Processamento e Navegação por Tópicos em Imagens de Páginas de Jornais Históricos

Gildácio J. de A. Sá
Universidade Estadual do Ceará– UECE
Fortaleza, CE, Brasil
gildacio.sa@gmail.com

José E. B. Maia
Universidade Estadual do Ceará– UECE
Fortaleza, CE, Brasil
jose.maia@uece.br

ABSTRACT

This paper presents the architecture and operation of a Historical Newspaper Page Image Topic Navigation System designed to facilitate the access and use of social and historical research to the historical newspaper collection. The system consists of four modules which are: Text Subimage Segmentation, Text Extraction and Preprocessing, Topic Network Extraction, and Document Viewing and Retrieval Interface. The algorithmic and technological approaches of each module are described and the initial test results are presented.

KEYWORDS

Historical NewsPaper, Induced Topic, NewsPaper Image Segmentation

1 INTRODUÇÃO

Este trabalho descreve e avalia os algoritmos e as tecnologias utilizados no projeto de um Sistema de Processamento e Navegação por Tópicos em Imagens de Páginas de Jornais Históricos.

Os jornais possuem a primazia de serem os primeiros veículos a apresentar determinada notícia, dada a sua contemporaneidade, onipresença e co-ocorrência temporal ao evento a ser noticiado. Isso os torna mais ricos em detalhes circunstanciais ao fato, ofertando às vezes o micro-detalhe que passa despercebido ao leitor mais desavisado, mas não ao estudioso social.

No entanto, devido à urgência de sua publicação, ele peca em não analisar o contexto mais profundo sobre o fato na maioria das notícias reportadas. A análise holística só acontece *a posteriori*, em livros ou outras publicações mais especializadas que, por sua vez, recorrem aos jornais para capturar os detalhes e as circunstâncias conjecturais pontuais e aplicam, então, seu contexto crítico ao fato.

Por serem reconhecidos como detentores desse nível de detalhe em relação ao fato que estão dando cobertura, os jornais históricos ou antigos preservam um rico momento daquela sociedade que precisa ser resgatado à análise. Por isso há grandes acervos de jornais históricos na forma de imagens de páginas [1] em todo o mundo. e que são de interesse de antropólogos, sociólogos e historiadores em geral [2]. No Brasil, na Biblioteca Nacional, existe a ***Hemeroteca da Biblioteca Nacional*** .

Entretanto, esse acervo é difícil de ser lido por máquina devido à baixa qualidade de impressão da época, à pouca padronização das páginas além da própria baixa qualidade fotográfica de alguns arquivos. Uma típica imagem de página de jornal histórico está mostrado na Figura 1. Muitas vezes os softwares de OCR (*Optical Character Recognition*) apenas capturam palavras desconexas sem que elas formem uma frase com sentido [3]. Por isso, nos acessos mais disponíveis atualmente, o leitor deve ler sequencialmente as imagens de páginas do jornal para encontrar um tópico do seu interesse.

Figure 1: Jornal *O NORDESTE*, da Arquidiocese de Fortaleza-CE em 10 de julho de 1943.

O Portal da História do Ceará [1] é uma iniciativa de pesquisador independente que garimpa, digitaliza e disponibiliza para acesso público documentos históricos do Ceará. O acervo atual conta com aproximadamente 300 mil documentos entre livros, jornais, revistas e outros documentos. O projeto descrito nesse artigo visa expandir e melhorar as formas de consulta ao Portal da História do Ceará com a ajuda de técnicas avançadas de Processamento de Linguagem Natural. A relevância de uma tal ferramenta está registrada em inúmeros trabalhos consultados, entre eles, [4, 5] , e projetos semelhantes existem em outras línguas e países [6, 7].

Tipicamente, se está interessado em responder questões do seguinte tipo: *Em qual edição ou edições desse jornal pode-se encontrar textos relevantes sobre um tópico X?*. A abordagem adotada para tal fim será chamada de ***Tópicos Induzidos***.

Trata-se de um processo semi-supervisionado, por agrupamento em tópicos, com classificação dentro do grau de cobertura de cada tópico. Os tópicos são induzidos a partir de sementes (*seeds - ou palavras-raiz*) formando um conjunto de tópicos parcialmente rotulados de forma a melhor agrupar uma coleção de documentos, segundo determinado critério de pertinência ou cobertura.

[1] https://www.ceara.pro.br

O sistema está concebido conforme o diagrama de blocos da Figura 2 inspirado em [8]. Os textos estão presentes em imagens que contém outros elementos, como por exemplo, figuras, gravuras e logomarcas. O primeiro bloco B1 tem a função de segmentar a imagem de página e extrair os segmentos de texto. A etapa seguinte (bloco B2) tem como função extrair os textos usando OCR e pré-processá-los, com a extração de seus radicais e demais formatações. Em seguida, utilizando técnicas de Processamento de Linguagem Natural (PLN) é executada a função de extração da estrutura de tópicos da coleção de textos (bloco B3).

Finalmente essa interface amigável irá apresentar ao usuário a rede dos tópicos encontrada com a possibilidade de recuperação dos documentos ou textos de interesse (bloco B4).

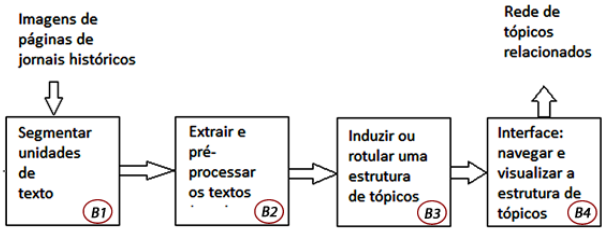

Figure 2: Diagrama de blocos funcionais do Sistema de Processamento e Navegação por Tópicos em Imagens de Páginas Jornais Históricos.

O restante deste trabalho está assim organizado. Na Seção 2 uma descrição dos algoritmos e tecnologias utilizadas no projeto são apresentados. A Seção 3 é sobre trabalhos relacionados e a Seção 4 descreve os experimentos já realizados e os resultados que anunciam a viabilidade do projeto. Na Seção 5 - Conclusao, algumas conclusões são enunciadas.

2 DESCRIÇÃO DO SISTEMA

Esta seção descreve os princípios, algoritmos e tecnologias utilizadas no projeto. A descrição segue a sequência da Figura 2. Por razões de praticidade os algoritmos referentes aos dois primeiros blocos não tiveram implementação própria. Após avaliada e estudados os algoritmos nela utilizados, foi escolhida a ferramenta comercial AbbyyFine Reader CE © [9] para realizar estas etapas.

Para cada módulo registra-se alguns dos algoritmos publicados e testados com vista a conhecer as limitações e a aplicabilidade, e define-se as soluções adotadas. Em função desses resultados, abordagens novas ou variações serão investigadas e aquelas com melhores resultados serão adotadas nas atualizações do sistema. O foco deste artigo é na visão sistêmica do sistema e nas suas características próprias. Entretanto, ainda que sem apresentar detalhes, ao leitor é fornecido um grande volume de referências bibliográfcias de algoritmos alternativos para cada uma das funções que compõem o sistema.

2.1 Segmentação de Subimagens de Texto

Inicialmente, as imagens são capturadas das páginas dos jornais escolhidos por processos de fotografia digital. Os cadernos de jornais históricos são estruturas grandes (45cmX60cm), pesadas e bastante frágeis devido ao tempo.

O problema da segmentação de imagens de páginas de jornais antigos em subimagens de texto e de não-texto é como ilustrado na Figura 3. Esta é uma tarefa de Visão Computacional e esta figura mostra algumas das principais dificuldades: variações no tamanho dos tipos (letras) e fronteiras quase indistinguiveis entre regiões.

Esta etapa é essencial para o sucesso da etapa de aplicação do OCR pois limita a captura às regiões da imagem contendo texto. Técnicas de descoberta de subimagens podem ser encontradas em [10]. Um *survey* de abordagens a esse problema pode ser encontrado em [11]. Outros trabalhos que abordam algoritmos específicos são [12–14].

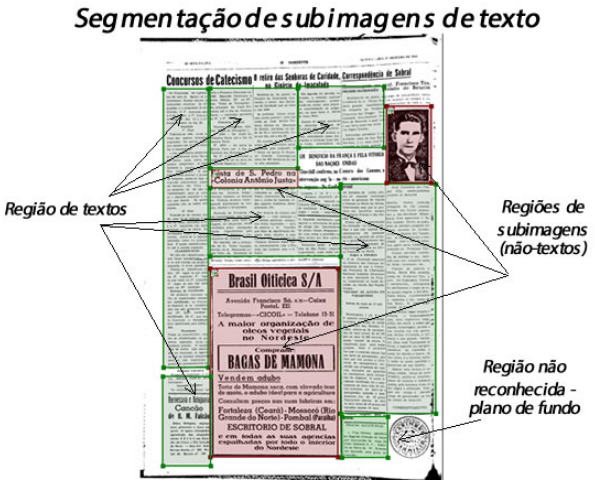

Figure 3: Jornal *O Nordeste* de 1/7/1943, pág. 4 - Identificação de regiões de texto e não-texto em uma imagem.

Um teste exploratório foi realizado com o algoritmo apresentado em [13], que é baseado em identificação de bordas. Entretanto, neste trabalho utilizou segmentação supervisionada com um algoritmo incluido na ferramenta AbbyyFine Reader CE ©. Além da segmentação em sí, a ferramenta fornece uma técnica de indexação na qual as subimagens são indexadas como subordinadas à imagem-mãe[2]. Estas subimagens são indexáveis pelas palavras contidas na região da página onde ela se encontra. O aplicativo pretende - quando solicitado - devolver as subimagens associadas à região do texto onde há palavras de contexto do tópico.

2.2 Extração e Pré-processamento de Texto

A entrada para essa etapa são as subimagens segmentadas e indexadas. A extração e pré-processamento dos textos descritos nessa seção representa os principais processos de preparação do *corpus*.

Uma vez que as subimagens de texto tenham sido extraidas corretamente, a etapa seguinte é a extração do texto em sí. Como ilustração, a Figura 4 mostra como um operador humano realiza essa operação quando trabalhando com um OCR manual. Note a

[2]Imagem-mãe refere-se à imagem de página que contem o segmento, ou região de não-texto, em análise

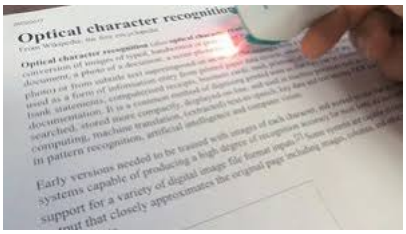

Figure 4: Ilustração do processo de digitalização com um OCR manual. Adaptado de [15].

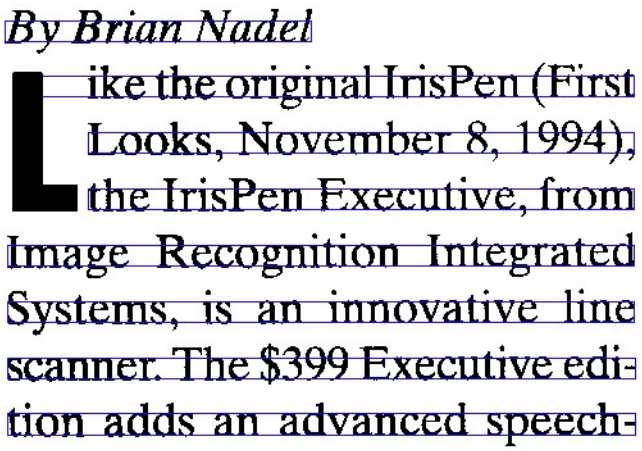

Figure 5: Um exemplo de segmentação de linha de texto util para guiar um OCR. Adaptado de [6].

destreza exigida do operador humano em percorrer as linhas de texto para obter uma boa recuperação do texto.

Para extrair textos de subimagens de texto de forma automática e eficiente é necessário segmentar as linhas de texto. Na Figura 5 procura-se ilustrar um tipo saída que um algoritmo de reconhecimento ou segmentação de linha de texto pode oferecer para guiar o OCR [6, 15].

A ferramenta AbbyyFine Reader CE © utilizada embute algoritmos para essa tarefas. O *OCR* é aplicado às imagens, gerando os arquivos dos tipos TXT, PDF e as subimagens das regiões não texto (quando possível) para cada página de cada edição do jornal. O *corpus* alvo dos algoritmos é formado por arquivos TXTś oriundos do processo de OCR nessas imagens.

Ainda assim, a tarefa não é totalmente automatizada. A utilização de scanners que usam o processo de ADF (*Automatic Document Feeder*, ou Alimentador automático de documento) é impossível devido ao fato de as folhas estarem encadernadas e de serem extremamente sensíveis a possível dobradura e dilaceração. Por outro lado, Scanners do tipo *FLAT* (vidro sobre o qual se coloca a peça a ser scaneada) são inadequados devido ao peso do caderno e à sua manipulação para a captura da página e torção para posicionamento no scanner. A fotografia digital é hoje o método menos agressivo frente à extrema debilidade daquelas páginas devido à ação do tempo, mau uso e degeneração devido às intempéries naturais (umidade, fungos e outros).

Finalmente, mesmo com as melhores ferramentas atuais, os textos que saem do OCR são inutilizaveis diretamente sem um trabalho de pré-processamento.

2.2.1 Pré-processamento. Os passos detalhados para a montagem do *corpus* são:

(1) Captura das imagens - fotografia digital - digitalização
(2) Preparação para o OCR
    - Aplicação de Nomenclatura estruturada
    - Ajustes dos pixels (binarização X grayscale)
    - Normalização das imagens (Ajustes para tamanho padrão)
(3) OCR - aplicação
    - Geração de arquivos TXT de trabalho
    - Geração dos arquivos PDF imagem da página
    - Separação das subimagens em subdiretórios - quando possível
(4) Limpeza dos arquivos texto TXT
    - Retirar *stopwords* (em português)
    - Identificar e selecionar substantivos e verbos [3]
        – Verbos - trazer para o infinitivo
        – Substantivos - trazer para a forma normal retirando
            * Plurais
            * Aumentativos e diminutivos
        – Arcaísmos - conversão para a estrutura atual [4]

Neste trabalho, a expressão *arcaísmo* refere-se à conversão a ser aplicada a uma palavra visando converte-la para sua escrita atual. Por exemplo, a palavra *pharmacia*, deve ser transformada em *farmácia*.

Por tratar-se de material impresso com mais de 50 anos, muitas palavras sofreram alteração em sua escrita. A conversão associada ao *arcaísmo* citado será baseada na forma como certas palavras eram escritas à época da publicação do jornal, a partir do resgate da palavra em dicionários da época dos jornais, devidamente estruturados e associados à base de dados. Esse é um diferencial desse trabalho em relação aos congêneres.

2.3 Extração da Estrutura de Tópicos

Uma vez concluída a preparação do *corpus C*, a abordagem de Tópicos Induzidos proposta e utilizada neste trabalho funciona conforme os passos descritos a seguir. Na literatura há um número grande de propostas para rotulação de tópicos. O leitor interessado em outras abordagens é direcionado a consultar as seguintes referências [16–19]

Inicialmente, supõem-se disponível como entrada o número de tópicos de interesse e uma palavra **semente** para cada tópico, criteriosamente escolhida. Essa definição do número de tópicos e das sementes pode ser influenciada pelo conhecimento do contexto ou do domínio de aplicação de interesse. Por exemplo, na história do

[3] Por convenção, utilizou-se palavras apenas desses dois tipos gramaticais. Adjetivos, advérbios, preposições, conjunções, interjeições e outros, para efeito desse trabalho, não agregariam semântica mínima necessária.
[4] É um grande diferencial deste trabalho em relação aos afins essa conversão. Não foi encontrado esse recurso nos trabalhos em português pesquisados. Palavras sofreram alterações consideráveis em sua escrita ao longo das reformas ortográficas e vivência da língua. Esse processo possibilita trazer para a mesma raiz palavras de grafias distintas, a partir de dicionários históricos usados como base.

Ceará vão aparecer certos tópicos e palavras que não apareceriam na história do Paraná. O processo está descrito no Algoritmo 1.

Primeiro, o algoritmo LDA [20] é aplicado ao corpus com o número de tópicos de interesse N ou um múltiplo seu fixado como parâmetro (poderia não ser). O retorno do LDA é o conjunto das palavras que compõe cada tópico com os seus pesos relativos no tópico. Note que uma mesma palavra pode está em diferentes tópicos porém normalmente com pesos diferentes. É possível que este procedimento não supervisionado retorne uma estrutura de tópicos na qual uma ou mais palavras *semente* não conste com peso relevante em nenhum dos tópicos. Nesse caso o processo é repetido com um número maior de tópicos até se obter uma configuração desejada. Este é o laço em i no Algoritmo 1.

A lógica do processo destes Tópicos Induzidos aqui introduzidos é o de uma expansão de conculta (*Query Expansion*) [21] em que cada tópico seja representado por um conjunto de K palavras que não se repetem em outros tópicos. Para obtê-las utilizou-se o seguinte algoritmo de busca guloso aplicado aos resultados do LDA: para cada palavra semente, ele procura em qual tópico aquela palavra tem maior peso. Esse tópico é rotulado com aquela palavra e o processo prossegue até rotular todos os tópicos. Ao final desse passo temos cada semente rotulando um ou mais tópicos do LDA. Cada conjunto de tópicos LDA rotulados pela mesma semente forma um **Tópico Induzido**.

Em seguida toma-se a próxima palavra de maior peso em cada tópico. Se não há repetição entre os tópicos, cada uma é adicionada ao seu grupo. Cada palavra atribuída a um tópico é subtraída dos demais tópicos onde ela aparece com pesos menores para garantir a não repetição. Se uma mesma palavras é a de maior peso em dois ou mais tópicos, ela é alocada ao tópico onde o peso é maior. O processo se repete até que cada tópico esteja representado por K palavras, onde K é um valor pré-definido. Neste trabalho cada tópico foi representado por $K = 10$ palavras. Este procedimento está concretizado no Algoritmo 1.

O conjunto de palavras representando cada tópico torna-se a **assinatura** do tópico que é utilizada para consulta. Nesta fase adota-se a representação TF-IDF para a conculta (*query*) e para o *corpus* e a consulta é realizada com base na similaridade cosseno. Os textos retornados são ordenados decrescente pela similaridade com a consulta e um limiar (*threshold*) mínimo na medida de similaridade é usado para limitar o número de resultados retornados. Supondo que $\mathbf{x}$ é o vetor que representa a assinatura do tópico e que $\mathbf{y}$ é um vetor que representa um documento, ambos no modelo TF-IDF, e que $||\mathbf{z}||$ é o módulo ou comprimento de $\mathbf{z}$, a similaridade cosseno entre esses dois documentos é definida por:

$$cos(\mathbf{x}, \mathbf{y}) = \frac{\mathbf{x}.\mathbf{y}}{||\mathbf{x}||.||\mathbf{y}||} \quad (1)$$

.

Para os experimentos deste trabalho foram definidos $N = 10$ tópicos. A lista dos tópicos utilizados está na Tabela 1. E a Tabela 2 apresenta a lista das palavras que compõem os tópicos 'eleição' e 'educação'. Esses dois tópicos serão utilizados na Seção 4 de resultados.

Table 1: Lista dos 10 tópicos utilizados nos experimentos.

educação	eleição	seca	ceará	cangaço
vaquejada	caatinga	fortaleza	cariri	escravidão

Table 2: Lista das palavras únicas que caracterizaram os tópicos 'eleição' e 'educação'.

eleição	fortaleza, estado, rio, brasil, governo, presidente, eleição, casa, capital, nordeste
educação	educacao, vida, padre, igreja, cidade, povo, deus, praça, pais, nacional.

Algoritmo 1: O algoritmo proposto para a geração do Tópicos Induzidos.

Entrada: Um corpus C de documentos texto, um conjunto de $S = \{s_1, ..., s_N\}$ de N palavras semente, uma para cada tópico, e o número K de palavras por tópico desejado.
Saída: N conjuntos de K palavras, cada conjunto contendo uma semente de tópico, sem repetição de palavras.
início
  para *i = 1 à 10* **faça**
    Obtenha o LDA com $i \times N$ tópicos;
    se *todas as sementes aparecem como palavras relevantes em algum dos tópicos* **então**
      Atribuir cada tópico do LDA a uma semente;
      Completar o conjunto de K palavras de cada tópico sem repetição com base nas palavras e nos pesos retornados do LDA.
    fim
    Saia (*break*);
  fim
  Retornar insucesso.
fim

2.4 Interface de Visualização e Recuperação de Documentos

O acervo catalogado no ***Portal da História do Ceará*** é composto de livros, revistas e jornais e conta atualmente com cerca de 300 mil documentos. Para as buscas (*queries*) em livros e revistas são utlizadas estruturas padrão de busca por palavra nos textos ou nos meta-dados (*selects* por chaves).

Para buscas em páginais de jornais, atualmente existe três métodos de navegação:

(1) ***Padrão*** - onde as páginas são colocadas em uma listagem por pesquisa simples;
(2) **Linha do Tempo** - as páginas são encontradas e classificadas por data de publicação
(3) **Mapa Mental** - onde as páginas são apresentadas navegando dentro de agrupamentos documentais, que são uma tentativa intuitiva anterior de definir tópicos *ad hoc*, e que neste trabalho está recebendo uma abordagem sistemática. .

Esses três métodos de consulta são apoiados em uma estrutura inicial de pesquisa primária - envolvendo chaves primárias básicas. A consulta por tópico é uma consulta semanticamente mais rica que será agregada às essas três existentes. Com a implementação desta consulta por tópicos, haverá diferença entre buscar pela **palavra** educação ou pelo **tópico** educação.

Existem diversas propostas na literatura para a visualização da estrutura de tópicos de uma coleção de documentos texto [22–24] e que deverão ser consideradas na evolução futura deste trabalho.

A implementação inicial atualmente em desenvolvimento é uma interface simples na qual um catálogo de tópicos é apresentado ao usuário e ele escolhe o tópico de pesquisa desejado. Embora menos flexível que uma busca por tópicos livremente definidos pelo usuário, esta abordagem tem a vantagem de oferecer como retorno uma maior qualidade da cobertura e do ranqueamento dos documentos.

3 TRABALHOS RELACIONADOS

Nesta seção apresenta-se uma breve revisão de trabalhos diretamente relacionados a este projeto. Sendo este um projeto sistêmico constituído pela composição de múltiplos algoritmos como segmentação de imagem, OCR, classificação de texto e modelos tópicos, nota-se que cada uma destas tarefas poderia gerar sua própria revisão de literatura. Assim optou-se por focar essa seção apenas em trabalhos sistêmicos [25, 26] e em trabalhos de modelos tópicos [17, 27], que é onde estão as principais contribuições deste projeto.

Em relação ao projeto sistêmico, os trabalhos mais próximos deste são [4, 25]. Em [25], Allen apresenta um arcabouço para processamento de imagens de páginas de jornais históricos muito semelhante àquele usado neste projeto nas etapas iniciais. Entretanto o autor não propõe formas de organizar ou visualizar o conhecimento nas etapas finais.

O trabalho [4] descreve um processo para a criação de uma interface para acesso aos arquivos de dois jornais antigos Suiços baseado em vários passos de processamento textual incluindo indexação, computação de n-gramas e reconhecimento de entidades e terminando com uma interface web para acesso pelos usuários finais. Algumas das técnicas de processamento textual também são usadas neste projeto.

Já em relação à modelagem tópica rotulada, os trabalhos mais próximos deste são [16, 28]. Em [28] é proposto o algoritmo de palavras-âncora *(anchor-words)* para a formulação de tópicos com significado. O método infere um modelo tópico encontrando um *convex hull* das palavras em co-ocorrência no espaço de alta dimensionalidade. Por outro lado, o trabalho [16] parte deste anterior e propõe uma versão que projeta os dados em um espaço bidimensional para obter uma solução aproximada que, segundo o autor, melhora a clareza dos tópicos e mostra aos usuários porque o algoritmo escolhe certas palavras.

O ponto fraco destes trabalhos, registrado na literatura, é que o algoritmo que escolhe as palavras-âncora frequentemente escolhe palavras inadequadas reduzindo muito a eficácia do método. A proposta deste projeto para contornar essa fraquesa é pela criteriosa escolha *ad hoc* antecipada da semente para cada tópico. Denomina-se este método de *tópicos induzidos*. Como descrito em seção anterior, tópicos induzidos trabalha sobre os resultados do algoritmo LDA tirando proveito do arcabouço teórico deste método.

4 EXPERIMENTOS E RESULTADOS

Esta seção descreve os experimentos de prova de conceito e uma discussão dos resultados. Foram realizadas duas consultas de teste, uma com a semente "eleição" e outra com a semente "educação".

Para exemplificar casos de interesse de pequisa nestes tópicos, no período do *data set* considerado, de 1922 a 1964, aconteceram 12 eleições a governador no estado do Ceará [5] e o interesse último do pesquisador na consulta por "eleições" poderia ser o de verificar qual foi o viés ideológico assumido pela igreja nesse período já que O NORDESTE foi um jornal editado pela Igreja Católica. No *ground truth* foram encontrados 20.684 textos onde aparece a palavra eleição no *data set*.

Já na segunda consulta, utilizando a palavra semente "educação", o interesse do pesquisador poderia ser investigar se seria possível inferir dos textos publicados qual era a corrente pedagógica predominante nos colégios confessionais da Igreja Católica. Na época estudada, a Igreja era detentora das mais importantes escolas no estado do Ceará. No *ground truth* foram encontrados 67.282 textos onde aparece a palavra educação no *data set*.

Após a descrição do data set, os resultados serão apresentados em duas subseções, uma para a consulta "eleição" e outra para a consulta "educação". Cada seção apresenta a matriz de confusão resultante da consulta e o ranqueamento obtido para os textos classificados como positivos. Estes resultados são complementados por uma discussão dos falsos positivos e falsos negativos obtidos.

4.1 Descrição do data set

O *corpus* utilizado para este trabalho foi integralmente construído pelos autores. Representa o resgate fotográfico (digital) de 36.617 imagens de páginas do Jornal ***O NORDESTE***, publicado pela Arquidiocese de Fortaleza - CE, durante o período de 1922 a 1964. As principais informações e estatísticas do *corpus* estão apresentadas na Tabela 3.

A baixa qualidade do material em papel, além da tipologia desgastada pelo tempo e do material propriamente dito usado para o papel jornalístico dificultaram a captura das palavras.

Para quantificar essas noções, um exemplo de página típica foi tomada aleatoriamente. A subimagem de texto utilizada possui 532 palavras, e o OCR conseguiu resgatar 318 palavras, representando quase 60%. Esse percentual melhora para 78% nas últimas edições (1960 e 1964) e reduz-se a menos de 50% nas edições da década de 1920, por força da qualidade do papel e do desgaste das letras.

Isso, no entanto, não quer dizer que 60% das palavras são úteis, pois deve-se aplicar ainda as correções de arcaísmos, acentuações invertidas, traços de quebra de linha e outros. Como a grafia da época era bem diferente, então o ajuste do *corpus* foi algo importante a ser feito, e representa forte diferencial deste trabalho em relação aos *corpora* pré-processados encontrados em *datasets* abertos.

Note da Tabela 3 que o tópico 'eleição' consta em 20684 textos e que o tópico 'educação' consta em 67282 textos, de um total de

[5] Segundo site do Governo do estado. Além dessas eleições, alguns governadores foram nomeados.

Table 3: Edições e estatísticas do *corpus* utilizado nos expe-rimentos de prova de conceito.

ano de publi-cação	num. de páginas	num. de textos no corpus	num. de textos em 'eleição'	num. de textos em 'educação'
1922	468	93116	148	1259
1923	233	48020	59	480
1924	298	105665	531	823
1925	556	244702	258	1880
1926	655	243847	286	1062
1927	1281	330273	708	2129
1928	898	163390	560	1419
1932	2500	294637	1336	5131
1933	2519	345025	2267	5785
1934	1537	211284	1095	3686
1935	2304	372532	1413	3400
1936	1634	269504	887	2543
1942	1742	225402	199	2877
1943	1459	255352	230	2888
1944	297	52417	40	651
1945	2374	423950	1442	4282
1946	2545	425777	1429	3728
1947	1492	231261	1432	1857
1952	2295	409666	817	3999
1953	981	170237	337	1557
1954	1216	183824	1269	1891
1955	1473	202448	1450	2472
1956	2364	359375	1018	3832
1957	690	111271	278	1123
1959	276	40545	165	511
1960	847	123985	396	2272
1961	1337	186827	455	2650
1964	346	42720	179	1095
Total	36617	6167052	20684	67282

6167052 textos do corpus, o que representa 0,34 % e 0,11 %, respectivamente. Isso impõe uma tarefa de categorização fortemente desbalanceada.

Os números na Tabela 3 referem-se a todo tipo de texto que foi segmentado. Isso inclui textos curtos ou longos, frases soltas, pequenos anúncios ou fragmentos. Ao final das etapas de aquisição e preprocessamento cada texto recebeu uma marca de identidade (ID) com a seguinte sintaxe: PyyyTxx representa o Texto número xx da Página yyy. As páginas foram numeradas em sequência cronológica e os textos na ordem em que aparecem na página. Assim, fica simples recuperar os textos que compõem uma página de interesse.

4.2 Consulta pela semente "eleição"

A primeira consulta de teste para avaliação do sistema foi pelo tópico 'eleição'. Para esta consulta, ajustou-se experimentalmente o limiar da similaridade cosseno para retornar um número pequeno de resultados. Estes são experimentos de classificação *one-class* altamente desbalanceados. Sendo assim, calculou-se aqui apenas a métrica *Precision*. As outras duas métricas comummente utilizadas para avaliar algoritmos em Recuperação de Informação, *Recall* e $F1 - measure$, não são úteis de se calcular nesse contexto. Ambas são evidentemente muito baixas.

A Tabela 4 mostra a matriz de confusão obtida neste teste para $cos(\mathbf{x}, \mathbf{y}) \geq 0,82$. Da Tabela 3 nota-se que 'eleição' consta em 20.684 textos do corpus e a matriz de confusão mostra que o procedimento recuperou 102 textos no total. O índice de performance *precision* resultou em $precision = 88/102 = 0,8627$ ou 86,27 %.

Por outro lado, Tabela 5 mostra a ordenação decrescente de relevância dos 10 primeiros textos recuperados como positivos (total da primeira coluna da matriz de confusão) neste teste. O *ground truth* nesta tabela foi realizado a *posteriori* lendo os textos recuperados. Vê-se que a maioria dos textos recuperados de fato tratam do tópico consultado. A precisão top-10 para a consulta eleição, entretanto foi de $precision = 7/10 = 0,7$ ou 70 %.

Table 4: Matriz de confusão para o experimento com o tópico 'eleição'.

		tópico predito		
		p	**n**	**total**
tópico verdade	**p′**	88 (VP)	20596 (FN)	20684
	n′	14 (FP)	6146354 (VN)	6146368
	total	102	6166950	6167052

Table 5: *Ground truth* e ranque gerado pelo método para os 10 textos preditos positivos melhor ranqueados para o tópico 'eleição'. Legenda: PyyTxx = Texto xx da Página yy.

ID do texto	ranque	*ground truth*
P00098T012	1	eleição
P00134T007	2	eleição
P11346T122	3	eleição
P20365T114	4	outro
P01016T026	5	eleição
P31201t100	6	eleição
P21954T048	7	eleição
P02565T048	8	eleição
P20450T029	9	outro
P10882T028	10	outro

4.3 Consulta pela semente "educação"

A segunda consulta de teste do sistema foi pelo tópico 'eleição'. Também para este caso, ajustou-se experimentalmente o limiar da

similaridade cosseno para retornar um número pequeno de resultados.

A Tabela 6 mostra a matriz de confusão obtida neste teste para $cos(\mathbf{x}, \mathbf{y}) \geq 0,88$. Da Tabela 3 nota-se que 'eleição' consta em 67.282 textos do corpus e a matriz de confusão mostra que o procedimento recuperou 235 textos no total. O índice de performance *precision* resultou em *precision* = 222/235 = 0, 9464 ou 94,64 %.

Por outro lado, Tabela 7 mostra a ordenação decrescente de relevância dos 10 primeiros textos recuperados como positivos (total da primeira coluna da matriz de confusão) neste teste. O *ground truth* nesta tabela foi realizado a *posteriori* lendo os textos recuperados. Vê-se que, também neste teste, a maioria dos textos recuperados de fato tratam do tópico consultado. A precisão top-10 para a consula eleição, entretanto foi de *precision* = 8/10 = 0, 8 ou 80 %.

Table 6: Matriz de confusão para o experimento com o tópico 'educação'.

		tópico predito		
		p	**n**	**total**
tópico verdade	**p′**	222 (VP)	67060 (FN)	67282
	n′	13 (FP)	6099979 (VN)	6099992
	total	235	6166817	6167052

Table 7: *Ground truth* e ranque gerado pelo método para os 10 textos preditos positivos melhor ranqueados para o tópico 'educação'. Legenda: PyyTxx = Texto xx da Página yy.

ID do texto	ranque	*ground truth*
P23077T111	1	educação
P09115T103	2	educação
P01299T045	3	outro
P11886T006	4	outro
P34368T205	5	educação
P09224T056	6	educação
P34366T098	7	educação
P18328T115	8	educação
P12676T048	9	educação
P11452T139	10	educação

Em resumo, o que os resultados das Tabelas 4 e 5 para o primeiro teste e das Tabelas 6 e 7 para o segundo teste mostram é que a abordagem adotada é promissora mas que há ainda ampla margem para melhoria de desempenho na acurácia geral do conjuntos de algoritmos.

5 CONCLUSÃO

A concepção e os principais métodos, algoritmos, tecnologias e conceitos adotados no projeto de um Sistema de Processamento e Navegação por Tópicos em Imagens de Páginas de Jornais Históricos foram descritos e os resultados de uma avaliação prova de conceito foram apresentados e analisados. Além da contribuição com o projeto sistêmico este trabalho propôs e avaliou preliminarmente uma abordagem semi-supervisionada própria para o problema da geração e organização dos assuntos por tópico.

Especificamente, partindo de uma palavra semente por tópico, o algoritmo alarga a cobertura do tópico pós-processando a saída do LDA, buscando de maior peso, sem repetição, para construir **assinaturas** para os tópicos. Com as assinaturas construídas, similaridade cosseno entre as representações TF-IDF da assinatura e do *corpus* são utilizados para recuperar os documentos mais relevantes.

Uma terceira contribuição deste trabalho foi a construção de um *data set* próprio, ou seja, um *corpus* para os testes. Ele representou o resgate fotográfico (digital) de 31.717 imagens de páginas do Jornal ***O NORDESTE***, publicado pela Arquidiocese de Fortaleza - CE, durante o período de 1922 a 1964.

A avaliação prova de conceito apresentou resultados animadores. Entretanto, entre as limitações deste trabalho está o pequeno volume de testes realizados, representado por dois tópicos. Apenas quando o sistema for colocado em operação aberta aos usuários reais é que se terá índices reais de desempenho.

A continuidade deste trabalho vai em duas direções. Primeiro, vão ser realizados um conjunto maior de testes utilizando, além da coleção do Jornal O NORDESTE (de 1922 à 1964), coleções de outros jornais disponíveis. Segundo, pretende-se trabalhar também na melhoria dos algoritmos utilizados nas etapas intermediárias do processo.

REFERENCES

[1] Chinmay Tumbe. Corpus linguistics, newspaper archives and historical research methods. *Journal of Management History*, 2019.
[2] Robert B Allen and Robert Sieczkiewicz. How historians use historical newspapers. *Proceedings of the American Society for Information Science and Technology*, 47(1):1–4, 2010.
[3] Myriam C Traub, Jacco Van Ossenbruggen, and Lynda Hardman. Impact analysis of ocr quality on research tasks in digital archives. In *International Conference on Theory and Practice of Digital Libraries*, pages 252–263. Springer, 2015.
[4] Yannick Rochat, Maud Ehrmann, Vincent Buntinx, Cyril Bornet, and Frédéric Kaplan. Navigating through 200 years of historical newspapers. *iPRES 2016*, page 186.
[5] Edwin Klijn. The current state-of-art in newspaper digitization. *D-Lib Magazine*, 14(2), 2008.
[6] Laurence Likforman-Sulem, Abderrazak Zahour, and Bruno Taconet. Text line segmentation of historical documents: a survey. *International Journal of Document Analysis and Recognition (IJDAR)*, 9(2-4):123–138, 2007.
[7] Barry Popik. Digital historical newspapers: A review of the powerful new research tools. *Journal of English Linguistics*, 32(2):114–123, 2004.
[8] Lyne Da Sylva. Nlp and digital library management. In *Emerging Applications of Natural Language Processing: Concepts and New Research*, pages 265–290. IGI Global, 2013.
[9] Andrey Shapenko, Vladimir Korovkin, and Benoit Leleux. Abbyy: the digitization of language and text. *Emerald Emerging Markets Case Studies*, 2018.
[10] Amer Dawoud and Mohamed S Kamel. Iterative multimodel subimage binarization for handwritten character segmentation. *IEEE Transactions on Image Processing*, 13(9):1223–1230, 2004.
[11] Showmik Bhowmik, Ram Sarkar, Mita Nasipuri, and David Doermann. Text and non-text separation in offline document images: a survey. *International Journal on Document Analysis and Recognition (IJDAR)*, 21(1-2):1–20, 2018.
[12] Maroua Mehri, Petra Gomez-Krämer, Pierre Héroux, Alain Boucher, and Rémy Mullot. Texture feature evaluation for segmentation of historical document images. In *Proceedings of the 2nd International Workshop on Historical Document*

Imaging and Processing, pages 102–109. ACM, 2013.

[13] Ankit Kumar Sah, Showmik Bhowmik, Samir Malakar, Ram Sarkar, Ergina Kavallieratou, and Nikos Vasilopoulos. Text and non-text recognition using modified hog descriptor. In *2017 IEEE Calcutta Conference (CALCON)*, pages 64–68. IEEE, 2017.

[14] Ilia V Safonov, Ilya V Kurilin, Michael N Rychagov, and Ekaterina V Tolstaya. Segmentation of scanned images of newspapers and magazines. In *Document Image Processing for Scanning and Printing*, pages 107–122. Springer, 2019.

[15] Jiří Martínek, Ladislav Lenc, and Pavel Král. Training strategies for ocr systems for historical documents. In *IFIP International Conference on Artificial Intelligence Applications and Innovations*, pages 362–373. Springer, 2019.

[16] David Mimno and Moontae Lee. Low-dimensional embeddings for interpretable anchor-based topic inference. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1319–1328, 2014.

[17] Xiao Fu, Kejun Huang, Nicholas D Sidiropoulos, Qingjiang Shi, and Mingyi Hong. Anchor-free correlated topic modeling. *IEEE transactions on pattern analysis and machine intelligence*, 41(5):1056–1071, 2018.

[18] Anni Järvelin, Heikki Keskustalo, Eero Sormunen, Miamaria Saastamoinen, and Kimmo Kettunen. Information retrieval from historical newspaper collections in highly inflectional languages: A query expansion approach. *Journal of the Association for Information Science and Technology*, 67(12):2928–2946, 2016.

[19] João Marcos Carvalho Lima and José Everardo Bessa Maia. A topical word embeddings for text classification. In *Anais do XV Encontro Nacional de Inteligência Artificial e Computacional*, pages 25–35. SBC, 2018.

[20] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022, 2003.

[21] Fabiano T. Silva and José E. B. Maia. Query expansion in text information retrieval with local context and distributional model. *Journal of Digital Information Management*, 17(6):313–320, 2019.

[22] Bénédicte Le Grand and Michel Soto. Topic maps visualization. In *Visualizing the Semantic Web*, pages 49–62. Springer, 2003.

[23] Changhong Zhang, Zeyu Li, and Jiawan Zhang. A survey on visualization for scientific literature topics. *Journal of Visualization*, 21(2):321–335, 2018.

[24] Mohammad Alharbi and Robert S Laramee. Sos textvis: An extended survey of surveys on text visualization. *Computers*, 8(1):17, 2019.

[25] Robert B Allen, Andrea Japzon, Palakorn Achananuparp, and Ki Jung Lee. A framework for text processing and supporting access to collections of digitized historical newspapers. In *Symposium on Human Interface and the Management of Information*, pages 235–244. Springer, 2007.

[26] Nathan Yarasavage, Robin Butterhof, and Christopher Ehrman. National digital newspaper program: a case study in sharing, linking, and using data. In *Proceedings of the 12th ACM/IEEE-CS joint conference on Digital Libraries*, pages 399–400. ACM, 2012.

[27] Tze-I Yang, Andrew Torget, and Rada Mihalcea. Topic modeling on historical newspapers. In *Proceedings of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*, pages 96–104, 2011.

[28] Sanjeev Arora, Rong Ge, and Ankur Moitra. Learning topic models–going beyond svd. In *2012 IEEE 53rd Annual Symposium on Foundations of Computer Science*, pages 1–10. IEEE, 2012.