A strategy for interpreting and visualizing the results of matrix-trifactorization-based coclustering algorithms

Ais B. R. Castro[1], Sarajane M. Peres[2], Waldyr L. de Freitas Junior[2], Paulo Pirozelli[3], Fábio G. Cozman[4], Anarosa A. F. Brandão[4]

[1] Curso de Ciências Moleculares — Universidade de São Paulo, São Paulo, SP

[2]Escola de Artes, Ciências e Humanidades — Universidade de São Paulo, São Paulo, SP

[3]Instituto de Estudos Avançados — Universidade de São Paulo, São Paulo, SP

[4]Escola Politécnica — Universidade de São Paulo, São Paulo, SP

{aisbrcastro,sarajane,waldyrjunior}@usp.br,
{paulo.pirozelli.silva,fgcozman,anarosa.brandao}@usp.br

Abstract. *Information yielded by unsupervised learning is often hard to interpret due to the lack of defined labels. To overcome this, we propose and illustrate a strategy for interpreting and visualizing the results of coclustering algorithms based on trifactorization. Our method consists of three steps: (1) vector space visualization; (2) cluster characterization by top documents/words; and (3) cocluster characterization by comparing top words between different clusters. The latter allows exploring the resulting clusters in a way which considers the relationship between attribute cluster and data cluster for every data cluster, instead of just the data cluster with the highest association with this attribute cluster. We illustrate the use of our method for the Non-negative Block Value Decomposition on a dataset of scientific abstracts.*

1. Introduction

Unsupervised data mining techniques are renowned for their capability to uncover inherent knowledge within data, forgoing extrinsic information like labels, annotations, rules, or policies linked to the domain or phenomena that generate the data. Thus, they represent a powerful computational tool for data analysis. Nevertheless, the outputs produced in this context typically require extra effort for interpretation and information extraction.

Various techniques enable unsupervised data analysis, and one such technique is the utilization of coclustering algorithms. Coclustering algorithms conduct data analysis by taking into account both the similarities among the data points and the similarities among the data attributes [Long et al. 2005, Wang et al. 2012]. In other words, they tackle the dual tasks of data clustering and attribute clustering simultaneously [Long et al. 2005, Wang et al. 2012]. By considering both dimensions of information, coclustering implements a type of partial similarity analysis that brings both flexibility and precision to the definition of clusters [Freitas Junior et al. 2020, Dhillon et al. 2003]. Furthermore, coclustering is able to provide a relationship between data clusters and attribute clusters called cocluster. Cocluster analysis is mainly useful for characterizing patterns in dyadic data [Hofmann et al. 1998], such as images [Chen et al. 2009] and texts [Shahnaz et al. 2006].

Within the context of textual data analysis, matrix-factorization-based coclustering algorithms have garnered attention due to their ability to yield promising results [Li and Ding 2006]. Similar to conventional clustering algorithms like k-means [Lloyd 1982], algorithms in this family also generate bases as output, which indicate the association of data (and attributes) with clusters. These bases serve as a quantization model for the data space (and attribute space). Such algorithms also bring the possibility of making an association between these bases. However, such an association is expressed in high-dimensional matrices with many seemingly unclear relationships between their values. Therefore, even though the algorithms produce structures that convey information about the data, additional post-processing efforts are necessary on this output. In coclustering algorithms there are more ways to extract information as compared to traditional clustering, but the challenge is also larger due to the presence of multiple information structures.

In this paper, we present a strategy to interpret the results of matrix-trifactorization-based coclustering algorithms, and illustrate it with one such algorithm — namely, Non-negative Block Value Decomposition (NBVD) [Long et al. 2005] — applied to abstracts of scientific papers. Our strategy, as applied to textual data, combines several ways of analyzing document and word clusters using the matrices generated by matrix trifactorization to gain additional information about the clustering results; its novelty lies in double-checking the association between document and word clusters contained in the trifactorization by analyzing how specific to each document cluster the words are.

To demonstrate the rationale and soundness of our approach, we applied it to abstracts of scientific papers from the Pirá dataset [Paschoal et al. 2021]. The Pirá dataset is specifically designed to support the development of question answering models, but it comprises a well-structured corpus of abstracts of scientific papers, which represent a complex domain where extraction of concise knowledge is challenging.

The remainder of this paper is organized as follows. Section 2 describes both the theoretical background surrounding this paper and also works closely linked to this one; Section 3 explains our strategy to interpret coclustering results, while Section 4 explains our use of the Pirá dataset and experimental setup for coclustering the documents used. Section 5 explains the results obtained, highlighting the effectiveness of this analysis; Section 6 concludes the paper.

2. Theoretical background

In this section, a brief summary of concepts essential to understanding this work is provided together with an overview of related works. Namely, the coclustering and the Non-Negative Block Value Decomposition problems are formally described; the use of validation indices in a coclustering context is shortly discussed; lastly, a categorization of related works is presented with some examples, and the distinctive aspects contributed by our work are highlighted.

2.1. Coclustering

Coclustering[1] is a technique for data analysis, similar to the clustering process; however, the clustering is applied simultaneously on the rows and columns of a data matrix

[1] The problem of clustering both dimensions of a matrix is referred to in the specialized literature using other names [Madeira and Oliveira 2004].

[Hartigan 1972]. Formally, consider a dataset represented by the matrix $X \in \mathbb{R}^{n \times m}$. The matrix X comprises a set of row vectors $N = \{x_{1\cdot}, \ldots, x_{n\cdot}\}$ and a set of column vectors $M = \{x_{\cdot 1}, \ldots, x_{\cdot m}\}$. The goal is to find $k \times l$ coclusters represented by submatrices of X, denoted by $X_{K_p L_q}$, with k subsets $K_p \subseteq N$, l subsets $L_q \subseteq M$, $p \in \{1, \ldots, k\}$ and $q \in \{1, \ldots, l\}$. In the coclustering problems, a cocluster $X_{K_p L_q}$ is formed by a data cluster K_p and an attribute cluster L_q.

As an example of the kind of information that coclustering techniques can offer, let us suppose that we wish to cocluster the datasets depicted in Figure 1 with $k = l = 3$. Even though this data is synthetic, suppose that we are working with a document-term matrix, in which rows represent documents and columns represent words. In this case, we will find that there are three clearly delimited document clusters (data clusters), three word clusters (attribute clusters) — with slight overlap between them in the second dataset —, and three coclusters of interest, corresponding to the three blue regions. (Note that the true number of coclusters is higher; however, coclusters associated with null data can be ignored in the context of textual data [Diaz and Peres 2019]). Our main conclusion would be that, although our set of documents covers three very distinct topics (sport, entertainment and education), in the second dataset there is some vocabulary overlap between document clusters, perhaps due to similar jargon or similar reference to proper nouns.

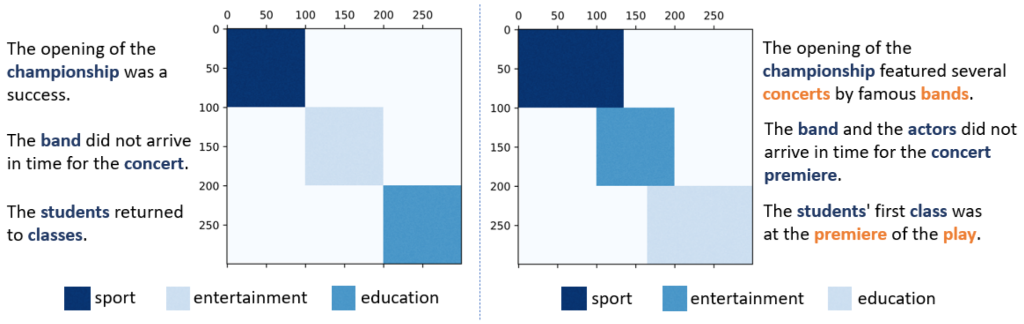

Figure 1. Two square matrices $X \in \mathbb{R}^{300 \times 300}$ displaying three coclusters of interest; darker shades represent higher values. Blue words represent the main topic of the document cluster, while orange words indicate topic overlaps between document clusters.

2.2. Non-negative Block Value Decomposition

Block Value Decomposition (BVD) searches for block structures in a data matrix and can be used for dyadic data analysis [Long et al. 2005]. This is a useful technique for coclustering solutions as it takes into account both dimensions of the data matrices (rows and columns) simultaneously. In this paper, we are interested in non-negative dyadic data (textual data); thus, the framework NBVD (Non-Negative Block Value Decomposition) was applied.

The NBVD optimization is performed by decomposing the matrix $X \in \mathbb{R}^{n \times m}$ into three other matrices (Problem $\mathcal{F}_1$): U as a row-coefficient matrix, S as a

In particular, a terminology frequently employed interchangeably with coclustering is biclustering. Nevertheless, there exist nuanced disparities in the definition of the optimization problem attributed to each terminology. For coclustering, the goal is to construct a partition of the vector space denoted by X. On the other hand, for biclustering, the partition is not mandatory, thereby allowing for the possibility that not all elements of the matrix need be assigned to clusters and coclusters [Pensa et al. 2010].

block structure matrix and V as a column-coefficient matrix, according to equation 1 [Freitas Junior 2023]:

$$\begin{aligned} \mathcal{F}_1(U,S,V) = \min_{U,S,V} \|X - USV^T\|_F^2 \\ \text{subject to: } U \geq 0;\ V \geq 0, \end{aligned} \tag{1}$$

in which $U \in \mathbb{R}_+^{n \times k}, S \in \mathbb{R}^{k \times l}$, $V \in \mathbb{R}_+^{m \times l}$ and F is the Frobenius norm for matrices. NBVD was chosen to perform the coclustering task in this paper due to its simplicity, performance and common use in textual analysis [Salah et al. 2018]. Algorithm 1 details the matrix update rules used in the factorization that solves Problem $\mathcal{F}_1$.

Algorithm 1 NBVD — Non-negative Block Value Decomposition [Freitas Junior 2023, Long et al. 2005]

```
function NBVD($X$, $k$, $l$, $itr^{max}$)
    Initialize: $U^{(0)} \leftarrow \mathcal{U}(0,1), V^{(0)} \leftarrow \mathcal{U}(0,1), S^{(0)} \leftarrow \mathcal{U}(0,1)$ and $t \leftarrow 0$.
    while ($t \leq itr^{max}$) and (unreached convergence) do
        $U^{(t+1)} \leftarrow U^{(t)} \odot \frac{XV^{(t)}S^{(t)^T}}{U^{(t)}S^{(t)}V^{(t)^T}V^{(t)}S^{(t)^T}}$
        $V^{(t+1)} \leftarrow V^{(t)} \odot \frac{X^TU^{(t+1)}S^{(t)}}{V^{(t)}S^{(t)^T}U^{(t+1)^T}U^{(t+1)}S^{(t)}}$
        $S^{(t+1)} \leftarrow S^{(t)} \odot \frac{U^{(t+1)^T}XV^{(t+1)}}{U^{(t+1)^T}U^{(t+1)}S^{(t)}V^{(t+1)^T}V^{(t+1)}}$
        $t \leftarrow t + 1$
    end while
    return $U^{(t)}, S^{(t)}, V^{(t)}$
end function
```

According to [Long et al. 2005], the row-coefficient matrix U represents the degree of association between rows and their corresponding clusters; the block structure matrix S offers a compact representation of the data matrix X; the column-coefficient matrix V represents the degree of association between columns and their corresponding clusters. Additionally, the product US represents the basis of the column space for X, while the product SV^T represents the basis of the row space for X.

2.3. Validation index

Assessing the quality of cocluster discovery remains an unresolved matter in the academic literature. Certain quality measures have been established for biclusters, and these measures can potentially be adapted for application in the context of coclustering. However, it is crucial to exercise caution when interpreting such measures, given that biclustering does not generate a partitioning of the data space. Consequently, in most cases, the existing literature evaluates the quality of coclustering by employing conventional clustering validation indices, primarily focusing on assessing the quality of data clusters. Some works even transpose the matrix to enable the application of the same measures for assessing column clusters. This approach provides a certain level of reliability to the assessment of the clustering quality for rows and columns, which has been deemed acceptable for coclustering evaluations.

In this study, our focus lies in evaluating the quality of the data clusters (documents) and subsequently interpreting the results based on the attainment of robust clusters within this validation framework. To accomplish this, we have opted for the Silhouette Index [Rousseeuw 1987], as it obviates the need for a ground truth for validation and offers a simplified interpretation by confining its values within the range of $[-1, 1]$. According to [Han et al. 2012] and following the specifications presented in [Luna et al. 2021], the Silhouette Index (SI) is calculated according to equation 2:

$$SI(X, C) = \frac{1}{|X|} \sum_{o \in X} \frac{b(o, C) - a(o, C)}{max\{b(o, C),\, a(o, C)\}}\,, \tag{2}$$

in which X is a data matrix represented as a set of vectors (N or M as defined earlier), C denotes either K_p or L_q as defined earlier, $a(o, C)$ is the average distance between o and all other objects in the cluster to which o belongs, and $b(o, C)$ is the minimum average distance between o to all objects in the clusters to which o does not belong.

2.4. Related work

Although the literature on trifactorization for coclustering is rich, interpreting and visualizing coclustering results is an underexplored aspect of it. Generally, in clustering, the interpretation of results is associated with the identification of relationships between topics and the underlying words within the analyzed texts. In the context of coclustering, a similar line of study is pursued, albeit with additional information. According to [Freitas Junior 2023], several evaluation strategies have been proposed, falling into at least three distinct categories: (a) analysis of prototype vectors, (b) analysis of the S matrix (cf. equation 1), and (c) visualization of graphical representations for clusters and coclusters. In this section, we provide a concise selection of recent works that also undertake this challenging task, aiming to exemplify such strategies.

In the implementation of the first strategy, the analysis focuses on prototype vectors derived from the products US and SV^T (or similar products, depending on the specific factorization method employed), with attention given to their values. High values within the coordinates of such vectors indicate significant information that characterizes the corresponding cluster, be it related to data or attributes. Studies dealing with textual data that adopt this interpretative approach [Brunialti et al. 2017, Salah et al. 2018] engage in discussions regarding the significance of a word in describing a cluster of documents or the extent to which a document can assign meaning to a word.

The second strategy implements an analysis that leverages the S matrix as a means to establish relationships between clusters of data and clusters of attributes. Studies investigating this matrix [Abe and Yadohisa 2019, Freitas Junior et al. 2020] hypothesize that the values within its cells reflect the degree to which a group of words can effectively describe a group of documents, and vice versa. This strategy presents similarities to the first one but incorporates information specifically associated with coclusters.

In the final strategy, the information derived from the factorized matrices and their resulting products is transformed into visual representations. Frequently, word clouds or ordered lists of words are employed as means of presenting clusters and coclusters to individuals (e.g., [Shahid et al. 2017, Hassani et al. 2021]), enabling them to make meaningful interpretations based on these visualizations.

The present study introduces an interpretation strategy that aligns with the strategies described here. Specifically, it combines analysis of prototype vectors, analysis of the S matrix and graphical representations for clusters in a three-step approach which seeks to elicit a rich understanding of the coclustering results. As a differential from the existing literature, our approach explores the relationship between attribute cluster and data cluster for every data cluster, as opposed to just the data cluster with the highest association to this attribute cluster.

3. Coclustering interpretation and visualization strategy

This section is dedicated to presenting our approach for interpreting and visualizing the outcomes of a matrix-trifactorization-based coclustering algorithm. While our strategy is applicable to any coclustering algorithm based on trifactorization, it is more effective when these algorithms introduce minimal changes and assumptions compared to the NBVD algorithm. In addition to organizing the steps for information interpretation, our strategy proposes a crucial contribution: a one-against-all analysis of the relationship between clusters of words and documents comprising the coclusters of interest (see Figure 1). We have not encountered this particular approach in the post-processing steps of coclustering applications presented in the existing literature. This contribution is presented as the third step in our strategy. To exemplify the effectiveness of this strategy, we present an application in Section 5.

Our interpretation strategy, as presented here, will focus on analyzing textual data and will assume an X matrix where rows represent documents and columns represent words. In order to gain useful insights from the data, the matrices U, S, and V serve as the foundation for analysis and interpretation. The assignment of documents and words to their respective clusters can be determined by utilizing the elements of matrices U and V, respectively, while employing a suitable normalization method (e.g., L2-norm-based, as demonstrated in [Long et al. 2005], or column-sum-to-one-based, as demonstrated in [Yoo and Choi 2010]). In such cases, the values in U and V are considered representative of the adherence of documents or words to their respective clusters [Diaz and Peres 2019]. Our strategy follows these assumptions and, for the sake of simplicity, enforces a strict association of each document or word with only one cluster. To establish the connection between document clusters and word clusters, the factor matrix S is used as a means to quantify the relationship between each pair of document cluster and word cluster. Our interpretation strategy is thus divided into three steps: (1) vector space visualization; (2) cluster characterization by top documents/words; (3) cocluster characterization by comparing top words between different clusters.

The first step is an analysis of the quality of the generated clustering. To this end, we calculate the Silhouette Index for the document labels; a positive value, ideally close to 1, indicates a good clustering result, suitable for interpretation. Next, our strategy suggests a visualization of the clustered space to make clear the disposition of each cluster, its respective centroid and its respective basis vector. This can be achieved by employing a dimensionality reduction technique such as Principal Component Analysis

(PCA), which linearly transforms a set of n-dimensional vectors into lower-dimensional vectors such that the new dimensions maximize variance of the data along these dimensions [Salih Hasan and Abdulazeez 2021]. Ideally, the directions of basis vectors will point toward their respective cluster and cluster centroid, as will be shown in Section 5.

The second step has the goal of using matrices U and V to characterize document/word clusters by their top documents/words. This is done by first labeling documents and words as in [Yoo and Choi 2010]: assigning document $x_{i\cdot}$ to cluster a^* if $a^* = \arg\max_a U_{ia}$ and assigning word $x_{\cdot j}$ to cluster b^* if $b^* = \arg\max_b V_{jb}$. Then, as in [Diaz and Peres 2019], the values in U and V are used in a similar way to calculate the representativeness of each document/word, with the most representative element being the one with the highest value in U or V. An important detail to consider is how to present this information. For word clusters, we can construct a table showing the top words (up to a certain number N of choice), whereas, for document clusters, we must abridge them somehow; an option is to show snippets (e.g., the first two sentences or sentences containing words belonging to a specific word cluster) of documents. Notice that, even though these representations are specific to textual data, this analysis can be adapted to accommodate different kinds of data as long as we can easily represent row and column vectors.

The third step aims to confirm or discard the representativeness of words in the document cluster associated with them through matrix S. To that end, our strategy recommends, for each word cluster, looking at the occurrence of top words in top documents of the associated document cluster and also in documents representative of the remaining document clusters. Furthermore, one could also include in this test the least representative words of the examined word cluster to verify whether they are in fact more related to their assigned cluster than to the remaining clusters. Note that, by looking at the relation between a word cluster and all document clusters — instead of just the associated document cluster — we double-check the information contained in S, getting a more complete picture of the coclustering results.

4. Research method

The strategy outlined in the previous section was employed according to the experimental method depicted in Figure 2.

In experimental step 1, the textual dataset underwent preprocessing using fundamental procedures in the field of natural language processing. Subsequently, the documents were vectorized using the TF-IDF scheme (as explained in [Rajaraman and Ullman 2011]), thereby generating a matrix representing the textual information. The NBVD algorithm was then employed to perform the triple factorization. Moving on to experimental step 2, our interpretation and visualization strategy was applied, resulting in the following artefacts:

- a document/word scatter plot that displays document/word vectors, document/word centroid vectors and basis vector directions;
- a cluster summary that indicates representative documents/words of each cluster and occurrence of top words in each document cluster;
- a bar plot that summarizes information from the cluster summary, allowing one

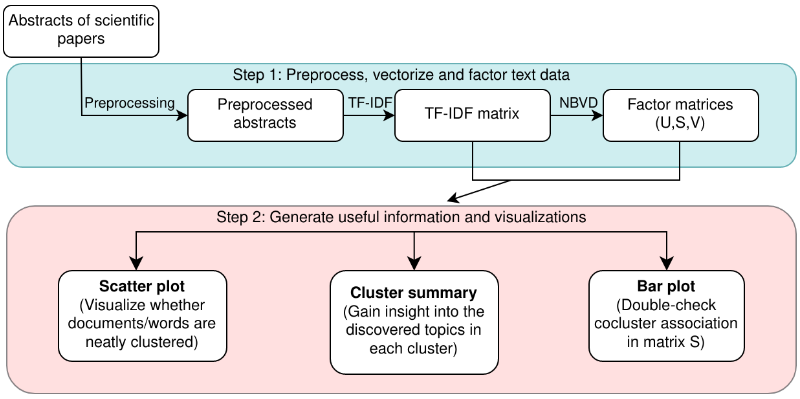
Abstracts of scientific papers
Step 1: Preprocess, vectorize and factor text data
Preprocessing
Preprocessed abstracts
TF-IDF
TF-IDF matrix
NBVD
Factor matrices (U,S,V)
Step 2: Generate useful information and visualizations
Scatter plot (Visualize whether documents/words are neatly clustered)
Cluster summary (Gain insight into the discovered topics in each cluster)
Bar plot (Double-check cocluster association in matrix S)

Figure 2. Experimental method which uses the proposed strategy to interpret and visualize coclustering results.

to quickly spot words characteristic of a cluster and words "out of place" in that cluster.

Additionally, to demonstrate the practicality of our interpretation and visualization strategy beyond static applications, we utilize the factored matrices as a model of the textual context being analyzed. This enables the inclusion of new documents that were not seen during the factorization process. In such cases, the new document undergoes the same preprocessing steps and is associated with a document cluster based on its similarity to the document bases (product SV^T). By establishing this association, it is possible to repeat all steps of our interpretation strategy in an analogous manner.

4.1. Pirá dataset

The experiment utilized a corpus of texts comprising 992 scientific article abstracts extracted from the Pirá dataset [Paschoal et al. 2021]. To facilitate the simulation of the inclusion of new documents, these abstracts were divided into two halves, with 496 abstracts in each. The Pirá dataset is a bilingual Portuguese-English resource specifically designed for question-answering tasks related to ocean data, biodiversity, and climate change. It encompasses scientific abstracts on Brazilian coastal subjects, and text excerpts derived from United Nations reports on the ocean; only the former were utilized. It is important to note that the content of the abstracts is highly technical in nature; consequently, the complexity of the material presents unique challenges for analysis and interpretation.

4.2. Experiment setup

For the purposes of preprocessing, we constructed a custom stopword list by extending NLTK's[2] stopword list for the English language to include words such as: function words not covered in the NLTK corpus (e.g., "thus", "even", "per"), metalinguistic and metadata words (e.g., "authors", "study", "publishing"), scientific units (e.g., "m", "kg", "yr"), spelled-out numbers and ordinal suffixes (e.g., "nine", "three", "th"), and general scientific jargon (e.g., "proposed", "observed", "analysis"). We started from the NLTK stopword list and gradually incorporated additional words, making sure not to add terms that might bear semantic significance. The full preprocessing procedure used is as follows:

[2] https://www.nltk.org

discarding excessively small (less than 300 characters) or duplicate abstracts, lowercasing words while attempting to preserve initialisms (e.g., FPSO[3]); substituting hyphens for underscores in hyphenated words; replacing all numbers with "1"; and removing stopwords.

To vectorize documents, *Scikit-learn's*[4] *Tf-idf_Vectorizer* was used with its default values, except for the minimum document frequency, which was set to four. Coclustering was performed according to Algorithm 1, using $k = l = 4$, and $itr^{max} = 2000$. Labels (and representativeness) for the original abstracts were calculated from the factor matrices U and V, while labels for new abstracts were calculated based on the cosine similarity of documents to the document cluster bases; the scatter plot, cluster summary and bar plot were generated using the obtained matrices, as discussed in Section 3.

Note that, as the optimization problem involved in matrix trifactorization is complex and deals with high-dimensional matrices, it might be interesting to run the coclustering algorithm multiple times (and to select the best result according to the Silhouette Index and the algorithm's minimization error), because — at least in the case of NBVD — it will not realistically converge to a global minimum, only to a local one [Long et al. 2005]. Additionally, the choice of $k = l = 4$ is meant to provide interesting clusters to analyze; it is, however, an arbitrary choice.

5. Results and discussion

Following the method outlined in the previous section, a set of 496 abstracts from the Pirá dataset was processed[5] to derive informative resources concerning the data. Subsequently, a separate set of 496 abstracts underwent processing utilizing the coclustering information obtained from the initial set. The objective was to evaluate the effectiveness of transferring the document collection's description to new documents pertaining to the same topics. Note that all analyses conducted in this section pertain to a specific run of the experiment, which achieved: SI scores of 0.022 and 0.012 for the clustering of original abstracts and words — respectively —, performed by the NBVD algorithm; and an SI score of 0.019 for the clustering of the additional set of abstracts, performed by calculating the cosine similarity between documents and cluster bases. Both of these results were considered acceptable within the context of a real-world textual dataset — which presents a complex problem — and with the evaluation metric yielding results greater than zero. Note that the chosen metric for the Silhouette Index was cosine similarity, as it is widely used in the context of comparing document vectors obtained from TF-IDF vectorization [Bafna et al. 2016].

Visualizations of the clustered vector space obtained in this NBVD run are shown in Figure 3. They represent the first step of our strategy, aiding in assessing the quality of the clustering and facilitating the observation of clusters that are more distinctly separated in the vector space. Centroids were normalized before dimensionality reduction in all but the first plot, so we must analyze their direction instead of position. The reason for this normalization is twofold: first, to distinguish the centroids from the clutter of document vectors; and second, to emphasize the significance of the basis vector directions, which

[3] Floating Production Storage and Offloading [unit], a floating vessel used by the offshore industry.

[4] https://scikit-learn.org

[5] All the code used in this project is available at https://github.com/C4AI/unsupervised-topic-model

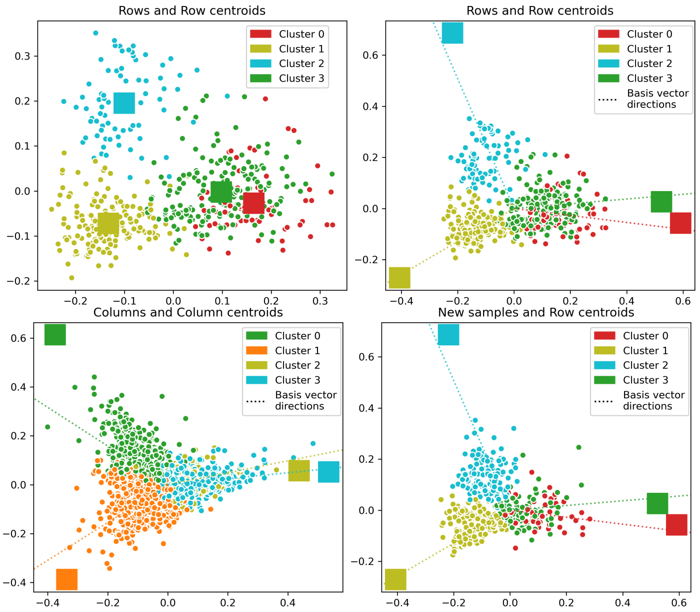

Figure 3. Reduced-dimension plot of data points (represented by circles) and cluster centroids (represented by squares). The third plot corresponds to clusters of words while the other plots correspond to clusters of documents. The fourth plot corresponds to an unseen set of documents.

indicate the perceived direction of important cluster features as determined by the coclustering algorithm and which should ideally point toward the cluster centroid. In the figure, we see that the trifactorization performed well in capturing information about documents, as the four basis vector directions point directly to their respective centroid. The fourth plot shows the distribution of the new set of abstracts using the established model, and it is noticeably similar to the second plot — representing the original set of abstracts —, which indicates that the organization established using the original document collection is still relevant for these new documents. The overlap between clusters 0 and 3 on first, second and fourth plots, and the overlap between clusters 2 and 3 on the third plot — both of which are exaggerated due to dimensionality reduction — will be explained by the next steps of our interpretation strategy.

By analyzing top documents/words, step two of our strategy enables us to semantically characterize the clusters shown in the previous step. Tables 1 and 2 illustrate this analysis. Note that, for a better presentation, abstracts were truncated to approximately 130 characters and are shown without preprocessing. By comparing the two tables, it is not hard to establish links between document clusters and word clusters. For instance: document clusters 0 and 3 seem to be related to oil and natural gas, as do word clusters 2 and 3; document cluster 2 and word cluster 0 both mention terminology pertaining to geology and oceanography, such as “basin”, “facies”, “carbonate platform” and “lacustrine [. . .] settings”; lastly, document cluster 1 and word cluster 1 contain concepts from ecology such as “species”, “marine”, “concentrations of total arsenic” and “fishery source”. As we will see next, this association is the same as the one found by the NBVD algorithm and contained in the block value matrix S.

Figure 4 is step three of our strategy: in it, we see a representation of each word

Table 1. Representative documents (the first two) for each cluster

	Cluster 0	Cluster 1	Cluster 2	Cluster 3
1st	In 2004, after a cycle of 11 yr in which the annual increase in crude oil production was 8.6% avg, the production decreased 3%, [. . .]	High concentrations of total arsenic (As), even above the Brazilian legislative threshold for marine sediments [. . .]	Coquinas constitute widespread deposits in lacustrine, estuarine, and shallow marine settings, [. . .]	Drilling operations in salt zones have gained importance in Brazil due to the discovery of large oil and gas reserves [. . .]
2nd	The Petroleo Brasileiro SA (Petrobras), a state-owned oil company utilizes its floating production storage and offloading (FPSO) [. . .]	Although significantly impacted, Guanabara Bay (GB), located in southeastern Brazil, is still an important fishery source [. . .]	This work intends to understand how the Ponta do Mel carbonate platform was implanted, to characterize its depositional model [. . .]	Drilling and completion in Campos Basin have been in constant evolution, from the first subsea wells and fixed platforms [. . .]

Table 2. Representative words (the first six) for each cluster

	Cluster 0	Cluster 1	Cluster 2	Cluster 3
1st	seismic	bay	oil	drilling
2nd	salt	coastal	production	well
3rd	basin	species	gas	offshore
4th	facies	areas	petrobras	technology
5th	carbonate	sediment	million	subsea
6th	continental	marine	FPSO	wells

cluster, characterized by its 12 most representative words and by the occurrence of these words in documents from each document cluster. Each bar represents how many of the 40 most representative documents of the corresponding document cluster contain the word in question. Note that each individual bar chart has different scaling. The document cluster associated with the current word cluster (through matrix S) is represented in green, emphasizing the assertion that this pair of document and word clusters should have the best match.

This visualization allows for quickly identifying well-separated clusters: if a word cluster's top words are substantially more frequent in the green document cluster, then the word cluster is divided well. If this applies to all word clusters, then the coclusters are neatly separated and there is a one-to-one association between document cluster and word cluster, which is desired. On the other hand, if a word cluster's top words are distributed evenly between document clusters, then there is significant overlap between document clusters. Furthermore, if a small number of words have higher frequency in red document clusters, it indicates that the word cluster could be associated with other document clusters and that there is overlap between word clusters and between document clusters. A large number of words with higher frequency in red document clusters would indicate a bad coclustering result which associates the word cluster to an incorrect document cluster.

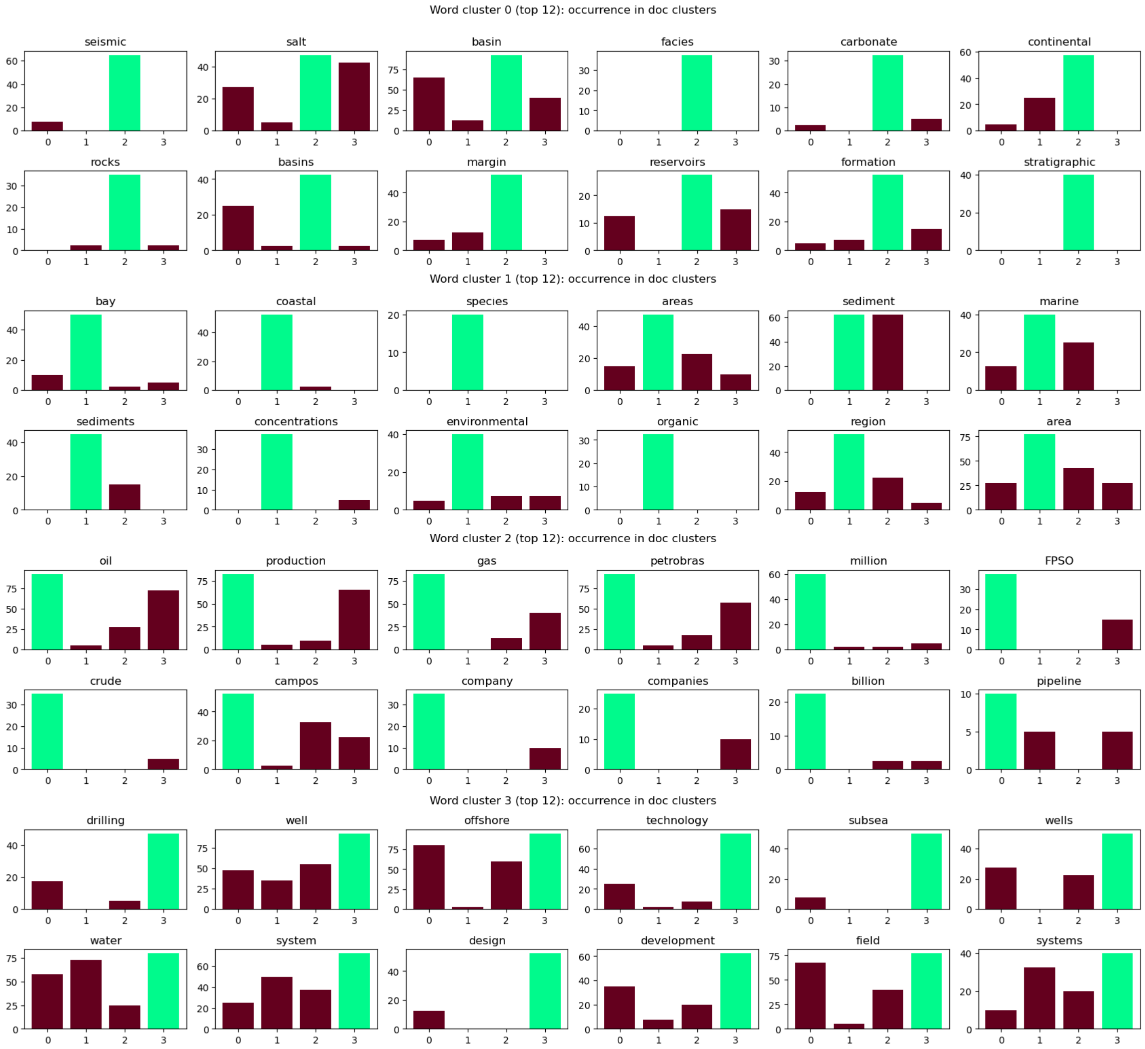

Figure 4. Bar charts that show how well the top words of each word cluster describe the corresponding document cluster (in green) compared to the remaining document clusters (in red).

Turning our attention back to the example in Figure 4, the general conclusion produced by our strategy is that the coclustering obtained is satisfactory, but there are several nuances. It is satisfactory because, for almost all top words shown, the green bar is higher; the only exception is “sediment” (from word cluster 1), which is divided between document clusters 1 and 2. To get a sense of the nuances to this positive result, we shall thoroughly analyze the plots for each word cluster.

- In word cluster 0, we see several words specific to the green document cluster (e.g., “seismic”, “facies”, “carbonate”); however, the occurrence in red document clusters is divided somewhat evenly, suggesting that it contains some words that are common across document clusters (e.g., “salt”, “basin”, “reservoirs”).
- In word cluster 1, we again see several words specific to the green document cluster (e.g., “bay”, “coastal”, “species”); additionally, document cluster 2 stands out as the second document cluster most related to this word cluster, indicating

some overlap between this word cluster and the previous one (e.g., “sediment”, “areas”, “marine”), which is associated with document cluster 2.

- In word cluster 2, there are some words specific to the green document cluster (e.g., “million”, “billion”, “crude”), but we observe that many words also occur frequently in document cluster 3 (e.g., “oil”, “production”, “petrobras”), indicating significant overlap between this word cluster and the next one, which is associated with document cluster 3.
- In word cluster 3, we see fewer words specific to the green document cluster (e.g., “drilling”, “technology”, “subsea”), indicating a cluster that is not defined as well as the other ones. Additionally, there is significant overlap between this word cluster and the previous one (e.g., “offshore”, “wells”, “field”), as noted above.

By noting that, in this case, word clusters can be mostly associated to one document cluster, we can summarize the previous observations by making observations about document clusters and word clusters simultaneously: cocluster $(2, 0)$ — the cocluster formed by associating document cluster 2 and word cluster 0 — has some overlap with cocluster $(1, 1)$, and cocluster $(0, 2)$ has significant overlap with cocluster $(3, 3)$ — as was evidenced by the figures from step one —, however there are enough distinctions between all coclusters to justify this organization of documents and words.

6. Conclusion

In this paper we have presented a strategy to give a deeper understanding of the results of trifactorization-based coclustering algorithms. First, by visualizing the cluster members, cluster centroids and basis vectors in a reduced-dimension space, we can visually determine whether a good clustering was obtained. Next, by examining the top documents and top words, we can attempt to summarize the organization created by the clusters. Lastly, we can use the cocluster structure to connect document clusters to word clusters, enabling us to further evaluate the quality of these clusters by seeing whether top words for one cluster are most relevant to their associated document cluster, and thus we can better understand the information each cluster captures.

There is considerable room for future research, such as: exploring different coclustering algorithms while selecting for performance and quality of clusters; exploring different text representations that capture more linguistic context than TF-IDF, such as word embeddings; and allowing overlaps between document/word clusters to indicate relations between topics. However, the strategy as currently established already enables the generation of information that supports both human understanding of the coclustering results and the training of language models for summarizing texts or generating paraphrases. Furthermore, the insights derived from significant words and text snippets can contribute to enhancing evaluation measures in text mining tasks.

Acknowledgment

This work was carried out at the Center for Artificial Intelligence (C4AI-USP), with support by the São Paulo Research Foundation (FAPESP grant #2019/07665-4) and by the IBM Corporation. Additionally, Ais B. R. Castro was supported by the “Programa Unificado de Bolsas de Estudos para Apoio e Formação de Estudantes de Graduação (PUB)” of the University of São Paulo; Paulo Pirozelli was supported by the FAPESP grant #2019/26762-0.

References

Abe, H. and Yadohisa, H. (2019). Orthogonal nonnegative matrix tri-factorization based on Tweedie distributions. *Advances in Data Analysis and Classification*, 13:825–853.

Bafna, P., Pramod, D., and Vaidya, A. (2016). Document clustering: Tf-idf approach. In *2016 International Conference on Electrical, Electronics, and Optimization Techniques (ICEEOT)*, pages 61–66. IEEE.

Brunialti, L. F., Peres, S. M., da Silva, V. F., and de Moraes Lima, C. A. (2017). The BinOvNMTF algorithm: Overlapping columns co-clustering based on non-negative matrix tri-factorization. In *Brazilian Conference on Intelligent Systems, BRACIS*, pages 330–335, Uberlândia, Brazil. IEEE - Conference Publishing Services.

Chen, Y., Dong, M., and Wan, W. (2009). Image co-clustering with multi-modality features and user feedbacks. In *Proceedings of the 17th ACM International Conference on Multimedia*, MM '09, page 689–692, New York, NY, USA. Association for Computing Machinery.

Dhillon, I. S., Mallela, S., and Modha, D. S. (2003). Information-theoretic co-clustering. In *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 89–98. ACM.

Diaz, A. K. R. and Peres, S. M. (2019). Biclustering and coclustering: concepts, algorithms and viability for text mining. *Revista de Informática Teórica e Aplicada*, 26(2):81–117.

Freitas Junior, W. L. (2023). Um comparativo quantitativo e qualitativo de algoritmos de coagrupamento baseados em fatoração de matrizes. Master's thesis, Universidade de São Paulo.

Freitas Junior, W. L., Peres, S. M., Freire, V., and Brunialti, L. F. (2020). OvNMTF Algorithm: an Overlapping Non-Negative Matrix Tri-Factorization for Coclustering. In *2020 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE.

Han, J., Pei, J., and Kamber, M. (2012). *Data Mining: Concepts and Techniques*. Morgan Kauffman, Waltham, 3rd edition.

Hartigan, J. A. (1972). Direct clustering of a data matrix. *Journal of the American Statistical Association*, 67(337):123–129.

Hassani, A., Amir, I., and Mansouri, N. (2021). Text mining using nonnegative matrix factorization and latent semantic analysis. *Neural Computing and Applications*, 33(20):13745–13766.

Hofmann, T., Puzicha, J., and Jordan, M. I. (1998). Learning from dyadic data. In *Advances in Neural Information Processing Systems 11, NIPS Conf., Denver, Colorado, USA*, pages 466–472.

Li, T. and Ding, C. (2006). The relationships among various nonnegative matrix factorization methods for clustering. In *Sixth International Conference on Data Mining (ICDM'06)*, pages 362–371.

Lloyd, S. P. (1982). Least squares quantization in PCM. *IEEE Transactions on Information Theory*, 28(2):129–136.

Long, B., Zhang, Z. M., and Yu, P. S. (2005). Co-clustering by block value decomposition. In *Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining*, pages 635–640. ACM.

Luna, M., Lima, A., Neubauer, T., Fantinato, M., and Peres, S. (2021). Vector space models for trace clustering: a comparative study. In *Anais do XVIII Encontro Nacional de Inteligência Artificial e Computacional*, pages 446–457, Porto Alegre, RS, Brasil. SBC.

Madeira, S. and Oliveira, A. (2004). Biclustering algorithms for biological data analysis: a survey. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 1(1):24–45.

Paschoal, A. F. A., Pirozelli, P., Freire, V., Delgado, K. V., Peres, S. M., José, M. M., Nakasato, F., Oliveira, A. S., Brandão, A. A. F., Costa, A. H. R., and Cozman, F. G. (2021). Pirá: A bilingual portuguese-english dataset for question-answering about the ocean. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, CIKM '21, page 4544–4553, New York, NY, USA. Association for Computing Machinery.

Pensa, R. G., Boulicaut, J.-F., Cordero, F., and Atzori, M. (2010). Co-clustering numerical data under user-defined constraints. *Statistical Analysis and Data Mining*, 3(1):38–55.

Rajaraman, A. and Ullman, J. D. (2011). *Mining of Massive Datasets*. Cambridge University Press, 1 edition.

Rousseeuw, P. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20:53–65.

Salah, A., Ailem, M., and Nadif, M. (2018). Word Co-Occurrence Regularized Non-Negative Matrix Tri-Factorization for Text Data Co-Clustering. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1).

Salih Hasan, B. M. and Abdulazeez, A. M. (2021). A review of principal component analysis algorithm for dimensionality reduction. *Journal of Soft Computing and Data Mining*, 2(1):20–30.

Shahid, N., Ilyas, M. U., Alowibdi, J. S., and Aljohani, N. R. (2017). Word cloud segmentation for simplified exploration of trending topics on twitter. *IET Software*, 11(5):214–220.

Shahnaz, F., Berry, M. W., Pauca, V., and Plemmons, R. J. (2006). Document clustering using nonnegative matrix factorization. *Information Processing & Management*, 42(2):373–386.

Wang, J., Zhao, Z., Zhou, J., Wang, H., Cui, B., and Qi, G. (2012). Recommending flickr groups with social topic model. *Information Retrieval*, 15(3-4):278–295.

Yoo, J. and Choi, S. (2010). Orthogonal nonnegative matrix tri-factorization for co-clustering: Multiplicative updates on Stiefel manifolds. *Information Processing & Management*, 46(5):559–570.