Retrieving and Processing Images from the Pages of a Historical Newspaper and Modeling the Text Topics

Gildácio J. de A. Sá, José E. B. Maia
Universidade Estadual do Ceará – UECE
Ciência da Computação - CCT
60714-903 - Fortaleza - Ceará - Brasil
{gildacio.sa@gmail.com} {jose.maia@uece.br}

Journal of Digital
Information Management

ABSTRACT: *Historical newspapers are a source of research for the human and social sciences. However, these image collections are difficult to read by machine due to the low quality of the print, the lack of standardization of the pages in addition to the low quality photograph of some files. This paper presents the processing model of a topic navigation system in historical newspaper page images. The general procedure consists of four modules which are: segmentation of text sub-images and text extraction, preprocessing and representation, induced topic extraction and representation, and document viewing and retrieval interface. The algorithmic and technological approaches of each module are described and the initial test results about a collection covering a range of 28 years are presented.*

Subject Categories and Descriptors: [I.2.7 Natural Language Processing];Text analysis: **[I.4 IMAGE PROCESSING AND COMPUTER VISION]** [H.3.3 Information Search and Retrieval]

General Terms: Text Processing, Image Processing, Newspaper Analysis, Information Retrieval

Keywords: Historical Newspapers, Lexical Standardization, Induced Topic Model, Information Retrieval, Natural Language Processing

DOI: 10.6025/jdim/2021/19/2/41-46

Received: Text Processing, Image Processing, Newspaper Analysis, Information Retrieval

Review Metrics: Review Scale: 0-6, Review Score: 4.55, Inter-reviewer consistency: 78.2%

1. Introduction

The job of retrieving the content of historical newspapers stored in photographic images and making their content accessible for viewing by topic is described. Since each historical newspaper is a closed collection of documents, pre-organizing content by labeled topics is an effective approach to cyber access to that content.

Newspaper texts have the primacy of being contemporary to the facts, being rich in micro-details and circumstantial details to the fact reported. However, due to the urgency of its publication, the critical and contextualized analysis of events only occurs a posteriori, being the object of social study. Thus, historical newspapers preserve a rich moment of that society that must be preserved and rescued for analysis. That is why there are large collections of historical newspapers in the form of page images around the world which are of interest to anthropologists, sociologists and historians in general [16, 12, 13]. However in this way of storing they are costly and tedious to consult.

Transforming a collection of historic newspapers for digital access and consultation is challenging in many ways. In the image in Figure 1, which shows one of these pages, you can see the main causes of this difficulty [5].

The original pages on paper have been worn out due to handling and in general are dry and brittle so that the photograph of the pages is the first step most used because it is the least invasive. Due to factors such as lack of standardization, text misalignment, page wear and poor quality of photographs in some cases, the OCR (Optical Character Recognition) [10] process often returns incomplete words, noises of various types and disconnected passages.

As for the natural language processing tasks, it presents other challenges. Note that a regional newspaper collection is a closed collection and as such may have a limited vocabulary resulting in searches using a broad

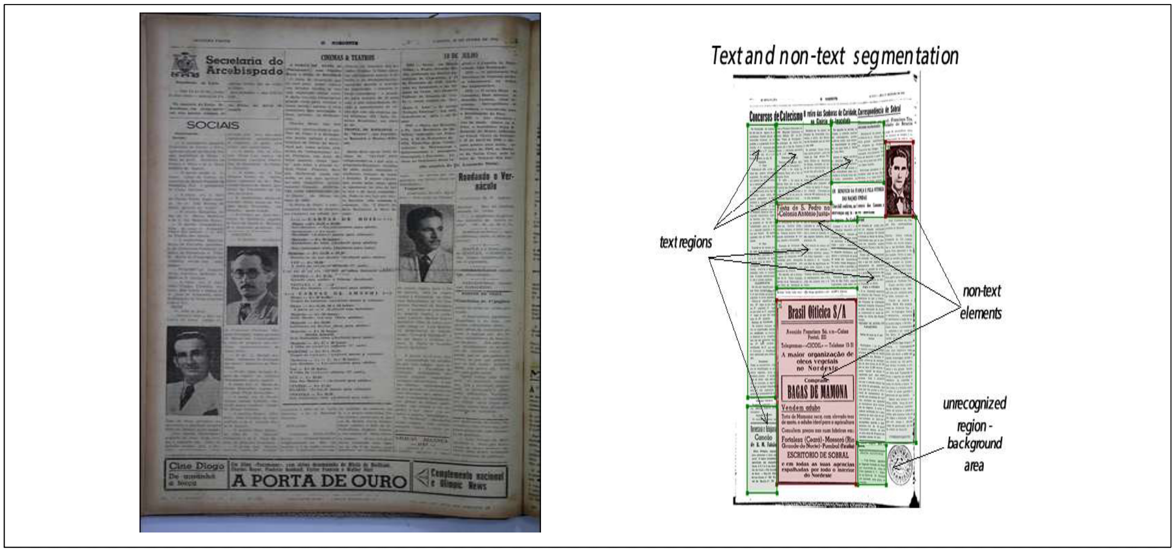
Text and non-text segmentation
text regions
non-text elements
unrecognized region - background area

Figure 1. Left: Journal O NORDESTE, from the Archdiocese of Fortaleza-CE-Brasil on July 10, 1943 (in Portuguese). Rigth: Journal O Nordeste of 7/1/1943, p. 4 - Example of identifying text and non-text regions in an image

universal vocabulary that may not retrieve existing relevant documents.

Because the collection of texts usually covers several decades of years, the different contexts in which the same subject appears over time creates a situation of multiples contexts (multimodality of contexts) in the statistical representation of a topic. Unsupervised topical modeling such as LDA (Latent Dirichlet Allocation) [4] is useful in exploratory search however we need a search directed by the user's interest represented by a query. The researcher has a specific query that may not be frequent in the collection but whose occurrence is important and will not appear as a topic in LDA topical modeling. These observations led us to develop a technique that we call induced topics in which the user provides one or more seed terms that should guide the characterization of the topic of interest. The induced topic procedure is performed as a post-processing stage of the LDA and returns a vector of words representing the topic signature which is used to retrieve the fragments of text of interest.

In continuity, Section 2 describes the general procedure and algorithms, Section 3 is about experiments and results and the conclusion comes in Section 4.

2. Methods

Corpus Construction: Figure 2 shows the functional blocks and information flows of the general procedure. It is assumed that a collection of images from indexed historical newspaper pages is available from which a collection of documents indexed via OCR [10] is obtained. This is a complex step involving image segmentation, border extraction, extraction of text sub-images with different types and sizes of characters and finally the optical character recognition [3]. The effectiveness of this phase can be as low as 60% due to the low overall quality of the original papers from which the page images were obtained. However, the performance of the entire process ahead depends on this step. This step was supported by the commercial software AbbyyFine Reader CE © [14].

The output of OCR faces a preprocessing step that includes lexical standardization [7], noises and stopwords removal, and stemming [9]. Lexical standardization was necessary because a newspaper's database spans decades of years during which the lexicon undergoes significant changes. The output is each text document represented as a list of lexical items with the entire collection forming a list of item lists. From this list, the TF-IDF (term frequency - inverse document frequency) [9] representation in the form of a term-document matrix is obtained for the entire collection (right branch in the Figure 2).

Induced Topics: The approach here is algorithmic. The reader interested in a principled approach to labeling topics should consult [2, 11, 17, 8]. The intuition behind the induced topic algorithm is to increasingly fragment the unsupervised topics obtained by LDA and regroup them by inspection, using the seed words, until you get an outline close to the target structure. This procedure is formally described in Algorithm 1.

Initially, the number of topics of interest N and one or more seed words for each topic are assumed to be available. The definition of the number of topics of interest and the seeds comes from knowledge of the context or domain of application. The search for a single induced topic is an instance. Note that the procedure can also be applied by defining more than one seed word for each topic.

First, the LDA algorithm [4] is applied to the corpus with the number of topics $M = n{\times}N$ as the input parameter. This fragmentation of LDA topics aims to model the statistical multimodality of context in the corpus. The return of the LDA is the set of words that make up each topic with their relative weights in the topic. Note that the same word can be on different topics but usually with different weights. It is possible that this unsupervised procedure returns an outline in which one or more seed words do not have a relevant weight in the LDA topics. In this case, the process is repeated with an increasing number of topics, increasing *n*, until a desired configuration is obtained.

The Induced Topic introduced here functions as a Query Expansion [9, 15] in which each topic is represented by a signature with *K* words with the words seeds not repeated in other topics. To obtain them, we used the following greedy search algorithm applied to the results of the LDA: for each seed word, it searches on which topics the word is the most important. These LDA topics are labeled with that word and the process continues until you label all topics.

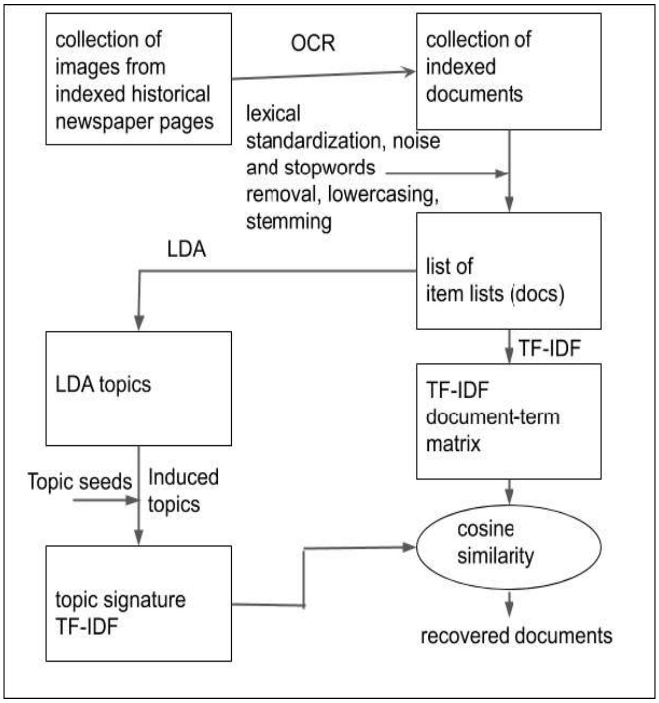
collection of images from indexed historical newspaper pages
OCR
collection of indexed documents
lexical standardization, noise and stopwords removal, lowercasing, stemming
LDA
list of item lists (docs)
TF-IDF
LDA topics
TF-IDF document-term matrix
Topic seeds
Induced topics
cosine similarity
topic signature TF-IDF
recovered documents

Figure 2. Flow diagram for retrieving and processing images from historical newspaper pages and modeling text topics

Then, each set of LDA topics labeled by the same seed forms an Induced Topic. Then the next most important word in each topic is taken. If there is no repetition between topics, each one is added to your group's signature. Each word assigned to a topic is subtracted from the other topics where it appears with lower weights to ensure mutual exclusion. If the same word has the highest weight in two or more groups, it is allocated to the group where the weight is greatest. The process is repeated until each topic is represented by *K* words, where *K* is a predefined value. In this work, each topic was represented by *K* = 10 words. Correction in this general process is included to handle exceptions.

Information retrieval: It begins by representing the signature of the topical query in TF-IDF in the space of terms in which the corpus was represented. The cosine similarity measure [9] given by

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

is then used to rank documents by importance. In this equation, x and y represent the TF-IDF vectors for the topical query and the document respectively, and x:y represents the dot product. An additional feature is that a minimum number of terms in each document can be set to filter documents of interest by size. This characteristic proved to be useful for a fragmented corpus obtained by OCR on the poor quality originals of historical newspapers.

3. Related Works

This section presents a brief review of works directly related to this project. This being a systemic project consisting of the composition of multiple algorithms such as image segmentation, OCR, text classification and topical models, it is noted that each of these tasks could generate its own literature review. So it was decided to focus this section only on systemic works [1, 19] and works on topical models [18, 6], which is where the main contributions of this project are.

Regarding the systemic project, the works closest to it are [1, 13]. In [1], Allen presents a framework for image processing of historical newspaper pages very similar to that used in this project in the early stages. However, the author does not propose ways to organize or visualize knowledge in the final stages.

The work [13] describes a process for creating an interface for accessing the archives of two old Swiss newspapers based on several textual processing steps including indexing, computation of n-grams and entity recognition and ending with a web interface for access by end users. Some of the text processing techniques are also used in this project.

In relation to labeled topical modeling, the works closest to this are [2, 11]. In [11] it is proposed the anchor-words algorithm for the formulation of meaningful topics. The method infers a topical model by finding a hull convex of the words in co-occurrence in the high dimensional space.

On the other hand, the work [2] starts from this previous one and proposes a version that projects the data in a two-dimensional space to obtain an approximate solution that, according to the author, improves the clarity of the topics and shows users why the algorithm chooses certain words.

The weakness of these works, recorded in the literature, is that the algorithm that chooses the anchor words often chooses inappropriate words, greatly reducing the effectiveness of the method. This project's proposal to overcome this weakness is due to the careful ad hoc early choice of seed for each topic.

This method is called induced topics. As described in the previous section, induced topics work on the results of the LDA algorithm taking advantage of the theoretical framework of this method.

Algorithm 1: Pseudo-code for the induced topic algorithm. The algorithm is based on human assisted fragmentation and regrouping of LDA topics.

Data: C, a corpus of text documents, a set of $S = \{s_1, ..., s_N\}$ of N seed words, one for each topic, and the desired number
of words per topic K.

Result: N sets of K words, each set containing a topic seed word, with mutual exclusion.

```
for $n \leftarrow 1$ to 10 do
    Get the LDA with $n \times N$ topics;
    if $each\_seed\_word\_is\_a\_relevant\_term\_in\_any\_of\_the\_topics$ then
        Regroup: Assign each LDA topic to a seed word;
        Complete $K$ words per topic with mutual exclusion using the terms and weights of the LDA topics;
        break();
    end if
end for
return Return failure; % Refine the words seeds.
```

4. Data, Experiments and Results

In this section, proof-of-concept experiments are described and evaluated. The entire actual archive of images from the pages of a historic regional newspaper in Portuguese has been recovered and processed.

To evaluate the performance of the approach introduced here, a two-step procedure was built. First, in view of the impossibility of labeling hundreds of thousands of short and larger texts by subject, a topical modeling with fifty topics was obtained by the human assisted induced topic procedure described in Algorithm 1. This labeling was taken as ground truth.

Then two simple topic queries using a single seed word are performed and evaluated using two performance metrics: *precision* and *top*_20_*precision*. Precision is calculated using the ground truth obtained by induced topics and top-20-precision was calculated based on the manual examination of the texts by an expert. Top-20-precision provides precision in the first 20 best-ranked documents.

Precision and recall are two relevant metrics in Information Retrieval. They are calculated from the data of a Confusion Matrix by the equations:

$$precision = \frac{tp}{tp + fp}$$

$$recall = \frac{tp}{tp + fn}$$

where tp stands for True Positive, fp stands for False Positive and fn stands for False Negative. However, note that due to the impossibility of manual verification (labeling) of a large amount of texts and also due to the minimum threshold mechanism in the value of the cosine similarity imposed in the generation of the confusion matrix, the calculation of the recall is not significant in the context of this experiment, although it is relevant to the application in question. Thus, racall values are not shown. The applied cosine similarity threshold was selected to generate a manageable volume of data.

4.1. Data set

The corpus used for this work, in Portuguese language, was entirely built by the authors. It represents the photographic (digital) rescue of 36,617 images of pages from the **O NORDESTE** newspaper, published by the Catholic Archdiocese of Fortaleza-CE-Brazil, during the period from 1922 to 1964. We use quotes for Portuguese words.

The low quality of the paper originals, in addition to the typology worn by time, made it difficult to capture the words. To quantify these notions, on a typical page the text subimage taken at random has 532 words, OCR was able to rescue 318 words, representing almost 60%. This percentage improved to 78% in the last editions (1960 and 1964) and reduced to less than 50% in the editions of the 1920s, due to the quality of the paper and the wear of the letters.

This, however, still does not mean that 60% of the words are useful, since the lexical corrections and standardization, inverted accents, line breaks and others must still be applied. As the spelling of the time was very different, then the adjustment of the corpus was something important to be done.

For the topic ‘eleição’ 20, 684 texts were labeled- and for the topic ‘educação’ 67, 282 texts were obtained, out of a total of 6167052 texts in the corpus, which represents 0.34% and 0.11%, respectively. This imposes a heavily unbalanced text categorization task. These numbers refer to all types of text that have been targeted. They include short or long texts, loose sentences, small advertisements or fragments of text.

4.2. Topical Queries

Topical query for ‘eleição’: The first topical queryused the seed word ‘eleição’. For this query, the topic signature for $K = 10$ resulted in $S_{el} =$ ‘eleição’,‘partido’,‘seção’,‘mesa’,‘pleito’,‘voto’, ‘presidente’,‘ *chapa’,‘titulo’,‘urna’}* and the cosine similarity threshold was experimentally adjusted to return a small number of results. Since these are highly unbalance one-class classification experiments, only the *precision* metric was calculated here. The other two metrics commonly used to evaluate information retrieval algorithms, Recall and F1-measure, are not useful to calculate in this context. Both are evidently very low.

Table 1 shows the confusion matrix obtained in this test for $cos(x, y) \geq 0.82$. The precision performance in dex resulted in $precision = 88{=}102 = 0{:}8627$ or 86.27 %. On the other hand, when considering only the first 20 texts retrieved in order of relevance (top 20) as positive (total of the first column of the confusion matrix) in this test. The ground truth in this table was carried out a posteriori by reading the recovered texts. It can be seen that most of the recovered texts actually deal with the consulted topic. The top 20 precision for the ‘eleição’ query, however, was $top\_20\_precision = 17/20 = 0.85$ or 85%.

Topical query for ‘educação’: The topic signature for $K = 10$ resulted in S_{ed} = *{‘educação’,‘faculdade’,* ‘instituto’,‘nacional’,‘universidade’,‘curso’,‘direito’, *‘brasil’,‘trabalho’,‘grupo’}*. The confusion matrix obtained in this test for $cos(x, y) \geq 0.88$ is shown in Table 1. The precision performance index resulted in $precision = 222/235 = 0.9464$ or 94.64 %.

The first 20 texts retrieved as positive (total of the first column of the confusion matrix) were ordered decreasing in relevance. The ground truth was carried out a posteriori by reading the recovered texts. It was seen that, also in this test, most of the recovered texts actually deal with the topic consulted. The top-20 precision for the ‘educação’ query, however, was $top\_20\_precision = 19{=}20 = 0.95$ or 95 %.

5. Conclusion

The concepts and methods adopted in the design of a System of Processing and Navigation by Topics in Images of Pages of Historical Newspapers were described and the results of a proof of concept evaluation were presented and analyzed. In addition to the contribution to the systemic project, this work proposed and preliminarily evaluated a semi-supervised approach to the problem of the generation and organization of subjects by topic.

Specifically, starting from one or more seed words per topic, the algorithm extends the topic coverage by processing the LDA output to build the topic signature which will be used as a topical query.

A third contribution of this work was to build a new data set of images of pages of a historical newspaper through the photographic (digital) rescue of 36,617 page images of the **O NORDESTE** newspaper, published by the Catholic Archdiocese of Fortaleza-CEBrazil in the period from 1922 to 1964. The proof-ofconcept evaluation produced encouraging results.

References

[1] Robert B, Allen., Japzon, Andrea., Achananuparp, Palakorn., Ki Jung, Lee. (2007). 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.

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

[3] Showmik, Bhowmik., Ram, Sarkar., Mita, Nasipuri., and David, Doermann. (2018). 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.

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

[5] Gildácio José de Almeida Sá and José Ever-ardo Bessa Maia. (2020). Processamento e naveg-ação por tópicos em imagens de páginas de jornais históricos. Anais do Computer on the Beach, 11(1): 432–439, 2020.

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

[7] Anni, Järvelin., Heikki, Keskustalo., Eero, Sor- munen, Miamaria, Saastamoinen., and Kimmo, Kettunen. (2016). 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.

[8] João Marcos Carvalho Lima., José Ever-ardo Bessa Maia. (2018). A topical word embeddings for text classifi-

fication. In *Anais do XV Encontro Nacional de Inteligência Artificial e Computacional*, pages 25–35. SBC, 2018.

[9] Christopher D, Manning., Prabhakar, Raghavan., and Hinrich, Schütze. (2008). *Introduction to information retrieval*. Cambridge university press, 2008.

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

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

[12] Barry, Popik. (2004). Digital historical newspapers: A review of the powerful new research tools. *Journal of English Linguistics*, 32 (2), 114–123, 2004.

[13] Yannick, Rochat., Maud, Ehrmann., Vincent, Buntinx., Cyril, Bornet., Frédéric, Kaplan. (2016). Navigating through 200 years of historical newspapers. *iPRES 2016*, page 186, 2016.

[14] Shapenko, Andrey., Korovkin, Vladimir., Leleux, Benoit. (2018). Abbyy: the digitization of language and text. *Emerald Emerging Markets Case Studies*, 2018.

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

[16] Tumbe, Chinmay. (2019). Corpus linguistics, newspaper archives and historical research methods. *Journal of Management History*, 2019.

[17] Wang, Hongbin., Wang, Jianxiong., Zhang, Yafei., Wang, Meng., Mao, Cunli. (2019). Optimization of topic recognition model for news texts based on lda. *Journal of Digital Information Management*, 17(5), 257, 2019.

[18] Yang, Tze-I., Torget, Andrew., Mihalcea, Rada. (2011). 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.

[19] Yarasavage, Nathan., Butterhof, Robin., Ehrman, Christopher. (2012). 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.