Biclustering and coclustering: concepts, algorithms and viability for text mining

Biagrupamento e coagrupamento: conceitos, algoritmos e viabilidade para mineração de texto

Alexandra Katiuska Ramos Diaz¹, Sarajane Marques Peres^{1*}

Abstract: Biclustering and coclustering are data mining tasks capable of extracting relevant information from data by applying similarity criteria simultaneously to rows and columns of data matrices. Algorithms used to accomplish these tasks simultaneously cluster objects and attributes, enabling the discovery of biclusters or coclusters. Although similar, the natures and aims of these tasks are different, and coclustering can be seen as a generalization of biclustering. An accurate study on algorithms related to biclustering and coclustering is essential to achieve effectiveness when solving real-world problems. Determining the values appropriate for the parameters of these algorithms is even more difficult when complex real-world data are analyzed. For example, when biclustering or coclustering is applied to textual data (i.e., in text mining), a representation through a vector space model is required. Such representation usually generates vector spaces with a high number of dimensions and high sparsity, which influences the performance of many algorithms. This tutorial aims to didactically present concepts related to the biclustering and coclustering tasks and how two basic algorithms address these concepts. In addition, experiments are presented in data contexts with a high number of dimensions and high sparsity, represented by both a synthetic dataset and a corpus of real-world news. In general and comparative terms, the results obtained show the algorithm used for coclustering (i.e., NBVD) as the most appropriate for the experiments' context. Although the biclustering algorithm (i.e., Cheng and Church) was responsible for producing less relevant results in textual data than NBVD, its application in data with a high number of dimensions and high sparsity provided a suitable study environment to understand its operation.

Keywords: Biclustering — Coclustering — Text Mining

Resumo: Biagrupamento e coagrupamento são tarefas de análise de dados que permitem a extração de informação relevante aplicando critérios de similaridade simultaneamente às linhas e às colunas de matrizes de dados. Algoritmos usados na resolução destas tarefas agrupam simultaneamente os objetos e os atributos, possibilitando a criação de bigrupos ou cogrupos. Embora semelhantes, essas tarefas possuem naturezas e objetivos diferentes, sendo que coagrupamento pode ser visto como uma generalização de biagrupamento. Um estudo detalhado sobre algoritmos associados à cada tarefa é importante para alcançar eficácia na resolução de problemas do mundo real. Escolher os valores adequados para os parâmetros desses algoritmos é mais difícil quando dados complexos do mundo real são analisados. Por exemplo, para aplicar biagrupamento ou coagrupamento em dados textuais (i.e., em mineração de textos) é necessário usar uma representação de textos em um modelo de espaço vetorial. Esse tipo de representação comumente leva à geração de espaços vetoriais caracterizados pela alta dimensionalidade e esparsidade, e afeta o desempenho de muitos dos algoritmos. Este tutorial tem por objetivo apresentar, de forma didática, os conceitos referentes às tarefas de biagrupamento e coagrupamento, e como dois algoritmos abordam esses conceitos. Além disso, são apresentadas experimentações em contextos de dados de alta dimensionalidade e alta esparsidade, gerados de maneira sintética e também representados por um corpus de notícias. De forma geral e em termos comparativos, os resultados obtidos indicam que o algoritmo para coagrupamento (i.e., NBVD) apresenta maior adequabilidade para o contexto de experimentação utilizado. Embora o uso do algoritmo de biagrupamento (i.e., de Cheng e Church) tenha apresentado resultados de mais baixa relevância no contexto dos dados textuais do que o NBVD, sua aplicação em dados de alta dimensionalidade e esparsidade propiciou um ambiente de estudo útil para entendimento do funcionamento do algoritmo.

Palavras-Chave: Biagrupamento — Coagrupamento — Mineração de Textos

¹ Escola de Artes, Ciências e Humanidades (EACH), Universidade de São Paulo (USP), Brasil

*Corresponding author: {sarajane}@usp.br

DOI: <http://dx.doi.org/10.22456/2175-2745.89063> • Received: 18/12/2018 • Accepted: 14/04/2019

CC BY-NC-ND 4.0 - This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License.

1. Introdução

Uma das tarefas de análise de dados mais populares é o “agrupamento” (do inglês *clustering*). O agrupamento auxilia o processo de descoberta de conhecimento, facilitando a identificação de padrões que descrevem grupos similares de dados [1, 2]. Os algoritmos de agrupamento operam sobre matrizes de dados e têm como objetivo o particionamento dos dados em subconjuntos segundo algum critério de similaridade. Os subconjuntos a serem formados contêm objetos (dados) que possuem valores semelhantes em seus atributos descritivos, e valores diferentes dos valores dos objetos pertencentes a outros subconjuntos. Alternativas à tarefa de agrupamento são as tarefas conhecidas como biagrupamento (do inglês *bi-clustering*)¹ e coagrupamento (do inglês *co-clustering*) [3]. A resolução de tais tarefas leva à extração de informações diferenciadas quando comparadas àquelas extraídas a partir da resolução da tarefa de agrupamento. Isso ocorre porque tanto no biagrupamento como no coagrupamento, os critérios de similaridade são aplicados simultaneamente às linhas e às colunas das matrizes de dados, agrupando simultaneamente os objetos e os atributos.

Para compreender melhor a diferença entre as três tarefas citadas, considere o exemplo mostrado na figura 1. Neste exemplo, um conjunto de dados textuais é formado por sete notícias (n#1 até n#7) e cinco palavras (p#1 até p#5). As relações entre notícias e palavras são denotadas por valores binários (ocorrência da palavra (1) e ausência da palavra (0) na notícia). As resoluções das tarefas de agrupamento, biagrupamento ou coagrupamento considerando tais dados têm o objetivo de encontrar notícias similares, porém cada tarefa tem uma atuação diferenciada sobre a matriz de dados.

De acordo com o exemplo, análises feitas a partir da tarefa de agrupamento resultam em dois grupos de notícias referentes aos tópicos “educação” e “política” (figura 1a). Embora essa análise seja prática para alguns contextos, as informações fornecidas podem não ser tão úteis devido à falta de especificidade. Por outro lado, resultados obtidos nas tarefas de biagrupamento e coagrupamento trazem informações diferenciadas. Por exemplo, podem existir notícias que tratam de temas relacionados tanto à “educação” quanto à “política”, tais como notícias sobre reformas educacionais. Em um caso hipotético, essas notícias podem formar um bigrupo/cogrupo com tópicos relacionados a “educação e política”, como está graficamente ilustrado nas figuras 1b e 1c. No mesmo exemplo, percebe-se uma diferença em relação à interpretação dos resultados do biagrupamento e do coagrupamento. No caso do biagrupamento, são evidenciados três bigrupos no resultado. Já o coagrupamento fornece um modelo dos dados sobre o qual é possível obter sete cogrupos (figura 1c), três equivalentes aos bigrupos e quatro referentes à organização

Figura 1. Exemplo de uma aplicação de mineração de textos, analisada a partir do agrupamento (a), biagrupamento (b) e coagrupamento (c).

de ausências de palavras nas notícias.

Embora o biagrupamento e o coagrupamento tragam mais flexibilidade para a análise de dados, estas tarefas apresentam dificuldades em relação ao tratamento de alta dimensionalidade e esparsidade dos dados, impactando nas análises dos bigrupos e cogrupos obtidos. Além disso, suas implementações envolvem um problema computacional complexo que se agrava no contexto de dados que estão em alta dimensionalidade e esparsidade como é o caso dos dados textuais².

Biagrupamento e coagrupamento são tarefas que têm sido amplamente aplicadas em diferentes áreas, tais como mineração de texto, *web mining*, bioinformática, *marketing*, ecologia, tecnologia de grupo, arqueologia e ciência da computação [4]. Recentemente, na área de aplicação em mineração de textos, tem sido comum encontrar estratégias especialmente preparadas para lidar com as dificuldades que esta área traz, adaptando algoritmos baseados em fatoração de matrizes [5, 6], heurísticos [7, 8], probabilísticos [9, 10], baseados em teoria de conjuntos *fuzzy* [11] e baseados em modelagem com estruturas de grafos [12, 13]. Neste tutorial, um algoritmo heurístico e um algoritmo baseado em fatoração de matrizes são explorados no contexto das tarefas de biagrupamento e coagrupamento, respectivamente.

O objetivo deste tutorial é apresentar conceitos, algoritmos e experimentações que permitam ao leitor compreender como as tarefas de biagrupamento e coagrupamento são implementadas e como a extração da informação pode ser feita a partir dos bigrupos e cogrupos encontrados. O contexto de análise de dados textuais, caracterizado por espaços de busca

¹Na literatura podem ser encontradas outras nomenclaturas para o problema de agrupar ambas dimensões, tais como: *bidimensional clustering*, *block clustering*, *direct clustering*, *multiway clustering*, *simultaneous clustering*, *two-mode clustering*, *two-sided clustering*, *subspace clustering* e *two-way clustering*.

²Tal complexidade é estudada no âmbito de trabalhos que apresentam otimizações em algoritmos aplicados às tarefas, contudo, a discussão sobre abordagens mais sofisticadas e especializadas está fora do escopo deste tutorial.

de alta dimensionalidade e alta esparsidade, é usado como motivador para aplicação de tais tarefas. A fim de organizar o conteúdo citado, este tutorial apresenta as definições e conceitos básicos relacionados à cada tarefa (seção 2), o algoritmo de Cheng e Church [14] para resolução da tarefa de biagrupamento (seção 3), o algoritmo NBVD (*Non-Negative Block Decomposition* [15]) para resolução da tarefa de coagrupamento (seção 4), um contexto de experimentação para cada algoritmo considerando dados sintéticos e do mundo real (seção 5), experimentos executados com o algoritmo de Cheng e Church (seção 6) e com o algoritmo NBVD (seção 7), e uma análise das diferenças entre os resultados obtidos por estes algoritmos (seção 8). Um conjunto de leituras adicionais sobre biagrupamento e coagrupamento aplicados a dados textuais é organizado na seção 9 com o intuito de oferecer caminhos alternativos, atuais e de nível avançado para exploração do tema. O tutorial é finalizado com um resumo destacando os principais pontos abordados no texto (seção 10).

2. Conceitos

Na literatura, as tarefas de biagrupamento e de coagrupamento são, por vezes, referenciadas de maneira intercambiável [16, 17, 18, 19, 20, 21, 22, 23]. Eventualmente, autores tratam problemas que originalmente tem natureza de coagrupamento por meio da aplicação de algoritmos de biagrupamento, e vice-versa. Mirkin[24] já usava o termo biagrupamento para designar a ação de simultaneamente agrupar tanto conjuntos de linhas quanto conjunto de colunas em uma matriz de dados – essa é uma definição também possível para explorar a ação de coagrupar dados. Nessa mesma obra, o autor se refere a biagrupamento em termos de subconjunto de dados e/ou partição de dados. Atualmente, ambos os problemas possuem definições bem estabelecidas e que diferem nas suas naturezas, ou objetivos básicos. Antes de discuti-las, é útil compreender como bigrupos e cogrupos são usados em análise de dados.

Considere o exemplo adaptado de Papadimitriou e Sun[25] representado na matriz $\mathcal{X}^{4 \times 5}$. A matriz $\mathcal{X}$ é composta por um conjunto de linhas (A, B, C, D) que representam estudantes e um conjunto de colunas (ciências, inglês, matemática, chinês e social) que representam as disciplinas que esses estudantes cursam. O conteúdo de cada célula da matriz $\mathcal{X}$ indica se o aluno teve bom desempenho em uma disciplina (valor 1) ou não (valor 0). A matriz $\mathcal{X}$ é:

$$\mathcal{X} = \begin{matrix} & \begin{matrix} \text{ciências} & \text{inglês} & \text{matemática} & \text{chinês} & \text{social} \end{matrix} \\ \begin{matrix} A \\ B \\ C \\ D \end{matrix} & \begin{pmatrix} 0 & 1 & 0 & 1 & 1 \\ 1 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 1 & 1 \\ 1 & 0 & 1 & 0 & 0 \end{pmatrix} \end{matrix}$$

Ao analisar a matriz $\mathcal{X}$, é possível observar a existência de dois bigrupos. O primeiro bigrupo é formado pelas linhas B, D e pelas colunas *ciências* e *matemática*, e representa os estudantes com bom desempenho nas disciplinas de *ciências* e *matemática*. O segundo bigrupo é formado pelas linhas A e C e pelas colunas *inglês*, *chinês* e *social*, e representa os estudantes com bom desempenho nas disciplinas de *inglês*, *chinês*

e *social*. A existência dos bigrupos pode ser observada mais claramente quando uma reorganização das linhas e colunas da matriz é realizada:

$$\mathcal{X} = \begin{matrix} & \begin{matrix} \text{ciências} & \text{matemática} & \text{inglês} & \text{chinês} & \text{social} \end{matrix} \\ \begin{matrix} D \\ B \\ C \\ A \end{matrix} & \begin{pmatrix} 1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 1 \\ 0 & 0 & 1 & 1 & 1 \end{pmatrix} \end{matrix}$$

Alternativamente, a depender das condições impostas a um algoritmo de biagrupamento, bigrupos referentes aos valores 0 também poderiam ser encontrados. Nesse caso, a organização final dos bigrupos se assemelharia à uma organização em cogrupos. Se a mesma matriz $\mathcal{X}$ fosse analisada sob a tarefa de coagrupamento, e sendo esse um problema que espera a determinação de uma partição dos dados, nenhum elemento seria deixado fora de algum grupo de linhas e colunas. Desta forma, dois grupos de linhas e dois grupos de colunas seriam formados, gerando quatro cogrupos. Com a reorganização da matriz $\mathcal{X}$, os cogrupos podem ser visualmente analisados:

$$\mathcal{X} = \begin{matrix} & \begin{matrix} \text{ciências} & \text{matemática} & \text{inglês} & \text{chinês} & \text{social} \end{matrix} \\ \begin{matrix} D \\ B \\ C \\ A \end{matrix} & \begin{pmatrix} 1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 1 \\ 0 & 0 & 1 & 1 & 1 \end{pmatrix} \end{matrix}$$

A partir da análise do exemplo da matriz $\mathcal{X}$, e seguindo a discussão apresentada pelos autores Pensa et al.[26], a diferença entre biagrupamento e coagrupamento é dada da seguinte maneira:

- Um bigrupo é um conjunto de objetos e um conjunto de atributos associados. Cada objeto pertencente a um bigrupo está, apenas, fortemente relacionado a qualquer outro objeto pertencente ao mesmo bigrupo, dado o conjunto de atributos associados. Na tarefa de biagrupamento, procura-se por bigrupos sobrepostos ou não³ que se organizam como **submatrizes de dados**.
- No coagrupamento, procura-se por um número pré-especificado de cogrupos que formam uma **bi-partição** (uma partição de objetos que está fortemente relacionada a uma partição de atributos). Cada objeto pertencente a um grupo de objetos está fortemente relacionado a qualquer outro objeto pertencente ao mesmo grupo, em relação ao subconjunto de atributos que contribuíram para o estabelecimento deste grupo; esses mesmos objetos estão também relacionados entre si considerando os demais atributos, porém essa relação tem menor intensidade ou é de natureza diferenciada. A relação inversa, referente a grupos de atributos diante de subconjunto de objetos, também ocorre.

³Pensa et al.[26] não espera a definição do número de bigrupos *a priori*, contudo, segundo a visão dos autores Madeira e Oliveira[3], o número de bigrupos deve ser conhecido *a priori*. Isso ilustra a variabilidade de visões existentes dentro desse contexto na literatura.

Ainda segundo Pensa et al.[26], um conjunto de bigrupos não fornece um modelo de dados, pois captura associações locais em vez de fornecer uma visão global dos dados, mas fornece grupos significativos que podem divergir da tendência global dos dados. Por outro lado, a tarefa de coagrupamento fornece um modelo descritivo dos dados, no qual a partição de objetos é descrita pela partição de atributos e vice-versa. Ainda, o coagrupamento pode ser visto como uma coleção de bigrupos que satisfazem a propriedade de que o conjunto de bigrupos forma uma bi-partição para o conjunto de objetos original; neste caso, o problema de coagrupamento é colocado como uma generalização do problema de biagrupamento em que os bigrupos cobrem, necessariamente, todos os elementos da matriz de dados. Madeira e Oliveira[3] discutem o problema de biagrupamento generalizado como a tarefa de encontrar um conjunto mínimo de bigrupos que cobre todos os elementos em uma matriz de dados. Segundo essa visão, esse problema é equivalente ao problema de coagrupamento.

Nas próximas seções são apresentados conceitos sobre cada uma das tarefas: biagrupamento e coagrupamento.

2.1 Biagrupamento

Segundo os autores Madeira e Oliveira[3], o biagrupamento é tratado como o problema de encontrar submatrizes de uma matriz de dados. Formalmente, seja $\mathcal{A}$ uma matriz de dimensão $n \times m$, com um conjunto de linhas X , em que $X = \{\mathbf{x}_1, \dots, \mathbf{x}_i, \dots, \mathbf{x}_n\}$, e um conjunto de colunas Y , em que $Y = \{\mathbf{y}_1, \dots, \mathbf{y}_j, \dots, \mathbf{y}_m\}$. A matriz $\mathcal{A}$ pode ser escrita como (X, Y) . Cada elemento da matriz $\mathcal{A}$ é representado por a_{ij} , em que $i \in \{1, \dots, n\}$ é o índice da linha e $j \in \{1, \dots, m\}$ é o índice da coluna, que corresponde a um valor que representa a relação entre a linha i e a coluna j , como ilustrado no exemplo:

$$\mathcal{A} = \begin{matrix} & \mathbf{y}_1 & \dots & \mathbf{y}_j & \dots & \mathbf{y}_m \\ \mathbf{x}_1 & a_{11} & \dots & a_{1j} & \dots & a_{1m} \\ \vdots & \vdots & & \vdots & & \vdots \\ \mathbf{x}_i & a_{i1} & \dots & a_{ij} & \dots & a_{im} \\ \vdots & \vdots & & \vdots & & \vdots \\ \mathbf{x}_n & a_{n1} & \dots & a_{nj} & \dots & a_{nm} \end{matrix}.$$

Se $I \subseteq X$ e $J \subseteq Y$ são subconjuntos de linhas e colunas respectivamente, $A_{IJ} = (I, J)$ denota a submatriz A_{IJ} de $\mathcal{A}$ que contém apenas os elementos a_{ij} pertencentes à submatriz com o conjunto de linhas I e o conjunto de colunas J . Um bigrupo é um subconjunto de linhas que exibem comportamento semelhante em um subconjunto de colunas e vice-versa. O bigrupo $A_{IJ} = (I, J)$ é um subconjunto de linhas e um subconjunto de colunas, em que $I = \{\mathbf{i}_1, \dots, \mathbf{i}_p\}$ é um subconjunto de linhas ($I \subseteq X$ e $p \leq n$), e $J = \{\mathbf{j}_1, \dots, \mathbf{j}_s\}$ é um subconjunto de colunas ($J \subseteq Y$ e $s \leq m$). O bigrupo (I, J) é definido como uma submatriz de dimensão $p \times s$ da matriz de dados $\mathcal{A}$.

O problema de biagrupamento trata de identificar um conjunto de bigrupos $B_{IJ} = (I, J)$, de forma que cada bigrupo B_{IJ} satisfaça algumas características específicas de homogeneidade. As características de homogeneidade influenciam o es-

tabelecimento de taxonomias que organizam diferentes tipos e diferentes estruturas de bigrupos⁴. Nas próximas seções, duas taxonomias de organização de bigrupos são apresentadas.

2.1.1 Tipos de bigrupos

Vários tipos de bigrupos foram descritos e categorizados na literatura, dependendo do comportamento que eles apresentam [27, 3]. Madeira e Oliveira[3] identificam quatro tipos de bigrupos:

1. Bigrupos com valores constantes: São aqueles que possuem todos seus elementos com o mesmo valor, de modo que seus elementos podem ser descritos como:

$$a_{ij} = \mu, \forall i, j \in I, J \quad (1)$$

sendo μ o valor constante para todos os elementos de A_{IJ} . Geralmente esse tipo de bigrupo não é encontrado em dados do mundo real devido à presença ruído. Para tratar casos com ruído, a definição da equação 1 é modificada para $a_{ij} = \mu + \eta_{ij}$, em que η_{ij} é o ruído associado aos valores de μ e a_{ij} . Na figura 2 é mostrado um exemplo de um bigrupo com valores constantes.

Figura 2. Exemplo de bigrupo com valores constantes.

Adaptado de Madeira e Oliveira[3]

3,0	3,0	3,0	3,0
3,0	3,0	3,0	3,0
3,0	3,0	3,0	3,0
3,0	3,0	3,0	3,0

2. Bigrupos com valores constantes nas linhas ou nas colunas: São aqueles que apresentam valores constantes nas linhas ou nas colunas, e podem ser representados como:

Valores constante nas linhas: Um bigrupo perfeito com valores constantes nas linhas é uma submatriz na qual todos os valores dentro do bigrupo podem ser obtidos usando as equações 2 ou 3.

$$a_{ij} = \mu + \chi, \forall i, j \in I, J \quad (2)$$

$$a_{ij} = \mu \times \chi, \forall i, j \in I, J \quad (3)$$

em que μ é o valor típico dentro do bigrupo e χ é o ajuste para a linha i , que pode ser obtido de forma aditiva (equação 2) ou multiplicativa (equação 3).

Valores constante nas colunas: Um bigrupo perfeito que apresenta um comportamento constante nas colunas é definido como uma submatriz na qual todos seus valores são obtidos usando as equações 4 ou 5.

⁴Embora sejam taxonomias à parte da literatura de coagrupamento, elas também são úteis para definição de objetivos e restrições em aplicações nas quais a modelagem da tarefa de coagrupamento será realizada, como no presente tutorial.

$$a_{ij} = \mu + \beta_j, \forall i, j \in I, J \quad (4)$$

$$a_{ij} = \mu \times \beta_j, \forall i, j \in I, J \quad (5)$$

em que μ é o valor típico dentro do bigrupo e β_j é o ajuste para a coluna j que pode ser obtido de forma aditiva (equação 4) ou multiplicativa (equação 5). Na figura 3 são mostrados exemplos de dois bigrupos, um com valores constantes nas linhas e outro com valores constantes nas colunas. Na figura 3a é mostrado um bigrupo com valores constantes nas linhas usando um ajuste aditivo, em que $\mu = 3,0$ e os valores de γ para cada linha i são $\{0, 3, 2, -1\}$. Similarmente, na figura 3b é mostrado um bigrupo com valores constantes nas colunas usando um ajuste multiplicativo, em que $\mu = 3,0$ e os valores de β para cada coluna j são $\{1, 2, 3, 0, 5\}$.

Figura 3. Exemplos de bigrupos com valores constantes nas linhas usando um ajuste aditivo (a) e nas colunas usando um ajuste multiplicativo (b). Adaptado de Madeira e Oliveira[3]

(a)				(b)			
3,0	3,0	3,0	3,0	3,0	6,0	9,0	1,5
6,0	6,0	6,0	6,0	3,0	6,0	9,0	1,5
5,0	5,0	5,0	5,0	3,0	6,0	9,0	1,5
2,0	2,0	2,0	2,0	3,0	6,0	9,0	1,5

3. Bigrupos com valores coerentes: São aqueles que não têm um valor constante para cada linha ou coluna porque recebem contribuições tanto por linha quanto por coluna. Eles podem ser representados como:

Valores coerentes baseados em um modelo aditivo: Um bigrupo perfeito com valores coerentes é definido como um subconjunto de linhas e um subconjunto de colunas cujos valores são preditos usando a expressão:

$$a_{ij} = \mu + \gamma_i + \beta_j, \forall i, j \in I, J \quad (6)$$

sendo μ o valor típico dentro do bigrupo, γ_i o ajuste para a linha i e β_j o ajuste para a coluna j . Os bigrupos com valores constante nas linhas (exemplo a figura 3a) ou nas colunas usando um ajuste aditivo podem ser considerados casos especiais deste modelo aditivo geral, já que a coerência dos valores pode ser observada nas linhas e nas colunas do bigrupo respectivamente. Desta forma, as equações 2 e 4 são casos especiais do modelo representado pela equação 6 quando $\beta_j = 0$ e $\gamma_i = 0, \forall i, j \in I, J$ respectivamente [3].

Valores coerentes baseados em um modelo multiplicativo: Os valores dentro do bigrupo podem ser obtidos usando a expressão:

$$a_{ij} = \mu' \times \gamma_i' \times \beta_j', \forall i, j \in I, J \quad (7)$$

sendo μ' o valor típico dentro do bigrupo, γ_i' o ajuste para a linha i e β_j' o ajuste para a coluna j . Os bigrupos com valores constantes nas linhas ou nas colunas (exemplo da figura 3b) de forma multiplicativa também podem ser considerados casos especiais deste modelo multiplicativo. Na figura 4 são mostrados exemplos de bigrupos com valores coerentes nas linhas e nas colunas. A figura 4a mostra um bigrupo com valores coerentes nas linhas e nas colunas usando um modelo aditivo, em que $\mu = 3,0$, os valores de γ para cada linha de i são $\{0, 3, 2, -1\}$ e os valores de β para cada coluna de j são $\{0, -1, 1, 3\}$. A figura 4b mostra um bigrupo com valores coerentes nas linhas e colunas usando um modelo multiplicativo, em que $\mu' = 3,0$, os valores de γ' para cada linha de i são $\{0, 1, 2, 3\}$ e os valores de β' para cada coluna de j são $\{1, 2, 3, 0, 5\}$.

Figura 4. Exemplos de bigrupos com valores coerentes nas linhas e nas colunas usando um modelo aditivo (a) e usando um modelo multiplicativo (b). Adaptado de Madeira e Oliveira[3]

(a)				(b)			
3,0	2,0	4,0	6,0	0,0	0,0	0,0	0,0
6,0	5,0	7,0	9,0	3,0	6,0	9,0	1,5
5,0	4,0	6,0	8,0	6,0	12,0	18,0	3,0
2,0	1,0	3,0	5,0	9,0	18,0	27,0	4,5

4. Bigrupos com evoluções coerentes: A evolução coerente procura por **propriedades** coerentes nos valores. Os bigrupos são visualizados como valores simbólicos que tem comportamentos coerentes, independentemente de seus valores numéricos exatos. O bigrupo pode ter a evolução coerente ou coevolução em todos os seus elementos, tanto nas linhas como nas colunas (figura 5a), apenas nas linhas (figura 5b) ou apenas nas colunas (figuras 5c e 5d), gerando uma similaridade com o conceito de bigrupos com valores constantes. Na figura 5d é mostrado um bigrupo com evolução coerente nas colunas onde a propriedade de coevolução é: $a_{i4} \leq a_{i2} \leq a_{i3} \leq a_{i1}, \forall i \in I$.

Figura 5. Exemplos de bigrupos com evoluções coerentes [3]

(a)				(b)			
S1	S1	S1	S1	S1	S1	S1	S1
S1	S1	S1	S1	S2	S2	S2	S2
S1	S1	S1	S1	S3	S3	S3	S3
S1	S1	S1	S1	S4	S4	S4	S4

(c)				(d)			
S1	S2	S3	S4	70	13	19	10
S1	S2	S3	S4	49	40	49	35
S1	S2	S3	S4	40	20	27	15
S1	S2	S3	S4	90	15	30	12

Os autores Mukhopadhyay, Maulik e Bandyopadhyay[27] apresentam uma classificação que organiza os bigrupos em seis tipos: (1) bigrupos com valores constantes, (2) bigrupos

com linhas constantes, (3) bigrupos com colunas constantes, (4) bigrupos com padrão aditivo, (5) bigrupos com padrão multiplicativo e (6) bigrupos com padrão aditivo e multiplicativo. Esses seis tipos correspondem aos três primeiros tipos de bigrupos (bigrupos com valores constantes, bigrupos com valores constantes nas linhas ou nas colunas e bigrupos com valores coerentes) apresentados na classificação de [Madeira e Oliveira\[3\]](#).

2.1.2 Estruturas de biagrupamento

Os algoritmos de biagrupamento assumem as seguintes situações: encontrar apenas um bigrupo (figura 6a), ou encontrar K bigrupos (figura 6b a 9b)⁵. [Madeira e Oliveira\[3\]](#) definem nove estruturas de biagrupamento:

a. Único bigrupo: É a estrutura mais simples na qual é visualizado um único bigrupo no conjunto de dados. Esta estrutura é mostrada na figura 6a.

b. Bigrupos com linhas e colunas exclusivas: Nesta estrutura de biagrupamento todas as linhas e todas as colunas do conjunto de dados pertencem exclusivamente a um dos K bigrupos considerados. As linhas que pertencem a um bigrupo exibem um comportamento semelhante e podem ser expressas considerando apenas as colunas que pertencem a esse mesmo bigrupo. Este tipo de estrutura é mostrada na figura 6b.

c. Bigrupos com estrutura de tabuleiro de xadrez: Uma estrutura de tabuleiro de xadrez permite a existência de K bigrupos não sobrepostos e não exclusivos, sendo que cada linha e cada coluna no conjunto de dados pertence a exatamente $\sqrt{K}$ bigrupos. Este tipo de estrutura é mostrado na figura 6c.

Figura 6. Estrutura de: (a) um único bigrupo; (b) bigrupos com linhas e colunas exclusivas; (c) bigrupos com estrutura de tabuleiro de xadrez. Adaptado de [Madeira e Oliveira\[3\]](#)

d. Bigrupos com linhas exclusivas: Esta estrutura assume que as linhas do conjunto de dados só podem pertencer a um bigrupo, enquanto as colunas podem pertencer a vários bigrupos, como apresentado na figura 7a.

e. Bigrupos com colunas exclusivas: Esta estrutura assume que as colunas do conjunto de dados só podem pertencer a um bigrupo, enquanto as linhas da matriz de dados podem pertencer a um ou mais bigrupos, como mostrado na figura 7b.

⁵Nessas figuras considere que cada um dos quadros representa um conjunto de dados, e cada quadrado ou retângulo dentro do quadro, com diferentes tons de azul, representa um bigrupo desse conjunto de dados. Dentro do

Figura 7. Estrutura de bigrupos: (a) com linhas exclusivas; (b) com colunas exclusivas. Adaptado de [Madeira e Oliveira\[3\]](#)

f. Bigrupos sem sobreposição com estrutura em árvore:

Segundo [Hartigan\[28\]](#), se em um biagrupamento qualquer dois bigrupos são disjuntos, porém suas projeções nas linhas ou nas colunas não o são, então eles formam uma estrutura em árvore. Na figura 8a, esse tipo de organização é ilustrada. As retas externas ao quadro indicam a hierarquia formada pelas projeções dos bigrupos nas linhas e nas colunas. Na projeção sobre linhas, a hierarquia é $\{a, b, d, c\}$, considerando um caminho pré-ordem. Na projeção sobre colunas, a hierarquia é $\{a, b, c, d\}$, também considerando um caminho pré-ordem.

Figura 8. Estrutura de bigrupos: (a) sem sobreposição com estrutura em árvore; (b) não exclusivos e sem sobreposição. Adaptado de [Madeira e Oliveira\[3\]](#) e [Hartigan\[28\]](#)

g. Bigrupos não exclusivos e sem sobreposição: Esta estrutura assume que tanto as linhas quanto as colunas do conjunto de dados podem pertencer a vários bigrupos, contudo, nenhum dos bigrupos presentes na estrutura se sobrepõem. Este tipo de estrutura é mostrada na figura 8b.

h. Bigrupos com sobreposição e com estrutura hierárquica: Esta estrutura assume que a hierarquia pode também ocorrer dentro de bigrupos. Na figura 9a, essa estrutura é ilustrada juntamente com a representação das projeções dos bigrupos nas linhas e colunas (retas externas ao quadro). Na projeção sobre linhas, a hierarquia é $\{b, c, d, e, f\}$, considerando um caminho pré-ordem. Na projeção sobre colunas, a hierarquia é $\{x, b, d, f, c, e\}$, em que x representa um nó pai artificial para b e c .

quadro, áreas em branco representam valores que não possuem nenhum tipo de relação ou significado para o biagrupamento desse conjunto.

i. Bigrupos com sobreposição e arbitrariamente posicionados: Esta estrutura permite a existência de K bigrupos possivelmente sobrepostos. Esses bigrupos são não exclusivos, ou seja, existem linhas e colunas que podem pertencer a vários bigrupos, e podem ser não exaustivos, o que significa que algumas linhas ou colunas podem não pertencer a nenhum bigrupo. Este tipo de estrutura é mostrada na figura 9b.

Figura 9. Estrutura de bigrupos: (a) com sobreposições e com estrutura hierárquica; (b) com sobreposições e arbitrariamente posicionados. Adaptado de Madeira e Oliveira[3]

As estruturas de biagrupamento **b**, **c**, **d**, **e**, **f** e **g** apresentam algumas características semelhantes: nelas é assumido que os bigrupos são exaustivos ou seja todas as linhas e todas as colunas do conjunto de dados pertencem a pelo menos um bigrupo; e nenhuma delas permite a sobreposição de bigrupos, ou seja, nenhuma dessas estruturas torna possível que um par particular (linha, coluna) pertença a mais de um bigrupo. No entanto, é mais provável que em dados reais, algumas linhas ou colunas não pertençam a nenhum bigrupo (bigrupos não exaustivos) e que os bigrupos se sobreponham em alguns locais, como é o caso das estrutura **h** e **i**.

Pontes, Giráldez e Aguilar-Ruiz[29] apresentam uma segunda taxonomia para a estrutura dos bigrupos que podem ser buscados pelos algoritmos. Embora simplificada, nesta taxonomia são apresentadas regras de formação de bigrupos que abrangem as mesmas estruturas presentes na taxonomia apresentada por Madeira e Oliveira[3]. Seis regras são apresentadas:

- a. Linha exaustiva:** Todas as linhas da matriz de dado devem pertencer a pelo menos um bigrupo.
- b. Coluna exaustiva:** Todas as colunas da matriz de dados devem pertencer a pelo menos um bigrupo.
- c. Não exaustiva:** Não é requerido que todas as linhas ou todas as colunas da matriz de dados estejam associadas aos bigrupos.
- d. Linha exclusiva:** Cada linha da matriz de dados pode pertencer a no máximo um bigrupo.
- e. Coluna exclusiva:** Cada coluna da matriz de dados pode pertencer a no máximo um bigrupo.

f. Não exclusivas: Bigrupos com sobreposições podem ser obtidos, tanto em relação às linhas quanto em relação às colunas da matriz de dados.

Note que se um algoritmo procura atender as estruturas **a** e **b**, simultaneamente, ele está procurando por uma partição da matriz de dados. De acordo com as definições usadas neste tutorial, esta seria uma situação na qual biagrupamento e coagrupamento se tornam tarefas equivalentes.

2.2 Coagrupamento

O coagrupamento é o problema de agrupar simultaneamente as linhas e as colunas da matriz de dados. O coagrupamento procura “blocos” (ou *cogrupos*) de linhas e colunas que estejam inter-relacionados [16, 17], em algum sentido.

Formalmente, o coagrupamento pode ser definido da seguinte forma⁶ [26]: dada uma matriz $\mathcal{X}$ de dimensão $m \times n$ ($\mathcal{X} \in \mathbb{R}^{m \times n}$), seja x_{ij} o elemento correspondente à linha i e à coluna j ; $\vec{x}_i$ e $\vec{y}_j$ indicam os vetores associados respectivamente à linha i e à coluna j . Um coagrupamento $\mathcal{C}^{k \times l}$ sobre $\mathcal{X}$ produz simultaneamente um conjunto de $k \times l$ cogrupos, ou seja uma partição $\mathcal{C}^r$ em k grupos de linhas associada a uma partição $\mathcal{C}^c$ em l grupos de colunas, que otimizam uma determinada função objetivo.

Uma definição alternativa, com a ideia de dados diádicos, provém do uso de métodos de fatoração de matrizes na modelagem e resolução do problema de coagrupamento⁷. Estes métodos abordam a matriz $\mathcal{X}$ como uma tabela de co-ocorrências, i.e., um caso simples de dados diádicos. Os dados diádicos referem-se a um domínio com dois conjuntos finitos de objetos $X = \{\mathbf{x}_1, \dots, \mathbf{x}_i, \dots, \mathbf{x}_n\}$ e $Y = \{\mathbf{y}_1, \dots, \mathbf{y}_j, \dots, \mathbf{y}_m\}$, nos quais observações são feitas para díades $(\mathbf{x}_i, \mathbf{y}_j)$. Uma díade é um valor escalar $w(\mathbf{x}_i, \mathbf{y}_j)$ que pode representar, por exemplo, a frequência de co-ocorrência, a força de preferência, associação, ou o nível de expressão [31]. Organizando os dados na matriz $\mathcal{X}$, cada díade $w(\mathbf{x}_i, \mathbf{y}_j)$ corresponderá a um elemento de $\mathcal{X}$, e a fatoração de matrizes será usada para agrupar simultaneamente linhas e colunas na matriz $\mathcal{X}$, interligando agrupamentos de linhas com agrupamentos de colunas a cada iteração [32, 15, 33, 34].

O problema de coagrupamento é similar ao problema de agrupamento (que agrupa as linhas de uma matriz de dados), mas traz algumas vantagens. De acordo com [35], os algoritmos para coagrupamento são mais eficazes para trabalhar com dados de alta dimensionalidade e alta esparsidade, característica de dados textuais (cf. interesse deste tutorial). Segundo os autores Dhillon, Mallela e Modha[32] e Long, Zhang e Yu[15], ao usar grupos de atributos como recursos subjacentes, e não o conjunto completo de atributos como fazem os algoritmos para agrupamento, o coagrupamento realiza uma redução

⁶Normalmente, definições de coagrupamento seguem a noção e notação de espaços vetoriais.

⁷A visão de dados diádicos é também a raiz da abordagem apresentada por Lee e Seung[30]. Esses autores apresentam a análise de objetos por meio da análise de suas partes. Nesse caso, as partes são equivalentes a grupos de atributos descritivos desses objetos.

de dimensionalidade implícita e adaptativa, assim como minimiza o problema de ruído nos dados.

O coagrupamento descreve grupos de objetos associando-os a grupos de atributos. Segundo Yoo e Choi[33], ao determinar simultaneamente os grupos de atributos e de objetos, o coagrupamento identifica como os grupos de atributos estão relacionados aos grupos de objetos, e traz conhecimento sobre relações entre pares de atributos diante de determinados grupos. Em mineração de textos, o coagrupamento pode levar diretamente à descoberta de palavras polissêmicas (palavras que podem ter múltiplos sentidos e vários tipos de uso em diferentes contextos [36]). Por exemplo, se as palavras (atributos) ‘dados’, ‘memória’, ‘software’ e ‘cache’ fazem parte de um grupo de atributos, a palavra ‘dados’ representa um elemento em uma arquitetura computacional; ao mesmo tempo, se a palavra ‘dados’ aparece em um outro grupo de atributos no qual também existem as palavras ‘jogo’, ‘regras’ e ‘sorte’, então ela pode se referir a um objeto usado durante a realização de uma partida de um jogo de azar. No caso do agrupamento, uma descoberta como essa exigiria um exame adicional dos resultados referentes aos grupos de objetos (documentos, no caso de mineração de textos) encontrados.

Ainda, o coagrupamento apresenta algumas diferenças em relação ao biagrupamento. Analizando as estruturas de biagrupamento da subseção 2.1.2 com uma visão de coagrupamento, os cogrupos mostrados poderiam variar em relação à visão de bigrupos. Por exemplo, na estrutura d existem três bigrupos (figura 10a), entretanto, se a mesma estrutura for analisada sob a visão de coagrupamento, mais cogrupos seriam observados. Um exemplo está representado na figura 10b, na qual seis cogrupos são contabilizados, considerando aqueles em cor branca, correspondentes aos cogrupos 4, 5 e 6. Note que o cogrupo 5 está dividido em duas regiões na figura. Isso ocorre porque, para a visualização, não é possível reorganizar as colunas sem que outros cogrupos passem a ficar também divididos em mais de uma região. Em termos de coagrupar documentos e palavras, esses cogrupos apresentariam pouco ou nenhum significado semântico se eles estivessem dizendo respeito a cogrupos com valores nulo (zero).

Figura 10. Visão de coagrupamento na estrutura d , discutida na seção 2.1.2

2.3 Qualidade de bigrupos e cogrupos

A avaliação dos resultados obtidos para biagrupamento e coagrupamento depende dos critérios de otimização usados nos algoritmos. Contudo, uma avaliação que independe

dos algoritmos usados e foca nos resultados finais, também é útil. Na tarefa de agrupamento, a avaliação da qualidade dos resultados é chamada de validação [2, 37, 38], e de maneira semelhante, validações devem ser realizadas sobre os resultados dos algoritmos de biagrupamento e coagrupamento. Inclusive, a informação extraída desses resultados, por vezes, reflete a existência de grupos (de linhas e de colunas) e, assim, a validação de agrupamento pode ser transferida para o contexto destas tarefas. Uma visão que defende que os algoritmos de coagrupamento produzem soluções mais acuradas para a tarefa de agrupamento é apresentada na literatura da área [33, 39, 34, 40], e portanto, medidas de validação de grupos são comumente aplicadas e os resultados do coagrupamento são comparados aos resultados produzidos para agrupamento.

2.3.1 Qualidade de bigrupos

Na literatura, são apresentadas medidas especificamente elaboradas para avaliação de bigrupos [3, 27]. A escolha de qual medida usar está relacionada com as características desejadas para os bigrupos em uma determinada aplicação. No trabalho de Castro et al.[41] são apresentadas as seguintes medidas:

Resíduo quadrático médio (RQM): Embora seja uma medida associada ao processo de minimização buscado no algoritmo de Cheng e Church [14], ela pode ser usada como uma medida geral de coerência ou de homogeneidade dos bigrupos. Diante desta medida, um bigrupo perfeito é aquele que apresenta RQM igual a 0.

Volume: O volume do bigrupo é dado pela quantidade de elementos alocados em um bigrupo. Geralmente, é esperado que o volume do bigrupo seja maximizado desde que uma medida de qualidade interna, como o RQM, seja minimizada.

Cobertura do conjunto de dados: A cobertura de um conjunto de dados corresponde à quantidade de elementos da matriz sob análise que pertence aos bigrupos gerados [41].

Grau de sobreposição: O grau de sobreposição entre os bigrupos é a interseção dos conjuntos de elementos pertencentes a cada um deles. Indica o quanto um determinado bigrupo está contido em outro bigrupo [41].

Grau de ocupação: O grau de ocupação dos bigrupos é a quantidade de valores não nulos presentes nos bigrupos [41].

2.3.2 Qualidade de agrupamento

Ao contrário do que ocorre para biagrupamento, a literatura não apresenta medidas especificamente elaboradas para avaliação de cogrupos. Portanto, os autores comumente usam avaliações que dizem respeito ao resultado de agrupamento que pode ser derivado do processo de coagrupamento (projeção dos cogrupos na dimensão de linhas e colunas), e avaliações qualitativas voltadas aos objetivos da aplicação sobre a qual o problema é modelado. Neste tutorial, as seguintes medidas de avaliação internas [42, 43], provenientes da literatura de agrupamento, são aplicadas:

Similaridade média intra-grupos: Essa medida calcula a similaridade média entre os objetos que pertencem a um mesmo grupo. [Buono e Pio\[34\]](#) definem a similaridade média⁸ de um determinado grupo G_z por meio da equação 8:

$$IntraSim(G_z) = \frac{\sum_{i=1}^{p_z} \sum_{j=i+1}^{p_z} sim(\mathbf{x}_i, \mathbf{x}_j)}{0.5 \times p_z \times (p_z - 1)}, \quad (8)$$

em que $\mathbf{x}_i$ e $\mathbf{x}_j$ representam os objetos pertencentes ao grupo G_z , p_z é o número de objetos no grupo G_z , e $sim(\mathbf{x}_i, \mathbf{x}_j)$ é uma função de similaridade⁹. A similaridade média intra-grupos para a partição obtida pelo algoritmo é definida pela equação 9:

$$IntraSim = \frac{\sum_{i=1}^k IntraSim(G_i) \times p_i}{n}, \quad (9)$$

em que n é o número total de objetos, k é o número de grupos de objetos e p_i é o número de objetos no grupo i .

Similaridade média dos centróides inter-grupos: Essa medida calcula a similaridade média entre objetos pertencentes a diferentes grupos. [Buono e Pio\[34\]](#) calcula a similaridade entre os centróides dos grupos, em vez de calcular a média entre todos os objetos. Esta medida é definida pela equação 10:

$$InterSim = \frac{\sum_{i=1}^k \sum_{j=i+1}^k sim(centroid_i, centroid_j)}{0.5 \times k \times (k - 1)}, \quad (10)$$

em que $centroid_i$ é o centroide do i – ésimo grupo, k é o número de grupos de objetos, e $sim(centroid_i, centroid_j)$ é a função de similaridade.

Índice Davies Bouldin (DB): É uma medida que avalia se os grupos são compactos e se eles exibem uma boa separabilidade entre si. Para isso é necessário definir uma medida geral da separação de grupos, baseada na relação entre a similaridade intra-grupos e a similaridade inter-grupos (a similaridade média de cada um dos grupos com o grupo mais próximo a ele) [45]. O índice DB foi definido pelos autores [Davies e Bouldin\[45\]](#) por meio da equação 11:

$$DB = \frac{1}{k} \sum_{\substack{i=1 \dots k \\ j=1 \dots k \\ i \neq j}} \max \left(\frac{disp(G_i) + disp(G_j)}{dist(G_i, G_j)} \right), \quad (11)$$

⁸No decorrer deste tutorial, a similaridade média intra-grupos e a similaridade média dos centróides inter-grupos foram nomeadas como distância intra-grupos e distância inter-grupos respectivamente, para fins de alinhamento com a implementação realizada para os experimentos. Assim, valores menores de distância intra-grupos indicam maior similaridade dos objetos pertencentes ao mesmo grupo. De forma similar, valores maiores de distância inter-grupos indicam maior separação dos centroides dos grupos.

⁹Neste tutorial, a função de similaridade foi implementada usando a distância euclidiana [44].

em que k é o número de grupos, $disp(G_i)$ é uma medida da dispersão do grupo G_i , $disp(G_j)$ é a uma medida da dispersão do grupo G_j e $dist(G_i, G_j)$ é a distância entre os centróides dos grupos G_i e G_j . Na versão mais simples e adotada neste tutorial, o cálculo de $dist(G_i, G_j)$ (a medida inter-grupo) é baseada nos centroides dos grupos, as medidas de dispersão (medida intra-grupo) são baseadas na distância média entre os objetos pertencentes a um grupo, e a distância euclidiana é adotada em ambos os cálculos. Valores pequenos para o índice DB indicam grupos compactos cujos centros estão bem separados uns dos outros.

2.4 A escolha dos algoritmos para estudo

No contexto deste tutorial, a fim de explorar o problema de aplicação de algoritmos para biagrupamento e coagrupamento em dados textuais, dois algoritmos, específicos para resolução de cada uma das tarefas, foram escolhidos para estudo. Essa escolha foi baseada nos seguintes critérios:

- escolha de um representante para cada uma das tarefas, com o fim de propiciar uma análise da relação entre a definição da natureza de cada tarefa com as características de um problema de análise de dados textuais, a saber: alta dimensionalidade de dados e esparsidade da matriz de dados;
- escolha de algoritmos que fossem capazes de encontrar tipos e/ou estruturas de bigrupos, ou cogrupos, de interesse para a tarefa de análise de dados textuais, a saber: bigrupos/cogrupos densos – dentro da matriz de dados esparsa; bigrupos/cogrupos com sobreposição de linhas e/ou colunas, em formatos arbitrários;
- escolha de algoritmos básicos, i.e., os algoritmos escolhidos representam a versão básica da estratégia algorítmica neles implementadas, para propiciar uma análise didática das virtudes e problemas desses algoritmos.

Os dois algoritmos escolhidos para estudo são: algoritmo de Cheng e Church [14] e algoritmo de decomposição de valores em blocos não negativos (do inglês *Non-negative Block Decomposition* - NBVD) [15]. Nas próximas seções, esses algoritmos são discutidos.

3. Algoritmo de Cheng e Church

A presente seção está baseada nos trabalhos de [Franca\[46\]](#), [Tanay, Sharan e Shamir\[47\]](#) e [Cheng e Church\[14\]](#). O algoritmo de Cheng e Church é um dos algoritmos de biagrupamento mais conhecidos porque foi o primeiro a aplicar biagrupamento a dados de *microarrays* de genes [3]. No seu trabalho seminal [14], Cheng e Church estavam interessados em encontrar bigrupos, ou *δ -biclusters*, baseados na minimização do RQM (Resíduo Quadrático Médio). Na visão de Cheng e Church, um bigrupo é um subconjunto de elementos de uma matriz que possuem uma alta taxa de similaridade, a qual é medida por meio de um cálculo que os autores denominam

como “coerência”. Os autores também destacam a resolução do problema de encontrar os bigrupos com a possibilidade de sobreposição. Tais grupos são encontrados a partir da projeção dos bigrupos sobre as linhas e colunas da matriz de dados. Na figura 11, está ilustrada a noção de sobreposição de bigrupos e da sobreposição de grupos em termos da projeção destes sobre linhas e colunas. Na figura, são consideradas três situações: a primeira e a segunda mostram bigrupos não sobrepostos mas cujas projeções, nas linhas e colunas, geram grupos com sobreposição; a terceira mostra bigrupos com sobreposição gerando as sobreposições nas projeções.

Cheng e Church usam o RQM de uma matriz como a medida de coerência ou de homogeneidade dos bigrupos. No algoritmo introduzido por esses autores, essa medida conduz o processo de otimização, ou busca, das submatrizes que representam os bigrupos em uma matriz de dados. A base da medida RQM é o resíduo de elementos individuais em um bigrupo. Dado os subconjuntos de linhas I e de colunas J , o resíduo do elemento a_{ij} é

$$r(a_{ij}) = a_{ij} - a_{iJ} - a_{IJ} + a_{IJ}, \quad (12)$$

em que $a_{iJ} = \frac{1}{|J|} \sum_{j \in J} a_{ij}$, $a_{IJ} = \frac{1}{|I|} \sum_{i \in I} a_{ij}$ e, por fim, $a_{IJ} = \frac{1}{|I||J|} \sum_{i \in I, j \in J} a_{ij}$. Nessas equações, a_{iJ} é a média da i -ésima linha no bigrupo, a_{IJ} é a média da j -ésima coluna no bigrupo, e a_{IJ} é a média de todos os elementos do bigrupo.

Para uma análise mais detalhada sobre essa medida, considere a matriz de dados $\mathcal{A}$, de dimensão 4×4 , que apresenta um bigrupo ilustrado pelo quadrado azul.

$$\mathcal{A} = \begin{pmatrix} 2 & 1 & 27 & 0 \\ 6 & 7 & 20 & 15 \\ 10 & 0 & 5 & 8 \\ 1 & 4 & 9 & 12 \end{pmatrix}$$

O bigrupo, denotado como $\mathbf{A}_1(I, J)$, é composto pelo subconjunto de linhas $I = \{1, 2\}$ e pelo subconjunto de colunas $J = \{1, 2\}$. O cálculo do resíduo de um elemento (a_{11}) a partir da equação 12, é:

$$\mathbf{A}_1(I, J) = \begin{pmatrix} 2 & 1 \\ 6 & 7 \end{pmatrix}$$

$$\begin{aligned} r(a_{11}) &= a_{11} - a_{1J} - a_{IJ} + a_{IJ} \\ &= 2 - 1,5 - 4,0 + 4,0 \\ &= 0,5. \end{aligned}$$

Para fins de comparação e compreensão do significado do resíduo, considere uma segunda submatriz $\mathbf{A}_2(I, J)$, composta pelo subconjunto de linhas $I = \{1, 2\}$ e pelo subconjunto de colunas $J = \{3, 4\}$ e o cálculo do seu resíduo:

$$\mathbf{A}_2(I, J) = \begin{pmatrix} 27 & 0 \\ 20 & 15 \end{pmatrix}$$

$$\begin{aligned} r(a_{11}) &= a_{11} - a_{1J} - a_{IJ} + a_{IJ} \\ &= 27 - 13,5 - 23,5 + 15,5 \\ &= 5,5. \end{aligned}$$

Nesses cálculos, observa-se as diferenças existentes entre os valores do resíduo de um elemento pertencente a um bigrupo ($r(a_{11}) = 0,5 \in \mathbf{A}_1(I, J)$) e o resíduo de um elemento pertencente a uma submatriz que não é considerada um bigrupo ($r(a_{11}) = 5,5 \in \mathbf{A}_2(I, J)$). O valor de resíduo $= 0,5$ indica uma maior coerência desse valor com respeito aos demais valores do bigrupo ao qual ele pertence, do que o valor de resíduo $= 5,5$, o qual indica que não existe uma coerência nos valores do bigrupo, ou se existir uma coerência, ela é baixa. Com base nessa noção de resíduo, Cheng e Church definem o problema de biagrupamento como sendo o problema de encontrar bigrupos, preferencialmente grandes, porém com RQM mínimo ou menor do que um limiar.

Seguindo o já exposto, o RQM de uma submatriz (ou de um bigrupo) pode ser obtido por meio da equação 13, assumindo que $\mathcal{A}$ é uma matriz de dados, a dupla (I, J) representa um *bicluster* ($\mathbf{A}$) em $\mathcal{A}$, em que I é o conjunto de linhas no *bicluster* e J é o conjunto de colunas, e uma submatriz de $\mathcal{A}$ é definida por $\mathbf{A}_{IJ}$. Assim,

$$\begin{aligned} H(I, J) &= \frac{1}{|I||J|} \sum_{i \in I, j \in J} (r(a_{ij}))^2 \\ &= \frac{1}{|I||J|} \sum_{i \in I, j \in J} (a_{ij} - a_{iJ} - a_{IJ} + a_{IJ})^2, \end{aligned} \quad (13)$$

sendo que $H(I, J)$ é o RQM da submatriz $\mathbf{A}_{IJ}$, $|I|$ é a quantidade de linhas no subconjunto de linhas I em $\mathbf{A}$, $|J|$ é a quantidade de colunas no subconjunto de colunas J e a_{ij} , a_{iJ} e a_{IJ} são equivalentes às definições da equação 12. A submatriz de $\mathbf{A}_{IJ}$ é chamada *δ -bicluster*, se $H(I, J) \leq \delta$ para algum $\delta \in \mathbb{R}_+$.

Seguindo o exemplo da matriz de dados $\mathcal{A}$ usada no exemplo sobre resíduo, e considerando o bigrupo nela contido $\mathbf{A}_1(I, J)$, os valores de a_{iJ} , a_{iI} e a_{IJ} para o subconjunto de linhas I e o subconjunto de colunas J de acordo com a equação 12 são: $a_{1J} = 1,5$, $a_{2J} = 6,5$, $a_{1I} = 4,0$, $a_{2I} = 4,0$ e $a_{IJ} = 4,0$. Os valores dos resíduos de cada elemento do

Figura 11. Exemplo de sobreposições passíveis de serem obtidas com o algoritmo de Cheng e Church

bigrupo $A_1(I, J)$ são:

$$\begin{aligned} r(a_{11}) &= 2 - 1,5 - 4,0 + 4,0 &= 0,5 \\ r(a_{12}) &= 1 - 1,5 - 4,0 + 4,0 &= -0,5 \\ r(a_{21}) &= 6 - 6,5 - 4,0 + 4,0 &= -0,5 \\ r(a_{22}) &= 7 - 6,5 - 4,0 + 4,0 &= 0,5, \end{aligned}$$

então, considerando a equação 13, o RQM do bigrupo $A_1(I, J)$ é: $A_1(I, J)$ é:

$$\begin{aligned} H(I, J) &= \frac{1}{4}((0,5)^2 + (-0,5)^2 + (-0,5)^2 + (0,5)^2) \\ &= 0,25 \end{aligned}$$

De forma similar, se tomamos a submatriz $A_2(I, J)$, que não é considerada um bigrupo, os valores de a_{iJ} , a_{iI} e a_{IJ} para o subconjunto de linhas I e o subconjunto de colunas J são: $a_{1J} = 13,5$, $a_{2J} = 17,5$, $a_{1I} = 23,5$, $a_{2I} = 7,5$ e $a_{IJ} = 15,5$. Os valores dos resíduos de cada elemento da submatriz $A_2(I, J)$ são:

$$\begin{aligned} r(a_{11}) &= 27 - 13,5 - 23,5 + 15,5 &= 5,5 \\ r(a_{12}) &= 0 - 13,5 - 7,5 + 15,5 &= -5,5 \\ r(a_{21}) &= 20 - 17,5 - 23,5 + 15,5 &= -5,5 \\ r(a_{22}) &= 15 - 17,5 - 7,5 + 15,5 &= 5,5 \end{aligned}$$

e, finalmente, o RQM de $A_2(I, J)$ é:

$$\begin{aligned} H(I, J) &= \frac{1}{4}((5,5)^2 + (-5,5)^2 + (-5,5)^2 + (5,5)^2) \\ &= 30,25 \end{aligned}$$

O exemplo mostra que quanto maior o valor do RQM, menor será a qualidade do bigrupo, já que o RQM diz como os dados do bigrupo estão inter-relacionados, isto é se há alguma coerência na evolução de seus valores ou se eles são aleatórios. Um valor alto de RQM significa que os dados da submatriz não possuem coerência ou homogeneidade (como é o caso da submatriz $A_2(I, J)$ com um RQM de 30,25). Já um valor baixo de RQM implica que a matriz possui elementos com

coerência e homogeneidade (bigrupo $A_1(I, J)$ com um RQM de 0,25). Por fim, um valor de RQM = 0 significaria que os dados da matriz fluem em uníssono como nos exemplos dos bigrupos ilustrados nas figuras 3a e 3b (bigrupos com valores constantes nas linhas e nas colunas) e na figura 4a (bigrupos com valores coerentes nas linhas e colunas).

No algoritmo Cheng e Church, busca-se por δ -biclusters que sejam tão grandes quanto possível (volume máximo), com a restrição de que seu RQM seja menor que um limiar δ , que é escolhido pelo usuário. Para alcançar esse objetivo, o algoritmo é construído sob uma estratégia gulosa que inicia com o maior bigrupo possível, iterativamente remove linhas e colunas minimizando $H(I, J)$ deste bigrupo, e então, também iterativamente, adiciona linhas e colunas que maximizem o volume do bigrupo sem aumentar o seu resíduo. No processo de remoção e adição de linhas e colunas, faz-se necessário o cálculo do RQM de uma linha i , $H(i, J)$, e de uma coluna j , $H(I, j)$. Esses resíduos são definidos, como:

$$\begin{aligned} H(i, J) &= \frac{1}{|J|} \sum_{j \in J} (a_{ij} - a_{iJ} - a_{1j} + a_{1J})^2, \\ H(I, j) &= \frac{1}{|I|} \sum_{i \in I} (a_{ij} - a_{iJ} - a_{1j} + a_{1J})^2. \end{aligned} \quad (14)$$

A visão geral da estratégia de Cheng e Church é apresentada no algoritmo 1. Como parâmetros de entrada, o algoritmo recebe: a matriz de dados $\mathcal{A}$ na qual os bigrupos serão buscados; o número K de bigrupos desejados; o limiar δ , sendo $\delta \in \mathbb{R}_+$; e a taxa α , sendo $\alpha \in \mathbb{R}_{\geq 1}$. A taxa α atua como uma margem que relaxa temporariamente o resíduo admitido para o bigrupo no processo de minimização do resíduo.

No algoritmo 1, a matriz de dados original $\mathcal{A}$ é tomada, em sua totalidade, como a inicialização do bigrupo que está sendo buscado, representado então pela matriz auxiliar A' (linha 2). Ao término de uma iteração do laço da linha 3, a matriz auxiliar A' recebe ruído (linha 10) nos elementos que já foram inseridos no último bigrupo criado. Assim, o próximo bigrupo buscado será inicialmente formado por todas as células da matriz A' , porém o processo de minimização do RQM impedirá que as células com valores ruidosos permaneçam no novo bigrupo. A possibilidade de sobreposição entre bi-

Algorithm 1 Algoritmo de Cheng e Church[14]

```

1: function ALGORITMO-CHENG-CHURCH( $\mathcal{A}$ ,  $K$ ,  $\delta \geq 0$ ,  $\alpha \geq 1$ )
2:   Inicialize  $A' = \mathcal{A}$ ;
3:   for cada bigrupo  $K$  do
4:     if o número de linhas ou colunas de  $A'$  é  $> 100$  then
5:        $C \leftarrow$  Remoção-Múltiplos-Nós( $A'$ ,  $\delta$ ,  $\alpha$ );  $\triangleright$  cf.
         algoritmo 2
6:     else
7:        $C \leftarrow$  Remoção-Simples( $A'$ ,  $\delta$ );  $\triangleright$  cf. algoritmo 3
8:        $D \leftarrow$  Adição-Nós( $\mathcal{A}$ ,  $C$ );  $\triangleright$  cf. algoritmo 4
9:       Reporte  $D$  como uma solução;
10:      Adicione ruído em  $A' \forall a \in D$ ;
11:   return  $K$   $\delta$ -bigrupos.

```

grupos é presente neste algoritmo por conta do procedimento de adição de linhas e colunas (linha 8), o qual possui como entrada o bigrupo em formação (C) e a matriz original de dados ($\mathcal{A}$).

Para a busca pelos bigrupos, o algoritmo executa três procedimentos. O procedimento Remoção-Múltiplos-Nós tem o objetivo de remover simultaneamente as linhas e colunas de um bigrupo que atendam à condição de que seu RQM seja maior que $\alpha \times H(I, J)$. Não havendo linhas ou colunas que atendam a tal condição, o procedimento Remoção-Simples é chamado. Esse processo está especificado no algoritmo 2.

Algorithm 2 Remoção de múltiplas linhas e colunas simultaneamente [14]

```

1: function REMOÇÃO-MÚLTIPLOS-NÓS( $A$ ,  $\delta$ ,  $\alpha$ )
2:   Inicialize  $A_{IJ} = A$ ;
3:   Calcule  $a_{iJ}$ ,  $a_{iJ}$ ,  $a_{IJ}$  e  $H(I, J)$ ;
4:   while  $H(I, J) > \delta$  do
5:     Remova simultaneamente todas as linhas  $i \in I$  que
       satisfazem:  $H(i, J) > \alpha \times H(I, J)$ ;
6:     Recalcule  $a_{iJ}$ ,  $a_{iJ}$ ,  $a_{IJ}$  e  $H(I, J)$ ;
7:     Remova simultaneamente todas as colunas  $j \in J$ 
       que satisfazem:  $H(I, j) > \alpha \times H(I, J)$ ;
8:     Recalcule  $a_{iJ}$ ,  $a_{iJ}$ ,  $a_{IJ}$  e  $H(I, J)$ ;
9:     if não houve nenhuma remoção nos passos das
       linhas 5 e 7 then
10:       $A_{IJ} =$  Remoção-Simples( $A_{IJ}$ ,  $\delta$ );  $\triangleright$  cf.
         algoritmo 3
11:   return  $A_{IJ}$ ;

```

Cheng e Church[14] destacam que o fato da estratégia não atualizar o RQM do bigrupo após a remoção de cada linha ou cada coluna pode resultar no demasiado encolhimento do bigrupo. Como consequência, alguns bigrupos grandes podem ser perdidos, apesar de que alguns deles podem ser novamente descobertos nas iterações seguintes do algoritmo 1¹⁰.

¹⁰Versões adaptativas do algoritmo Cheng e Church permitem alterar α durante as iterações do algoritmo com base no RQM e no tamanho dos

Para ilustrar a execução deste procedimento, considere o exemplo com a matriz de dados $\mathcal{A}$, de dimensão 3×4 , cujo cálculo do RQM resulta em 0,57:

$$\mathcal{A} = \begin{pmatrix} 4 & 5 & 1 & 2 \\ 2 & 3 & 2 & 2 \\ 3 & 2 & 2 & 3 \end{pmatrix}.$$

Para selecionar uma submatriz cujo RQM seja baixo, é necessário calcular os RQMs de todas as submatrizes que podem ser geradas como consequência de qualquer remoção de linha ou coluna. Os valores resultantes desses cálculos são mostrados na tabela 1.

Tabela 1. Cálculo do RQM depois de remover a linha ou a coluna da matriz de dados $\mathcal{A}$

	Coluna (removida)				Linha (removida))		
	1	2	3	4	1	2	3
RQM	0,64	0,29	0,49	0,59	0,17	0,68	0,42

A maior redução do RQM é obtida com a remoção da linha 1. Como resultado da remoção desta linha, tem-se a submatriz (ou o bigrupo) com RMQ de 0,17:

$$C = \begin{pmatrix} 2 & 3 & 2 & 2 \\ 3 & 2 & 2 & 3 \end{pmatrix}.$$

O procedimento Remoção-Simples pode ser chamado em duas situações: na linha 10 do algoritmo 2 ou na linha 7 do algoritmo 1 caso a matriz de dados tenha um tamanho pequeno. Esse procedimento remove, iterativamente, a linha ou a coluna de mais alto RQM existente no bigrupo. Essas remoções ocorrem até que o RQM do bigrupo seja suficientemente pequeno. Tal procedimento é detalhado no algoritmo 3.

Algorithm 3 Remoção de uma linha ou uma coluna por vez. Adaptado de Cheng e Church[14]

```

1: function REMOÇÃO-SIMPLES( $A$ ,  $\delta$ )
2:   Inicialize  $A_{IJ} = A$ ;
3:   Calcule  $a_{iJ}$ ,  $a_{iJ}$ ,  $a_{IJ}$  e  $H(I, J)$ ;
4:   while  $H(I, J) > \delta$  do
5:     Encontre a linha  $i \in I$  com maior  $H(i, J)$ ;
6:     Encontre a coluna  $j \in J$  com maior  $H(I, j)$ ;
7:     Remova de  $A_{IJ}$  a linha ou coluna com o maior
       RQM dos dois passos anteriores;
8:     Recalcule  $a_{iJ}$ ,  $a_{iJ}$ ,  $a_{IJ}$  e  $H(I, J)$ ;
9:   return  $A_{IJ}$ 

```

Para valores adequados de α , o número de iterações requeridas pelo o algoritmo 2 é pequeno e sua execução é efi-

bigrupos que estão sendo descobertos, melhorando o aspecto originalmente estático do algoritmo.

ciente [14]¹¹. Assim, é útil executar a remoção de múltiplos nós antes, de forma a reduzir o tamanho da matriz sobre a qual a remoção simples será aplicada. Entretanto, implicitamente trata-se de um problema de objetivos antagônicos: por um lado, uma matriz reduzida para a aplicação de remoções de nós simples é interessante; por outro lado, o algoritmo deseja maximizar o volume dos bigrupos. Portanto, após a exclusão de linhas e colunas, o δ -bicluster resultante pode não ter o seu volume maximizado, i.e., algumas linhas e colunas da matriz de dados original podem ser adicionadas, sem aumentar o valor de seu RQM, executando do procedimento Adição-Linhas-Colunas (Algoritmo 4).

Algorithm 4 Adição de linhas e colunas. Adaptado de Cheng e Church[14]

```

1: function ADIÇÃO-NÓS( $A, A_{IJ}$ )
2:   repeat
3:     Calcule  $a_{iJ}, a_{i\bar{j}}, a_{IJ}$  e  $H(I, J)$ 
4:     Insira as colunas  $\bar{j} \notin J$  que satisfazem:  $H(I, \bar{j}) \leq H(I, J)$ 
5:     Recalcule  $a_{iJ}, a_{i\bar{j}}, a_{IJ}$  e  $H(I, J)$ 
6:     Insira as linhas  $\bar{i} \notin I$  que satisfazem:  $H(\bar{i}, I) \leq H(I, J)$ 
7:     Insira as linhas  $\bar{i} \notin I$  que satisfazem:
        $\frac{1}{|J|} \sum_{j \in J} (-a_{ij} + a_{iJ} - a_{i\bar{j}} + a_{IJ})^2 \leq H(I, J)$ 
8:   until  $I$  e  $J$  não sofrerem mudanças (i.e. não ocorrerem
       adições em  $A_{IJ}$ )
9:   return  $A_{IJ}$ 

```

Na linha 7 do algoritmo 4, as linhas da matriz original que ainda não pertencem ao bigrupo são analisadas a partir dos valores do elemento (a_{ij}) e da média da i -ésima linha do bigrupo (a_{iJ}) multiplicados por -1 . Desta forma, o algoritmo se torna capaz de encontrar coerências negativas (opostas)¹².

O objetivo do procedimento Adição-Nós é, portanto, inserir linhas e colunas ao bigrupo para maximizar o seu volume. Entretanto, a inserção de linhas e colunas é limitada pelo valor atual do RQM do bigrupo, o qual pode ser menor do que δ . Além disso, a inserção de uma linha ou coluna pode diminuir o RQM. Logo, mesmo após a execução deste procedimento, o volume do bigrupo pode não ser máximo.

Ao fim da execução da primeira iteração do laço iniciado na linha 3 do algoritmo 1, o algoritmo encontrou um bigrupo que atende à restrição do valor mínimo para o RQM e com a maximização não ótima do volume. Para encontrar os bigrupos subsequentes, o algoritmo acrescenta ruído à matriz auxiliar A' , substituindo os elementos pertencentes ao bigrupo por números aleatórios dentro do domínio de variação dos

demais elementos da matriz de dados. A introdução do ruído redireciona a busca do algoritmo por bigrupos diferentes.

Para ilustrar a execução do algoritmo de Cheng e Church completo, considere a matriz de dados $\mathcal{A}$, de dimensão 4×4 :

$$\mathcal{A} = \begin{pmatrix} 2 & 3 & 27 & 0 \\ 6 & 7 & 20 & 15 \\ 10 & 0 & 5 & 8 \\ 1 & 4 & 9 & 12 \end{pmatrix}.$$

Considera também que, algoritmo de Cheng e Church (Algoritmo 1) recebe como parâmetros de entrada a matriz $\mathcal{A}$, $K = 2$ (número de bigrupos), $\alpha = 1, 2$ (valor tomado a partir das experimentações da seção 6.2) e $\delta = 0, 1$; na linha 2 a matriz auxiliar é inicializada ($A' = \mathcal{A}$), sendo portanto composta pelo subconjunto de linhas $I = \{1, 2, 3, 4\}$ e pelo subconjunto de colunas $J = \{1, 2, 3, 4\}$. A partir da linha 3 são realizadas duas iterações ($K = 2$) na procura dos dois bigrupos desejados. Na primeira iteração, dado que a matriz A' tem dimensão 4×4 e as dimensões são ≤ 100 , aplicar o procedimento de Remoção-Múltiplos-Nós (linhas 4 e 5) fará com que a matriz A' diminua enormemente, então o algoritmo seguirá para o procedimento de Remoção-Simples (linha 7).

No procedimento de Remoção-Simples (Algoritmo 3), dado que o RQM da matriz A' é $H(I, J) = 25, 87$, a Remoção-Simples iterará até que seu RQM seja menor ou igual a δ . Calculando os valores de $H(\bar{i}, J)$ e $H(I, \bar{j})$, $\forall \bar{i}, \bar{j} \in I, J$, tem-se: $H(1, J) = 55, 6$, $H(2, J) = 3, 35$, $H(3, J) = 31, 01$, $H(4, J) = 13, 54$, $H(I, 1) = 19, 10$, $H(I, 2) = 1, 5$, $H(I, 3) = 56, 29$ e $H(I, 4) = 26, 6$. Fazendo a remoção da coluna 3, que apresenta o maior resíduo de coluna, a matriz resultante é:

$$C = \begin{pmatrix} 2 & 3 & 0 \\ 6 & 7 & 15 \\ 10 & 0 & 8 \\ 1 & 4 & 12 \end{pmatrix}.$$

Desde que a submatriz C , agora composta pelo subconjunto de linhas $I = \{1, 2, 3, 4\}$ e pelo subconjunto de colunas $J = \{1, 2, 3\}$, apresenta um RQM ainda maior do que δ ($H(I, J) = 9, 48$), o procedimento de Remoção-Simples continua iterando e removendo linhas e colunas, seguindo com o cálculo dos valores de $H(\bar{i}, J)$ e de $H(I, \bar{j})$, $\forall \bar{i}, \bar{j} \in I, J$: $H(1, J) = 12, 12$, $H(2, J) = 4, 18$, $H(3, J) = 13, 34$, $H(4, J) = 8, 29$, $H(I, 1) = 11, 40$, $H(I, 2) = 6, 8$ e $H(I, 3) = 10, 24$.

Removendo a linha 3 por ter maior RQM de linha ($H(3, J) = 13, 34$), obtém-se a nova submatriz C , composta por $I = \{1, 2, 3\}$ e $J = \{1, 2, 3\}$ e que apresenta um RQM de 6, 71:

$$C = \begin{pmatrix} 2 & 3 & 0 \\ 6 & 7 & 15 \\ 1 & 4 & 12 \end{pmatrix}.$$

O RQM da submatriz C é maior do que δ e a Remoção-Simples continua. Os próximos valores de $H(\bar{i}, J)$ e de $H(I, \bar{j})$,

¹¹A sensibilidade do algoritmo no que diz respeito à relação entre parametrização e tempo de execução é ilustrada nos experimentos relatados na seção 6.

¹²Coerências negativas estão fora do escopo de teste deste tutorial por serem úteis para problemas em bioinformática, contexto para o qual o algoritmo de Cheng e Church foi originalmente proposto.

$\forall i, j \in I, J$ são: $H(1, J) = 13, 13$, $H(2, J) = 2, 54$, $H(3, J) = 4, 47$, $H(I, 1) = 4, 46$, $H(I, 2) = 2, 54$ e $H(I, 3) = 13, 13$.

A Remoção-Simples pode remover a linha 1 ou a coluna 3, já que ambas possuem o mesmo valor de resíduo. Removendo a coluna 3 tem-se:

$$C = \begin{pmatrix} 2 & 3 \\ 6 & 7 \\ 1 & 4 \end{pmatrix},$$

sendo seu RQM igual 0,22 e os valores $H(i, J)$ e de $H(I, j)$, $\forall i, j \in I, J$: $H(1, J) = 0, 11$, $H(2, J) = 0, 11$, $H(3, J) = 0, 44$, $H(I, 1) = 0, 22$ e $H(I, 2) = 0, 22$.

Removendo a linha 3, a submatriz resultante **C** apresenta um RQM igual a 0, e é dada por:

$$C = \begin{pmatrix} 2 & 3 \\ 6 & 7 \end{pmatrix}.$$

Uma vez que o RQM da submatriz **C** satisfaz a condição: $H(I, J) \leq \delta$, o procedimento de Remoção-Simples finaliza, retornando a submatriz **C** ao algoritmo de Cheng e Church (Algoritmo 1). Este, por sua vez, segue para o procedimento de Adição-Linhas-Colunas (linha 8).

No procedimento de Adição-Linhas-Colunas (Algoritmo 4), são analisadas as linhas e colunas da matriz $\mathcal{A}$ que não pertencem à submatriz **C**, sendo:

$$\mathcal{A} = \begin{pmatrix} 2 & 3 & 27 & 0 \\ 6 & 7 & 20 & 15 \\ 10 & 0 & 5 & 8 \\ 1 & 4 & 9 & 12 \end{pmatrix} \quad \text{e} \quad C = \begin{pmatrix} 2 & 3 \\ 6 & 7 \end{pmatrix}.$$

As colunas 3 e 4 da matriz $\mathcal{A}$ são analisadas calculando seus respectivos $H(I, j)$: $H(I, 3) = 30, 25$ e $H(I, 4) = 30, 25$. Como nenhum destes valores é menor do que $H(I, J) = 0$ (resíduo da submatriz **C**), nenhuma coluna é adicionada. De forma similar são analisadas as linhas 3 e 4 da matriz $\mathcal{A}$ que não pertencem à submatriz **C**. Sendo seus resíduos $H(3, J) = 30, 25$ e $H(4, J) = 1, 0$ respectivamente, nenhuma destas linhas serão adicionadas pois nenhum desses valores satisfazem a equação da linha 6 do algoritmo 4. Desta forma, a submatriz resultante **D** é igual à submatriz **C**, pois nenhuma adição de linhas ou colunas ocorreu:

$$D = \begin{pmatrix} 2 & 3 \\ 6 & 7 \end{pmatrix}.$$

Encontrado o primeiro bigrupo (submatriz **D**), segue a adição de ruído¹³ (em cinza) na matriz A' nas linhas e colunas que pertencem ao bigrupo retornado (linha 10 do algoritmo 1). O algoritmo segue para a próxima iteração em busca do se-

gundo bigrupo, com a matriz

$$A' = \begin{pmatrix} 10 & 24 & 27 & 0 \\ 21 & 11 & 20 & 15 \\ 10 & 0 & 5 & 8 \\ 1 & 4 & 9 & 12 \end{pmatrix}.$$

Na procura do segundo bigrupo, o algoritmo de Cheng e Church começa a segunda iteração na matriz A' executando o procedimento de Remoção-Simples de linhas e colunas até que o $H(I, J)$ do bigrupo seja menor ao δ . Este procedimento é mostrado aqui usando retas sobre linhas e colunas removidas da matriz:

$$\begin{pmatrix} 10 & 24 & 27 & 0 \\ 21 & 11 & 20 & 15 \\ 10 & 0 & 5 & 8 \\ 1 & 4 & 9 & 12 \end{pmatrix} \rightarrow \begin{pmatrix} 21 & 11 & 20 & 15 \\ 10 & 0 & 5 & 8 \\ 1 & 4 & 9 & 12 \end{pmatrix} \rightarrow \begin{pmatrix} 11 & 20 & 15 \\ 0 & 5 & 8 \\ 4 & 9 & 12 \end{pmatrix} \rightarrow \begin{pmatrix} 0 & 5 & 8 \\ 4 & 9 & 12 \end{pmatrix}.$$

O processo termina ao encontrar um bigrupo com RQM menor do que δ . No presente caso, é retornada a submatriz **C** com um $H(I, J) = 3, 94 \times 10^{-31}$:

$$C = \begin{pmatrix} 0 & 5 & 8 \\ 4 & 9 & 12 \end{pmatrix}.$$

Continuando com o algoritmo de Cheng e Church, segue-se para o procedimento de Adição-Linhas-Colunas que satisfazem a condição da linha 4 e 6 do algoritmo 1. Depois da análise da coluna 1 e das linhas 1 e 2 da matriz $\mathcal{A}$, visto que nenhuma delas melhora o RQM, a submatriz **D** resultante é igual à submatriz **C**, sendo ela o segundo bigrupo encontrado:

$$D = \begin{pmatrix} 0 & 5 & 8 \\ 4 & 9 & 12 \end{pmatrix}.$$

Enfim, a execução do algoritmo de Cheng e Church retorna dois bigrupos com RQM mínimo. O RQM do primeiro bigrupo é 0 e do segundo bigrupo é $3, 94 \times 10^{-31}$. Os bigrupos na matriz $\mathcal{A}$ são:

$$\mathcal{A} = \begin{pmatrix} 2 & 3 & 27 & 0 \\ 6 & 7 & 20 & 15 \\ 10 & 0 & 5 & 8 \\ 1 & 4 & 9 & 12 \end{pmatrix}.$$

A complexidade dos algoritmos que compõem o algoritmo de Cheng e Church são: $O(m \times \log n)$ para o algoritmo 2 e $O(n \times m)$ para os algoritmos 3 e 4, sendo n e m o número de linhas e colunas, respectivamente [14]. Esse cálculo de complexidade ainda deve considerar que os algoritmos 2, 3 e 4 são invocados K vezes a partir do algoritmo 1, sendo K o número de bigrupos procurados. Note que, retomando a

¹³Para o exemplo, números aleatórios dentro do domínio de variação dos demais elementos da matriz são gerados para inserção de ruído.

lógica implementada nesta estratégia de biagrupamento, o parâmetro α é determinante na aceleração da execução, uma vez que ele está atrelado à decisão sobre invocar, ou não, o algoritmo 3 a partir do algoritmo 2.

4. Non-negative Block Decomposition

O algoritmo de decomposição de valores em blocos não negativos (*Non-negative Block Decomposition* - NBVD) foi proposto no trabalho de Long, Zhang e Yu[15]. Este algoritmo fatora a matriz de dados $\mathcal{X}$, vista como uma matriz de dados diádicos (subseção 2.2), em três componentes: a matriz de coeficientes de linhas (U), a matriz de valores de blocos (S) e a matriz de coeficientes de colunas (V). Segundo Long, Zhang e Yu[15], os coeficientes de linhas indicam o quanto cada linha é aderente aos grupos encontrados no conjunto de linhas, os coeficientes de colunas indicam o quanto cada coluna é aderente aos grupos encontrados no conjunto de colunas, e S é uma representação explícita e compacta da estrutura de blocos ocultos da matriz de dados.

Formalmente Long, Zhang e Yu[15] definem $\mathcal{X}$ como uma matriz de dados diádicos de dimensão $n \times m$, composta por dois conjuntos de objetos $X = \{\mathbf{x}_1, \dots, \mathbf{x}_i, \dots, \mathbf{x}_n\}$ e $Y = \{\mathbf{y}_1, \dots, \mathbf{y}_j, \dots, \mathbf{y}_m\}$, sendo cada díade $w(\mathbf{x}_i, \mathbf{y}_j)$, um valor escalar correspondente a um elemento da matriz $\mathcal{X}$. Os autores proponentes deste algoritmo estavam interessados em agrupar simultaneamente X em k grupos disjuntos e Y em l grupos disjuntos. Este tipo de agrupamento equivale a encontrar a estrutura de blocos da matriz $\mathcal{X}$, ou seja, encontrar $k \times l$ blocos em $\mathcal{X}$ tais que os elementos dentro de cada bloco sejam semelhantes entre si e os elementos de diferentes blocos sejam dissimilares uns aos outros.

Ainda segundo Long, Zhang e Yu[15], como os elementos dentro de cada bloco são semelhantes entre si, faz sentido aceitar que existe um centro que representa cada bloco. Portanto, uma pequena matriz $k \times l$, adequadamente estabelecida, pode ser considerada como a representação compacta para a matriz de dados original que é formada com uma estrutura de blocos $k \times l$. Diante desta premissa, de forma similar à tarefa de agrupamento, os dados originais podem ser aproximados por combinações lineares dos centros do grupo, e é possível aproximar ou “reconstruir” a matriz de dados original pelas combinações lineares dos centros de blocos. Então, os autores formulam o problema de coagrupamento de dados diádicos como o problema de otimização da decomposição da matriz de dados, ou seja, a decomposição de valores de blocos (*Block Value Decomposition* - BVD). A definição formal para esse contexto é dada por:

Problema 1 A decomposição dos valores em blocos de uma matriz de dados $\mathcal{X} \in \mathbb{R}^{n \times m}$ é dada pela minimização de

$$f(U, S, V) = \min \| \mathcal{X} - USV^T \|^2, \quad (15)$$

em que $U \in \mathbb{R}_+^{n \times k}$, $S \in \mathbb{R}^{k \times l}$, $V \in \mathbb{R}_+^{m \times l}$, $\| \cdot \|$ denota a norma de Frobenius para matrizes, $k \ll n$ e $l \ll m$.

Como mostrado na equação 15, o problema BVD aproxima a matriz de dados original pela matriz resultante da multiplicação das matrizes U , S e V , o que resulta em uma matriz reconstruída. Note que a minimização da norma *Frobenius* no problema 1 equivale à minimização da distância euclidiana entre a matriz original $\mathcal{X}$ e a sua aproximação (ou matriz reconstruída) USV^T . Esse processo é ilustrado na figura 12.

Figura 12. Fatoração da matriz original de dados X nas matrizes U , S e V . Adaptado de Yoo e Choi[33]

Na figura 12, os elementos de cor com maior intensidade indicam a presença de valores maiores nas matrizes. Nas matrizes U , S e V , esses valores maiores indicam adequabilidades maiores entre linhas, colunas, grupos de linhas, grupos de colunas e blocos.

Yoo e Choi[33] exemplificam o problema de coagrupamento por fatoração de matrizes usando o contexto de análise de textos. Seguindo o exemplo desses autores e o exposto na figura 12, $\mathcal{X}$ é uma matriz de “documentos $\times$ palavras” ($\mathcal{X} \in \mathbb{R}^{6 \times 8}$), que apresenta seis documentos e oito palavras, U indica três grupos de documentos (partição de documentos), V indica dois grupos de palavras (partição de palavras) e S indica os blocos que relacionam os grupos de ambas partições. Um algoritmo para resolução do problema de BVD determina as matrizes U , S e V , de forma que U é a matriz de “documentos $\times$ grupos de documentos” ($U \in \mathbb{R}^{6 \times 3}$), V^T é uma matriz de “grupo de palavras $\times$ palavras” ($V^T \in \mathbb{R}^{2 \times 8}$) e a matriz S representa como os “grupos de documentos” estão relacionados aos “grupos de palavras” ($S \in \mathbb{R}^{3 \times 2}$). Dentro do processo de minimização do problema 1, os grupos de linhas influenciam na formação dos grupos de colunas e vice-versa, então, do exemplo é possível dizer que os grupos de documentos 1 e 3 contribuem para a formação do grupo de palavras 1, os grupos de documentos 2 e 3 contribuem para a formação do grupo de palavras 2, e vice-versa.

Segundo os autores Long, Zhang e Yu[15], as combinações dos componentes da fatoração também possuem uma interpretação intuitiva, já que SV^T contém a base para o espaço de linhas de $\mathcal{X}$ e US é a matriz que contém a base para o espaço de colunas de $\mathcal{X}$ ¹⁴. Seguindo o exemplo de análise de

¹⁴O raciocínio aplicado para a interpretação de SV^T é análogo àquele usado na interpretação do problema de otimização para o algoritmo *k-means* no que diz respeito à interpretação das matrizes que associam pertinência de linhas a grupos e os respectivos centroides. Considere a formulação do problema do *k-means* com base em fatoração de matrizes, da seguinte forma: $f(UC) = \min \| \mathcal{X} - UC \|^2$, em que $U \in \{0, 1\}^{n \times k}$, $C \in \mathbb{R}_+^{k \times m}$, $\sum_{p=1}^k a_{ip} = 1 \forall i$, $\| \cdot \|$ denota a norma de Frobenius para matrizes e $k \ll n$. Sabe-se que U é uma matriz indicadora que associa cada linha de $\mathcal{X}$ a um dos grupos k e C

textos, a matriz SV^T representa a matriz de “grupos de documentos $\times$ palavras”, em que cada linha captura a ideia de uma base para a representação de grupos de documentos (protótipo para grupos de documentos), e a matriz US representa a matriz de “documentos $\times$ grupos de palavras”, em que cada coluna captura a ideia de uma base para a representação de grupos de palavras (protótipos para grupos de palavras). Esses conceitos estão ilustrados nas figuras 13 e 14, respectivamente.

Figura 13. Reconstrução da matriz de dados $\mathcal{X}$ pelas multiplicações da matriz indicadora de grupos de linhas (U) e a matriz dos protótipos das linhas (SV^T) [48]

Figura 14. Reconstrução da matriz de dados $\mathcal{X}$ pelas multiplicações da matriz dos protótipos das colunas (US) e a matriz indicadora de grupos de colunas (V^T). Adaptado de Yoo e Choi[33] e Brunialti[48]

Long, Zhang e Yu[15] inserem uma restrição de não negatividade da matriz original na definição formal do BVD, formulando a decomposição de valores em blocos não negativos (NBVD). Essa restrição é motivada pelo fato de muitas aplicações serem caracterizadas por representações de dados em matrizes não-negativas (tabelas de co-ocorrência, matrizes de similaridade), e também pela possibilidade de transformar a representação de um problema em uma matriz não-negativa, como é o caso de aplicações sobre expressão gênica.

é a matriz de centroides – base de representação de grupos de linhas. Assim como SV^T , C tem dimensão $k \times m$.

Problema 2 A decomposição dos valores em blocos não negativos de uma matriz de dados $\mathcal{X} \in \mathbb{R}_+^{n \times m}$ é dada pela minimização de:

$$f(U, S, V) = \min \| \mathcal{X} - USV^T \|^2 \quad (16)$$

em que $U \in \mathbb{R}_+^{n \times k}$, $S \in \mathbb{R}_+^{k \times l}$, $V \in \mathbb{R}_+^{l \times m}$, $\| \cdot \|$ denota a norma de Frobenius para matrizes, $k \ll n$ e $l \ll m$.

Para resolução do problema 2, o cálculo das matrizes U , S e V pode ser feito de forma iterativa com base nas regras de atualização multiplicativa: $U \odot \frac{\mathcal{X}VS^T}{USV^TVS^T}$, $V \odot \frac{\mathcal{X}^TUS}{VS^TUTUS}$ e $S \odot \frac{U^T\mathcal{X}V}{U^TUSV^TV}$. O operador $\odot$ representa a multiplicação de matrizes ponto a ponto, também conhecida como *produto de Hadamard*, e a sequência de matrizes, por exemplo $\mathcal{X}VS^T$, é a multiplicação linha por coluna das matrizes $\mathcal{X}$, V e S^T .

O algoritmo 5 descreve o processo iterativo para minimização do problema 2 (NBVD). No algoritmo 5, as matrizes U , S e V são inicializadas por uma distribuição uniforme ($\mathcal{U}(0, 1)$) de números reais no intervalo $[0, 1]$, T é o número máximo de iterações, t é o contador de iterações, e $U^{(t)}$, $S^{(t)}$ e $V^{(t)}$ as matrizes U , S , V na iteração t . O algoritmo itera até cumprir a condição de convergência ou até chegar ao número máximo de iterações. A condição de convergência considerada é que a diferença do erro de reconstrução em duas iterações consecutivas seja menor ou igual a um ζ :

$$\| \mathcal{X} - U^{(t)}S^{(t)}V^{(t)T} \|^2 - \| \mathcal{X} - U^{(t+1)}S^{(t+1)}V^{(t+1)T} \|^2 \leq \zeta.$$

Algorithm 5 Algoritmo de decomposição de valores em blocos não negativos (NBVD). Baseado de Long, Zhang e Yu[15] e Brunialti[48]

```

1: function NBVD( $\mathcal{X}, k, l, T$ )
2:   Inicialize:  $U^{(0)} = \mathcal{U}(0, 1)$ ,  $V^{(0)} = \mathcal{U}(0, 1)$ ,  $S^{(0)} = \mathcal{U}(0, 1)$ ;
3:   while  $t \leq T$  e (não convergiu) do
4:      $U^{(t+1)} \leftarrow U^{(t)} \odot \frac{\mathcal{X}V^{(t)}S^{(t)T}}{U^{(t)}S^{(t)}V^{(t)T}V^{(t)}S^{(t)T}}$ ;
5:
6:      $V^{(t+1)} \leftarrow V^{(t)} \odot \frac{\mathcal{X}^TU^{(t+1)}S^{(t)}}{V^{(t)}S^{(t)T}U^{(t+1)T}U^{(t+1)}S^{(t)}}$ ;
7:
8:      $S^{(t+1)} \leftarrow S^{(t)} \odot \frac{U^{(t+1)T}\mathcal{X}V^{(t+1)}}{U^{(t+1)T}U^{(t+1)}S^{(t)}V^{(t+1)T}V^{(t+1)}}$ ;
9:
10:     $t = t + 1$ ;
11:  return  $U^{(t)}, V^{(t)}, S^{(t)}$ 

```

As regras multiplicativas implementam um processo de minimização de erro [33], sendo que a função de erro é baseada na distância Euclidiana entre as matrizes $\mathcal{X}$ e USV^T . O algoritmo converge para um mínimo global quando as frações das equações de atualização de U , V e S resultam em 1, i.e., quando $\mathcal{X} = USV^T$ ¹⁵.

¹⁵O processo iterativo de minimização do erro implementado no algoritmo 5 é semelhante ao processo iterativo de minimização do problema

Para obter as partições das linhas (U), colunas (V) e a relação entre estas duas partições (S), é necessário fazer um pós-processamento das matrizes U , S e V , uma vez que são matrizes compostas por valores reais cujo significado para interpretação do resultado de coagrupamento não é possível de ser obtido de forma direta.

Para obter os elementos pertencentes a cada grupo na partição das linhas da matriz $\mathcal{X}$, sob a interpretação adotada neste tutorial, a matriz U é analisada em termos dos valores maiores ou iguais a um limiar escolhido de acordo com o contexto de cada grupo de linhas. Esse limiar foi determinado como a média dos valores presentes na matriz U para cada grupo de linhas, ou seja, a média dos elementos das colunas da matriz U . Considerando o contexto de análise de textos, e lembrando que a matriz U representa uma matriz de “documentos $\times$ grupo de documentos”, ao escolher os documentos associados aos valores maiores ou iguais ao limiar de cada grupo, estão sendo indicados os documentos que possuem maior aderência a cada grupo¹⁶. Na figura 15 é mostrado um exemplo para ilustrar essa interpretação. Nesta figura, a matriz $\mathcal{X}$ é apresentada como uma matriz de dimensão 7×9 de sete documentos por nove palavras, e as matrizes U , S e V são apresentadas com números reais. Na figura, os tons de azul mais escuros denotam os cogrupos de interesse. A matriz U mostra a aderência dos documentos 5, 6 e 7 ao grupo representado pela primeira coluna, uma vez que nas células (5, 1), (6, 1) e (7, 1), os valores são maiores ou iguais à média de valores para a primeira coluna (3, 19). Note que o documento 5 também é aderente ao grupo representado pela coluna 2, o que indica que há sobreposição nos grupos de documentos.

Da mesma forma, a matriz V^T é analisada para obter os elementos pertencentes a cada grupo na partição das colunas da matriz $\mathcal{X}$. No exemplo de análise de textos, desde que a matriz V^T representa a matriz de “grupo de palavras $\times$ palavras”, ao escolher as palavras associadas aos valores maiores ou iguais ao limiar de cada grupo (agora olhando sob a perspectiva das linhas), estão sendo indicadas as palavras que possuem maior aderência a cada grupo de palavras. Na figura 15, a análise da matriz V^T revela quais palavras são associadas a quais grupos de palavras e também mostra que para essa partição não existem sobreposições de grupos.

Para obter as relações existentes entre as duas partições (de linhas e de colunas) da matriz $\mathcal{X}$, e portanto realizar a extração do conhecimento sobre os cogrupos de interesse, a matriz S deve ser analisada. Foram considerados os valores maiores ou iguais a um limiar. Esse limiar foi determinado como a média dos valores na matriz S para cada grupo de

linhas, ou seja, a média dos elementos das linhas da matriz S . Na análise de textos, a matriz S representa a matriz de “grupos de documentos $\times$ grupos de palavras”, então esta escolha assegura a relação de pelo menos um grupo de documentos com um grupo de palavras e permite a sobreposição de cogrupos. No caso do exemplo da figura 15, o terceiro grupo de documentos (coluna 3 na matriz U) está mais fortemente associado ao primeiro grupo de palavras (linha 1 na matriz V^T), uma vez que a célula (3, 1) na matriz S possui um valor maior ou igual à média (0, 38) dos valores para a terceira linha dessa matriz.

De acordo com o exemplo ilustrado na figura 15, na figura 16 é mostrado o processo de reconstrução obtido a partir da multiplicação das matrizes U , S e V , mostrando a semelhança entre a matriz reconstruída e a matriz original. O erro de reconstrução é obtido pela diferença entre a matriz original e a matriz reconstruída, por meio da combinação das matrizes resultantes U , S e V , i.e., $\|X - USV^T\|^2$. O erro de reconstrução pode ser usado como uma medida de qualidade sobre o coagrupamento obtido.

Seguindo a estratégia de interpretação aqui descrita, a aceitação de uma relação como um cogrupos de interesse depende dos valores associados à matriz S no processo de minimização e também do domínio de aplicação em questão e organização de cogrupos buscada. A organização proposta para bigrupos (seção 2.1.2) pode ser estendida para este contexto, fazendo a ressalva para o fato que a tarefa de coagrupamento está associada à distribuição de todos os elementos da matriz original em um ou mais cogrupos (conforme definição de coagrupamento apresentada na seção 2.2).

Para ilustrar a execução do algoritmo de NBVD, considere que o algoritmo receberá como entrada $k = 2$, $l = 2$, e a matriz de dados $\mathcal{X}$:

$$\mathcal{X} = \begin{pmatrix} 2 & 3 & 27 & 0 \\ 6 & 7 & 20 & 15 \\ 10 & 0 & 5 & 8 \\ 1 & 4 & 9 & 12 \end{pmatrix}.$$

Na figura 17 são mostradas instâncias para as matrizes U , S e V resultantes da execução do algoritmo NBVD sobre $\mathcal{X}$. O processo para interpretar as matrizes da decomposição é ilustrado por meio das cores.

Para obter uma visualização mais clara dos cogrupos resultantes, a matriz de dados $\mathcal{X}$ pode ser reorganizada trocando as posições das colunas 2 e 4:

$$\mathcal{X} = \begin{pmatrix} 2 & 0 & 27 & 3 \\ 6 & 15 & 20 & 7 \\ 10 & 8 & 5 & 0 \\ 1 & 12 & 9 & 4 \end{pmatrix}$$

A complexidade do algoritmo NBVD é $O(t \times (k + l) \times n \times m)$, sendo t o número de iterações, k e l o número de grupos de linhas e de colunas, respectivamente, e n e m o

definido para o algoritmo k -means. A regra de atualização da matriz U (matriz indicadora no k -means) corresponde às regras de atualização das matrizes U e V para o NBVD. Já a regra de atualização para a matriz C (matriz de protótipos do k -means) corresponde à regra de atualização da matriz S no NBVD.

¹⁶Esse raciocínio é análogo àquele aplicado para interpretação da matriz U (matriz indicadora) no algoritmo clássico k -means. Neste algoritmo, a matriz U assume apenas valores no conjunto $\{0, 1\}$, sendo que em cada linha da matriz, apenas um elemento assume o valor 1, indicando que uma linha só pode estar associada a um grupo de linhas.

Figura 15. Pós-processamento das matrizes U , S e V em um conjunto de dados sintéticos **Figura 16.** Reconstrução das matrizes U , S e V do conjunto de dados sintéticos da figura 15

número de linhas e colunas, respectivamente [15]. Em termos de melhorias no tempo de execução, implementações usando processamento distribuído podem ser construídas para computação das operações sobre matrizes. Em termos de melhorias na qualidade de coagrupamento, algumas alterações para o algoritmo NBVD foram propostas. O trabalho de Ding et al.[49] propõem a fatoração ortogonal tripla de matrizes não-negativas (*Orthogonal Non-negative Matrix Tri-factorization* - ONMTF), na qual restrições de ortonormalidades são admitidas para encontrar protótipos mais centralizados nos cogrupos. O trabalho de Wang et al.[39] propõe a fatoração tripla rápida de matrizes não-negativas (*Fast Non-negative Matrix Tri Factorization* - FNMTF) cuja formulação permite que o problema de otimização possa ser desacoplado, o que resulta em subproblemas de tamanho menor. Os trabalhos de Brunialti[48] e Brunialti et al.[40] propõem duas estratégias, a fatoração tripla de matrizes não-negativas sobrepostas (*Overlapped Non-negative Matrix Tri-factorization* - OvNMTF) e a fatoração binária tripla de matrizes não-negativas com sobreposição (*Overlapped Binary Non-negative Matrix Tri-factorization* - BinOvNMTF), ambas baseadas nas estratégias BVD e FNMTF respectivamente. A formulação proposta por esses autores permite encontrar grupos de atributos independentes para cada grupos de dados.

5. Método para experimentação

Para contribuir com o entendimento dos algoritmos, é útil analisar casos de uso. Esta seção descreve as estratégias de experimentação usadas para ilustrar o comportamento dos algoritmos e de seus parâmetros em conjuntos de dados que apresentam alta dimensionalidade e alta esparsidade.

5.1 Conjuntos de dados sintéticos

Os conjuntos de dados sintéticos são baseados nas diferentes estruturas de bigrupos descritas na seção 2.1.2, as quais, neste tutorial, são também consideradas sob a visão de coagrupamento¹⁷. Sobre a interpretação dos conjuntos de dados sintéticos, segundo a visão de coagrupamento, mais cogrupos podem ser observados em cada conjunto de dados (veja figura 10b). No entanto, para fins de análise de dados, interessa descobrir os cogrupos que são equivalentes aos bigrupos. Desta forma, as análises dos resultados seguem esse objetivo: encontrar os cogrupos de interesse.

Nas experimentações que incluem análise de esparsidade foi necessário realizar um procedimento para inserção controlada de esparsidade em cada um dos conjuntos de dados sintéticos, uma vez que eles não apresentavam essa característica. Diferentes níveis de esparsidade foram considerados.

¹⁷Matrizes de 150 linhas e 150 colunas (conjunto de dados em alta dimensão) foram sintetizadas de maneira a representar cada estrutura de bigrupo. Elas podem ser obtidas em: <<http://each.uspnet.usp.br/sarajane/>>.

Figura 17. Exemplo da execução do algoritmo NBVD na matriz de dados $\mathcal{X}$

Para uma melhor compreensão dos conjuntos de dados sintéticos esparsos, considere o conjunto de dados sintéticos da estrutura i , que apresenta bigrupos com sobreposição e arbitrariamente posicionados (figura 9b). Nesse conjunto são inseridos zeros aleatoriamente, dependendo do nível de esparsidade. Define-se o nível de esparsidade como a porcentagem do conjunto de dados que apresenta valores zeros. Então, se o tamanho do conjunto de dados é 150×150 , e o número total de células desse conjunto de dados é 22.500, um conjunto de dados esparsos com um nível de esparsidade de 10 significa que 10% do número total de células são valoradas com zero. Os níveis de esparsidade considerados para o experimento são: 0, 10, 20, 30, 40, 50, 60, 70, 80, 90, 98. Na figura 18, versões esparsas do conjunto de dados da estrutura i são apresentadas. As figuras representam as matrizes de dados, sendo que quanto mais forte o tom de azul em uma célula, maior é o valor nela contido.

Figura 18. Efeito da inserção de esparsidade no conjunto de dados da estrutura i , com níveis de esparsidade de 0 (figura a), 30 (figura b), 50 (figura c), 70 (figura d) e 90 (figura e)

5.2 Conjunto de dados reais - corpus de notícias

O corpus de notícias é composto por uma coleção 1.157 notícias obtidas do período de abril de 2017 até setembro de 2017, a partir do Portal EBC - Empresa Brasil de Comunicação¹⁸, as quais estão organizadas em dez canais, conforme mostrado na tabela 2.

O pré-processamento de textos é uma atividade fundamental para representação computacional de dados textuais, permitindo que modelos computacionais possam ser aplicados para aquisição de conhecimento [50]. Os principais procedimentos de pré-processamento de textos foram aplicados ao corpus de notícias¹⁹: remoção de números e caracteres

¹⁸Portal EBC: <<http://www.ebc.com.br/>>

¹⁹Uma descrição mais detalhada dos procedimentos de pré-processamento de textos está fora do escopo deste tutorial. Porém, eles podem ser encontrados, em [51, 50, 52].

Tabela 2. Organização do corpus de notícias

Id	Canal da notícia	Total de notícias	% do total
1	Cidadania	96	8,30
2	Cultura	124	10,72
3	Educação	257	22,21
4	Esportes	200	17,29
5	Infantil	14	1,21
6	Política	86	7,43
7	Tecnologia	125	10,80
8	Economia	41	3,54
9	Internacional	89	7,69
10	Geral	125	10,80
	Total	1157	100%

especiais, tokenização, remoção de stopwords, mapeamento para representações vetoriais (binária, *term frequency* (tf) e *term frequency-inverse document frequency* (tf-idf) [53, 2] e seleção heurística de características (palavras).

As matrizes de dados obtidas foram submetidas aos algoritmos de biagrupamento e coagrupamento. O conjunto de dados reais, independente da forma de representação de dados usada, apresenta alta esparsidade. O nível de esparsidade nas matrizes de dados é de aproximadamente 98%.

5.3 Método de experimentação: algoritmo de Cheng e Church

O algoritmo de Cheng e Church possui três parâmetros: δ , α e K . A qualidade dos resultados apresentados pelo algoritmo, assim como a sua eficiência em termos de tempo de execução, dependem de boas escolhas para esses parâmetros. Para estudo desses parâmetros, no contexto do presente tutorial, também é importante levar em consideração diferentes níveis de esparsidade que podem estar presentes nos dados. Esse algoritmo foi estudado nos conjuntos de dados sintéticos referentes às estruturas de bigrupos ($a, b, c, d, e, f, g, h, i$) nas suas versões originais e nas versões com os diferentes níveis de esparsidade, e no corpus de notícias. Foram realizados seis experimentos, com os seguintes objetivos:

- *Exp.#1* - estudar os parâmetros α e δ com a finalidade de escolher um intervalo de valores para estes parâmetros que retornem resultados promissores. Para esta experimentação foram considerados os conjuntos de dados sintéticos sem esparsidade e o número de bigrupos (K) já conhecido para cada um dos conjuntos.

- *Exp.#2* - estudar os valores dos parâmetros α e δ , escolhidos no *Exp.#1*, com o fim de escolher os valores para estes parâmetros que se mostrem adequados em um ambiente com esparsidade. Para esta experimentação foram considerados os conjuntos de dados sintéticos com diferentes níveis de esparsidade e o número de bigrupos (K) já conhecido em cada um dos conjuntos.
- *Exp.#3* - estudar o comportamento do algoritmo conforme o número de bigrupos (K) procurados aumenta ou diminui. Para esta experimentação foram considerando os valores α e δ escolhidos no *Exp.#2* e os conjuntos de dados sintéticos sem esparsidade.
- *Exp.#4* - estudar o comportamento do algoritmo conforme os níveis de esparsidade tendem a crescer, quando aplicados a conjuntos de dados com diferentes níveis de esparsidade. Nesta experimentação foram considerados os valores de α e δ escolhidos no *Exp.#2* e o número de bigrupos (K) já conhecido em cada conjunto de dados.
- *Exp.#5* - estudar valores para os parâmetros α , δ e K que mostram resultados promissores no corpus de notícias.
- *Exp.#6* - estudar o comportamento do algoritmo aplicado ao corpus de notícias, considerando valores de α e δ fixos e diferentes números de bigrupos (K). Os valores de α , δ e K considerados foram os escolhidos nos *Exp.#2*, *Exp.#5* e a partir do número de canais já conhecido do corpus, respectivamente.

Para avaliação dos resultados nos conjuntos de dados sintéticos foram consideradas as medidas de biagrupamento e as medidas internas de agrupamento e a análise do tempo de execução. Para a relação entre os valores dos parâmetros e o tempo de execução foi considerada a diferença dos tempos do início e fim do experimento em cada combinação destes valores, aplicados em cada conjunto de dados sintéticos.

As experimentações com o corpus de notícias foram avaliadas quantitativamente e qualitativamente. Na avaliação quantitativa foram utilizadas as medidas de biagrupamento e as medidas internas de agrupamento calculadas na projeção nas linhas dos bigrupos. Estas medidas suportaram a análise do comportamento do algoritmo de Cheng e Church diante da variação de valores para seus parâmetros de entrada.

Espaços de busca para os parâmetros do algoritmo O estudo do algoritmo de Cheng e Church iniciou com a exploração de valores para os seus parâmetros de entrada α , δ e K , considerando os conjuntos de dados sintéticos. O espaço de busca para tais parâmetros foi determinado da seguinte forma:

- Margem para o RQM de linhas e colunas no bigrupos (parâmetro α): a escolha do parâmetro α interfere diretamente na eficiência da execução dos procedimentos de Remoção-Múltiplos-Nós e Remoção-Simples (algoritmos 2 e 3). Um valor alto para α pode levar a

execuções ineficientes do procedimento de Remoção-Simples, uma vez que dificulta a retirada de linhas ou colunas no procedimento de Remoção-Múltiplos-Nós. Por outro lado, um valor pequeno para α implica na remoção de muitas linhas ou colunas no procedimento de Remoção-Múltiplos-Nós. Assim, é interessante manter α tão alto quanto for possível, porém sem deixar a execução do algoritmo ineficiente. Os proponentes do algoritmo usam $\alpha = 1, 2$ [14]. Sendo assim, neste estudo os valores assumidos para explorar α foram: 1, 1,2, 1,5, 2, 2,5 e 3.

- Limiar mínimo para o RQM do bigrupos (parâmetro δ): uma heurística sugerida por Cheng e Church[14] foi aplicada para a determinação do espaço de busca para o δ . Nessa heurística, o valor do RQM mínimo entre os RQMs de cada bigrupos conhecido no conjunto de dados é usado como valor para δ . Para criar o espaço de busca para esse parâmetro, também foram usados valores abaixo do RQM mínimo e acima do RQM mínimo, obtidos por meio da divisão e multiplicação do RQM mínimo por 10, respectivamente. Assim, $\delta_1 \leq \frac{\min(\text{RQM})}{10}$, $\delta_2 = \min(\text{RQM})$ e $\delta_3 \geq 10 * \min(\text{RQM})$. Como exemplo da determinação do espaço de busca para o parâmetro δ , considere o conjunto de dados com a estrutura de bigrupos com linhas exclusivas (estrutura d), mostrada na figura 19, e a tabela 3 com o espaço de busca de δ para esse caso.

Figura 19. Bigrupos com linhas exclusivas (estrutura d)

Tabela 3. Espaço de busca para o parâmetro δ considerando o conjunto de dados da estrutura d , ilustrado na figura 19

Bigrupos	RQM	$\min(\text{RQM})$	valores para δ
1	0,078		
2	0,077	0,077	0,007, 0,077, 0,772
3	0,083		

Nas experimentações em todos os conjuntos de dados, os três valores de δ são nomeados como δ_1 , δ_2 e δ_3 , mas o valor do parâmetro varia dependendo de cada conjunto de dados. Como é mostrado na tabela 4, os referidos valores foram calculados para cada conjunto de dados das estruturas a , b , c , d , e , f , g , h , e i , da mesma forma como a estrutura d . Na tabela, o parâmetro δ_2 representa o RQM mínimo encontrado nos bigrupos de cada conjunto de dados.

- Número de bigrupos (parâmetro K): este parâmetro

Tabela 4. Espaço de busca para o parâmetro δ em cada conjunto de dados das estruturas

Conjuntos de dados	δ_1	δ_2	δ_3
<i>a</i>	0,008	0,081	0,814
<i>b</i>	0,007	0,078	0,782
<i>c</i>	0,007	0,077	0,775
<i>d</i>	0,007	0,077	0,772
<i>e</i>	0,007	0,077	0,779
<i>f</i>	0,007	0,078	0,789
<i>g</i>	0,007	0,079	0,794
<i>h</i>	0,008	0,080	0,805
<i>i</i>	0,007	0,077	0,776

assume valores correspondentes ao número de bigrupos conhecido para o conjunto de dados no *Exp.#1* e, para a variação explorada no *Exp.#3*, o parâmetro assume valores que variaram de 1 até no máximo o número de cogrupos²⁰ que podem ser observados nos conjuntos de dados, com passos de tamanho 1, 2 ou 3, a depender de cada conjunto de dados. O espaço de busca para esse parâmetro está listado na tabela 5.

Tabela 5. Espaço de busca para o parâmetro K

Conj. de dados	#bigrupos	valores para K
<i>a</i>	1	1 2 3
<i>b</i>	3	1 2 3 4 6
<i>c</i>	9	1 2 3 4 6 9
<i>d</i>	3	1 2 3 4 6
<i>e</i>	3	1 2 3 4 6
<i>f</i>	4	1 2 3 4 5
<i>g</i>	4	1 2 3 4 5
<i>h</i>	5	1 2 3 4 5 9
<i>i</i>	4	1 2 3 4 6 9

5.4 Método de experimentação: algoritmo NBVD

O algoritmo NBVD recebe como parâmetros de entrada o conjunto de dados para análise, os valores considerados nos critérios de parada (1.000 iterações ou a diferença do erro de reconstrução em duas iterações consecutivas, o que ocorrer primeiro), e os parâmetros k e l . O algoritmo foi aplicado nos conjuntos de dados sintéticos referentes às estruturas de bigrupos (*a*, *b*, *c*, *d*, *e*, *f*, *g*, *h*, e *i*) e em um corpus de notícias (Seção 5.2). Com respeito aos conjuntos de dados sintéticos, o algoritmo foi estudado em um ambiente sem esparsidade (aplicado a estes conjuntos nas suas formas originais) e em um ambiente com esparsidade (aplicado a estes conjuntos com inserção de esparsidade em diferentes níveis). As experimentações em conjunto de dados sintéticos estão relatadas no *Exp.#1* e *Exp.#2*, respectivamente e as experimentações nas matrizes de dados extraídas do corpus de notícias estão relatadas no *Exp.#3*. O objetivo de cada experimento é:

- *Exp.#1* - estudo dos parâmetros k e l , e seu impacto nos resultados obtidos quando é aplicado nos conjuntos de dados sintéticos sem esparsidade.
- *Exp.#2* - estudo do comportamento do algoritmo conforme os níveis de esparsidade aumentam, quando aplicados aos conjuntos de dados sintéticos. Nesta experimentação foram considerados os valores para k e l que retornaram resultados satisfatórios do *Exp.#1*.
- *Exp.#3* - estudo do comportamento do algoritmo aplicado ao corpus de notícias, considerando valores para k e l a partir do número de canais já conhecidos.

Os resultados das experimentações nos conjuntos de dados sintéticos foram analisados usando medidas de qualidade de biagrupamento, medidas internas de agrupamento na projeção nas linhas dos cogrupos, analisando o erro de reconstrução e realizando uma análise visual. A análise visual foi feita por meio da comparação visual entre a matriz original e a matriz reconstruída obtida pelo algoritmo.

Para avaliação dos resultados do coagrupamento nas matrizes extraídas do corpus de notícias foram utilizadas duas estratégias: avaliação quantitativa e avaliação qualitativa dos cogrupos resultantes do coagrupamento. Na avaliação quantitativa foram utilizadas as medidas de biagrupamento, medidas internas de agrupamento calculadas na projeção nas linhas dos cogrupos e o erro de reconstrução. A avaliação qualitativa foi realizada por meio da interpretação da semântica inerente aos cogrupos obtidos.

Espaço de busca para os parâmetros k e l : Para a execução do *Exp.#1*, foram testados valores para os parâmetros de k e l escolhidos a partir do conhecimento *a priori* existente sobre o número de cogrupos de interesse, e considerando um número máximo de grupos de linhas e grupos de colunas a ser testado no experimento. Assim, dado um determinado número de cogrupos de interesse presente em cada conjunto de dados, o número máximo de grupos de linhas (k) usado para formar estes cogrupos é dado por $k_{max} = 2P + 1$, em que P é o número de cogrupos de interesse conhecido para cada estrutura de conjunto de dados. O número máximo de grupos de colunas a ser testado foi determinado para ser igual ao número obtido para os grupos de linhas. O número mínimo foi determinado para 2, visto ser o menor número de grupos de linhas que faz sentido para a tarefa sob resolução.

Na figura 20, os gráficos ilustram a heurística usada para determinar k . Os cogrupos de interesse estão destacados da cor azul e a distribuição esperada para o número máximo de grupos de linhas é mostrada na cor rosa, no lado esquerdo de cada representação de conjunto de dados. No lado direito da figura, números indicam outra possível configuração para os grupos de linhas. No conjunto de dados representado na parte (c) da figura, três interpretações possíveis para grupos de linhas são apresentadas.

Uma análise mais detalhada das situações representadas na figura 20a mostra que, para o caso (a), 3 pode ser um valor

²⁰O número de cogrupos foi usado para ampliar o espaço de busca, já que a visão de coagrupamento estende a visão de biagrupamento.

Figura 20. Representação do número máximo de grupos de linhas

adequado, no entanto, o valor 2 para k também pode ser usado visto que também permite formar o cogrupos de interesse. No caso de $k = 2$, dois grupos de linhas descobertos com $k = 3$ são unidos (cf. disposição das linhas na cor verde). Similarmente, $k_{max} = 5$ para o caso (b), porém, poderiam ser buscados apenas os três grupos de linhas de forma que os grupos de linhas $\{1,3,5\}$ sejam vistos como um único grupo de linhas. Para o caso (c), $k_{max} = 5$ (linha da cor rosa), porém poderiam ser buscados quatro grupos de linhas (linha da cor verde) de forma que os grupos de linhas $\{1,5\}$ sejam vistos como um único grupo. Também poderiam ser encontrados três grupos de linhas se os grupos de linhas $\{1,5\}$ e os grupos de linhas $\{2,3\}$ fossem vistos como um único grupo respectivamente (linha da cor vermelha). Nesta última distribuição de grupos de linhas há perda de informação em relação ao cogrupos azul escuro, permitindo descrever parcialmente a organização dos dados.

Com base na heurística discutida para escolha de k_{max} , o espaço de busca dos parâmetros k e l foi determinado considerando o intervalo de $[2, k_{max}]$ com passos 1, 2 ou 3 a depender do conjunto de dados sob análise. A tabela 6 lista os valores testados para k e l .

Tabela 6. Espaço de busca para os parâmetro k e l , em cada conjunto de dados das estruturas

Conj. dados	#cogrupos interesse	k	l
<i>a</i>	1	2, 3	2, 3
<i>b</i>	3	2, 3, 4, 5, 6, 7	2, 3, 4, 5, 6, 7
<i>c</i>	9	2, 3, 6, 9, 12, 15, 19	2, 3, 6, 9, 12, 15, 19
<i>d</i>	3	2, 3, 4, 5, 6, 7	2, 3, 4, 5, 6, 7
<i>e</i>	3	2, 3, 4, 5, 6, 7	2, 3, 4, 5, 6, 7
<i>f</i>	4	2, 3, 4, 5, 6, 7, 8, 9	2, 3, 4, 5, 6, 7, 8, 9
<i>g</i>	4	2, 3, 4, 5, 6, 7, 8, 9	2, 3, 4, 5, 6, 7, 8, 9
<i>h</i>	5	2, 3, 4, 5, 6, 7, 8, 9, 10, 11	2, 3, 4, 5, 6, 7, 8, 9, 10, 11
<i>i</i>	4	2, 3, 4, 5, 6, 7, 8, 9	2, 3, 4, 5, 6, 7, 8, 9

Para as experimentações do algoritmo nos conjuntos de dados com diferentes níveis de esparsidade (*Exp.#2*), foram usados, para k e l , valores maiores ou iguais àqueles que geraram erros de reconstrução baixos no *Exp.#1*. Esses valores são destacados em **negrito** na tabela 6.

Nas experimentações do algoritmo com o corpus de notícias (*Exp.#3*) foram escolhidos valores para os parâmetros k

e l segundo as análises do *Exp.#2*, de forma a buscar resultados satisfatórios para dados com esparsidade, fixando o valor máximo para k e l em 10, em acordo com o número de canais conhecidos nos corpus. Os valores escolhidos foram para k e l foram 3, 5, 7 e 10.

6. Resultados e discussões: algoritmo de Cheng e Church

O algoritmo de Cheng e Church foi estudado nos conjuntos de dados sintéticos sem esparsidade (*Exp.#1* e *Exp.#3*), nos conjuntos de dados sintéticos com esparsidade (*Exp.#2* e *Exp.#4*) e no conjunto de dados do mundo real (corpus de notícias) (*Exp.#5* e *Exp.#6*).

6.1 Exp.#1: estudo do parâmetro α e δ

A análise do estudo do parâmetro α mostra que o algoritmo apresenta um consumo menor de tempo quando trabalha com valores de α menores ou iguais a 1,5 em relação ao comportamento com outros valores do parâmetro (figura 21)²¹. Valores altos para o parâmetro α permitem uma quantidade menor de remoção de múltiplas linhas e colunas, deixando bigrupos com volumes maiores para o procedimento de remoção simples, que por sua vez, é custoso quando os tamanhos da matriz de dados são grandes. Este custo, associado à quantidade de bigrupos que se quer descobrir influencia no tempo total de execução do algoritmo. Note, pela figura 21, que o tempo de execução do algoritmo aumenta conforme aumenta o valor do parâmetro α e a quantidade de bigrupos procurados. O conjunto de dados *a* possui apenas um bigrupos de interesse. O conjunto de dados *c* possui nove bigrupos de interesse. Embora ambos sejam representados por matrizes com o mesmo número de elementos, o tempo de execução do algoritmo é diferente para cada um deles. Para os demais conjuntos de dados, o comportamento é similar, levando a execuções mais lentas conforme aumenta o valor de α .

Figura 21. Variação do tempo de execução do algoritmo de Cheng e Church em relação ao parâmetro α para cada conjunto de dados sintéticos, para δ_3

Adicionalmente foi observado que valores ideais para o parâmetro α variam a depender do objetivo de qualidade de

²¹Cada linha de tendência mostra o comportamento do algoritmo nos conjuntos de dados correspondentes às estruturas de bigrupos *a, b, c, d, e, f, g, h, e i*. Embora apenas o caso com δ_3 esteja sendo mostrado, para δ_1 e δ_2 o comportamento do algoritmo é similar.

resultados buscado, conforme mostrado na tabela 7. Para esta análise, os valores ideais para cada medida de avaliação foram calculados considerando a estrutura de bigrupo conhecida *a priori* para cada conjunto de dados. Na tabela, portanto, são apresentados os valores de α que levam aos melhores resultados de avaliação quantitativa, sob diferentes indicadores de avaliação, quando comparados com os valores ideais para cada medida. Apesar de ter sido observada uma variabilidade para a decisão do melhor valor para o parâmetro α , conclui-se desta análise que valores maiores do que 1,5 não levam a uma boa distribuição para os grupos projetados nas linhas da matriz (segunda a quarta coluna na tabela). Por outro lado, considerando a principal medida de qualidade usado pelo algoritmo de Cheng e Church (sexta coluna na tabela), valores menores do que 1,2 não são adequados.

Tabela 7. Valores para o parâmetro α que levam aos melhores resultados de avaliação quantitativa quando comparados aos valores ideais para cada medida de avaliação, em cada estrutura de bigrupos. Não houve influência do valor de α para a medida de “ocupação” [41]

	Dist inter	Dist intra	Índ DB	Vol	RQM	Cob	Sob
a	-	∇	-	1	$\geq 1,2$	1	∇
b	1,5	1,5	1,5	1	$\geq 1,2$	1	∇*
c	1	1,2	1	1	1,2	1	$\geq 1,2^*$
d	1,2	1,2	1,2	1	$\geq 1,2$	1	$\geq 1,2$
e	1,2	1	1,2	1	$\geq 1,2$	1	1,2
f	1	1	1	1	$\geq 1,2$	1	∇*
g	1,5	1,5	1	1,2	$\geq 1,2$	1,2	∇*
h	1	1,5	1,5	1,2	$\geq 1,2$	1,2	1,5
i	1,2	1	1,2	1,2	$\geq 1,2$	1,2	1,2

* considerando apenas a combinação com δ_1

∇ = qualquer valor para α

Distância inter-grupos (Dist inter), distância intra-grupos (Dist intra), índice Davies Bouldin (Índ DB), média dos volumes (Vol), média dos resíduos quadráticos médios dos bigrupos (RQM), porcentagem da cobertura do conjunto de dados (Cob), porcentagem do grau de sobreposição (Sob).

Em uma análise detalhada sob os efeitos do parâmetro α sobre a maximização do volume, foi observado que, embora com um $\alpha = 1$ o algoritmo apresente resultados bem próximos aos resultados esperados (volume ideal considerando o conhecimento *a priori* sobre os bigrupos), as execuções sob esse parâmetro também apresentam maior variabilidade (desvio padrão alto) nessa medida. Resultados próximos ao ideal e com menor variabilidade entre as execuções foram obtidos considerando $\alpha = 1,2$. Como conclusão do estudo realizado para esse parâmetro, valores de α iguais a 1, 1,2 e 1,5 são indicados como os mais promissores, com destaque para $\alpha = 1,2$, como já indicado pelos proponentes do algoritmo.

Os valores usados no estudo do parâmetro δ dependem de cada conjunto de dados sob análise, conforme apresentado na seção 5. Os gráficos da figura 22 mostram resultados obtidos para cada valor de δ , considerando um conjunto de execuções do algoritmo de Cheng e Church sobre todos os conjuntos de dados sintéticos. Para cada conjunto de dados, o tempo de execução e a média dos volumes e dos RQMs dos bigrupos

encontrados são obtidos. Nos gráficos da figura, essas medidas são novamente agregadas e uma média geral é apresentada diante de cada valor de δ , sendo que: as variações nos tempos de execução são mostradas na figura 22a, os valores médios dos volumes são mostrados na figura 22b, e os valores médios dos RQMs são mostrados na figura 22c. Da figura 22 pode-se deduzir que o uso de δ_1 , embora leve a valores de RQMs pequenos, não maximiza o volume, como objetivado pelo algoritmo, e ainda leva a um tempo de execução maior. Por outro lado, valores de δ_2 e δ_3 trazem resultados bons em termos de volumes e RQMs, e tempos de execução menores, parecendo portanto, serem boas escolhas.

Figura 22. Variação do tempo de execução do algoritmo (a), dos volumes dos bigrupos resultantes (b) e dos RQMs dos bigrupos resultantes (c) em relação ao parâmetro δ para cada conjunto de dados sintético

Considerando a análise sob ponto de vista de valores ideais para δ e dos valores esperados para as medidas de acordo com o conhecimento *a priori* sobre os conjuntos, observa-se que

da mesma forma que ocorre para o parâmetro α , o algoritmo apresenta uma variabilidade nos valores dos parâmetros de δ quando se trata de atender diferentes objetivos de qualidade. Na tabela 8 são mostrados os valores de δ que levam aos melhores resultados comparados com os resultados desejados para cada conjunto de dados. Note que δ_1 contribui pouco para o alcance da qualidade desejada para os bigrupos.

Tabela 8. Valores para o parâmetro δ que levam aos melhores resultados de avaliação quantitativa comparativa com os resultados desejados, para cada estrutura de bigrupos. Não houve influência do valor de δ para a medida de “ocupação”

	Dist inter	Dist intra	Índ DB	Vol	RQM	Cob	Sob
a	-	δ_2, δ_3	-	δ_2, δ_3	δ_2, δ_3	δ_2, δ_3	$\forall$
b	δ_1	δ_1	δ_1	δ_3	δ_3	δ_3	δ_1
c	δ_3	δ_1	δ_1	δ_3	δ_2	δ_3	δ_1^*
d	δ_2, δ_3	δ_2, δ_3	δ_2, δ_3	δ_3	δ_2, δ_3	δ_3	$\forall^*$
e	δ_3	δ_2	δ_3	δ_3	δ_2, δ_3	δ_3	δ_1, δ_3
f	δ_3	δ_3	δ_2	δ_2	δ_2, δ_3	δ_2	δ_1
g	δ_3	δ_3	δ_2	δ_2	δ_2, δ_3	δ_2	δ_1
h	δ_3	δ_2	δ_2, δ_3	δ_2	δ_2, δ_3	δ_2	δ_3
i	δ_3	δ_2	δ_3	δ_3	δ_2, δ_3	δ_3	δ_3

$\forall$ = qualquer valor para δ

* qualquer valor de α com exceção de $\alpha = 1$.

distância inter-grupos (Dist inter), distância intra-grupos (Dist intra), índice Davies Bouldin (Índ DB), média dos volumes (Vol), desvio padrão dos volumes (DP), média dos resíduos quadráticos médios dos bigrupos (RQM), porcentagem da cobertura do conjunto de dados (Cob), porcentagem do grau de sobreposição (Sob).

Como considerações finais para o *Exp.#1* tem-se que para os conjuntos de dados sintéticos originais, a escolha ideal para os parâmetros é: $\alpha = 1, 1, 2$ e $1, 5$; $\delta = \min(\text{RQM})$ ou $> \min(\text{RQM})$. Valores pequenos para a margem do RQM (α) de linhas e colunas em um bigrupo combinados a limiares próximos ao RQM mínimo esperado para os bigrupos (δ), levam a execuções mais eficientes em tempo de execução, ao alcance de boas avaliações para agrupamentos obtidos com projeções de bigrupos, e a medidas mais próximas ao ideal para as avaliações de bigrupos.

6.2 Exp.#2: estudo do parâmetro α e δ em conjunto de dados com níveis de esparsidade

No algoritmo de Cheng e Church o parâmetro α age como um parâmetro de qualidade, porém conforme o valor desse parâmetro aumenta, o algoritmo se torna mais custoso. Os resultados das experimentações realizadas em conjuntos de dados com diferentes níveis de esparsidade confirmam esse comportamento, como pode ser observado nos gráficos organizados nas figuras 23 e 24. Estas figuras trazem informações sobre tempo de execução do algoritmo de Cheng e Church, diante da variação em α , considerando δ_2 (figura 23) e δ_3 (figura 24), com nível de esparsidade em 98%.

Em geral, analisando todos os testes realizados, foi observado que o tempo de execução do algoritmo sofre nenhuma ou pouca influência do aumento da esparsidade quando $\alpha = 1$ ou $\alpha = 1, 2$. A única exceção ocorre no conjunto de dados da estrutura de bigrupos c. Essa estrutura revela maior sensibili-

Figura 23. Variação do tempo de execução do algoritmo de Cheng e Church em relação ao parâmetro α , em cada conjunto de dados sintéticos, com nível de esparsidade alta, usando δ_2

Figura 24. Variação do tempo de execução do algoritmo de Cheng e Church em relação ao parâmetro α , em cada conjunto de dados sintéticos, com nível de esparsidade alta, usando δ_3

dade do algoritmo em diferentes aspectos pois possui muitos bigrupos e tem a característica de ter todos os valores da matriz de dados associados a um bigrupo de forma exclusiva. Da mesma forma como ocorreu no experimento sem níveis de esparsidade (*Exp.#1*), os resultados referentes ao volume e cobertura associados aos bigrupos sofreram mais variações para $\alpha = 1$. Assim, uma boa escolha para o parâmetro α em dados com níveis de esparsidade é 1, 2, pois proporciona bigrupos de qualidade em tempo de execução menores.

Para analisar a influência da esparsidade nos valores do parâmetro δ , observe os gráficos organizados nas figuras 25, 26 e 27. Esses gráficos ilustram a implicação dos valores de δ no tempo de execução (figura 25), no RQM (figura 26) e no volume (figura 27) dos bigrupos, considerando conjuntos de dados com 80% de esparsidade.

De acordo as experimentações, não é uma boa escolha usar valores de δ pequenos (δ_1), já que estes tendem a retornar volumes baixos. Para δ_2 e δ_3 , apesar das medidas assumirem valores levemente similares entre si, os resultados usando δ_3 tendem a ser melhores do que usando δ_2 . Seguindo essa análise e considerando os objetivos do algoritmo de obter bigrupos com volumes máximos e RQMs pequenos, combinados com um tempo de execução aceitável, a melhor escolha é usar um limiar acima ao RQM mínimo para δ , ou seja, δ_3 .

Figura 25. Variação dos tempos de execução em relação ao parâmetro δ , em cada conjunto de dados sintéticos, com níveis de esparsidade e 80%, usando $\alpha = 1, 2$

Figura 26. Variação dos RQM em relação ao parâmetro δ , em cada conjunto de dados sintéticos, com níveis de esparsidade e 80%, usando $\alpha = 1, 2$

6.3 Exp.#3: estudo do comportamento do algoritmo frente à variação do parâmetro K

A fim de analisar o comportamento do algoritmo quanto ao número de bigrupos procurados nos conjuntos de dados sintéticos na suas formas originais, foram testados diferentes valores para o parâmetro K (tabela 5). Nestas experimentações foram considerados $\alpha = 1, 2$ e δ_3 , escolhidos nos experimentos *Exp.#1* e *Exp.#2*. Esta análise está baseada nos objetivos do algoritmo de Cheng e Church e, por consequência, estão sendo avaliados o volume e o RQM dos bigrupos, assim como o tempo de execução do algoritmo. Além disso, são analisadas as medidas da porcentagem de cobertura e a porcentagem do grau de sobreposição, já que elas estão relacionadas com o volume dos bigrupos. As variações dessas medidas, diante dos diferentes valores para K , são mostradas nos gráficos da figura 28, considerando os conjuntos de dados referentes às estruturas de bigrupos *a*, *b*, e *i*. Na figura os valores ideais das medidas sob avaliação para cada um dos conjuntos de dados estão destacados por um círculo. Os conjuntos de dados referentes às demais estruturas de bigrupos apresentam comportamentos similares aos apresentados na figura 28.

Sendo a natureza do algoritmo, a de encontrar um bigrupo por vez, é de se esperar que seu esforço seja maior na procura de valores altos do parâmetro K . Exemplos notórios deste comportamento estão presentes na execução do algoritmo sobre as estruturas *b* e *i*, em que valores altos para o parâmetro K , como $K = 6$ na estrutura *b* e $K = 9$ na estrutura *i*, levam

Figura 27. Variação dos volumes em relação ao parâmetro δ , em cada conjunto de dados sintéticos, com níveis de esparsidade e 80%, usando $\alpha = 1, 2$

a tempos de execução altos comparados aos tempos gastos quando os valores ideais de K são aplicados.

Com respeito ao RQM e ao volume, se o número de bigrupos a procurar é maior do que o ideal, o algoritmo retorna bigrupos com RQMs menores combinados a volumes pequenos. Além disso, a procura por um número de bigrupos maior do que o necessário, leva ao aumento na porcentagem de cobertura e na porcentagem do grau de sobreposição. Esta última medida sobe pois estes bigrupos apresentam uma maior semelhança entre si. Porém, se o parâmetro K é menor do que o ideal, os bigrupos apresentam RQMs e volumes levemente maiores, combinados com menores valores para a porcentagem de cobertura e a porcentagem do grau de sobreposição. Embora o algoritmo retorne bigrupos com maior volume, existem mais elementos que não pertencem a nenhum bigrupo (cobertura pequena). Os bigrupos encontrados possuem baixa semelhança entre si – o que não é uma característica ruim, pois pode significar maior precisão na formação dos bigrupos.

6.4 Exp.#4: estudo do comportamento do algoritmo em conjuntos de dados com diferentes níveis de esparsidade

A fim de fornecer uma visão geral do comportamento do algoritmo Cheng e Church sobre dados de alta dimensionalidade e esparsidade, nesse experimento, os valores dos parâmetros $\alpha = 1, 2$, $\delta = \delta_3$, escolhidos a partir dos experimentos *Exp. #1* e *Exp. #2*, foram mantidos fixos, e o valor de K foi estabelecido de acordo com o número de bigrupos ideal para cada conjunto de dados, seguindo as conclusões obtidas no *Exp. #3*. Então, o algoritmo foi aplicado a cada um dos conjuntos de dados sintéticos alterados para apresentarem diferentes níveis de esparsidade. Os bigrupos resultantes dessas execuções do algoritmo foram avaliados sob as medidas de qualidade para biagrupamento e as medidas de validação internas para agrupamento.

Com base nas análises das tendências das medidas em relação ao incremento da esparsidade, foi possível concluir que o algoritmo de Cheng e Church é sensível à presença de zeros nos conjuntos de dados. O algoritmo tem a finalidade de procurar bigrupos com valores coerentes menores que um limiar, mas segundo o trabalho de Madeira e Oliveira[3] bigru-

Figura 28. Tempo de execução, volumes e RQMs dos bigrupos, porcentagem de cobertura e porcentagem do grau de sobreposição em relação ao número de bigrupos (K) para os conjuntos de dados das estruturas a , b e i na suas formas originais

pos com valores constantes e bigrupos com valores constantes nas linhas ou colunas são casos especiais de bigrupos com valores coerentes. Isto leva o algoritmo a considerar importantes os elementos nulos no biagrupamento, retornando bigrupos formados principalmente com elementos nulos²². Observe, na figura 29, que o grau de ocupação (quantidade de elementos não nulos no bigrupos) decresce conforme a esparsidade do conjunto de dados sob análise aumenta.

Figura 29. Variação do grau de ocupação dos bigrupos em relação ao incremento dos níveis de esparsidade, em cada conjunto de dados sintéticos após a aplicação do algoritmo usando $\alpha = 1, 2$, δ_3 e K determinado como o número de bigrupos presente nos dados

Este comportamento é mais evidente a partir dos níveis

²²Diante desta limitação, com a finalidade de direcionar a busca dos bigrupos aos elementos não nulos, uma estratégia de substituição de elementos nulos por números aleatórios pode ser usada. No trabalho de [Martín-Fernández, Palarea-Albaladejo e Olea\[54\]](#), os autores apresentam diferentes maneiras de lidar com zeros nos conjuntos de dados.

de esparsidade de 60% ou 70%, em que o algoritmo tende a biagrupar os elementos nulos, desconsiderando a informação potencialmente relevante para análise de dados. Por consequência os bigrupos retornados apresentam volumes maiores e valores de RQMs próximos a zero (maior quantidade de elementos nulos) alcançando, portanto, os objetivos originais do algoritmo. Os gráficos organizados na figura 30 mostram a variação das medidas de volume e RQM obtidas para os testes em diferentes níveis de esparsidade dos conjuntos de dados sintéticos. Além disso, em níveis de esparsidade alta, estes bigrupos são semelhantes entre si, apresentando valores altos no índice DB e maior grau de sobreposição. Os gráficos organizados na figura 31 ilustram os comportamentos obtidos nos experimentos em relação a essas duas medidas de qualidade.

6.5 Exp.#5: determinação dos parâmetros δ e K para o corpus de notícias

Segundo a análise realizada nas experimentações nos conjunto de dados sintéticos, o valor adequado para o parâmetro α é 1, 2, considerando dados com alta esparsidade, portanto este valor foi aplicado neste experimento. Para o parâmetro K , foram escolhidos valores em torno do número de canais nos quais as notícias do corpus foram organizadas originalmente. Os valores testados foram: 3, 5, 7 e 10.

Para a determinação do valor do parâmetro δ , o procedimento sugerido pelos autores [Cheng e Church\[14\]](#) foi aplicado. Seguindo esse procedimento, 1000 submatrizes com diferentes dimensões (quantidade de linhas e colunas) foram

Figura 30. Variação da média dos volumes (a) e dos RQMs (b) em relação aos incrementos dos níveis de esparsidade, em cada conjunto de dados sintéticos após da aplicação do algoritmo usando $\alpha = 1, 2$, δ_3 e K determinado como o número de bigrupos presente nos dados

Figura 31. Variação do índice DB (a) e do grau de sobreposição dos bigrupos (b) em relação ao incremento dos níveis de esparsidade, em cada conjunto de dados sintéticos após da aplicação do algoritmo usando $\alpha = 1, 2$, δ_3 e K determinado como o número de bigrupos presente nos dados

obtidas a partir de uma representação vetorial do corpus de notícias. Essas submatrizes são consideradas como bigrupos aleatórios. Sobre elas foram calculados os respectivos RQMs. Esse procedimento permite observar o intervalo de valores pelo qual o RQM varia no conjunto de dados reais, sendo $[0,00021, 0,00256]$ o intervalo obtido nesta instância de matrizes. Dado os pequenos valores para os RQMs, os valores para δ foram determinados também como valores pequenos.

Na tabela 9 são apresentados resultados do uso do algoritmo de Cheng e Church sobre a matriz de dados com representação tfidf. Os resultados são mostrados em termos de qualidade de bigrupos e tempo de execução do algoritmo, para um α fixo, K variando em torno da quantidade de canais, e δ variando conforme indicação do procedimento heurístico. Os resultados mostram que $\delta = 0$ leva à descoberta de bigrupos triviais. A análise destes bigrupos revelou que todos os seus elementos são zeros, ou seja, o algoritmo acha bigrupos com zeros constantes. Valores altos para o parâmetro δ como 0, 1, 0, 5 e 1 leva à descoberta bigrupos com volumes grandes. Nesses testes, todos os elementos do conjunto de dados foram associados a todos os bigrupos. O valor de δ determinado para 0.001 levou o algoritmo a descobrir bigrupos com grande volume, porém, com taxa de sobreposição menor indicando que há diferença entres os bigrupos. Seguindo este raciocínio, e retomando o intervalo de RQMs obtidos no procedimento heurístico, para a análise final com o corpus de notícias, valo-

res de $0,00021 \leq \delta \approx 0,001$ são os mais indicados.

Tabela 9. Medidas de biagrupamento calculadas a partir da aplicação do algoritmo em uma representação vetorial do corpus de notícias usando $\alpha = 1, 2$,

$\delta = \{0, 1 \times 10^{-3}, 0, 1, 0, 5, 1\}$ e $K = \{3, 5, 7, 10\}$

K	δ	Vol	RQM	Cob (%)	Ocup (%)	Sob (%)	Tempo (seg)
3	0	5,7e+04	0	14,16	0	3,74	128
	1e-03	3,9e+05	3e-04	100	6,57	1,08	8115
	0,1	1,2e+06	8e-04	100	6,57	100	18
	0,5	1,2e+06	8e-04	100	6,57	100	18
	1	1,2e+06	8e-04	100	6,57	100	18
5	0	4,5e+04	0	17,45	0	11,58	149
	1e-03	2,4e+05	2e-04	100	6,57	2,90	16202
	0,1	1,2e+06	8e-04	100	6,57	100	32
	0,5	1,2e+06	8e-04	100	6,57	100	32
	1	1,2e+06	8e-04	100	6,57	100	231
7	0	4,1e+04	0	20,57	0	20,00	178
	1e-03	1,8e+05	1e-04	100	6,57	19,99	24587
	0,1	1,2e+06	8e-04	100	6,57	100	45
	0,5	1,2e+06	8e-04	100	6,57	100	44
	1	1,2e+06	8e-04	100	6,57	100	206
10	0	3,7e+04	0	22,71	0	24,73	206
	1e-03	1,5e+05	1e-04	100	6,57	14,67	32786
	0,1	1,2e+06	8e-04	100	6,57	100	58
	0,5	1,2e+06	8e-04	100	6,57	100	57
	1	1,2e+06	8e-04	100	6,57	100	58

Média dos volumes (Vol), média dos resíduos quadráticos médios dos bigrupos (RQM), porcentagem da cobertura do conjunto de dados (Cob %), porcentagem do grau de ocupação (Ocup %) e porcentagem do grau de sobreposição (Sob %).

6.6 Exp.#6: Algoritmo de Cheng e Church aplicado no corpus de notícias

A fim de estudar o comportamento do algoritmo de Cheng e Church no conjunto de dados de notícias, foram realizados testes com matrizes de dados na representação binária, *term frequency* (tf) e *term frequency-inverse document frequency* (tfidf) [53, 2]. Os resultados destes testes são apresentados na tabela 10 em termos de valores médios para medidas de qualidade de bigrupos e de qualidade de grupos (projeções de bigrupos nas linhas da matriz de dados).

De acordo com os resultados médios obtidos, a representação binária permite encontrar bons grupos de notícias segundo as medidas de qualidade de grupo (índice DB e distância inter-grupos), e também apresenta bigrupos que possuem o mais baixo grau de sobreposição. Contudo, os valores baixos para o grau de ocupação e RQMs indicam que os bigrupos possuem muitos valores iguais a zero. Esse resultado representa o alcance do objetivo do algoritmo (minimização do RQM), e a obtenção de bigrupos compostos por maioria (ou totalidade) de elementos iguais zeros representa o encontro de bigrupos coerentes. Contudo, para fins de análise de informação textual, esse resultado não traz significado relevante.

Para as matrizes de dados em representação tf e tfidf, os resultados são semelhantes considerando pares de matrizes (matrizes de dados que passaram pelos mesmos procedimentos de pré-processamento). Resultados diferentes foram observados nos RQMs, sendo que a representação tfidf leva a resultados levemente melhores do que a representação tf.

Tabela 10. Médias, desvios padrão e limites inferiores e superiores dos intervalos de confiança para média, calculados sobre as medidas de qualidade por representação de dados obtidas a partir da aplicação do algoritmo de Cheng e Church sobre o corpus de notícias usando $\alpha = 1, 2$, $\delta = 0,0008$ e $K = \{3, 5, 7, 10\}$. Para cálculo dos intervalos de confiança foram considerados: amostras de tamanho 20; distribuição *t-student* e nível de confiança 95%

Medidas de qualidade	Binária				TF				TF-IDF			
	μ	σ	l_{inf}	l_{sup}	μ	σ	l_{inf}	l_{sup}	μ	σ	l_{inf}	l_{sup}
Ocupação (%)	0.009	0.017	0.00	0.01	0.91	0.15	0.87	0.95	0.95	0.14	0.91	0.99
Sobreposição (%)	33.52	18.46	28.37	38.67	48.21	14.17	44.25	52.16	49.17	13.13	45.51	52.83
Cobertura (%)	52.69	9.02	50.18	55.21	78.35	8.05	76.11	80.60	79.46	7.46	77.38	81.54
Volume	$1.17e^6$	$4.96e^5$	$1.03e^6$	$1.31e^6$	$2.18e^6$	$9.73e^5$	$1.91e^6$	$2.45e^6$	$2.24e^6$	$1.00e^6$	$1.96e^6$	$2.52e^6$
RQM	$1.25e^{-4}$	$1.25e^{-4}$	$4.33e^{-5}$	$1.25e^{-4}$	$1.90e^{-4}$	$7.68e^{-5}$	$1.69e^{-4}$	$2.11e^{-4}$	$2.30e^{-4}$	$7.14e^{-5}$	$2.10e^{-4}$	$2.50e^{-4}$
Índice DB	18.86	4.08	17.73	20.00	44.82	6.07	43.13	46.51	46.59	6.33	44.83	48.36
Dist. inter-grupos	1.79	0.26	1.72	1.86	0.26	0.24	0.00	0.33	0.27	0.27	0.20	0.35
Dist. intra-grupos	$3.62e^3$	$7.18e^2$	$3.42e^3$	$3.82e^3$	$1.54e^3$	$4.34e^2$	$1.42e^3$	$1.66e^3$	$1.57e^3$	$4.55e^2$	$1.45e^3$	$1.70e^3$

Na análise qualitativa dos bigrupos foram considerados aqueles descobertos com uma matriz de dados em representação tfidf e com 3.324 palavras. A análise foi realizada por meio da inspeção visual dos bigrupos diante das nuvens de palavras construídas a partir das 30 palavras mais frequentes nos documentos pertencentes à projeção de cada bigrupo sobre as linhas da matriz de dados, considerando o subconjunto de palavras pertencentes à projeção de cada bigrupo sobre as colunas da matriz (característica que difere a informação trazido pelos bigrupos da informação que classicamente se tem quando a tarefa de agrupamento é realizada). Nas nuvens, quanto maior o tamanho da palavra, maior a sua frequência no grupo de documentos e, portanto, maior a sua proeminência na caracterização semântica daquele grupo. A figura 32 traz três nuvens de palavras, cada uma correspondente a um bigrupo encontrado pelo algoritmo de Cheng e Church.

Figura 32. Visualização em nuvem de palavras das top-30 palavras mais frequentes em cada bigrupo, gerados usando $K = 3$, $\alpha = 1, 2$ e $\delta = 0,0008$

(a) B.#1 “educação”

(b) B.#2 “política”

(c) B.#3 “esporte (futebol)”

um rótulo a cada bigrupo, o qual reflete o tópico (assunto) tratado pelos documentos (ou pela maioria dos documentos) associados a aquele bigrupo. De acordo com a interpretação dos autores deste tutorial, os seguintes rótulos podem ser atribuídos aos bigrupos:

- “educação” (bigrupo 1 - B.#1), com 755 notícias caracterizadas por 1956 palavras;
- “política” (bigrupo 2 - B.#2), com 562 notícias caracterizadas por 1410 palavras;
- “esporte (futebol)” (bigrupo 3 - B.#3), com 494 notícias caracterizadas por 2.094 palavras.

Especificamente analisando a configuração dos bigrupos formados no que diz respeito à projeção nas linhas da matriz de dados, grandes sobreposições se formaram entre os grupos. Por exemplo, o grupo de notícias referentes à “educação” e “política” possui 160 notícias em comum; já o grupo referente à “educação” e “esporte” tem uma sobreposição de 494 notícias; “política” e “esporte” possuem uma sobreposição de 75 notícias; e finalmente, os três grupos possuem 75 notícias em comum. A organização dos grupos de documentos é visualmente explicada na figura 33.

Figura 33. Visualização para a organização dos documentos em grupos projetados dos bigrupos

Diante da análise das nuvens de palavras é possível atribuir

Interessante ressaltar que todas as notícias pertencentes ao grupo “esporte” pertencem também ao grupo “educação”.

Do ponto de vista de agrupamento de documentos, a projeção do bigrupo “esporte” representa um subgrupo da projeção do bigrupo “educação”, o que leva à noção de agrupamento hierárquico. Em termos de significado, alguém pode entender que as notícias referentes à esporte especializam as notícias referentes à educação, e portanto, trata-se de uma visão do assunto esporte sob aspectos educacionais, ou ainda que as notícias sobre esporte possuem um viés referente à educação.

Entretanto, biagrupamento fornece informação ainda mais específica, uma vez que a relação entre as palavras pode ser analisada. Sob esse ponto de vista, existem sobreposições que estão organizadas da seguinte forma: existem seis grupos de palavras, conforme distribuído na barra superior da figura 34 e caracterizados na parte inferior da figura de acordo com os tópicos estabelecidos segundo a interpretação dos autores deste tutorial; 36 palavras formam o grupo de palavras que é comum na caracterização dos bigrupos referentes aos tópicos “educação” e “política”; há um grande grupo formado por 1502 palavras que é comum à caracterização dos bigrupos “educação” e “esporte”; e um grupo de 586 palavras que é comum aos bigrupos “política” e “esportes”; e finalmente, há uma sobreposição em todos os bigrupos em relação a um grupo de 6 palavras. Note que as notícias caracterizadas como notícias de esportes podem ser colocadas em dois perfis: o primeiro perfil é de notícias que também versam sobre assuntos relacionados à política; o segundo perfil é de notícias que também versam sobre educação. Esses perfis são visíveis apenas na visão de projeção dos bigrupos nas colunas da matriz, pois na projeção dos bigrupos sobre as linhas, a relação apenas dos tópicos “esporte” e “política” não pode ser observada.

Figura 34. Visualização para a organização das palavras em grupos projetados dos bigrupos

A sobreposição das palavras pode ser observada na nuvem de palavras da figura 32. Por exemplo, as palavras “jogo”, “vasco”, “januário” são representativas tanto para o bigrupo “política” quanto para o bigrupo “esporte” (embora tenham maior proeminência no bigrupo “esporte”). A análise, por esse ponto de vista, mostra com mais ênfase as similaridades existentes entre os assuntos tratados em cada bigrupo. Para

um entendimento ainda mais específico sobre o tipo de organização de documentos obtida neste resultado, considere as seguintes notícias extraídas dos bigrupos B.#2 (política) e B.#1 $\cap$ B.#3 $\cap$ B.#2 (educação $\cap$ esporte $\cap$ política):

- notícia intitulada “*Justiça interdita São Januário por seis meses a pedido do Ministério Público*” pertencente ao bigrupo B.#2: trata da interdição do estádio de futebol de São Januário pelo Ministério Público do Estado de Rio de Janeiro;
- notícia intitulada “*Vasco e Grêmio jogam em São Januário*” pertencente ao bigrupo B.#3: trata do jogo das equipes do Vasco da Gama e Grêmio no estado de São Januário.

Analisando o resultado do biagrupamento sob essa perspectiva, situações como essas são comuns: notícias similares, que poderiam ser colocadas em um único grupo por um algoritmo de agrupamento clássico, são organizadas de maneira mais precisa em bigrupos diferentes e em suas sobreposições. Neste caso, o estádio de São Januário é objeto de dois tópicos de discussão diferentes (“política” e “esporte”), entretanto representa um conceito que permite criar uma ligação entre tópicos. Esse exemplo, embora útil para ilustrar o poder de expressividade conseguido nos resultados de biagrupamento, também permite revelar que o resultado produzido pelo algoritmo de Cheng e Church apresenta imprecisões. É difícil explicar, neste contexto, a relação da notícia com o tópico de “educação”. Muito provavelmente, a alta esparsidade dos dados está tornando a separabilidade dos bigrupos mais complexa para o algoritmo, como ilustrado nos experimentos com os conjuntos de dados sintéticos.

7. Resultados e discussões: algoritmo NBVD

O algoritmo NBVD foi estudado nos conjuntos de dados sintéticos sem esparsidade (Exp.#1), nos conjuntos de dados sintéticos com esparsidade (Exp.#2) e no conjunto de dados do mundo real (corpus de notícias) Exp.#3).

7.1 Exp.#1: estudo dos parâmetros k e l

O algoritmo NBVD foi executado sobre os conjuntos de dados sintéticos em sua forma original, considerando os procedimentos descritos na seção 5. A avaliação da qualidade das partições obtidas foi feita com base nos erros de reconstrução e na inspeção visual da plotagem das matrizes reconstruídas. Foi observado que quanto maiores são os valores atribuídos a k e l , menores são os erros de reconstrução e maior a qualidade visual das plotagens das matrizes reconstruídas.

Seguindo essa constatação, valores altos para k e l podem ser bons, mas sendo o algoritmo NBVD baseado em cálculos matriciais, que tendem a apresentar alta complexidade computacional com valores altos de k e l , a escolha de valores altos pode inviabilizar o uso do algoritmo em conjuntos de dados volumosos e de alta dimensionalidade. A escolha para

os valores de k e l deve ser guiada principalmente com base no custo computacional imposto pelo problema sob resolução, considerando a maximização dos valores de k e l ou o uso de algum conhecimento a priori sobre a quantidade de cogrupos existente no conjunto de dados ou esperado pelo contexto da aplicação em questão. Os gráficos da figura 35 ilustram uma boa reconstrução (para valores $k, l \geq 4$) e reconstrução ruim (para valores $k, l \leq 4$) no conjunto de dados da estrutura i . Na figura, o caso (a) é a matriz original, (b) a reconstrução ruim e (c) a reconstrução boa.

Figura 35. Representação visual de uma reconstrução ruim (b) e boa (c) do conjunto de dados da estrutura i

7.2 Exp.#2: estudo do comportamento do algoritmo com presença de esparsidade

Como próximo passo no entendimento do comportamento do algoritmo NBVD, foram realizadas execuções do algoritmo nos conjuntos de dados sintéticos alterados para apresentarem diferentes níveis de esparsidade. A análise dos resultados permitiu concluir que o algoritmo NBVD se adapta para níveis de esparsidade baixos e altos. A natureza do algoritmo lhe permite encontrar uma equivalência entre a importância dos elementos não nulos e dos elementos nulos dentro dos conjuntos de dados. O gráfico mostrado na 36 relaciona o valor do erro de reconstrução com os níveis de esparsidade (cada linha no gráfico diz respeito à execução do algoritmo sobre um conjunto de dados diferente). O algoritmo alcança erros mais baixos quando existe prevalência ou de elementos não nulos ou de elementos nulos.

Figura 36. Variação do erro de reconstrução em relação ao aumento dos níveis de esparsidade, em cada conjunto de dados sintético

Para altos níveis de esparsidade, com exceção do nível máximo com 98% de esparsidade, foi observado que embora os cogrupos encontrados contenham muitos valores iguais a zero, estes ainda são agrupados com base nas informações

relevantes dos conjuntos de dados e não com base nos elementos nulos (i.e. não são formados cogrupos equivalentes a bigrupos constantes e com valores iguais a zero). Um exemplo de como a perda de informação é observada na reconstrução da matriz original é apresentado na figura 37.

Figura 37. Reconstrução visual dos resultados no conjunto de dados da estrutura d para diferentes os níveis de esparsidade 30 (figura a), 70 (figura b), 90 (figura c) e 98 (figura d)

Adicionalmente, o algoritmo NBVD apresenta um comportamento estável no que diz respeito à avaliação do volume dos cogrupos e da cobertura do conjunto de dados. Esse comportamento é ilustrado nos gráficos da figura 38.

Figura 38. Variação do (a) volume dos cogrupos e (b) porcentagem de cobertura do conjunto de dados em relação ao aumento dos níveis de esparsidade, em cada conjunto de dados sintético

Nos experimentos também foi observado que nos casos extremos de desbalanceamento entre a prevalência de valores nulos e não nulos, o algoritmo tem dificuldade em separar os cogrupos, retornando cogrupos com porcentagens altas no grau de sobreposição e valores de índice DB mais altos. Os gráficos da figura 39 ilustram os comportamentos obtidos nos experimentos em relação a essas duas medidas de qualidade.

Figura 39. Variação da (a) porcentagem do grau de sobreposição e (b) do índice DB em relação ao aumento dos níveis de esparsidade, em cada conjunto de dados sintético

7.3 Exp.#3: Algoritmo NBVD aplicado ao corpus de notícias

Nos experimentos do algoritmo de NBVD sobre o corpus de notícias, foi observado que, embora exista uma variabilidade no resultado das medidas de qualidade para as diferentes representações de dados utilizada, a representação tf-idf foi a que permitiu alcançar os melhores índices em seis medidas. Além disso, altos erros de reconstrução obtidos com a representação binária desmotivam o seu uso. Os valores médios obtidos para cada uma das medidas em cada uma das representações de dados estão listados na tabela 11.

Para análise qualitativa de resultados, foi escolhida a partição obtida sobre a matriz com representação tf-idf com 3.324 palavras resultantes da heurística de seleção. Nas análises qualitativas, o algoritmo mostrou ser capaz de gerar informação sobre os dados, e de fornecer como cada grupo de notícias se organiza em termos de grupos de palavras. O algoritmo apresentou uma boa capacidade de separação dos assuntos presentes em cada grupo de notícias, permitindo fazer rotulações sobre os tópicos representados em cada cogruppo.

A análise qualitativa realizada teve a finalidade de discutir a relação existente entre os grupos de palavras e os grupos de notícias encontrados pelo algoritmo. A relação dos grupos de linhas com os grupos de palavras, estabelecida na matriz S , para o caso de $k = l = 3$, está mostrada na tabela 12.

Na tabela 12 é observado a formação de três cogrupos de interesse bem definidos. Observe que o grupo de notícias 1 (GN#1) está fortemente relacionado com o grupo de palavras 3 (GP#3), formando o cogruppo 1 (C#1). Do mesmo modo o grupo de notícias 2 (GN#2) está fortemente relacionado com o grupo de palavras 1 (GP#1), formando o segundo cogruppo (C#2). E, por fim, o grupo de notícias 3 (GN# 3) está fortemente relacionado com o grupo de palavras 2 (GP#2), gerando o terceiro cogruppo (C#3). Adicionalmente, por meio do processo de pós-processamento detalhado na seção 4, são obtidas as notícias pertencentes a cada grupo de notícias e as palavras pertencentes a cada grupo de palavras.

Uma forma de visualização que permite a interpretação do significado que está associado ao cogruppo é nuvem de palavras construída sobre cada um dos grupos de notícias e dos grupos de palavras referentes ao cada cogruppo. Na nuvem, o tamanho das palavras é definido de acordo com a quantidade de ocorrências da palavra nos documentos associados a cada cogruppo. Além disso, para evidenciar a importância das palavras que mais bem definem cada cogruppo, apenas as 30 palavras de maior ocorrência no cogruppo são usadas. As nuvens são apresentadas na figura 40. Por meio de inspeção visual das palavras na nuvem associada a cada cogruppo, uma rotulação possível para o resultado apresentado na figura 40 é: o grupo C#1 está associado a notícias de “Cidadania”; o grupo C#2 está associado a notícias de “Educação”; o grupo C#3 está associado a notícias de “Política”.

Em uma análise mais específica da figura 40 é possível observar que existem palavras que aparecem em diferentes cogrupos. Um exemplo é a palavra “social” que aparece

Figura 40. Visualização em nuvem de palavras das top-30 palavras para cada cogruppo gerado pelo algoritmo

nos três grupos de palavras (GP#1, GP#2, GP#3). Esse fato significa que existem notícias, tanto no GP#1 como no GP#2 e no GP#3, nas quais a palavra “social” ocorre, sendo que a ocorrência mais frequente é no GP#3 (veja o tamanho da palavra em cada uma das nuvens). Entretanto, o algoritmo de coagrupamento, ao fazer a análise simultânea de grupos de linhas e palavras, proporciona que contextos diferentes para cada palavra sejam descobertos. Ao analisar a ocorrência da palavra “social” em notícias de cada cogruppo percebe-se que embora o seu significado não mude, ela adquire uma função especial em cada contexto, por exemplo:

- em uma notícia do cogruppo de “Cidadania”, a palavra “social” está associada a “nome social” usado nos registros civis para transexuais;
- em uma notícia do cogruppo de “Política”, a palavra “social” aparece associada ao “programa de integração social” que é o benefício outorgado pelo governo aos trabalhadores de empresas privadas;
- em uma notícia do cogruppo de “Educação”, a palavra “social” aparece associada a “blindagem social”, termo utilizado em solicitações feitas por professores à prefeitura frente à situação das escolas.

8. Revisitando os resultados

Para traçar um paralelo da aplicação dos algoritmos de Cheng e Church e NBVD sobre os dados com altas dimensionalidade e esparsidade, a análise dos bigrupos (seção 6.6) e dos cogrupos (seção 7.3) são revisitadas nesta seção²³. Em

²³ Uma análise das interseções dos intervalos de confiança para as médias de cada medida de qualidade apresentados nas tabelas 10 e 11 mostra que os

Tabela 11. Médias, desvios padrão e limites inferiores e superiores dos intervalos de confiança para média, calculados sobre as medidas de qualidade por representação de dados, aplicando o algoritmo NBVD no corpus de notícias usando $k = l = \{3, 5, 7, 10\}$. Para cálculo dos intervalos de confiança foram considerados: amostras de tamanho 60; distribuição *t-student* e nível de confiança 95%

Medidas de qualidade	Binária				TF				TF-IDF			
	μ	σ	l_{inf}	l_{sup}	μ	σ	l_{inf}	l_{sup}	μ	σ	l_{inf}	l_{sup}
Ocupação (%)	3.99	0.71	3.77	4.22	2.90	0.46	2.76	3.05	2.88	0.45	2.74	3.03
Sobreposição (%)	36.57	16.33	31.38	41.77	33.07	16.01	27.97	38.16	31.07	16.14	25.93	36.20
Cobertura (%)	23.23	4.49	21.80	24.66	31.28	4.69	29.79	32.78	31.52	5.00	29.93	33.11
Volume	$4.61e^5$	$2.11e^5$	$3.94e^5$	$5.28e^5$	$6.09e^5$	$2.72e^5$	$5.23e^5$	0.00	$5.99e^5$	$2.65e^5$	$5.15e^5$	$6.84e^5$
RQM	0.045	$6.11e^{-3}$	0.043	0.047	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Índice DB	13.86	6.88	11.67	16.05	23.52	6.89	21.33	25.71	23.04	6.86	20.86	25.23
Dist. inter-grupos	2.42	0.44	2.27	2.56	0.30	0.06	0.29	0.32	0.30	0.06	0.29	0.32
Dist. intra-grupos	$4.33e^3$	$1.07e^3$	$3.99e^3$	$4.67e^3$	$1.18e^3$	$3.34e^2$	$1.08e^3$	$1.29e^3$	$1.18e^3$	$3.33e^2$	$1.07e^3$	$1.29e^3$
Erro reconstrução	284,62	53,29	267,78	301,47	86,94	18,95	80,95	92,93	86,94	18,96	80,95	92,94

Tabela 12. Matriz *S* gerada após da aplicação do algoritmo de NBVD no corpus com $k = l = 3$

	GP#1	GP#2	GP#3
GN#1	0,0	0,0	2,08
GN#2	1,84	0,0	0,0
GN#3	0,0	0,44	0,0

ambos os casos, os resultados são provenientes de execuções que geraram três bi/cogrupos, sendo que no caso de coagrupamento, mais cogrupos foram gerados, porém apenas três foram considerados de interesse. De maneira resumida, as características dos bi/cogrupos são:

- bigrupos (*Cheng e Church*):
 - tópicos: educação, política, esporte;
 - volume (total): $1,1 \times 10^6$
 - cobertura (%): 65,2
 - sobreposição (%): 31,7
 - ocupação (%): 1,1
- cogrupos (*NBVD*):
 - tópicos: cidadania, educação, esporte;
 - volume (total): $3,1 \times 10^5$;
 - cobertura (%): 22,7
 - sobreposição (%): 6,04
 - ocupação (%): 3,9

As medidas quantitativas ilustram algumas diferenças básicas entre os dois resultados. Visto que o resultado oferecido pelo NBVD pode ser pós-processado, cogrupos contendo a relação documento-palavras representadas por valores abaixo

algoritmos analisados possuem comportamentos distintos. Exceções foram observadas nas medidas: sobreposição na representação binária; distâncias inter-clusters nas representações tf e tfidf.

da média na matriz *S* foram desconsiderados. Então, a quantidade de elementos da matriz de dados (volume e cobertura) representada nos cogrupos é menor do que a quantidade representada nos bigrupos. Em uma primeira análise, essa pode ser uma característica ruim do resultado final do coagrupamento. Porém, analisando também as medidas de sobreposição e ocupação, percebe-se que a situação oferecida pelo coagrupamento indica uma definição mais precisa dos cogrupos, e com uma maior quantidade de elementos significativos (ocupação - quantidade de elementos não nulos). Essas duas medidas pesam em favor dos resultados de coagrupamento, pois embora a sobreposição entre co/bigrupos possa ser interessante, quando ela é muito grande dificulta a análise do significado inerente à organização dos dados. Claramente, a menor quantidade de elementos nulos nos cogrupos também oferece mais condições de expressividade ao resultado obtido.

Nas figuras 41 e 42, são apresentadas abstrações gráficas²⁴ da organização dos bigrupos e dos cogrupos, respectivamente, obtidos nos resultados analisados neste tutorial. Nas figuras, é possível observar a representação das medidas dos bi/cogrupos e como os tópicos descobertos estão relacionados. Especialmente no caso de coagrupamento, o cogrupos de notícias relacionadas à política aparece representado em três regiões, significando que ele possui interseção total (em termos de notícias e palavras) com os dois outros cogrupos e também possui uma parte sem interseção. Em outras palavras é possível inferir que há notícias que tratam de cidadania com um viés mais forte para política, assim como há notícias relacionadas à educação com viés político, além daquelas que tratam do tema de forma isolada. Análises similares podem ser feitas em relação às demais sobreposições.

A abstração para bigrupos mostra que as fronteiras entre os tópicos são mal definidas, já que as sobreposições são maiores. O grande volume de notícias e palavras associadas a cada grupo é um indício de que uma análise qualitativa nesse

²⁴Estas visualizações são ditas abstrações da organização original dos dados já que, na realidade, as notícias (linhas) e as palavras (colunas) estão dispersas na matriz de dados, sem estarem necessariamente organizadas de forma adjacente para cada tópico, como mostrado nas figuras.

Figura 41. Abstração gráfica para a organização dos bigrupos **Figura 42.** Abstração gráfica para a organização dos cogrupos

contexto é complexa. De fato, da experiência vivenciada pelos autores deste tutorial, pode ser dito que a tarefa de analisar os biagrupamentos se mostrou bem mais difícil de ser realizada, e produziu resultados menos expressivos.

9. Leituras adicionais

Neste tutorial, duas classes de algoritmos foram abordadas por meio da discussão de dois algoritmos básicos. O primeiro algoritmo constitui-se como uma estratégia de biagrupamento baseada em procedimentos heurísticos; o segundo representa uma estratégia de coagrupamento baseada em fatoração de matrizes. Recentemente, algoritmos pertencentes a essas classes têm sido aplicados na mineração de textos. Alguns exemplos são: [7, 8, 55, 56, 57, 41] - da classe de algoritmos que usam procedimentos heurísticos; e [5, 58, 59, 34, 60, 61] - da classe de algoritmos que fazem uso de fatoração de matrizes.

Na literatura, a apresentação destes algoritmos geralmente é feita em termos de sua capacidade de melhorar os resultados da clássica tarefa de agrupamento de textos [7, 8, 55, 56, 5, 58, 59, 34, 60, 61]. Eventualmente, os autores inserem o algoritmo em um método para resolução da tarefa de classificação de textos [57], apresentam o algoritmo em um contexto prático, como o do problema de expansão de consultas [41], ou ainda mostram a efetividade do algoritmo também em outros tipos de problemas, por exemplo, naqueles provenientes da área de bioinformática [58].

Em termos de avaliação da qualidade dos resultados alcançados pelos algoritmos, não é comum encontrar o uso das medidas de qualidade de bigrupos referentes a volume,

cobertura, sobreposição e ocupação na literatura de mineração de textos. Encontra-se, entretanto, o uso de resíduo quadrático médio [41, 62], índice Davies-Bouldin [63], similaridade média intra e inter grupos [34] e inspeção visual [62]. Para análise do desempenho do algoritmo, encontra-se artigos nos quais os autores destacam análises de tempo de execução [60, 64]. Além disso, o uso de índices de validação externa é bastante frequente na área de mineração de textos. Para uso desses índices, é necessário conhecer *a priori* a organização dos textos em classes. No estudo desenvolvido para construção deste tutorial, foi possível levantar alguns índices de validação externa recentemente aplicados²⁵: informação mútua normalizada [58, 9, 7], acurácia [65, 33], pureza [7, 8, 66], precisão, revocação e *F-measures* [5, 11, 67] e índice de Rand ajustado [9, 68, 59].

Quando o intuito de um estudo é apresentar algoritmos novos ou adaptar algoritmos existentes para biagrupamento ou coagrupamento de textos, é interessante contrapor os resultados obtidos àqueles disponíveis na literatura correlata. Nesses casos, conjuntos de dados textuais de referência podem ser usados para facilitar comparações. Os conjuntos de dados usados nos artigos já mencionados nesta seção são: 20-Newsgroups [11, 68, 7, 63, 59, 34, 60, 57, 67, 55, 56, 65, 41], Reuters [9, 68, 34, 60, 57, 67, 69, 33, 64], CLASSIC3 [9, 59, 7, 63, 60, 67], CLUTO toolkit [9, 59, 34, 66, 60, 55], CSTR [9, 59, 34, 65, 33], WebACE [65, 11, 59, 34, 60], CLASSIC4 [65, 56, 9, 59, 55], MovieLens [7, 67, 41], Yahoo-K1 [9, 57], TDT2 [9, 65], Citeseer [5, 61], CORA [5, 8], BASEHOCK [58], CNAE [58], DBLP [61], URCS technical reports [61], TREC [11], Jester [67] e Brazilian Newspaper Folha [62].

10. Considerações finais

Este tutorial ofereceu ao leitor:

- Visão conjunta sobre os conceitos das tarefas de biagrupamento e coagrupamento. Tais conceitos são geralmente apresentados na literatura de forma isolada, de maneira que para leitores iniciantes o entendimento das diferenças entre as tarefas se torna difícil.
- Detalhamento sobre o funcionamento de um algoritmo para solução da tarefa de biagrupamento – o algoritmo de Cheng e Church. Gráficos e exemplos numéricos foram usados para ilustrar, de forma didática, os passos básicos do algoritmo.
- Detalhamento sobre o funcionamento de um algoritmo para solução da tarefa de coagrupamento – o NBVD. Gráficos e exemplos numéricos foram usados para ilustrar, de forma didática, os passos básicos do algoritmo.

²⁵Os artigos citados para ilustrar o uso de índices não estão restritos a artigos que apresentem algoritmos baseados em heurísticas ou baseado em fatoração de matrizes. Outras classes de algoritmos de biagrupamento e coagrupamento foram incluídas no escopo de análise a fim de trazer mais informações sobre índices de validação. Além disso, cada um dos artigos não está restrito ao uso de apenas um índice de validação.

Além disso, uma maneira de extrair os cogrupos das matrizes fatoradas no algoritmo também foi apresentada de forma a possibilitar que o leitor facilmente possa aplicar o algoritmo em problemas de seu interesse e verificar o resultado de maneira aplicada.

- Apresentação de medidas de avaliação de qualidade de biagrupamento e de agrupamento, com uso tanto no contexto de biagrupamento quanto de coagrupamento. Desde que não é comum encontrar medidas de avaliação específicas para coagrupamento, é útil verificar que formas de avaliação preparadas para outros contextos podem ser transferidas e aplicadas para analisar os resultados dos algoritmos que resolvem tal tarefa. Por outro lado embora medidas específicas de avaliação da qualidade de biagrupamento estejam disponíveis na literatura e tenham sido aplicadas aqui, as medidas usadas em agrupamento foram úteis nas análises.
- Método de experimentação para cada um dos algoritmos e experimentações com conjuntos de dados sintéticos e reais, de altas dimensões. As experimentações foram estendidas para conjuntos de dados sintéticos alterados com a inserção de níveis de esparsidade, de maneira a ilustrar como os algoritmos se comportam neste tipo de problema. No caso dos conjuntos de dados reais, um corpus de notícias foi usado, ilustrando o desempenho dos algoritmos em uma tarefa de grande interesse atual - a mineração de textos.
- Apresentação de análises qualitativas. Em relação às experimentações com o corpus de notícias, as análises qualitativas obtidas dos cogrupos mostraram o maior potencial de extração de informação presente nas soluções para a tarefa de coagrupamento, do que as análises qualitativas obtidas nos bigrupos. A abstração gráfica criada especialmente para este tutorial serve como uma forma de facilitar a inspeção visual dos resultados.
- Uma breve revisão de literatura de biagrupamento e coagrupamento aplicados a análise de dados textuais foi apresentada. Esta revisão constitui um indicativo de que tem havido esforços recentes para aprimorar a resolução dessas tarefas em um contexto de grande interesse atualmente.

No que diz respeito à aplicação dos algoritmos sobre dados de alta dimensionalidade e com esparsidade, e com base nas experimentações realizadas, o algoritmo de coagrupamento NBVD se mostra mais bem adaptado ao problema do que o algoritmo de biagrupamento proposto por Cheng e Chung, uma vez que consegue encontrar cogrupos mais bem definidos (com menos sobreposições) e com uma quantidade maior de elementos não nulos. O algoritmo de NBVD apresentou uma capacidade maior de separação de temas, permitindo fazer rotações mais precisas dos tópicos relacionados aos cogrupos resultantes, e portanto trazendo resultados mais relevantes. O

algoritmo de Cheng e Church apresentou maior sensibilidade em relação à variação de níveis de esparsidade nos conjuntos de dados sintéticos. Essa sensibilidade está relacionada à natureza do algoritmo – encontrar bigrupos coerentes com RQMs baixos – já que os elementos nulos, abundantes em dados esparsos, podem ser considerados como casos especiais de valores coerentes quando organizados dentro de bigrupos. Essas constatações, no entanto, não são generalizáveis para as respectivas tarefas que os algoritmos testados resolvem. Outros algoritmos especialmente desenvolvidos para resolver a tarefa de biagrupamento podem apresentar resultados mais interessantes, enquanto o contrário também pode ocorrer com outros algoritmos desenvolvidos para coagrupamento.

Agradecimentos

Os autores agradecem à Universidade de São Paulo pela disponibilidade do laboratório de pesquisa em Inteligência Artificial (GrIA) utilizado no desenvolvimento deste projeto, à Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) pela concessão de bolsa de mestrado e ao pesquisador André Paulino de Lima pelas valorosas críticas que possibilitaram o aprimoramento deste texto.

Contribuição dos Autores

A autora Alexandra Katiuska Ramos Diaz é a principal responsável pela execução do trabalho e escrita do artigo. A autora Sarajane Marques Peres orientou o desenvolvimento da dissertação de mestrado que originou este artigo e participou ativamente nas revisões deste texto.

References

- [1] FELDMAN, R.; SANGER, J. *The text mining handbook: advanced approaches in analyzing unstructured data*. New York: Cambridge University Press, 2007.
- [2] SILVA, L. A. da; PERES, S.; BOSCARIOLI, C. *Introdução À Mineração de Dados - Com Aplicação Em R*. São Paulo, Brasil: Elsevier Editora Ltda., 2016.
- [3] MADEIRA, S. C.; OLIVEIRA, A. L. Biclustering algorithms for biological data analysis: a survey. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, IEEE Computer Society Press, v. 1, n. 1, p. 24–45, 2004.
- [4] NADIF, M. et al. *Co-Clustering: Models, Algorithms and Applications*. London: John Wiley & Sons, 2013.
- [5] XUAN, J. et al. Doubly nonparametric sparse nonnegative matrix factorization based on dependent Indian buffet processes. *IEEE Transactions on Neural Networks and Learning Systems*, IEEE, v. 29, n. 5, p. 1835–1849, 2018.
- [6] HUANG, S.; XU, Z.; LV, J. Adaptive local structure learning for document co-clustering. *Knowledge-Based Systems*, Elsevier, v. 148, p. 74–84, 2018.

- [7] FRANÇA, F. O. de. A hash-based co-clustering algorithm for categorical data. *Expert Systems with Applications*, Elsevier, v. 64, p. 24–35, 2016.
- [8] HONDA, K. et al. MMMs-induced k-member co-clustering for k-anonymization of cooccurrence information. In: IEEE. *Proceedings of the International Joint Conference on Neural Networks*. Vancouver, BC, Canadá, 2016. p. 2961–2966.
- [9] AILEM, M.; ROLE, F.; NADIF, M. Sparse poisson latent block model for document clustering. *IEEE Transactions on Knowledge and Data Engineering*, IEEE, v. 29, n. 7, p. 1563–1576, 2017.
- [10] SALAH, A.; ROGOVSKI, N.; NADIF, M. Stochastic co-clustering for document-term data. In: SIAM. *Proceedings of the SIAM International Conference on Data Mining*. Miami, Florida, USA, 2016. p. 306–314.
- [11] MEI, J.-P. et al. Large scale document categorization with fuzzy clustering. *IEEE Transactions on Fuzzy Systems*, IEEE, v. 25, n. 5, p. 1239–1251, 2017.
- [12] AILEM, M.; ROLE, F.; NADIF, M. Graph modularity maximization as an effective method for co-clustering text data. *Knowledge-Based Systems*, Elsevier, v. 109, p. 160–173, 2016.
- [13] AILEM, M.; ROLE, F.; NADIF, M. Co-clustering document-term matrices by direct maximization of graph modularity. In: ACM. *Proceedings of the 24th ACM International Conference on Information and Knowledge Management*. Melbourne, VIC, Australia, 2015. p. 1807–1810.
- [14] CHENG, Y.; CHURCH, G. M. Biclustering of expression data. In: AAAI. *Proceedings of the Intelligent Systems for Molecular Biology*. La Jolla, San Diego, CA, USA, 2000. v. 8, n. 2000, p. 93–103.
- [15] LONG, B.; ZHANG, Z. M.; YU, P. S. Co-clustering by block value decomposition. In: ACM. *Proceedings of the 11th ACM SIGKDD International Conference on Knowledge Discovery in Data Mining*. Chicago, Illinois, USA, 2005. p. 635–640.
- [16] BANERJEE, A. et al. A generalized maximum entropy approach to bregman co-clustering and matrix approximation. *Journal of Machine Learning Research*, v. 8, n. Aug, p. 1919–1986, 2007.
- [17] ANAGNOSTOPOULOS, A.; DASGUPTA, A.; KUMAR, R. Approximation algorithms for co-clustering. In: ACM. *Proceedings of the 27th ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems*. Vancouver, BC, Canada, 2008. p. 201–210.
- [18] PENZA, R. G.; BOULICAUT, J.-F. Constrained co-clustering of gene expression data. In: SIAM. *Proceedings of the SIAM International Conference on Data Mining*. Atlanta, Georgia, USA, 2008. p. 25–36.
- [19] DEODHAR, M. et al. A scalable framework for discovering coherent co-clusters in noisy data. In: ACM. *Proceedings of the 26th Annual International Conference on Machine Learning*. Montreal, Quebec, Canada, 2009. p. 241–248.
- [20] LEE, M. et al. Biclustering via sparse singular value decomposition. *Biometrics*, Wiley Online Library, v. 66, n. 4, p. 1087–1095, 2010.
- [21] DEODHAR, M.; GHOSH, J. SCOAL: A framework for simultaneous co-clustering and learning from complex data. *ACM Transactions on Knowledge Discovery from Data*, ACM, v. 4, n. 3, p. 11, 2010.
- [22] CHARRAD, M.; AHMED, M. B. Simultaneous clustering: A survey. In: SPRINGER. *Proceedings of the International Conference on Pattern Recognition and Machine Intelligence*. Moscow, Russia, 2011. p. 370–375.
- [23] ORZECOWSKI, P.; BORYCZKO, K. Hybrid biclustering algorithms for data mining. In: SPRINGER. *Proceedings of the European Conference on the Applications of Evolutionary Computation*. Porto, Portugal, 2016. p. 156–168.
- [24] MIRKIN, B. Mathematical classification and clustering: From how to what and why. In: BALDERJAHN, I.; MATHAR, R.; SCHADER, M. (Ed.). *Proceedings of the Classification, Data Analysis, and Data Highways*. Berlin, Heidelberg: Springer Berlin Heidelberg, 1998. p. 172–181.
- [25] PAPADIMITRIOU, S.; SUN, J. Disco: Distributed co-clustering with map-reduce: A case study towards petabyte-scale end-to-end mining. In: IEEE. *Proceedings of the 2008 8th IEEE International Conference on Data Mining*. Las Vegas, Nevada, USA, 2008. p. 512–521.
- [26] PENZA, R. G. et al. Co-clustering numerical data under user-defined constraints. *Statistical Analysis and Data Mining*, Wiley Online Library, v. 3, n. 1, p. 38–55, 2010.
- [27] MUKHOPADHYAY, A.; MAULIK, U.; BANDYOPADHYAY, S. On biclustering of gene expression data. *Current Bioinformatics*, Bentham Science Publishers, v. 5, n. 3, p. 204–216, 2010.
- [28] HARTIGAN, J. A. Direct clustering of a data matrix. *Journal of the American Statistical Association*, Taylor & Francis Group, v. 67, n. 337, p. 123–129, 1972.
- [29] PONTES, B.; GIRÁLDEZ, R.; AGUILAR-RUIZ, J. S. Biclustering on expression data: A review. *Journal of Biomedical Informatics*, Elsevier, v. 57, p. 163–180, 2015.
- [30] LEE, D. D.; SEUNG, H. S. Learning the parts of objects by non-negative matrix factorization. *Nature*, Nature Publishing Group, v. 401, n. 6755, p. 788–791, 1999.
- [31] HOFMANN, T.; PUZICHA, J.; JORDAN, M. I. Learning from dyadic data. In: NIPS. *Proceedings of the Advances in Neural Information Processing Systems*. Denver, Colorado, USA, 1999. p. 466–472.

- [32] DHILLON, I. S.; MALLELA, S.; MODHA, D. S. Information-theoretic co-clustering. In: *ACM. Proceedings of the ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. [S.l.], 2003. p. 89–98.
- [33] YOO, J.; CHOI, S. Orthogonal nonnegative matrix tri-factorization for co-clustering: Multiplicative updates on stiefel manifolds. *Information Processing & Management*, Elsevier, v. 46, n. 5, p. 559–570, 2010.
- [34] BUONO, N. D.; PIO, G. Non-negative matrix tri-factorization for co-clustering: An analysis of the block matrix. *Information Sciences*, Elsevier, v. 301, p. 13–26, 2015.
- [35] IENCO, D. et al. Parameter-less co-clustering for star-structured heterogeneous data. *Data Mining and Knowledge Discovery*, Springer, v. 26, n. 2, p. 217–254, 2013.
- [36] HOFMANN, T. Probabilistic latent semantic analysis. In: MORGAN KAUFMANN. *Proceedings of the 15th Conference on Uncertainty in Artificial Intelligence*. Stockholm, Sweden, 1999. p. 289–296.
- [37] HALKIDI, M.; BATISTAKIS, Y.; VAZIRGIANNIS, M. Cluster validity methods: part I. *ACM Sigmod Record*, ACM, v. 31, n. 2, p. 40–45, 2002.
- [38] HALKIDI, M.; BATISTAKIS, Y.; VAZIRGIANNIS, M. Clustering validity checking methods: part II. *ACM Sigmod Record*, ACM, v. 31, n. 3, p. 19–27, 2002.
- [39] WANG, H. et al. Fast nonnegative matrix tri-factorization for large-scale data co-clustering. In: IJCAI/AAAI. *Proceedings of the International Joint Conference on Artificial Intelligence*. Barcelona, Catalonia, Spain, 2011. v. 22, n. 1, p. 1553.
- [40] BRUNIALTI, L. F. et al. The BinOvNMTF algorithm: Overlapping columns co-clustering based on non-negative matrix tri-factorization. In: SBC. *Proceedings of the 2017 6th Brazilian Conference on Intelligent Systems*. Uberlândia, MG, Brazil, 2017. p. 330–335.
- [41] CASTRO, P. A. de et al. Query expansion using an immune-inspired biclustering algorithm. *Natural Computing*, Springer, v. 9, n. 3, p. 579–602, 2010.
- [42] HALKIDI, M.; BATISTAKIS, Y.; VAZIRGIANNIS, M. On clustering validation techniques. *Journal of Intelligent Information Systems*, Springer, v. 17, n. 2-3, p. 107–145, 2001.
- [43] DESGRAUPES, B. Clustering indices. *University of Paris Ouest-Lab ModalX*, v. 1, p. 34, 2013.
- [44] SOLER, J. et al. Data clustering and similarity. In: AAAI. *Proceedings of the 26th International Florida Artificial Intelligence Research Society Conference*. St. Pete Beach, Florida, USA.
- [45] DAVIES, D. L.; BOULDIN, D. W. A cluster separation measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, IEEE, n. 2, p. 224–227, 1979.
- [46] FRANCA, F. O. de. *Biclusterização na Análise de Dados Incertos*. Tese (Doutorado) — Universidade Estadual de Campinas, 2010.
- [47] TANAY, A.; SHARAN, R.; SHAMIR, R. Biclustering algorithms: A survey. *Handbook of Computational Molecular Biology*, Citeseer, v. 9, n. 1-20, p. 122–124, 2005.
- [48] BRUNIALTI, L. F. *Fatoração de matrizes no problema de coagrupamento com sobreposição de colunas*. Dissertação (Mestrado) — Escola de Artes, Ciências e Humanidades da Universidade de São Paulo, 2016.
- [49] DING, C. et al. Orthogonal nonnegative matrix t-factorizations for clustering. In: *ACM. Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. Philadelphia, PA, USA, 2006. p. 126–135.
- [50] DIAZ, A. K. R. et al. *Uma análise comparativa das ferramentas de pré-processamento de dados textuais: NLTK, PreText e R*. Relatório Técnico: PPgSI-001/2018. Universidade de São Paulo. São Paulo, 2018.
- [51] DIAZ, A. K. R. *Biagrupamento heurístico e coagrupamento baseado em fatoração de matrizes: um estudo em dados textuais*. Dissertação (Mestrado) — Universidade de São Paulo, São Paulo, SP, Brasil, 2018.
- [52] PAGNOSSIM, J. L. M. *Uma abordagem híbrida para sistemas de recomendação de notícias*. Dissertação (Mestrado) — Escola de Artes, Ciências e Humanidades da Universidade de São Paulo, 2018.
- [53] BAEZA-YATES, R.; RIBEIRO-NETO, B. *Modern Information Retrieval: The Concepts and Technology behind Search (ACM Press Books)*. USA: Addison-Wesley Professional, 2011.
- [54] MARTIN-FERNANDEZ, J. A.; PALAREA-ALBALADEJO, J.; OLEA, R. A. Dealing with zeros. *Compositional data analysis: Theory and applications*, John Wiley & Sons, Ltd Hoboken, NJ, p. 43–58, 2011.
- [55] NOURASHRAFEDDIN, S.; MILIOS, E.; ARNOLD, D. V. An evolutionary algorithm for feature selective double clustering of text documents. In: IEEE. *Proceedings of the 2013 IEEE Congress on Evolutionary Computation (CEC)*. Cancun, Mexico, 2013. p. 446–453.
- [56] FRANCA, F. O. D. Scalable overlapping co-clustering of word-document data. In: IEEE. *Proceedings of the 2012 11th International Conference on Machine Learning and Applications*. Boca Raton, FL, USA, 2012. v. 1, p. 464–467.
- [57] YAN, Y.; CHEN, L.; TJHI, W.-C. Fuzzy semi-supervised co-clustering for text documents. *Fuzzy Sets and Systems*, Elsevier, v. 215, p. 74–89, 2013.
- [58] WANG, S.; HUANG, A. Penalized nonnegative matrix tri-factorization for co-clustering. *Expert Systems with Applications*, Elsevier, v. 78, p. 64–73, 2017.

- [59] ALLAB, K.; LABIOD, L.; NADIF, M. SemiNMF-PCA framework for sparse data co-clustering. In: ACM. *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*. Indianapolis, IN, USA, 2016. p. 347–356.
- [60] LI, Z.; WU, X. Single multiplicatively updated matrix factorization for co-clustering. In: ACM. *Proceedings of the 29th Annual ACM Symposium on Applied Computing*. Gyeongju, Korea, 2014. p. 97–104.
- [61] MA, H.; ZHAO, W.; SHI, Z. A nonnegative matrix factorization framework for semi-supervised document clustering with dual constraints. *Knowledge and information systems*, Springer, v. 36, n. 3, p. 629–651, 2013.
- [62] CASTRO, P. A. de et al. Applying biclustering to text mining: an immune-inspired approach. In: *Proceedings of the Artificial Immune Systems*. Santos, SP, Brazil: Springer, 2007. p. 83–94.
- [63] PRIAM, R.; NADIF, M.; GOVAERT, G. Generalized topographic block model. *Neurocomputing*, Elsevier, v. 173, p. 442–449, 2016.
- [64] MIMAROGLU, S.; UEHARA, K. Bit sequences and biclustering of text documents. In: IEEE. *Proceedings of the 7th IEEE International Conference on Data Mining Workshops*. Omaha, Nebraska, USA, 2007. p. 51–56.
- [65] GU, Q.; DING, C.; HAN, J. On trivial solution and scale transfer problems in graph regularized nmf. In: IJCAI. *Proceedings of the International Joint Conference on Artificial Intelligence*. Barcelona, Catalonia, Spain, 2011. v. 22, n. 1, p. 1288.
- [66] CAO, Y.; HUANG, M.; ZHU, X. Clustering sentiment phrases in product reviews by constrained co-clustering. In: SPRINGER. *Proceedings of the National CCF Conference on Natural Language Processing and Chinese Computing*. Nanchang, China, 2015. p. 79–89.
- [67] GARG, V. K.; CHAUDHARI, S.; NARANG, A. Multi-regularization for fuzzy co-clustering. In: SPRINGER. *Proceedings of the International Conference on Neural Information Processing*. Daegu, Korea, 2013. p. 67–75.
- [68] RHO, V.; PENSA, R. G. Concept-enhanced multi-view co-clustering of document data. In: SPRINGER. *Proceedings of the International Symposium on Methodologies for Intelligent Systems*. Warsaw, Poland, 2017. p. 457–467.
- [69] HUANG, F. et al. Semi-supervised hierarchical co-clustering. In: SPRINGER. *Proceedings of the International Conference on Rough Sets and Knowledge Technology*. Chengdu, China, 2012. p. 310–319.