0
TV Canção Nova	2019	https://www.youtube.com/watch?v=51mAzsnfhSE
TV Atalaia	2019	https://a8se.com/tv-atalaia/se-no-ar/video/2019/03/156800-conheca-o-trabalho-da-hemeroteca.html
TV Atalaia	2020	https://a8se.com/tv-atalaia/balanco-geral/video/2020/01/173772-trabalhos-do-governo-estadual-estao-sendo-digitalizados.html
UNIT	2020	https://www.facebook.com/Unit.br/videos/2551827001762591/
CBSE	2020	https://www.cbm.se.gov.br/cbmse-firma-parceria-para-mapeamento-historico-digital/
TV Sergipe	2020	https://g1.globo.com/se/sergipe/edicao/2020/02/18/videos-bom-dia-sergipe-desta-terca-feira-18-de-fevereiro.ghtml#video-8331631-id
Palestra Bienal do Livro	2019	https://www.youtube.com/watch?v=PQ3YSp29svA

Fonte: Elaborado pelo autor (2020).

APÊNDICE D – PRODUÇÕES RESULTANTES DA TESE

Título	Evento/periódico	Status
Classificação Documental por meio de Processamento de Linguagem Natural	19° Seminfo https://eventos.set.edu.br/index.php/sempesq/article/view/6914	Publicado
Um modelo de mapeamento sistemático para a educação	Cadernos da Fucamp http://www.fucamp.edu.br/editora/index.php/cadernos/article/view/1180	Publicado
Aquisição de imagem para hemeroteca digital	Interfaces Científica – Exatas e Tecnológica https://periodicos.set.edu.br/index.php/exatas/article/view/5882	Publicado
Novas tecnologias aplicadas à pesquisa em história da educação	Cadernos de História da Educação	Aceito
Design Science in Digital Innovation: a literature review	XVI Simpósio Brasileiro de Sistemas de Informação	Aceito

Fonte: Elaborado pelo autor (2020).

ANEXOS

ANEXO A – ACEITE DO ARTIGO NOVAS TECNOLOGIAS APLICADAS À PESQUISA EM HISTÒRIA DA EDUCAÇÃO

[CHE] Comunicação importante Caixa de entrada ×

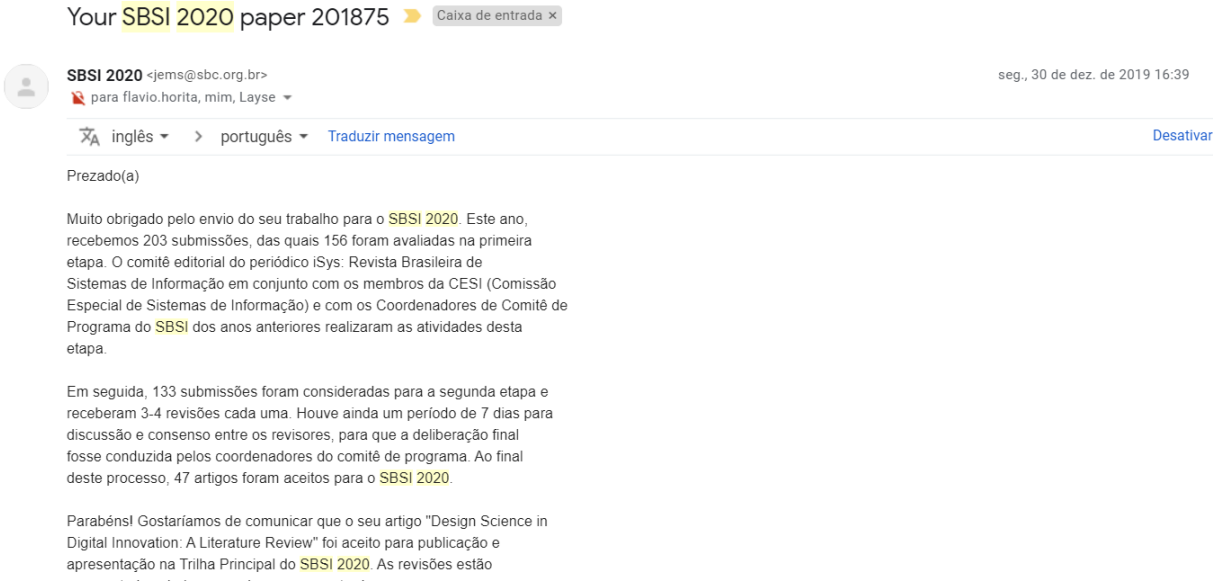
che@faced.ufu.br ter., 11 de fev. 09:27 (há 12 dias)
para mim, ester_fraga

Prezados Fabio Gomes Rocha e Ester Fraga Vilas-Bôas Carvalho do Nascimento,

Temos a grata satisfação de comunicar que o trabalho intitulado "NOVAS TECNOLOGIAS APLICADAS À PESQUISA EM HISTÓRIA DA EDUCAÇÃO" foi APROVADO para constar nos Cadernos de História da Educação, com previsão de ser publicado em um dos números de 2021. Todavia, no sentido de atender as normas do periódico, pedimos a gentileza de que enviem, como resposta deste e-mail, uma versão em inglês do mesmo, na qual, em nota de rodapé, a partir do título, constem o nome e o e-mail do responsável pela elaboração da referida versão, para que possamos publicar o artigo em versão bilíngue. Para tanto, informamos que o ideal é que esta versão em inglês seja encaminhada até 10/04/2020.

Com nossos cumprimentos,
Comissão Editorial – Cadernos de História da Educação

ANEXO B – ACEITE DO ARTIGODESIGN SCIENCE IN DIGITAL INNOVATION: A LITERATURE REVIEW



Your SBSI 2020 paper 201875 Caixa de entrada ×

SBSI 2020 <jems@sbc.org.br> seg., 30 de dez. de 2019 16:39
para flavio.horita, mim, Layse

inglês › português Traduzir mensagem Desativar

Prezado(a)

Muito obrigado pelo envio do seu trabalho para o SBSI 2020. Este ano, recebemos 203 submissões, das quais 156 foram avaliadas na primeira etapa. O comitê editorial do periódico iSys: Revista Brasileira de Sistemas de Informação em conjunto com os membros da CESI (Comissão Especial de Sistemas de Informação) e com os Coordenadores de Comitê de Programa do SBSI dos anos anteriores realizaram as atividades desta etapa.

Em seguida, 133 submissões foram consideradas para a segunda etapa e receberam 3-4 revisões cada uma. Houve ainda um período de 7 dias para discussão e consenso entre os revisores, para que a deliberação final fosse conduzida pelos coordenadores do comitê de programa. Ao final deste processo, 47 artigos foram aceitos para o SBSI 2020.

Parabéns! Gostaríamos de comunicar que o seu artigo "Design Science in Digital Innovation: A Literature Review" foi aceito para publicação e apresentação na Trilha Principal do SBSI 2020. As revisões estão