UNIVERSIDADE DE SÃO PAULO
ESCOLA DE ARTES, CIÊNCIAS E HUMANIDADES
PROGRAMA DE PÓS-GRADUAÇÃO EM SISTEMAS DE INFORMAÇÃO

ALEXANDRA KATIUSKA RAMOS DIAZ

**Biagrupamento heurístico e coagrupamento baseado em fatoração de
matrizes: um estudo em dados textuais**

São Paulo

2018

ALEXANDRA KATIUSKA RAMOS DIAZ

**Biagrupamento heurístico e coagrupamento baseado em fatoração de
matrizes: um estudo em dados textuais**

Dissertação apresentada à Escola de Artes,
Ciências e Humanidades da Universidade de
São Paulo para obtenção do título de Mestre
em Ciências pelo Programa de Pós-graduação
em Sistemas de Informação.

Área de concentração: Metodologia e
Técnicas da Computação

Versão corrigida contendo as alterações
solicitadas pela comissão julgadora em 16 de
outubro de 2018. A versão original encontra-
se em acervo reservado na Biblioteca da
EACH-USP e na Biblioteca Digital de Teses
e Dissertações da USP (BDTD), de acordo
com a Resolução CoPGr 6018, de 13 de
outubro de 2011.

Orientador: Profa. Dra. Sarajane Marques
Peres

São Paulo

2018

Autorizo a reprodução e divulgação total ou parcial deste trabalho, por qualquer meio convencional ou eletrônico, para fins de estudo e pesquisa, desde que citada a fonte.

CATALOGAÇÃO-NA-PUBLICAÇÃO

(Universidade de São Paulo. Escola de Artes, Ciências e Humanidades. Biblioteca)
CRB-8 4936

Ramos Diaz, Alexandra Katiuska

Biagrupamento heurístico e coagrupamento baseado em
fatoração de matrizes: um estudo em dados textuais / Alexandra
Katiuska Ramos Diaz ; orientadora, Sarajane Marques Peres. –
2018.

198 p. : il.

Dissertação (Mestrado em Ciências) - Programa de Pós-
Graduação em Sistemas de Informação, Escola de Artes,
Ciências e Humanidades, Universidade de São Paulo.

Versão corrigida

1. Inteligência artificial. 2. Mineração de dados. 3.
Fatorização de matrizes. 4. Matrizes esparsas. I. Peres,
Sarajane Marques, orient. II. Título.

CDD 22.ed.– 006.3

Dissertação de autoria de Alexandra Katiuska Ramos Diaz, sob o título “**Biagrupamento heurístico e coagrupamento baseado em fatoração de matrizes: um estudo em dados textuais**”, apresentada à Escola de Artes, Ciências e Humanidades da Universidade de São Paulo, para obtenção do título de Mestre em Ciências pelo Programa de Pós-graduação em Sistemas de Informação, na área de concentração Metodologia e Técnicas da Computação, aprovada em 16 de outubro de 2018 pela comissão julgadora constituída pelos doutores:

Profa. Dra. Violeta Sun

Escola de Artes, Ciências e Humanidades da Universidade de São Paulo
Presidente

Prof. Dr. Alexandre da Silva Freire

Escola de Artes, Ciências e Humanidades da Universidade de São Paulo

Prof. Dr. Ademir Aparecido Constantino

Departamento de Informática da Universidade Estadual de Maringá

Prof. Dr. Guilherme Palermo Coelho

Faculdade de Tecnologia da Universidade Estadual de Campinas

Aos meus avós que não estão mais ao meu lado, mas seu carinho e amor prevalecem em meu coração. Aos meus pais, Ana Maria e Luis Alberto pelo carinho, apoio incondicional e ensinamentos transmitidos no decorrer da minha vida. Às minhas irmãs Lucero, Tania e Brisa pela força e perseverança transmitidas. Aos meus pitufos Dafne e Piero Juan José pela ternura e felicidade proporcionada. Eu acredito que o amor é fortalecido pela distância, e embora eu não os tenha fisicamente por perto, vocês estão sempre me apoiando. Para todos vocês, meu amor e admiração eterna.

Agradecimentos

À minha orientadora, Profa. Dra. Sarajane Marques Peres, pelos ensinamentos, apoio, paciência e dedicação oferecidos em todo este tempo. Sua contribuição foi fundamental para a realização deste trabalho. Não posso deixar de agradecer por contribuir para o meu aprendizado do português.

A meu extraordinário companheiro de viagem, Pavel. Sem ele muitas coisas não teriam sido possíveis. Obrigada por sempre me acompanhar em todas as aventuras que vivi, por me apoiar nos bons e maus momentos e por sempre me transmitir sua coragem, persistência e força.

Aos meus professores pelos ensinamentos transmitidos ao longo do curso de mestrado.

Aos companheiros e amigos que ganhei com o mestrado. Agradeço pelo companheirismo, amizade e apoio oferecidos ao longo deste tempo.

Às pessoas que estiveram ao meu lado por todo este tempo, obrigada. Vossas companhias e apoios foram vitais nesta fase da minha vida.

*“A imaginação é mais importante que o conhecimento. O conhecimento é limitado. A
imaginação circunda o mundo.”*
(Albert Einstein)

Resumo

RAMOS DIAZ, Alexandra Katiuska. **Biagrupamento heurístico e coagrupamento baseado em fatoração de matrizes**: um estudo em dados textuais. 2018. 198 f. Dissertação (Mestrado em Ciências) – Escola de Artes, Ciências e Humanidades, Universidade de São Paulo, São Paulo, 2018.

Biagrupamento e coagrupamento são tarefas de mineração de dados que permitem a extração de informação relevante sobre dados e têm sido aplicadas com sucesso em uma ampla variedade de domínios, incluindo aqueles que envolvem dados textuais – foco de interesse desta pesquisa. Nas tarefas de biagrupamento e coagrupamento, os critérios de similaridade são aplicados simultaneamente às linhas e às colunas das matrizes de dados, agrupando simultaneamente os objetos e os atributos e possibilitando a criação de bigrupos/cogrupos. Contudo suas definições variam segundo suas naturezas e objetivos, sendo que a tarefa de coagrupamento pode ser vista como uma generalização da tarefa de biagrupamento. Estas tarefas, quando aplicadas nos dados textuais, demandam uma representação em um modelo de espaço vetorial que, comumente, leva à geração de espaços caracterizados pela alta dimensionalidade e esparsidade, afetando o desempenho de muitos dos algoritmos. Este trabalho apresenta uma análise do comportamento do algoritmo para biagrupamento Cheng e Church e do algoritmo para coagrupamento de decomposição de valores em blocos não negativos (*Non-Negative Block Value Decomposition* - NBVD), aplicado ao contexto de dados textuais. Resultados experimentais quantitativos e qualitativos são apresentados a partir das experimentações destes algoritmos em conjuntos de dados sintéticos criados com diferentes níveis de esparsidade e em um conjunto de dados real. Os resultados são avaliados em termos de medidas próprias de biagrupamento, medidas internas de agrupamento a partir das projeções nas linhas dos bigrupos/cogrupos e em termos de geração de informação. As análises dos resultados esclarecem questões referentes às dificuldades encontradas por estes algoritmos nos ambiente de experimentação, assim como se são capazes de fornecer informações diferenciadas e úteis na área de mineração de texto. De forma geral, as análises realizadas mostraram que o algoritmo NBVD é mais adequado para trabalhar com conjuntos de dados em altas dimensões e com alta esparsidade. O algoritmo de Cheng e Church, embora tenha obtidos resultados bons de acordo com os objetivos do algoritmo, no contexto de dados textuais, propiciou resultados com baixa relevância.

Palavras-chaves: Biagrupamento. Coagrupamento. Mineração de textos. Matrizes de dados esparsos. Algoritmos heurísticos. Fatoração de matrizes não-negativas.

Abstract

RAMOS DIAZ, Alexandra Katiuska. **Heuristic biclustering and coclustering based on matrix factorization**: a study on textual data. 2018. 198 p. Dissertation (Master of Science) – School of Arts, Sciences and Humanities, University of São Paulo, São Paulo, 2018.

Biclustering e coclustering are data mining tasks that allow the extraction of relevant information about data and have been applied successfully in a wide variety of domains, including those involving textual data - the focus of interest of this research. In biclustering and coclustering tasks, similarity criteria are applied simultaneously to the rows and columns of the data matrices, simultaneously grouping the objects and attributes and enabling the discovery of biclusters/coclusters. However their definitions vary according to their natures and objectives, being that the task of coclustering can be seen as a generalization of the task of biclustering. These tasks applied in the textual data demand a representation in a model of vector space, which commonly leads to the generation of spaces characterized by high dimensionality and sparsity and influences the performance of many algorithms. This work provides an analysis of the behavior of the algorithm for biclustering Cheng and Church and the algorithm for coclustering non-negative block decomposition (NBVD) applied to the context of textual data. Quantitative and qualitative experimental results are shown, from experiments on synthetic datasets created with different sparsity levels and on a real data set. The results are evaluated in terms of their biclustering oriented measures, internal clustering measures applied to the projections in the lines of the biclusters/coclusters and in terms of generation of information. The analysis of the results clarifies questions related to the difficulties faced by these algorithms in the experimental environment, as well as if they are able to provide differentiated information useful to the field of text mining. In general, the analyses carried out showed that the NBVD algorithm is better suited to work with datasets in high dimensions and with high sparsity. The algorithm of Cheng and Church, although it obtained good results according to its own objectives, provided results with low relevance in the context of textual data.

Keywords: Biclustering, Coclustering. Text Mining. Matrix of sparse data. Heuristic algorithms. Non-negative matrix factorization.

Lista de figuras

Figura 1 – Exemplo de uma aplicação de mineração de dados	26
Figura 2 – Exemplo de bigrupo com valores constantes	38
Figura 3 – Exemplos de bigrupos com valores constantes nas linhas usando um ajuste aditivo (a) e nas colunas usando um ajuste multiplicativo (b) . .	39
Figura 4 – Exemplos de bigrupos com valores coerentes nas linhas e nas colunas usando um modelo aditivo (a) e usando um modelo multiplicativo (b) .	41
Figura 5 – Exemplos de bigrupos com evoluções coerentes	41
Figura 6 – Estrutura de um único bigrupo	42
Figura 7 – Estrutura de bigrupos com linhas e colunas exclusivas	43
Figura 8 – Estrutura de bigrupos com estrutura de tabuleiro de xadrez	43
Figura 9 – Estrutura de bigrupos com linhas exclusivas	43
Figura 10 – Estrutura de bigrupos com colunas exclusivas	44
Figura 11 – Estrutura de bigrupos sem sobreposição com estrutura em árvore . . .	44
Figura 12 – Estrutura de bigrupos não exclusivos e sem sobreposição	45
Figura 13 – Estrutura de bigrupos com sobreposição com estrutura hierárquica . .	45
Figura 14 – Estrutura de bigrupos com sobreposição e arbitrariamente posicionados	46
Figura 15 – Visão de coagrupamento na estrutura d , discutida na seção 2.1.1	48
Figura 16 – Exemplo de sobreposições obtidas com o algoritmo de Cheng e Church	51
Figura 17 – Fatoração da matriz original de dados X nas matrizes U , S e V	65
Figura 18 – 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)	66
Figura 19 – 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)	67
Figura 20 – Pós-processamento das matrizes U , S e V em um conjunto de dados sintéticos	69
Figura 21 – Reconstrução das matrizes U , S e V do conjunto de dados sintéticos da figura 20	70
Figura 22 – Exemplo da execução do algoritmo NBVD na matriz de dados $\mathcal{X}$. . .	71

Figura 23 – Resumo dos estudos primários por ano de publicação.	81
Figura 24 – Conjunto de dados sintéticos baseados nas diferentes estruturas de bigrupos	93
Figura 25 – Visão de coagrupamento nos conjuntos de dados sintéticos	94
Figura 26 – Exemplo do conteúdo de uma notícia do corpus EBC	96
Figura 27 – Efeito da inserção de esparsidade no conjunto de dados da estrutura i .	98
Figura 28 – Esquema gráfico de estudo aplicado ao algoritmo Cheng e Church . . .	100
Figura 29 – Esquema gráfico de estudo aplicado ao algoritmo NBVD	102
Figura 30 – Bigrupos com linhas exclusivas (estrutura d)	106
Figura 31 – Variação do tempo de execução do algoritmo de Cheng e Church em relação ao parâmetro α em cada conjunto de dados sintéticos	108
Figura 32 – Variação do tempo de execução do algoritmo de Cheng e Church em relação ao parâmetro α no conjunto de dados da estrutura d	109
Figura 33 – Variação da média do volume e da média do RQM dos bigrupos em relação do parâmetro α após a aplicação do algoritmo de Cheng e Church, para cada conjunto de dados sintéticos	110
Figura 34 – Representação dos bigrupos resultantes do algoritmo de Cheng e Church usando o conjunto de dados da estrutura d	112
Figura 35 – Variação dos tempos de execução totais em base a δ , do algoritmo de Cheng e Church, em cada conjunto de dados	113
Figura 36 – Variação do tempo de execução do algoritmo de Cheng e Church em relação ao parâmetro δ no conjunto de dados da estrutura d	114
Figura 37 – Variação da média das médias dos volumes dos bigrupos resultantes, em base a δ , após a aplicação do algoritmo de Cheng e Church, em cada conjunto de dados	115
Figura 38 – Variação da média dos RQMs dos bigrupos resultantes, em base a δ , após a aplicação do algoritmo de Cheng e Church, em cada conjunto de dados	115
Figura 39 – Variação da média do tempo de execução sobre os diferentes níveis de esparsidade em relação ao parâmetro α , após a aplicação do algoritmo de Cheng e Church nos conjuntos de dados sintéticos	118

Figura 40 – Variação do tempo de execução do algoritmo de Cheng e Church em relação ao parâmetro α , no conjunto de dados da estrutura d com diferentes níveis de esparsidade	119
Figura 41 – Variação do tempo de execução do algoritmo de Cheng e Church em relação ao parâmetro α , em cada conjunto de dados sintéticos, com níveis de esparsidade baixa e alta	120
Figura 42 – Variação da média dos tempos de execução, das médias dos RQM e das médias dos volumes sobre os diferentes níveis de esparsidade em relação ao parâmetro δ , após a aplicação do algoritmo de Cheng e Church nos conjuntos de dados sintéticos	122
Figura 43 – Variação dos tempos de execução, dos RQM e dos volumes do algoritmo de Cheng e Church em relação ao parâmetro δ , em cada conjunto de dados sintéticos, com níveis de esparsidade baixa e alta	123
Figura 44 – Variação do tempo de execução, dos volumes, dos RQMs dos bigrupos, a porcentagem da cobertura e a porcentagem do grau de sobreposição em relação ao número de bigrupos (K) após a aplicação do algoritmo de Cheng e Church nos conjuntos de dados das estruturas a , b , e i	125
Figura 45 – Variação das medidas internas de agrupamento em relação ao incremento dos níveis de esparsidade, em cada conjunto de dados sintéticos após a aplicação do algoritmo de Cheng e Church	127
Figura 46 – Variação da média dos volumes e dos RQMs em relação aos incremento dos níveis de esparsidade, em cada conjunto de dados sintéticos, após a aplicação do algoritmo de Cheng e Church	128
Figura 47 – Variação da porcentagem de cobertura, do grau de sobreposição e 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 de Cheng e Church	129
Figura 48 – Mapa de calor da porcentagem do grau de ocupação dos resultados obtidos após a aplicação do algoritmo de Cheng e Church nas matrizes obtidas do corpus EBC	133
Figura 49 – Mapa de calor da porcentagem do grau de sobreposição dos resultados obtidos após a aplicação do algoritmo de Cheng e Church nas matrizes obtidas do corpus EBC	134

Figura 50 – Mapa de calor do índice DB dos resultados obtidos após a aplicação do algoritmo de Cheng e Church nas matrizes obtidas do corpus EBC . . .	135
Figura 51 – Diagrama de caixa do volume e da porcentagem de cobertura dos resultados obtidos após a aplicação do algoritmo de Cheng e Church nas matrizes obtidas do corpus EBC	135
Figura 52 – Diagrama de caixa dos RQMs, distância intra-grupos e distância inter-grupos dos resultados obtidos após a aplicação do algoritmo de Cheng e Church nas matrizes obtidas do corpus EBC	136
Figura 53 – Visualização em nuvem de palavras das top-30 palavras para cada bigrupo gerado pelo algoritmo de Cheng e Church, com $K = 3$	139
Figura 54 – Comparação dos grupos de notícias com os canais conhecidos, após a aplicação do algoritmo de Cheng e Church com $K = 3$	140
Figura 55 – Visualização em nuvem de palavras das top-30 palavras da sobreposição de notícias pertencentes ao GN# 1 e GN# 3	141
Figura 56 – Conteúdo da notícia 11022, pertencente ao GN# 1 e ao GN# 3, após a aplicação do algoritmo de Cheng e Church	142
Figura 57 – Ocorrência da palavra “jogo”, “vasco”, “januário” nas notícias 11034 e 11509	143
Figura 58 – Exemplo de sobreposição de bigrupos na notícia 4425, após a aplicação do algoritmo de Cheng e Church	144
Figura 59 – Visualização em nuvem de palavras das top-30 palavras para cada bigrupo gerados pelo algoritmo de Cheng e Church, com $K = 7$	145
Figura 60 – Representação do número máximo de grupos de linhas	148
Figura 61 – Variação do erro de reconstrução do algoritmo de NBVD em relação aos parâmetros k e l nos conjuntos de dados das estruturas d e i	150
Figura 62 – Variação das medidas internas de agrupamento em relação ao incremento dos níveis de esparsidade, em cada conjunto de dados sintéticos após a aplicação do algoritmo de NBVD	153
Figura 63 – Variação da média dos volumes e dos RQMs em relação aos incremento dos níveis de esparsidade, em cada conjunto de dados sintéticos, após a aplicação do algoritmo de NBVD	154

Figura 64 – Variação da porcentagem de cobertura, do grau de sobreposição e do grau de ocupação dos bigrupos em relação aos incremento dos níveis de esparsidade, em cada conjunto de dados sintéticos após a aplicação do algoritmo de NBVD	155
Figura 65 – Variação do erro de reconstrução em relação aos incremento dos níveis de esparsidade, em cada conjunto de dados sintéticos, após a aplicação do algoritmo de NBVD	156
Figura 66 – Reconstrução visual dos resultados do algoritmo de NBVD no conjunto de dados da estrutura d com diferentes níveis de esparsidade	157
Figura 67 – Mapa de calor do erro de reconstrução dos resultados obtidos a partir da aplicação do algoritmo de NBVD nas matrizes obtidas do corpus EBC159	
Figura 68 – Mapa de calor da porcentagem do grau de ocupação dos resultados obtidos a partir da aplicação do algoritmo de NBVD nas matrizes obtidas do corpus EBC	160
Figura 69 – Mapa de calor da porcentagem do grau de sobreposição dos resultados obtidos a partir da aplicação do algoritmo de NBVD nas matrizes obtidas do corpus EBC	160
Figura 70 – Mapa de calor do índice DB dos resultados obtidos a partir da aplicação do algoritmo de NBVD nas matrizes obtidas do corpus EBC	161
Figura 71 – Diagrama de caixa do volume e da porcentagem de cobertura dos resultados obtidos após a aplicação do algoritmo de NBVD nas matrizes obtidas do corpus EBC	162
Figura 72 – Diagrama de caixa do RQM, distância inter-grupos e distância intra-grupos dos resultados obtidos após a aplicação do algoritmo de NBVD nas matrizes obtidas do corpus EBC	162
Figura 73 – Visualização em nuvem de palavras das top-30 palavras para cada cogrupo gerado pelo algoritmo NBVD	165
Figura 74 – Comparação dos grupos de notícias com os canais conhecidos	166
Figura 75 – Exemplo gráfico de sobreposição de linhas no corpus EBC	166
Figura 76 – Conteúdo da notícia 10999, pertencente ao GN# 2 e ao GN# 3	167
Figura 77 – Exemplo gráfico de sobreposição de colunas no corpus EBC	168
Figura 78 – Ocorrência da palavra “social” nas notícias 6008 e 2425	169

Figura 79 – Visualização em nuvem de palavras das top-30 palavras para cada cogrupo gerado pelo algoritmo NBVD, com $k = 7$ e $l = 7$	170
Figura 80 – Comparação dos grupos de notícias com os canais conhecidos, com $k = 7, l = 7$	172
Figura 81 – Visualização em nuvem de palavras das top-30 palavras para cada cogrupo gerado pelo algoritmo NBVD, com $k = 10$ e $l = 10$	173
Figura 82 – Comparação dos grupos de notícias com os canais conhecidos, com $k = 10$ e $l = 10$	174
Figura 83 – Exemplo de sobreposição de cogrupos na notícia 10992, após a aplicação do algoritmo de NBVD	176
Figura 84 – Estrutura de bigrupos de formato arbitrários: equivalente à estrutura de cogrupos encontrada no corpus EBC	177

Lista de algoritmos

Algoritmo 1 – Algoritmo de Cheng e Church	55
Algoritmo 2 – Remoção de múltiplas linhas e colunas simultaneamente	56
Algoritmo 3 – Remoção de uma linha ou uma coluna por vez	58
Algoritmo 4 – Adição de linhas e colunas	58
Algoritmo 5 – Algoritmo de decomposição de valores em blocos não negativos (NBVD) .	68

Lista de tabelas

Tabela 1 – Cálculo do RQM depois de remover a linha ou a coluna da matriz de dados $\mathcal{A}$	57
Tabela 2 – Medidas do biagrupamento	74
Tabela 3 – Lista de estudos selecionados	80
Tabela 4 – Resumo dos estudos primários por tipo de publicação e por ano de publicação.	81
Tabela 5 – Classificação dos estudos de acordo com os conjuntos de dados neles utilizados	83
Tabela 6 – Classificação dos estudos de acordo com a modalidade de aprendizado neles utilizada	84
Tabela 7 – Classificação dos estudos de acordo com sua natureza	84
Tabela 8 – Classificação dos estudos de acordo com as medidas de avaliação neles utilizadas	87
Tabela 9 – Classificação dos estudos de acordo com as tarefas utilizadas em mineração de texto	88
Tabela 10 – Classificação dos estudos de acordo com suas aplicações	88
Tabela 11 – Organização do corpus EBC	95
Tabela 12 – Conjunto de dados reais	97
Tabela 13 – Níveis de esparsidade das matrizes de dados extraídas do corpus EBC .	98
Tabela 14 – Espaço de busca para o parâmetro δ considerando o conjunto de dados com a estrutura d	106
Tabela 15 – Espaço de busca para o parâmetro δ	107
Tabela 16 – Espaço de busca para o parâmetro K	107
Tabela 17 – Avaliação quantitativa dos biagrupamentos: ponto de vista do parâmetro α	111
Tabela 18 – Valores para o parâmetro α que levam aos melhores resultados de avaliação quantitativa comparados aos resultados desejados, para cada estrutura de bigrupos.	112
Tabela 19 – Avaliação quantitativa dos biagrupamentos: ponto de vista do parâmetro δ	116

Tabela 20 – Valores para o parâmetro δ que levam aos melhores resultados de avaliação quantitativa comparativa com os resultados desejados, para cada estrutura de bigrupos	117
Tabela 21 – RQMs estimados de 1000 submatrizes geradas aleatoriamente para cada combinação de tamanho	131
Tabela 22 – Medidas de biagrupamento com diferentes valores do parâmetro δ e K	132
Tabela 23 – Médias das medidas por representação, obtidas depois da aplicação do algoritmo de Cheng e Church nas matrizes obtidas do corpus EBC . .	138
Tabela 24 – Melhores resultados obtidas após a aplicação do algoritmo de Cheng e Church na Mat. 12 do corpus EBC	138
Tabela 25 – Espaço de busca para os parâmetro k e l	149
Tabela 26 – Combinação de valores dos parâmetros de k e l que levam aos melhores resultados de erro de reconstrução para cada conjunto de dados sintético	150
Tabela 27 – Representação visual da reconstrução dos conjunto de dados sintéticos, após a aplicação do algoritmo de NBVD	151
Tabela 28 – Valores dos parâmetros de k e l que obtiveram bons resultados do algoritmo de NBVD nos conjuntos de dados sintéticos	152
Tabela 29 – Médias das medidas por representação, obtidas depois da aplicação do algoritmo de NBVD nas matrizes obtidas do corpus EBC	163
Tabela 30 – Melhores resultados obtidas depois da aplicação do algoritmo de NBVD nas matrizes obtidas do corpus EBC	164
Tabela 31 – Matriz S gerada por após a aplicação do algoritmo de NBVD no corpus com $k = 3, l = 3$	164
Tabela 32 – Top-30 palavras para cada grupo de palavras	168
Tabela 33 – Matriz S gerada após a aplicação do algoritmo de NBVD no corpus com $k = 7, l = 7$	169
Tabela 34 – Matriz S gerada por após a aplicação do algoritmo de NBVD no corpus com $k = 10, l = 10$	172
Tabela 35 – <i>String</i> de busca	196
Tabela 36 – Formulário de extração da informação dos estudos	197

Lista de abreviaturas e siglas

BVD	<i>Block Value Decomposition</i>
CE	Critério de Exclusão
CI	Critério de Inclusão
DB	Índice Davies Bouldin
EBC	Empresa Brasil de Comunicação
NBVD	<i>Non-negative Block Value Decomposition</i>
NMF	<i>Non-negative Matrix Factorization</i>
RS	Revisão Sistemática
RSS	<i>Residual sum of squares</i>
RQM	Resíduo Quadrático médio
TF	<i>Term Frequency</i>
TF-IDF	<i>Term Frequency - Inverse Document Frequency</i>

Lista de símbolos

$\mathcal{A}, \mathcal{X}$	matriz de dados/objetos
a, b	elemento de uma matriz
α	taxa ≥ 0
$\beta, \gamma, \eta, \mu, \lambda, \zeta$	valores reais
$\mathbf{A}, \mathbf{B}, \mathbf{C}, \mathbf{D}$	submatriz de dados/objetos - um bigrupo
$\mathcal{C}^{k \times l}$	coagrupamento (partição) com $k \times l$ cogrupos
$\mathcal{C}^r$	partição de linhas
$\mathcal{C}^c$	partição de colunas
δ	limiar real e positivo
I	subconjunto de linhas
J	subconjunto de colunas
i, j, t	contador/índice
$\mathbf{i}, \mathbf{x}$	uma linha de uma matriz
$\mathbf{j}, \mathbf{y}$	uma coluna de uma matriz
K	número de bigrupos/cogrupos
k	número de grupos de linhas
l	número de grupos de colunas
G	grupo de linhas
n, p	número de linhas em uma matriz
m, s	número de colunas em uma matriz
T	número total de iterações
$\mathbb{R}_+$	conjunto dos números reais não negativos

$\mathbb{R}_{+*}$	conjunto dos números reais positivos e maiores que 0
$\overrightarrow{x_i}$	vetor linha
X	conjunto de linhas
$\overrightarrow{y_j}$	vetor coluna
Y	conjunto de colunas
$ \cdot $	quantidade de elementos no conjunto $\cdot$
$(\mathbf{x}_i, \mathbf{y}_j)$	díade composta pelo elemento $\mathbf{x}_i$ e o elemento $\mathbf{y}_j$
$w(\mathbf{x}_i, \mathbf{y}_j)$	valor escalar de uma díade composta pelo elemento $\mathbf{x}_i$ e o elemento $\mathbf{y}_j$
O	conjunto de centros na criação de cogrupos
$\mathcal{U}(0, 1)$	distribuição uniforme de zero a um
s	melhores pontuações em termos de TF-IDF

Sumário

1	Introdução	24
1.1	<i>Apresentação do problema</i>	27
1.2	<i>Questão de Pesquisa</i>	28
1.3	<i>Objetivos</i>	29
1.4	<i>Método</i>	29
1.5	<i>Organização do documento</i>	32
2	Conceitos fundamentais	33
2.1	<i>Biagrupamento e coagrupamento</i>	33
2.1.1	Biagrupamento	36
2.1.2	Coagrupamento	46
2.2	<i>Dois algoritmos: um para biagrupamento e um para coagrupamento . .</i>	49
2.2.1	Algoritmo de Cheng e Church	50
2.2.2	Algoritmo de decomposição de valores em blocos não negativos (NBVD)	63
2.3	<i>Avaliação em biagrupamento e coagrupamento</i>	72
2.4	<i>Análise de textos</i>	77
2.5	<i>Estado da arte de biagrupamento e coagrupamento em dados textuais</i>	79
2.5.1	Uma análise geral da área em relação ao problema de biagrupamento e coagrupamento	79
2.5.2	Análise dos algoritmos de biagrupamento e coagrupamento	82
2.6	<i>Considerações finais</i>	90
3	Método para experimentação	92
3.1	<i>Conjuntos de dados sintéticos</i>	92
3.2	<i>Conjunto de dados reais - corpus EBC</i>	94
3.2.1	Descrição do conjunto de dados	95
3.2.2	Pré-processamento	95
3.3	<i>Estratégia de estudo geral</i>	97
3.4	<i>Método para experimentação: Cheng e Church</i>	99
3.5	<i>Método para experimentação: NBVD</i>	101

3.6	<i>Configurações para experimentação</i>	103
3.7	<i>Considerações finais</i>	104
4	Resultados e Análises do algoritmo de Cheng e Church	105
4.1	<i>Estudo do algoritmo de Cheng e Church</i>	105
4.1.1	Espaços de busca para os parâmetros do algoritmo	105
4.1.2	<i>Exp. #1</i> : estudo do parâmetro α	107
4.1.3	<i>Exp. #1</i> : estudo do parâmetro δ	113
4.1.4	<i>Exp. #2</i> : estudo do parâmetro α em conjunto de dados com níveis de esparsidade	117
4.1.5	<i>Exp. #2</i> : estudo do parâmetro δ em conjunto de dados com níveis de esparsidade	121
4.1.6	<i>Exp. #3</i> : estudo do comportamento do algoritmo frente à variação do parâmetro K	124
4.1.7	<i>Exp. #4</i> : estudo do comportamento do algoritmo em conjuntos de dados com níveis de esparsidade	126
4.2	<i>Algoritmo de Cheng e Church aplicado no corpus EBC</i>	130
4.2.1	<i>Exp. #5</i> : determinação dos parâmetros α , δ e K	130
4.2.2	<i>Exp. #6</i> : Avaliações quantitativas	132
4.2.3	<i>Exp. #6</i> : Avaliações qualitativas	139
4.3	<i>Considerações finais</i>	145
5	Resultados e Análises do algoritmo de NBVD	147
5.1	<i>Estudo do algoritmo de NBVD</i>	147
5.1.1	Espaço de busca para os parâmetros do algoritmo	147
5.1.2	<i>Exp. #1</i> : estudo dos parâmetros k e l	149
5.1.3	<i>Exp. #2</i> : estudo do comportamento do algoritmo nos conjuntos de dados com níveis de esparsidade	152
5.2	<i>Algoritmo de NBVD aplicado no corpus EBC</i>	158
5.2.1	<i>Exp. #3</i> : Avaliações quantitativas	158
5.2.2	<i>Exp. #3</i> : Avaliações qualitativas	164
5.3	<i>Considerações finais</i>	176
6	Conclusão	178

6.1	<i>Limitações do trabalho</i>	180
6.2	<i>Trabalhos futuros</i>	181
	Referências¹	182
	Apêndice A – Protocolo da revisão Sistemática	194
A.1	<i>Justificativa para a RS</i>	194
A.2	<i>Questões de pesquisa</i>	194
A.3	<i>Escolha de fontes de dados e estratégias de busca</i>	195
A.4	<i>Definição de estratégias para a seleção de estudos primários e avaliação de sua qualidade</i>	195
A.5	<i>Método de extração de dados dos estudos selecionados</i>	197
A.6	<i>Condução da RS</i>	197

¹ De acordo com a Associação Brasileira de Normas Técnicas. NBR 6023.

1 Introdução

Descobrir conhecimento significa receber ou identificar informações relevantes, processá-las e agregá-las ao conhecimento prévio de um usuário, mudando o estado de seu conhecimento atual, a fim de que um determinado problema, ou situação, possa ser resolvido (SILVA; PERES; BOSCARIOLI, 2016). Neste sentido, observa-se que o processo de descoberta de conhecimento está fortemente relacionado à forma pela qual a informação é processada. Atualmente, com a disseminação das funcionalidades de produção e propagação de conteúdo online, o volume de informação disponível na forma textual é muito grande e configura-se como um ambiente propício à descoberta de conhecimento de forma automática.

Para automatizar o processo de descoberta de conhecimento no contexto de dados textuais, é necessário que os dados estejam representados de maneira apropriada para sua manipulação. A representação mais frequentemente usada para documentos textuais é baseada no modelo de espaço vetorial (FELDMAN; SANGER, 2007; TURNEY; PANTEL, 2010; AMINE; ELBERRICHI; SIMONET, 2010; MARCACINI; REZENDE, 2010; MI; GAO; TAN, 2015). Nesse modelo, cada documento é representado por um vetor, e cada coordenada desse vetor corresponde a uma dimensão descritiva de uma coleção de documentos (SALTON; WONG; YANG, 1975). A abordagem mais utilizada na área de mineração de dados para processamento de textos, baseada no modelo de espaço vetorial, é a representação *bag-of-words*. Nessa abordagem, cada palavra encontrada na coleção de documentos pode se tornar uma dimensão no espaço vetorial e cada vetor que representa um documento terá um componente para cada palavra (FELDMAN; SANGER, 2007). No entanto, essa abordagem é caracterizada por gerar representações de alta dimensionalidade e esparsidade (grande quantidade relativa de elementos nulos ou ausentes), gerando espaços complexos que dificultam a extração da informação.

A tarefa de mineração de dados chamada “agrupamento” (do inglês *clustering*) auxilia o processo de descoberta de conhecimento, facilitando a identificação de padrões (características comuns dos elementos) que descrevem grupos similares de dados (FELDMAN; SANGER, 2007; SILVA; PERES; BOSCARIOLI, 2016). As técnicas de agrupamento operam sobre matrizes de dados (no caso de dados textuais, considerando a representação *bag-of-words*), e têm como objetivo particionar dados em subconjuntos, seguindo algum

critério de similaridade entre os dados. Os subconjuntos a serem formados a partir dessas matrizes de dados contêm objetos descritos por atributos que possuem valores semelhantes entre si e diferentes dos valores dos objetos de outros subconjuntos de dados.

Contudo, as diferentes técnicas de agrupamento (BERKHIN, 2006) apresentam algumas dificuldades quando os dados sob análise apresentam alta dimensionalidade e são esparsos, falhando em encontrar uma partição ótima, devido ao problema conhecido como a maldição da dimensionalidade (*curse of dimensionality*) (PENSA; IENCO; MEO, 2014), e impactando nas análises desejadas a partir dos resultados do algoritmo.

Alternativas à tarefa de agrupamento são aquelas conhecidas como biagrupamento (do inglês *biclustering*) e coagrupamento (do inglês *co-clustering*), as quais são capazes de extrair informações diferenciadas da matriz de dados, quando comparada à informação extraída com o 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 atributos¹ (MADEIRA; OLIVEIRA, 2004).

Um exemplo ilustrativo no contexto de tarefa de mineração de dados textuais é mostrado nos gráficos da 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-palavras são denotadas por valores binários, denotando se a palavra ocorre (1) ou não (0) na notícia. A resolução da tarefa de agrupamento, biagrupamento ou coagrupamento considerando tais dados teria o objetivo de encontrar notícias similares. Um exemplo de aplicação na qual os grupos, bigrupos ou cogrupos de notícias poderiam ser usados é sistemas de recomendação, no qual notícias pertencentes aos grupos, bigrupos e cogrupos poderiam compor a base para a formulação de estratégias de recomendação baseadas em conteúdo.

De acordo com o exemplo, as análises feitas a partir da tarefa de agrupamento resultam em dois grupos de notícias referentes a tópicos relacionados a “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 com a resolução das tarefas de biagrupamento e coagrupamento trazem

¹ Na 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*, e *two-way clustering* (ANAGNOSTOPOULOS; DASGUPTA; KUMAR, 2008).

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

Fonte: Alexandra Katiuska Ramos Diaz, 2018

informações diferenciadas. Por exemplo, em um contexto real poderiam existir notícias que tratam de temas relacionados tanto à “educação” quanto à “política”, tais como notícias sobre reformas educacionais, benefícios outorgados pelo governo aos estudantes ou professores, etc. Em um caso hipotético, essas notícias poderiam formar um bigrupo/cogrupo com tópicos relacionados a “educação e política”, como está graficamente ilustrado nas figuras 1b e 1c.

Ainda considerando o 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 de ausências de palavras nas notícias². A análise de ambos os resultados depende do contexto de aplicação.

Embora o biagrupamento e o coagrupamento tragam uma maior flexibilidade para a análise de dados, estas tarefas ainda apresentam dificuldades em relação ao tratamento da 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 textuais. Tal complexidade é estudada no âmbito dos trabalhos que propõem otimizações para ambas tarefas, mas esses trabalhos não apresentam estudos que envolvam técnicas de diferentes naturezas e se limitam a realizar explorações comparativas de seus algoritmos com outros da mesma família de algoritmos.

² Conceitos que clarificam essa contagem de cogrupos são apresentados no capítulo 2

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 (NADIF et al., 2013). 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 heurísticos (FRANÇA, 2016; HONDA et al., 2016), baseados em fatoração de matrizes (XUAN et al., 2018; HUANG; XU; LV, 2018), probabilísticos (AILEM; ROLE; NADIF, 2017; SALAH; ROGOVSCHI; NADIF, 2016) e baseados na teoria de conjuntos *fuzzy* (MEI et al., 2017, 2017).

Até o presente momento, não foi encontrado no âmbito deste trabalho, um estudo que tivesse como objetivo específico fornecer uma análise detalhada do comportamento dos algoritmos básicos de biagrupamento e de coagrupamento e como eles lidam com o problema da esparsidade e de alta dimensionalidade. Assim, a questão de pesquisa que delimita este trabalho está relacionada a exploração do comportamento de algoritmos de biagrupamento e coagrupamento quando eles são aplicados a dados textuais, que possuem uma representação esparsa com alta dimensionalidade

1.1 Apresentação do problema

Em mineração de textos, é importante trabalhar para melhoria dos resultados de análises subjetivas que trazem significado interessante para o contexto do qual os dados derivam. Uma das possibilidades para alcançar essa melhoria é fazer uma análise de similaridade parcial, ou seja, não usar todos os atributos descritivos do dado no cálculo da similaridade. Essa é uma análise capaz de fornecer resultados diferenciados e potencialmente úteis para alguns contextos ou objetivos específicos.

A análise de similaridade parcial é uma das características inerentes às tarefas de biagrupamento e coagrupamento, já que estas encontram subconjuntos de objetos e atributos do conjunto de dados que refletem relações significativas entre si. Dessa forma, a análise dos dados é feita apenas em sub-regiões do conjunto de dados original, realizando de forma intrínseca o procedimento de eliminar atributos redundantes, ou de pouca relevância, de acordo com o conjunto de objetos que está sendo avaliado no momento (FRANCA, 2010). No entanto, os resultados obtidos com a resolução destas tarefas ainda podem ser

afetados em contextos de alta dimensionalidade e grande quantidade de dados ausentes (alta esparsidade) podendo, muitas vezes, ocasionar problemas na determinação dos grupos de objetos. Embora ambas as tarefas utilizem as similaridades parciais nos processos de biagrupar ou coagrupar dados, suas definições variam segundo suas naturezas e objetivos, sendo que o coagrupamento pode ser visto como uma generalização do biagrupamento em que os bigrupos cobrem, necessariamente, todos os elementos da matriz de dados (PENSA et al., 2010).

Segundo Madeira e Oliveira (2004), a tarefa de biagrupamento envolve um problema computacional complexo. Na forma mais simples, quando $\mathcal{A}$ é uma matriz binária, um bigrupo corresponde a um *biclique* de um grafo bipartido, supondo que as linhas da matriz $\mathcal{A}$ são os vértices à esquerda e as colunas os vértices à direita, e os valores dos elementos indicam se existe uma aresta ligando uma linha a uma coluna caso o valor seja 1, ou 0 caso contrário. Encontrar o bigrupo de tamanho máximo, nessa situação, é equivalente a encontrar o *biclique* com o máximo de arestas em um grafo bipartido, que é um problema NP-difícil (PEETERS, 2003). Assim, é possível inferir que problemas mais complexos nos quais a matriz $\mathcal{A}$ apresenta valores numéricos reais que serão levados em conta para calcular a qualidade de um bigrupo e/ou de um cogrupos, não apresentarão menor complexidade do que a forma mais simples (MADEIRA; OLIVEIRA, 2004; BULTEAU et al., 2016). Por esse motivo, grande parte dos algoritmos de biagrupamento e coagrupamento são métodos imprecisos, baseados em heurísticas.

O problema tratado nesta pesquisa é a análise de dois algoritmos que implementam soluções para as tarefas de biagrupamento e coagrupamento. Ambos são capazes de fornecer soluções sub-ótimas para o problema em tempo aceitável, e que sejam adequadas para a realização da busca de soluções em espaços caracterizados pela esparsidade e a alta dimensionalidade dos dados. O objetivo é apresentar uma análise do comportamento destes algoritmos nesse ambiente de experimentação.

1.2 Questão de Pesquisa

A questão de pesquisa que delimita este estudo exploratório é: quais são as dificuldades enfrentadas pelo algoritmo de biagrupamento Cheng e Church e pelo algoritmo de

coagrupamento *Non-Negative Block Decomposition* (NBVD) quando eles são aplicados a dados textuais, que possuem uma representação esparsa e de alta dimensionalidade.

1.3 Objetivos

O objetivo principal deste trabalho é apresentar uma análise do comportamento do algoritmo de biagrupamento Cheng e Church e do algoritmo de coagrupamento NBVD, aplicado ao contexto de dados textuais.

Assim, foram estabelecidos os seguintes objetivos específicos:

- Analisar o comportamento dos algoritmos em conjuntos de dados sintéticos em um ambiente sem esparsidade e em ambientes com diferentes níveis de esparsidade.
- Analisar o comportamento dos algoritmos no contexto da resolução da tarefa de biagrupamento e coagrupamento em mineração de texto em um conjunto de dados real, verificar se os algoritmos são capazes de encontrar tópicos e explicar o significado dos grupos de documentos e bigrupos/cogrupos descobertos.

1.4 Método

Esta pesquisa exigiu um estudo exploratório e uma revisão sistemática para a obtenção de conhecimentos sobre as áreas de biagrupamento e coagrupamento, assim como dos algoritmos e métodos de avaliação utilizados no contexto dos dados textuais.

Diversos algoritmos para resolução das tarefas de biagrupamento e coagrupamento podem ser encontrados na literatura. Nos trabalhos de Madeira e Oliveira (2004), Tanay, Sharan e Shamir (2005), Busygin, Prokopyev e Pardalos (2008) e Pontes, Giráldez e Aguilar-Ruiz (2015), os autores apresentam diversos algoritmos de biagrupamento classificando-os de acordo com: as estratégias algorítmicas nas quais são baseados (PONTES; GIRÁLDEZ; AGUILAR-RUIZ, 2015); o tipo de bigrupos que eles podem encontrar; as estruturas de bigrupos que são descobertos; os métodos de otimização usados (MADEIRA; OLIVEIRA, 2004); e suas aplicações tanto na área de expressão gênica (TANAY; SHARAN; SHAMIR, 2005) como na área de mineração de dados (BUSYGIN; PROKOPYEV; PARDALOS, 2008). Em (NADIF et al., 2013), os autores detalham o problema de coagrupamento e apresentam alguns algoritmos mostrando suas áreas de aplicação e os dados sobre os quais

eles trabalham (dados binários, dados categóricos, tabelas de contingência e de dados contínuos). Existem ainda outros trabalhos aplicados especialmente em dados textuais, os quais serão apresentados na seção 2.5. Cabe ressaltar que em muitos destes estudos, os termos de biagrupamento e coagrupamento são usados de forma intercambiável.

Esses algoritmos podem ser organizados em categorias a partir do estabelecimento de critérios que podem definir uma taxonomia. Porém, mais importante do que isso, é escolher o algoritmo que melhor se adequa à resolução de um problema em específico. Geralmente, pesquisadores se baseiam em conhecimento especializado sobre o domínio do problema e sobre os dados para fazer comparações referentes às capacidades de cada algoritmo. Exemplos de iniciativas que analisam e propõem algoritmos são: o trabalho de Franca (2010), que explora as propriedades do processo de biagrupamento em dados incertos (dados ruidosos ou dados faltantes); os trabalhos que fazem comparações de algoritmos de biagrupamento específicos na área de expressão gênica (BHATTACHARYA; CHOWDHURY; DE, 2012; LI et al., 2012; PRELIĆ et al., 2006); e os trabalhos que propõem novos algoritmos demonstrando que as estratégias neles implementadas são melhores em determinados contextos por meio da comparação com algoritmos já existentes (CASTRO et al., 2007a; YOO; CHOI, 2010; WANG et al., 2011; BRUNIALTI et al., 2017).

No contexto deste trabalho, 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, biagrupamento e coagrupamento, com o fim de propiciar uma análise de 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 as versões básicas das estratégias algorítmicas neles implementadas, para propiciar uma análise didática das virtudes e problemas desses algoritmos. Ambos algoritmos escolhidos

possuem versões aprimoradas, que fazem uso de recursos mais avançados a fim de adequá-los a características específicas encontradas em problemas reais. Exemplos dessa situação são apresentados na descrição básica dos algoritmos.

Seguindo os critérios anteriores, os dois algoritmos escolhidos para o estudo foram: o algoritmo de Cheng e Church, e o algoritmo de decomposição de valores em blocos não negativos (do inglês *Non-negative Block Decomposition* - NBVD).

Estes algoritmos foram analisados em um primeiro momento em um ambiente controlado representado por uma coleção de conjuntos de dados sintéticos, e na sequência aplicados a um conjunto de dados reais.

Os conjuntos de dados sintéticos foram gerados assumindo as diferentes estruturas de bigrupos (MADEIRA; OLIVEIRA, 2004). Estes conjuntos de dados são de alta dimensionalidade, criados tanto sem esparsidade quanto com diferentes níveis de esparsidade. O conjunto de dados reais utilizado foi o corpus EBC (PAGNOSSIM, 2018), que é uma coleção de notícias na língua portuguesa extraídos do portal da Empresa Brasil de Comunicação (EBC) ³. O portal EBC é uma plataforma na internet que integra conteúdos dos veículos (Agência Brasil, Radioagência Nacional, Rádios EBC, TV Brasil, TV Brasil Internacional) da Empresa Brasil de Comunicação e da sociedade em um único local.

O conteúdo do corpus EBC passou pelas rotinas de pré-processamento comuns na área de mineração de texto (DIAZ et al., 2018): “tokenização”, eliminação de “stopwords”, representação dos documentos baseadas na representação “bag-of-words” (“documentos $\times$ palavras”) usando as atribuições de pesos como a binária, TF e TF-IDF com diferentes seleções de atributos. Estas rotinas de pré-processamento estão detalhadas no capítulo 3.

O resultados dos algoritmos foram avaliados e analisados sob quatro estratégias, escolhidas com base na revisão sistemática realizada:

- aplicação das medidas de biagrupamento;
- aplicação das medidas internas de agrupamento, calculadas sobre a projeção dos bigrupos e cogrupo nas linhas dos conjuntos de dados;
- inspeção visual sobre representações gráficas de matrizes reconstruídas em relação aos conjuntos de dados sintéticos;

³ <http://www.ebc.com.br>

- avaliação de conteúdo semântico por meio da análise de grupos de palavras representativas para cada bigrupo e cogrupo resultante do biagrupamento e coagrupamento, respectivamente, aplicados no corpus EBC.

A estrutura da análise e os métodos de experimentação dos algoritmos escolhidos estão detalhados no capítulo 3.

1.5 Organização do documento

A fim de discutir a pesquisa delineada em biagrupamento e coagrupamento de textos, este documento é composto por seis capítulos incluindo esta introdução. Os demais capítulos estão organizados da seguinte forma:

- No capítulo 2 são apresentados os conceitos fundamentais necessários para a compreensão desta pesquisa, as definições referentes aos campos de estudo das tarefas de biagrupamento e coagrupamento, a descrição dos algoritmos e das medidas para avaliação destes, assim como as análises da revisão sistemática com foco nos dados textuais.
- O capítulo 3 apresenta o método de experimentação usado para as análises dos algoritmos, junto com esquemas gráficos que ajudam sua compreensão e, além disso, descreve os conjuntos de dados e métodos de avaliação utilizados.
- Os capítulos 4 e 5 apresentam os resultados e análise obtidos a partir das experimentações do algoritmo de Cheng e Church e do algoritmo NBVD, respectivamente. Os resultados dizem respeito tanto à aplicação desses algoritmos nos conjuntos de dados sintéticos quanto no corpus EBC. O capítulo ainda traz as considerações finais obtidas a partir destas análises.
- E finalmente, o capítulo 6 apresenta as conclusões, limitações deste trabalho e sugestões de trabalhos futuros.

2 Conceitos fundamentais

Este capítulo apresenta os principais conceitos e definições necessários para a compreensão dos estudos descritos nos demais capítulos desta dissertação. O conteúdo deste capítulo segue organizado da seguinte forma: definições referentes aos campos de estudo das tarefas de coagrupamento e biagrupamento são apresentadas na seção 2.1; na seção 2.2 algoritmos básicos que implementam a solução para tarefas de biagrupamento (algoritmo de Cheng e Church (CHENG; CHURCH, 2000)) e coagrupamento (NBVD - *Non-negative Block Value Decomposition* (LONG; ZHANG; YU, 2005)) são detalhadamente descritos; então, na seção 2.3 são discutidas algumas medidas comumente usadas na literatura para avaliação dos resultados de biagrupamento e coagrupamento; na seção 2.4 são discutidos os principais aspectos referentes à análise de dados textuais, com ênfase aos fatores que geram representações com matrizes esparsas para esse contexto de análise; finalmente, com a seção 2.5, o capítulo é finalizado mediante a apresentação do estado da arte na área de análise de textos via algoritmos de biagrupamento e coagrupamento, obtidos a partir da realização de uma revisão sistemática da literatura.

2.1 Biagrupamento e coagrupamento

Na literatura, as tarefas de biagrupamento e de coagrupamento são, por vezes, referenciadas de maneira intercambiável (BANERJEE et al., 2007; ANAGNOSTOPOULOS; DASGUPTA; KUMAR, 2008; PENSA; BOULICAUT, 2008; DEODHAR et al., 2009; LEE et al., 2010; DEODHAR; GHOSH, 2010; CHARRAD; AHMED, 2011; ORZECOWSKI; BORYCZKO, 2016). Eventualmente, autores tratam problemas que originalmente têm natureza de coagrupamento por meio da aplicação de algoritmos de biagrupamento, e vice-versa. Em Mirkin (1998), o autor 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 definir 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. Contudo, atualmente, ambos os problemas possuem definições bem estabelecidas e que diferem nas suas naturezas, ou objetivos, básicos.

Seguindo a discussão apresentada pelos autores Pensa et al. (2010), a diferenciação entre biagrupamento e coagrupamento pode ser vista da seguinte maneira:

- Um bigrupo é um conjunto de objetos e um conjunto de atributos a eles associados. Cada objeto pertencente a um bigrupo está fortemente correlacionado a qualquer outro objeto pertencente ao mesmo bigrupo, em relação ao conjunto de atributos associados. Na tarefa de biagrupamento, as estratégias algorítmicas procuram por um número, algumas vezes¹ não especificado (*a priori*), de bigrupos sobrepostos ou não.
- Na tarefa de coagrupamento, estratégias algorítmicas procuram 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 ao 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; ao mesmo tempo eles estão diferentemente relacionados a qualquer outro objeto pertencente ao mesmo grupo, em relação a todos os demais atributos. A relação inversa, referente a grupos de atributos diante de subconjunto de objetos também ocorre.

Ainda segundo Pensa et al. (2010), um conjunto de bigrupos não fornece um modelo de dados, pois captura associações locais em vez de fornecer uma visão global dos dados, e fornecem 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 bi-grupos forma uma bipartição para o conjunto de objetos original; neste caso, o problema de coagrupamento pode ser visto como uma generalização do problema de biagrupamento em que os bigrupos cobrem todos os elementos da matriz de dados. Madeira e Oliveira (2004) discutem o problema de biagrupamento generalizado, o qual consiste em encontrar um conjunto mínimo de bigrupos, mutuamente exclusivos ou sobrepostos, que cobrem todos os elementos em uma matriz de dados. Segundo os mesmos autores, esse problema é equivalente ao problema de coagrupamento.

¹ Pensa et al. (2010) não espera a definição do número de grupos *a priori*, contudo, segundo a visão dos autores Madeira e Oliveira (2004), o número de bigrupos deve ser conhecido *a priori*. Isso ilustra a variabilidade de visões existentes dentro desse contexto na literatura.

Para ilustrar essa discussão, considere o exemplo adaptado de (PAPADIMITRIOU; SUN, 2008) que apresenta a matriz $\mathcal{X}$ com dimensões 4×5 ($n = 4$ é o número de linhas; $m = 5$ é o número de colunas). A matriz $\mathcal{X}$ é composta por um conjunto de linhas (A , B , C , D) que representam estudantes (estudante A , estudante B , estudante C e estudante D) 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:

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

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 como em:

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

Alternativamente, a depender das condições impostas a um algoritmo de biagrupamento, bigrupos referentes à organização dos valores 0 também poderiam ser encontrados. Nesse caso, a organização final dos bigrupos se assemelha 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, 2 grupos de linhas e 2

grupos de colunas poderiam ser formados. Um exemplo de resultado de coagrupamento neste exemplo seria:

- um cogruppo formado pelo grupo de linhas B e D e pelo grupo de colunas *ciências* e *matemática*, representando os estudantes com bom desempenho nas disciplinas *ciências* e *matemática*;
- um cogruppo formado pelo grupo linhas B e D e pelo grupo de colunas *inglês*, *chinês* e *social*, representando os estudantes de desempenho ruim nas disciplinas *inglês*, *chinês* e *social*;
- um cogruppo formado pelo grupo de linhas A e C e pelo grupo de colunas *ciências* e *matemática*, representando os estudantes de desempenho ruim nas disciplinas *ciências* e *matemática*;
- um cogruppo formado pelo grupo de linhas A e C e pelo grupo de colunas *inglês*, *chinês* e *social*, representando os estudantes com bom desempenho nas disciplinas *inglês*, *chinês* e *social*.

Com a reorganização da matriz $\mathcal{X}$, os cogrupos podem ser visualmente evidenciados, como em:

$$\mathcal{X} = \begin{matrix} & \begin{matrix} \textit{ciências} & \textit{matemática} & \textit{inglês} & \textit{chinês} & \textit{social} \end{matrix} \\ \begin{matrix} D \\ B \\ C \\ A \end{matrix} & \left(\begin{array}{cc|cc|c} \hline 1 & & & & \\ \hline 1 & & & & \\ \hline 0 & & & & \\ 0 & & & & \\ \hline \end{array} \right) \end{matrix}$$

Nas próximas seções são apresentados detalhes específicos dos contextos referentes a cada uma das tarefa de biagrupamento e coagrupamento.

2.1.1 Biagrupamento

Segundo os autores Madeira e Oliveira (2004), o biagrupamento é tratado como o problema de encontrar submatrizes de uma matriz de dados. Sob essa visão, a definição formal do problema de biagrupamento pode ser dada da seguinte maneira:

- 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, e 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 das linhas e colunas respectivamente, $\mathbf{A}_{IJ} = (I, J)$ denota a submatriz $\mathbf{A}_{IJ}$ 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 $\mathbf{A}_{IJ} = (I, J)$ 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) pode ser 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 $\mathbf{B}_{IJ} = (I, J)$, de forma que cada bigrupo $\mathbf{B}_{IJ}$ satisfaça algumas características específicas de homogeneidade.

As características de homogeneidade influenciam o estabelecimento de taxonomias que nomeiam os diferentes tipos e estruturas de bigrupos existentes². Nas próximas seções, duas taxonomias de organização de bigrupos são apresentadas.

² Embora 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 trabalho.

Tipos de bigrupos

Vários tipos de bigrupos foram descritos e categorizados na literatura, dependendo do comportamento que eles apresentam (MUKHOPADHYAY; MAULIK; BANDYOPADHYAY, 2010; MADEIRA; OLIVEIRA, 2004). Madeira e Oliveira (2004) 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 $\mathbf{A}_{IJ}$. Geralmente esse tipo de bigrupo não é encontrado em dados do mundo real devido à frequente presença de ruído nesse tipo de dado. Então, a definição da equação 1 deve ser 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

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

Fonte: Adaptado de Madeira e Oliveira (2004)

2. **Bigrupos com valores constantes nas linhas ou nas colunas:** São aqueles bigrupos que apresentam valores constantes nas linhas ou nas colunas. Eles podem ser representados como:

- Bigrupos com 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 e 3.

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

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

em que μ é o valor típico dentro do bigrupo e γ_i é o ajuste para a linha $\mathbf{i}$, sendo que este ajuste pode ser obtido de forma aditiva (Equação 2) ou multiplicativa (Equação 3).

- Bigrupos com 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 e 5.

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

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

em que μ é o valor típico dentro do bigrupo e β_j é o ajuste para a coluna $\mathbf{j}$ 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 $\mathbf{i}$ 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 $\mathbf{j}$ 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)

(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

Fonte: Adaptado de Madeira e Oliveira (2004)

3. **Bigrupos com valores coerentes:** São aqueles bigrupos 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:

- Bigrupos com 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:

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

sendo μ o valor típico dentro do bigrupo, γ_i o ajuste para a linha $\mathbf{i}$ e β_j o ajuste para a coluna $\mathbf{j}$.

Os bigrupos com valores constantes nas linhas (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 $\gamma_i = 0$, $\forall \mathbf{i}, \mathbf{j} \in I, J$ respectivamente (MADEIRA; OLIVEIRA, 2004).

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

$$a_{ij} = \mu' \times \gamma'_i \times \beta'_j, \forall \mathbf{i}, \mathbf{j} \in I, J \quad (7)$$

sendo μ' o valor típico dentro do bigrupo, γ'_i o ajuste para a linha $\mathbf{i}$ e β'_j o ajuste para a coluna $\mathbf{j}$. Além disso, os bigrupos com valores constantes nas linhas ou nas colunas (Figura 3b) de forma multiplicativa também podem ser considerados casos especiais deste modelo multiplicativo, já que as equações 3 e 5 são casos especiais do modelo representado pela equação 7, quando $\beta'_j = 1$ e $\gamma'_i = 1$, $\forall \mathbf{i}, \mathbf{j} \in I, J$ respectivamente.

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 $\mathbf{i}$ são $\{0, 3, 2, -1\}$ e os valores de β para cada coluna de $\mathbf{j}$ 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 $\mathbf{i}$ são $\{0, 1, 2, 3\}$ e os valores de β' para cada coluna de $\mathbf{j}$ são $\{1, 2, 3, 0.5\}$.

4. **Bigrupos com evoluções coerentes:** A evolução coerente difere da medida de coerência por não buscar valores coerentes dentro do bigrupo, mas sim propriedades coerentes nos valores. Os bigrupos são visualizados como valores simbólicos que têm

Figura 4 – Exemplos de bigrupos com valores coerentes nas linhas e nas colunas usando um modelo aditivo (a) e usando um modelo multiplicativo (b)

(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

Fonte: Adaptado de Madeira e Oliveira (2004)

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

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

(d)			
70	13	19	10
49	40	49	35
40	20	27	15
90	15	30	12

Fonte: Madeira e Oliveira (2004)

Os autores Mukhopadhyay, Maulik e Bandyopadhyay (2010) apresentam uma classificação que organiza os bigrupos em seis tipos de bigrupos: (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 de bigrupos 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 (2004).

Estruturas de biagrupamento

Os algoritmos de biagrupamento assumem as seguintes situações: encontrar apenas um bigrupo, como na situação mostrada na figura 6, ou encontrar K bigrupos, como é o caso dos exemplos mostrados nas figuras³ 7 a 14. Nesses casos, os algoritmos de análise dos dados assumem mais de um bigrupo no conjunto de dados e podem obter como resultado diferentes estruturas. Madeira e Oliveira (2004) 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 6.

Figura 6 – Estrutura de um único bigrupo

Fonte: Adaptado de Madeira e Oliveira (2004)

- 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 7.

³ Nessas 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 quadro, áreas em branco representam valores que não possuem nenhum tipo de relação ou significado para o biagrupamento desse conjunto.

Figura 7 – Estrutura de bigrupos com linhas e colunas exclusivas

Fonte: Adaptado de Madeira e Oliveira (2004)

- 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 8.

Figura 8 – Estrutura de bigrupos com estrutura de tabuleiro de xadrez

Fonte: Adaptado de Madeira e Oliveira (2004)

- 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 9.

Figura 9 – Estrutura de bigrupos com linhas exclusivas

Fonte: Adaptado de Madeira e Oliveira (2004)

- e. **Bigrupos com colunas exclusivas:** 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 10.

Figura 10 – Estrutura de bigrupos com colunas exclusivas

Fonte: Adaptado de Madeira e Oliveira (2004)

- f. **Bigrupos sem sobreposição com estrutura em árvore:** Segundo Hartigan (1972), se em um biagrupamento quaisquer dois bigrupos são disjuntos, porém suas projeções nas linhas ou nas colunas não o são, então eles formam uma estrutura em árvore. Na figura 11, 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 percurso pré-ordem. Na projeção sobre colunas, a hierarquia é $\{a, b, c, d\}$.

Figura 11 – Estrutura de bigrupos sem sobreposição com estrutura em árvore

Fonte: Adaptado de Madeira e Oliveira (2004) e Hartigan (1972)

- g. **Bigrupos não exclusivos e sem sobreposição:** 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 12.

Figura 12 – Estrutura de bigrupos não exclusivos e sem sobreposição

Fonte: Adaptado de Madeira e Oliveira (2004)

Madeira e Oliveira (2004) afirmam que existem duas especializações da estrutura de bigrupo hierárquica: a estrutura de árvore apresentada na figura 11, na qual os bigrupos formam uma árvore; e a estrutura de tabuleiro de xadrez, representada na figura 8, na qual os bigrupos em si representam uma árvore.

- h. Bigrupos com sobreposição e com estrutura hierárquica:** Assume que a hierarquia pode também ocorrer dentro de bigrupos. Na figura 13, 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 .

Figura 13 – Estrutura de bigrupos com sobreposição com estrutura hierárquica

Fonte: Adaptado de Madeira e Oliveira (2004) e Hartigan (1972)

- i. Bigrupos com sobreposição e arbitrariamente posicionados:** Esta estrutura permite a existência de K bigrupos possivelmente sobrepostos. Além disso 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 14.

Figura 14 – Estrutura de bigrupos com sobreposição e arbitrariamente posicionados

Fonte: Adaptado de Madeira e Oliveira (2004)

As estruturas de biagrupamento b , c , d , e , f , e g apresentam algumas características semelhantes:

- Em todas elas 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.
- 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 estruturas h e i .

2.1.2 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 (BANERJEE et al., 2007; ANAGNOSTOPOULOS; DASGUPTA; KUMAR, 2008), em algum sentido. Trata-se de um problema similar ao problema de agrupamento (que agrupa as linhas de uma matriz de dados), mas traz algumas vantagens (IENCO et al., 2013):

- Os algoritmos para coagrupamento são mais eficazes para trabalhar com dados que apresentam alta dimensionalidade e alta esparsidade, característica de dados

textuais (cf. Seção 2.4), portanto, lidam melhor com o clássico problema da maldição da dimensionalidade. Segundo os autores Dhillon, Mallela e Modha (2003) e Long, Zhang e Yu (2005), 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 de dimensionalidade implícita e adaptativa, assim como minimiza o problema de ruído nos dados.

- O coagrupamento fornece uma descrição dos grupos de objetos associando-os a grupos de atributos. Segundo Yoo e Choi (2010), 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 pode trazer conhecimento sobre relações entre pares de atributos diante de determinados grupos. Num contexto de 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 (HOFMANN, 1999)). 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 ‘aprendizado’, ‘reconhecimento’ e ‘visual’, então ela pode se referir ao objeto sob análise em um sistema de visão computacional. No caso o agrupamento, uma descoberta como essa exigiria um exame adicional dos resultados referentes aos grupos de objetos (documentos, no caso de mineração de textos) encontrados.

Além de apresentar vantagens sobre o agrupamento, o coagrupamento apresenta algumas diferenças em relação ao biagrupamento. Analisando as estruturas de biagrupamento da subseção 2.1.1 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 15a), entretanto, se a mesma estrutura for analisada sob a visão de coagrupamento, seis cogrupos serão observados (Figura 15b). Esses cogrupos a mais, a depender da aplicação em que o problema de coagrupamento está sendo usado, apresentam valores de pouco ou nenhum interesse. Tais cogrupos estão representados na figura 15b pela cor branca e correspondem 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 15 – Visão de coagrupamento na estrutura d , discutida na seção 2.1.1

Fonte: Alexandra Katiuska Ramos Diaz, 2018

Formalmente, o coagrupamento pode ser definido da seguinte forma⁴ (PENSA et al., 2010): 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 vetores⁵ 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 que usa a idéia de dados diádicos provém do uso de métodos de fatoração de matrizes na modelagem e resolução do problema de coagrupamento. Estes métodos abordam a matriz $\mathcal{X}$ como uma tabela de co-ocorrências, ou seja, 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 $(\mathbf{x}_i, \mathbf{y}_j)$. Normalmente, 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 (HOFMANN; PUZICHA; JORDAN, 1999). 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

⁴ Normalmente, definições de coagrupamento seguem a noção e notação de espaços vetoriais.

⁵ Para fins de simplificação da notação, as variáveis referentes a linhas e colunas da matriz de dados não serão grafadas como vetores (com flechas sobre as letras), embora elas possam ser vistas como vetores a depender do algoritmo utilizado para processamento.

na matriz $\mathcal{X}$, interligando agrupamentos de linhas com agrupamentos de colunas a cada iteração (DHILLON; MALLELA; MODHA, 2003; LONG; ZHANG; YU, 2005; YOO; CHOI, 2010; BUONO; PIO, 2015).

A visão de dados diádicos é também a raiz da abordagem apresentada por Lee e Seung (1999). 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.

2.2 Dois algoritmos: um para biagrupamento e um para coagrupamento

Os dois algoritmos escolhidos para estudo são:

- Algoritmo de Cheng e Church: O algoritmo de Cheng e Church (CHENG; CHURCH, 2000) é considerado um trabalho seminal porque introduziu a necessidade da busca de submatrizes em uma matriz de dados, a fim de encontrar relações interessantes e úteis entre os dados. É um algoritmo guloso que minimiza o valor de uma medida chamada Resíduo Quadrático Médio (RQM), para encontrar bigrupos (CHENG; CHURCH, 2000). Ele é usado na área de expressão gênica para extrair informação biológica relevante. Nesta área é utilizado como algoritmo de referência para validação de diferentes métodos propostos na área (TANAY; SHARAN; SHAMIR, 2002; IHMELS; BERGMANN; BARKAI, 2004; DEODHAR et al., 2009; BRYAN; CUNNINGHAM; BOLSHAKOVA, 2005; PRELIĆ et al., 2006) e em comparações para análises de métodos de biagrupamento (PRELIĆ et al., 2006; HOCHREITER et al., 2010). Adicionalmente foram propostas algumas extensões do algoritmo como o trabalho de Yang et al. (2002). Neste trabalho os autores apresentam um novo algoritmo baseado no algoritmo de Cheng e Church permitindo encontrar bigrupos de forma simultânea. O algoritmo proposto é testado em conjuntos de dados de expressão gênica e em conjuntos de dados textuais (YANG et al., 2003); melhorias para lidar com valores ausentes por meio de um limiar são propostos por Yang et al. (2005). O algoritmo de Cheng e Church está descrito na seção 2.2.1.
- Algoritmo de decomposição de valores em blocos não negativos (do inglês *Non-negative Block Decomposition* - NBVD): O algoritmo de decomposição de valores em blocos não negativos (LONG; ZHANG; YU, 2005) pertence à classe de algoritmos de

fatoração de matrizes não-negativas (*Non-negative Matrix Factorization* - NMF) que foi estudada como um método para análise de dados capaz de extrair conhecimento sobre um objeto a partir do estudo de suas partes (LEE; SEUNG, 1999). Todos os algoritmos de NMF para resolução das tarefas de coagrupamento (DING et al., 2006; YOO; CHOI, 2010; WANG et al., 2011; BUONO; PIO, 2015) encontram uma partição de linhas a partir de uma partição de colunas para então encontrar a melhor partição de colunas a partir da partição de linhas (e vice-versa). Os algoritmos seguem iterando desta forma até que a aproximação entre a fatoração e a matriz fatorada atinja um mínimo local ou global (por meio de alguma medida). O algoritmo de NBVD está descrito na seção 2.2.2.

2.2.1 Algoritmo de Cheng e Church

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 (MADEIRA; OLIVEIRA, 2004). No seu trabalho seminal (CHENG; CHURCH, 2000), Cheng e Church estavam interessados em encontrar bigrupos, ou δ -*biclusters*, baseados na minimização do RQM (Resíduo Quadrático Médio). A presente seção está baseada nos trabalhos de Franca (2010), Tanay, Sharan e Shamir (2005), Cheng e Church (2000). 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”. Essa medida permite encontrar bigrupos com valores coerentes baseados em um modelo aditivo. Os autores também destacam que ao resolver o problema de encontrar os bigrupos obtém-se uma solução para o problema de encontrar grupos, de linhas ou colunas, 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 16, está ilustrada a noção de sobreposição de bigrupos e da sobreposição de grupos em termos da projeção destes sobre linhas e colunas.

Cheng e Church usam o RQM de uma matriz como a medida de coerência ou de homogeneidade dos bigrupos. Essa medida conduzirá o processo de otimização, ou busca, das submatrizes que representam os bigrupos em uma matriz de dados. A base da medida

Figura 16 – Exemplo de sobreposições obtidas com o algoritmo de Cheng e Church

Fonte: Alexandra Katiuska Ramos Diaz, 2018

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

$$r(a_{ij}) = a_{ij} - a_{iJ} - a_{IJ} + a_{IJ}, \quad (8)$$

em que

$$a_{iJ} = \frac{1}{|J|} \sum_{\forall j \in J} a_{ij},$$

$$a_{IJ} = \frac{1}{|I|} \sum_{\forall i \in I} a_{ij},$$

$$a_{IJ} = \frac{1}{|I| |J|} \sum_{\forall i \in I, j \in J} a_{ij},$$

sendo a_{iJ} a média da i -ésima linha no bigrupo, a_{IJ} a média da j -ésima coluna no bigrupo, e a_{IJ} a média de todos os elementos do bigrupo.

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

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

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 8, é:

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

$$\begin{aligned} a_{1J} &= \frac{2+1}{2} &= 1,5 \\ a_{I1} &= \frac{2+6}{2} &= 4,0 \\ a_{IJ} &= \frac{2+1+6+7}{4} &= 4,0 \end{aligned}$$

$$\begin{aligned} r(a_{11}) &= a_{11} - a_{1J} - a_{I1} + a_{IJ} \\ &= 2 - 1,5 - 4,0 + 4,0 \\ &= 0,5. \end{aligned}$$

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:

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

$$\begin{aligned} a_{1J} &= \frac{27+0}{2} &= 13,5 \\ a_{I1} &= \frac{27+20}{2} &= 23,5 \\ a_{IJ} &= \frac{27+0+20+15}{4} &= 15,5 \end{aligned}$$

$$\begin{aligned} r(a_{11}) &= a_{11} - a_{1J} - a_{I1} + a_{IJ} \\ &= 27 - 13,5 - 23,5 + 15,5 \\ &= 5,5. \end{aligned}$$

A partir destes 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 9, 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,

$$\begin{aligned} H(I, J) &= \frac{1}{|I| |J|} \sum_{\forall i \in I, j \in J} (r(a_{ij}))^2 \\ &= \frac{1}{|I| |J|} \sum_{\forall i \in I, j \in J} (a_{ij} - a_{iJ} - a_{IJ} + a_{IJ})^2, \end{aligned} \quad (9)$$

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 8. 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 8 são: $a_{1J} = 1,5$, $a_{2J} = 6,5$, $a_{I1} = 4,0$, $a_{I2} = 4,0$ e $a_{IJ} = 4,0$. Os valores dos resíduos de cada elemento do bigrupo $\mathbf{A}_1(I, J)$ 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 9, o RQM do bigrupo $\mathbf{A}_1(I, J)$ 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 $\mathbf{A}_2(I, J)$, 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: $a_{1J} = 13, 5$, $a_{2J} = 17, 5$, $a_{1I} = 23, 5$, $a_{I2} = 7, 5$ e $a_{IJ} = 15, 5$. Os valores dos resíduos de cada elemento da submatriz $\mathbf{A}_2(I, J)$ 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 $\mathbf{A}_2(I, J)$ 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 $\mathbf{A}_2(I, J)$ com um RQM de 30, 25). Já um valor baixo de RQM implica que a matriz possui elementos com coerência e homogeneidade (bigrupo $\mathbf{A}_1(I, J)$ com um RQM de 0, 25). Por fim, um valor de RQM = 0 significaria que os dados da matriz flutuam 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 $\mathbf{i}$, $H(\mathbf{i}, J)$, e de uma coluna $\mathbf{j}$, $H(I, \mathbf{j})$. Esses resíduos são definidos, como:

$$\begin{aligned} H(\mathbf{i}, J) &= \frac{1}{|J|} \sum_{\forall j \in J} (a_{\mathbf{i}j} - a_{\mathbf{i}J} - a_{Ij} + a_{IJ})^2, \\ H(I, \mathbf{j}) &= \frac{1}{|I|} \sum_{\forall i \in I} (a_{i\mathbf{j}} - a_{iJ} - a_{Ij} + a_{IJ})^2. \end{aligned} \quad (10)$$

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.

Algoritmo 1 Algoritmo de Cheng e Church

```

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:        $\mathbf{C} \leftarrow$  Remoção-Múltiplos-Nós( $A'$ ,  $\delta$ ,  $\alpha$ ); ▷ cf. algoritmo 2
6:     else
7:        $\mathbf{C} \leftarrow$  Remoção-Simples( $A'$ ,  $\delta$ ); ▷ cf. algoritmo 3
8:      $\mathbf{D} \leftarrow$  Adição-Nós( $\mathcal{A}$ ,  $\mathbf{C}$ ); ▷ cf. algoritmo 4
9:     Reporte  $\mathbf{D}$  como uma solução;
10:    Adicione ruído em  $A' \forall a \in \mathbf{D}$ ;
11:   return  $K$   $\delta$ -bigrupos.
```

Fonte: Cheng e Church (2000)

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 bigrupos 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 ($\mathbf{C}$) e a matriz original de dados.

Para a busca pelos bigrupos, o algoritmo executa três procedimentos: remoção de múltiplas linhas e colunas (Algoritmo 2), originalmente nomeadas como nós pelos autores do algoritmo, da matriz A' (linha 5); remoção simples (Algoritmo 3) de linhas ou colunas aplicada sobre a submatriz resultante do passo anterior (linha 10 do Algoritmo 2) ou sobre a matriz A' caso seu número de linhas ou colunas seja menor que 100^6 (linha 7 do Algoritmo 1); adição de linhas e colunas (Algoritmo 4) à submatriz $\mathbf{C}$ resultante do passo anterior e considerando as linhas e colunas existentes na matriz original $\mathcal{A}$ (linha 8).

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 (cf. equação 10) 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.

Algoritmo 2 Remoção de múltiplas linhas e colunas simultaneamente

```

1: function REMOÇÃO-MÚLTIPLOS-NÓS( $A, \delta, \alpha$ )
2:   Inicialize  $\mathbf{A}_{IJ} = A$ ;
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  $\mathbf{i} \in I$  que satisfazem:  $H(\mathbf{i}, J) > \alpha \times H(I, J)$ ;
6:     Recalcule  $a_{iJ}, a_{Ij}, a_{IJ}$  e  $H(I, J)$ ;
7:     Remova simultaneamente todas as colunas  $\mathbf{j} \in J$  que satisfazem:  $H(I, \mathbf{j}) > \alpha \times H(I, J)$ ;
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:       $\mathbf{A}_{IJ} = \text{Remoção-Simples}(\mathbf{A}_{IJ}, \delta)$ ; ▷ cf. algoritmo 3
11:   return  $A_{IJ}$ ;

```

Fonte: Cheng e Church (2000)

Cheng e Church (2000) destacam que, o fato da estratégia de não atualizar o RQM do bigrupo após a remoção de cada nó (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 1. Versões adaptativas do algoritmo Cheng e Church permitem alterar α durante as iterações do algoritmo, com base no RQM e no tamanho dos bigrupos que estão sendo descobertos.

⁶ Os autores definem que uma matriz com número de linhas ou colunas menor do que 100 é pequena e, portanto, tratável para o procedimento de remoção simples.

Um procedimento alternativo, discutido por Cheng e Church (2000) é realizar a busca das linhas e colunas a serem removidas, uma a uma, e recalcular o RQM após cada remoção. No entanto, um procedimento para remover a linha ou a coluna de forma a alcançar o maior decréscimo do RQM exigirá a computação de RQM para todas as possibilidades de remoção de uma linha ou uma coluna. Este procedimento não seria eficiente. Ainda, um método de busca baseado em estratégias do tipo subida de encosta simples facilmente cairia em um máximo local, não atendendo às necessidades do algoritmo de busca por bigrupos.

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

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

Para selecionar uma submatriz cuja 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 deste 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

Fonte: Alexandra Katiуска Ramos Diaz, 2018

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:

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

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 tem o objetivo de remover, 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.

Para valores adequados de α , o custo de execução do algoritmo 2 é menor que do algoritmo 3. Assim, é útil executar a remoção de múltiplos nós antes, de forma a

Algoritmo 3 Remoção de uma linha ou uma coluna por vez

```

1: function REMOÇÃO-SIMPLES( $A, \delta$ )
2:   Inicialize  $\mathbf{A}_{IJ} = A$ ;
3:   Calcule  $a_{iJ}$ ,  $a_{Ij}$ ,  $a_{IJ}$  e  $H(I, J)$ ;
4:   while  $H(I, J) > \delta$  do
5:     Encontre a linha  $\mathbf{i} \in I$  com maior  $H(\mathbf{i}, J)$ ;
6:     Encontre a coluna  $\mathbf{j} \in J$  com maior  $H(I, \mathbf{j})$ ;
7:     Remova de  $\mathbf{A}_{IJ}$  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  $\mathbf{A}_{IJ}$ 

```

Fonte: Cheng e Church (2000)

reduzir o tamanho da matriz sobre a qual a remoção simples será aplicada, ainda que haja o risco de redução excessiva do bigrupo. 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 volume⁷ 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, por meio da execução do procedimento Adição-Linhas-Colunas, descrito no algoritmo 4.

Algoritmo 4 Adição de linhas e colunas

```

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

```

Fonte: Cheng e Church (2000)

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).

⁷ O volume de um bigrupo é definido como a quantidade de elementos alocados no bigrupo, para mais detalhes veja a seção 2.3.

O objetivo do procedimento Adição-Nós é, portanto, inserir linhas e colunas ao bicluster 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 encontrado 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 4.1.4) 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 $H(\mathbf{i}, J)$ e $H(I, \mathbf{j})$, $\forall \mathbf{i}, \mathbf{j} \in I, J$, tem-se:

- $H(\mathbf{1}, J) = 55, 6$
- $H(\mathbf{2}, J) = 3, 35$
- $H(\mathbf{3}, J) = 31, 01$
- $H(\mathbf{4}, J) = 13, 54$
- $H(I, \mathbf{1}) = 19, 10$
- $H(I, \mathbf{2}) = 1, 5$
- $H(I, \mathbf{3}) = \mathbf{56, 29}$
- $H(I, \mathbf{4}) = 26, 6,$

e, fazendo a remoção da coluna 3, que apresenta o maior resíduo de coluna ($H(I, \mathbf{3}) = 56, 29$), a matriz resultante é:

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

Desde que a submatriz $\mathbf{C}$, 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 $H(\mathbf{i}, J)$ e de $H(I, \mathbf{j}), \forall \mathbf{i}, \mathbf{j} \in I, J$:

- $H(\mathbf{1}, J) = 12, 12$
- $H(\mathbf{2}, J) = 4, 18$
- $H(\mathbf{3}, J) = \mathbf{13, 34}$
- $H(\mathbf{4}, J) = 8, 29$
- $H(I, \mathbf{1}) = 11, 40$
- $H(I, \mathbf{2}) = 6, 8$
- $H(I, \mathbf{3}) = 10, 24.$

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

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

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

- $H(\mathbf{1}, J) = \mathbf{13, 13}$
- $H(\mathbf{2}, J) = 2, 54$
- $H(\mathbf{3}, J) = 4, 47$
- $H(I, \mathbf{1}) = 4, 46$
- $H(I, \mathbf{2}) = 2, 54$
- $H(I, \mathbf{3}) = \mathbf{13, 13}$

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:

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

sendo seu RQM igual 0, 22 e os valores $H(\mathbf{i}, J)$ e de $H(I, \mathbf{j})$, $\forall \mathbf{i}, \mathbf{j} \in I, J$:

- $H(\mathbf{1}, J) = 0, 11$
- $H(I, \mathbf{1}) = 0, 22$
- $H(\mathbf{2}, J) = 0, 11$
- $H(I, \mathbf{2}) = 0, 22$
- $H(\mathbf{3}, J) = \mathbf{0, 44}$

Removendo a linha 3, a submatriz resultante $\mathbf{C}$ apresenta um RQM igual a 0, e é dada por:

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

Uma vez que o RQM da submatriz $\mathbf{C}$ satisfaz a condição: $H(I, J) \leq \delta$, o procedimento de Remoção-Simples finaliza, retornando a submatriz $\mathbf{C}$ 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 $\mathbf{C}$, sendo:

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

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

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

Encontrado o primeiro bigrupo (submatriz $\mathbf{D}$), segue a adição de ruído⁸ (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 segundo 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 $\mathbf{C}$ com um $H(I, J) = 3,94e - 31$:

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

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 $\mathbf{D}$ resultante é igual à submatriz $\mathbf{C}$, sendo ela o segundo bigrupo encontrado:

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

Finalizada, 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 $3,94e - 31$. Os bigrupos na matriz $\mathcal{A}$ são:

⁸ Para o exemplo, números aleatórios dentro do domínio de variação dos demais elementos da matriz são gerados para inserção de ruído.

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

Na literatura foram propostos alguns outros algoritmos baseados no algoritmo de Cheng e Church com o propósito de melhorar alguns aspectos do algoritmo original. Os autores Yang et al. (2005) propõem o algoritmo FLOC (*Flexible Overlapped biclustering*), o qual lida com a adição de ruído nos valores dos bigrupos encontrados na matriz de dados, sendo capaz de descobrir simultaneamente um conjunto de K bigrupos. Esses bigrupos são possivelmente sobrepostos e baseados em movimentos probabilísticos. Já o trabalho de Qu et al. (2005) visa melhorar o processo de adição de linhas e colunas introduzindo um parâmetro λ com o objetivo de estender o escopo da pesquisa e aumentar o número de nós (linhas ou colunas) que podem ser adicionados aos bigrupos.

2.2.2 Algoritmo de decomposição de valores em blocos não negativos (NBVD)

O algoritmo de decomposição de valores em blocos não negativos (NBVD) foi proposto no trabalho de Long, Zhang e Yu (2005). Este algoritmo fatora a matriz de dados $\mathcal{X}$, vista como uma matriz de dados diádicos (Subseção 2.1.2), 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 (2005), 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 (2005) 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 diáde $w(\mathbf{x}_i, \mathbf{y}_j)$, 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 $k \times l$ submatrizes de $\mathcal{X}$ tais que os elementos dentro de cada submatriz sejam semelhantes entre si e os elementos de diferentes submatrizes sejam dissimilares uns aos outros.

Ainda segundo Long, Zhang e Yu (2005), 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 pode-se aproximar os dados originais 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 (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*

$$f(U, S, V) = \min \| \mathcal{X} - USV^T \|^2, \quad (11)$$

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 Frobenius⁹ para matrizes, $k \ll n$ e $l \ll m$.

Como é mostrado na equação 11, o problema de BVD visa aproximar 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 17.

Na figura 17, 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.

⁹ Uma norma mapeia um arranjo de valores para um único valor e está mais comumente associada a um vetor. Uma das normas mais frequentes no contexto de análise de dados é a norma Euclidiana, que no contexto de matrizes é conhecida como norma de *Frobenius* e é definida como: $\| \mathcal{X} \| = \sqrt{\sum_{i,j} a_{ij}^2}$.

Figura 17 – Fatoração da matriz original de dados X nas matrizes U , S e V

Fonte: Yoo e Choi (2010)

Yoo e Choi (2010) exemplificam o contexto do problema de coagrupamento por fatoração de matrizes usando o contexto de análise de textos. Seguindo o exemplo desses autores e o exposto na figura 17, $\mathcal{X}$ 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 (2005) 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 $\mathcal{X}$ ¹⁰. Seguindo o exemplo de análise de 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

¹⁰ O 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 é a matriz de centroides – base de representação de grupos de linhas. Assim como SV^T , C tem dimensão $k \times m$.

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 18 e 19, respectivamente.

Figura 18 – 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)

Fonte: Brunialti (2016)

Long, Zhang e Yu (2005) 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.

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:

$$f(U, S, V) = \min ||\mathcal{X} - USV^T||^2 \quad (12)$$

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 $l \ll m$.

Figura 19 – 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)

Fonte: Adaptado de Yoo e Choi (2010) e Brunialti (2016)

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 apresentadas nas expressões 13, 14 e 15. Nessas equações, 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 clássica (linha por coluna) das matrizes $\mathcal{X}$, V e S^T .

$$U \odot \frac{\mathcal{X}VS^T}{USV^TVS^T} \quad (13)$$

$$V \odot \frac{\mathcal{X}^TUS}{VS^TU^TUS} \quad (14)$$

$$S \odot \frac{U^T\mathcal{X}V}{U^TUSV^TV} \quad (15)$$

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 :

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

Algoritmo 5 Algoritmo de decomposição de valores em blocos não negativos (NBVD)

```

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:      $V^{(t+1)} \leftarrow V^{(t)} \odot \frac{\mathcal{X}^T U^{(t+1)} S^{(t)}}{V^{(t)} S^{(t)T} U^{(t+1)T} U^{(t+1)} S^{(t)}}$ ;
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  $U^{(t)}, V^{(t)}, S^{(t)}$ 

```

Fonte: Baseado de Long, Zhang e Yu (2005) e Brunialti (2016)

As regras multiplicativas implementam um processo de minimização de erro (YOO; CHOI, 2010), 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 (equação 13, 14 e 15) resultam em 1, i.e., quando $\mathcal{X} = USV^{t11}$.

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 trabalho, 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

¹¹ O processo iterativo de minimização do erro implementado no algoritmo 5 é semelhante ao processo iterativo de minimização do problema definido para o algoritmo *k-means*. A regra de atualização da matriz U (matriz indicadora no *k-means*) é equivalente 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*) equivale à regra de atualização da matriz S no NBVD.

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 grupo¹². Na figura 20 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.

Figura 20 – Pós-processamento das matrizes U , S e V em um conjunto de dados sintéticos

Fonte: Alexandra Katiuska Ramos Diaz, 2018

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

¹² Esse 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.

olhando sob a perspectiva das linhas), estão sendo indicadas as palavras que possuem maior aderência a cada grupo de palavras. Na figura 20, 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. Além disso permite a sobreposição de cogrupos. No caso do exemplo da figura 20, por exemplo, 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 20, na figura 21 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.

Figura 21 – Reconstrução das matrizes U , S e V do conjunto de dados sintéticos da figura 20

Seguindo a estratégia de interpretação aqui descrita, a aceitação de uma relação como um cogruppo 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 2.1.1) 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 2.1.2).

Para ilustrar a execução do algoritmo de NBVD, considere que o algoritmo recebe como entrada $k = 2$, $l = 2$, e a mesma matriz de dados $\mathcal{X}$ usada na seção 2.2.1:

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

Na figura 22 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.

Figura 22 – Exemplo da execução do algoritmo NBVD na matriz de dados $\mathcal{X}$

<table border="1" style="border-collapse: collapse; text-align: left;"> <tr><td>13,85</td><td>0,02</td></tr> <tr><td>6,7</td><td>6,71</td></tr> <tr><td>0,01</td><td>4,52</td></tr> <tr><td>2,01</td><td>4,77</td></tr> </table>	13,85	0,02	6,7	6,71	0,01	4,52	2,01	4,77	×	<table border="1" style="border-collapse: collapse; text-align: left;"> <tr><td>0,9</td><td>0,001</td></tr> <tr><td>0,45</td><td>2,75</td></tr> </table>	0,9	0,001	0,45	2,75	×	<table border="1" style="border-collapse: collapse; text-align: left;"> <tr><td>0,07</td><td>0,29</td><td>2,14</td><td>0,02</td></tr> <tr><td>0,34</td><td>0,16</td><td>0,03</td><td>0,79</td></tr> </table>	0,07	0,29	2,14	0,02	0,34	0,16	0,03	0,79	≈	<table border="1" style="border-collapse: collapse; text-align: left;"> <tr><td>2</td><td>3</td><td>27</td><td>0</td></tr> <tr><td>6</td><td>7</td><td>20</td><td>15</td></tr> <tr><td>10</td><td>0</td><td>5</td><td>8</td></tr> <tr><td>1</td><td>4</td><td>9</td><td>12</td></tr> </table>	2	3	27	0	6	7	20	15	10	0	5	8	1	4	9	12
13,85	0,02																																									
6,7	6,71																																									
0,01	4,52																																									
2,01	4,77																																									
0,9	0,001																																									
0,45	2,75																																									
0,07	0,29	2,14	0,02																																							
0,34	0,16	0,03	0,79																																							
2	3	27	0																																							
6	7	20	15																																							
10	0	5	8																																							
1	4	9	12																																							
<table border="1" style="border-collapse: collapse; text-align: left;"> <tr><td>5,64</td><td>4,01</td></tr> </table>	5,64	4,01																																								
5,64	4,01																																									
U		S		V^T		X																																				

Fonte: Alexandra Katiuska Ramos Diaz, 2018

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:

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

Os resultados obtidos a partir deste exemplo mostram a variabilidade entre os bigrupos obtidos aplicando o algoritmo de Cheng e Church e os cogrupos obtidos aplicando o algoritmo de NBVD. Esta variabilidade está ligada às características próprias de cada algoritmo. Por exemplo, o algoritmo de Cheng e Church está focado em procurar bigrupos com uma alta taxa de similaridade (valores de RMQs baixos) enquanto o algoritmo de NBVD procura soluções que minimizem o erro de reconstrução.

Existem trabalhos baseados em fatoração de matrizes que visam melhorar alguns aspectos do algoritmo de NBVD. O trabalho de Ding et al. (2006) propõem a fatoração ortogonal tripla de matrizes não-negativas (*Orthogonal Non-negative Matrix Tri-factorization* - ONMTF), na qual duas restrições de ortonormalidades são acrescentadas para as matrizes de grupos de linhas (U), e grupos de colunas (V), com a finalidade de restringir o problema de fatoração para um número menor de possíveis soluções. Esta restrição tende a encontrar protótipos mais centralizados nos cogrupos. O trabalho de Wang et al. (2011) 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 muito menor, exigindo muito menos multiplicações de matrizes, porém ele só funciona para matrizes binárias. Os trabalhos de Brunialti (2016) e Brunialti et al. (2017) 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 analisar os grupos de atributos formados independentemente dos grupos de dados. Essa estratégia torna-se útil ao trazer benefícios para a análise de agrupamento, como interpretações mais detalhadas ou descoberta de informações incorporadas nos grupos de atributos.

2.3 Avaliação em biagrupamento e coagrupamento

O processo de avaliação do resultado obtido a partir de um algoritmo de agrupamento é comumente chamado de validação (SILVA; PERES; BOSCARIOLI, 2016; HALKIDI; BATISTAKIS; VAZIRGIANNIS, 2002a; HALKIDI; BATISTAKIS; VAZIR-

GIANNIS, 2002b). De maneira semelhante, a avaliação dos resultados produzidos pelos algoritmos de biagrupamento e coagrupamento também pode ser considerada como uma validação. Inclusive, a visão dos resultados de biagrupamento e coagrupamento, por vezes, pode refletir a existência de grupos (de linhas e de colunas) e, assim, a validação de agrupamento pode ser transferida para tais resultados.

A avaliação da qualidade dos bigrupos e os cogrupos detectados é um problema importante e desafiador, independentemente do tipo de dado que está sendo utilizado no experimento. Devido aos diferentes critérios utilizados pelos algoritmos e aos diferentes objetivos relacionados às tarefas de biagrupamento e coagrupamento, uma avaliação única e padronizada não é viável.

As medidas de avaliação dos resultados dos algoritmos de biagrupamento estão bem estabelecidas na literatura da área. Na literatura existem medidas que ajudam a identificar ou a avaliar a qualidade de um bigrupo (MADEIRA; OLIVEIRA, 2004; MUKHOPADHYAY; MAULIK; BANDYOPADHYAY, 2010). A escolha de qual medida usar esta fortemente relacionada com as características dos bigrupos que cada algoritmo pretende encontrar. Madeira e Oliveira (2004) lista algumas destas medidas: a variância e o resíduo quadrático médio. A variância, proposta no algoritmo Block Clustering (HARTIGAN, 1972), é usada para avaliar bigrupos com valores constantes, de forma que um bigrupo perfeito é aquele que apresenta variância igual a 0. O resíduo quadrático médio é a medida mais conhecida para a identificação de bigrupo com valores coerentes, sendo um bigrupo perfeito aquele que apresenta o resíduo quadrático médio igual a 0. Esta medida é usada no presente trabalho. Além do resíduo quadrático médio existem outras medidas consideradas no trabalho de Castro et al. (2010) como o volume, a cobertura do conjunto de dados, o grau de sobreposição e o grau de ocupação. Na sequência deste texto cada uma destas medidas é discutida:

Volume: Medida usada no trabalho de Cheng e Church (2000). O volume¹³ do bigrupo é denotado pela equação 16:

$$Volume = |I| |J|, \quad (16)$$

em que $|I|$ é o número de linhas do bigrupo, e $|J|$ o número de colunas do bigrupo.

¹³ Uma nomenclatura mais adequada para compreensão dessa medida é o uso do conceito de área do bigrupo.

Geralmente um bigrupo que apresenta maior volume (grande número de linhas e colunas) permite uma análise mais aprofundada dos dados.

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 (CASTRO et al., 2010).

Grau de sobreposição: O grau de sobreposição entre os bigrupos é a interseção dos conjuntos de elementos em cada um deles. Indica o quanto um determinado bigrupo está contido em outro bigrupo (CASTRO et al., 2010).

Grau de ocupação: O grau de ocupação dos bigrupos é a quantidade de valores não nulos presentes nos bigrupos (CASTRO et al., 2010).

Para fins de compreensão dos significados das medidas de avaliação de biagrupamento considere um exemplo baseado na matriz $\mathcal{X}$. A matriz $\mathcal{X}$ é um conjunto de dados com 42 elementos distribuídos em 7 linhas e 6 colunas.

$$\mathcal{X} = \begin{bmatrix} 0.0 & 0.0 & 8.0 & 3.0 & 1.0 & 2.0 \\ 0.0 & 8.0 & 8.0 & 2.0 & 10.0 & 4.0 \\ 8.0 & 6.0 & 6.0 & 6.0 & 9.0 & 7.0 \\ 14.0 & 6.0 & 0.0 & 6.0 & 1.0 & 0.0 \\ 20.0 & 6.0 & 0.0 & 5.0 & 5.0 & 5.0 \\ 0.0 & 4.0 & 5.0 & 0.0 & 5.0 & 5.0 \\ 2.0 & 1.0 & 5.0 & 5.0 & 0.0 & 5.0 \end{bmatrix}$$

Note que na matriz $\mathcal{X}$ existem três bigrupos hipotéticos, os quais apresentam sobreposição de dois elementos entre o bigrupo laranja e o bigrupo azul e também sobreposição de dois elementos entre o bigrupo azul e o bigrupo verde. Os valores resultantes para cada medida de qualidade de biagrupamento estão apresentados na tabela 2.

Tabela 2 – Medidas do biagrupamento

Medidas	
Resíduo quadrático médio (média)	3,99
Volume (média)	10,0
Cobertura (%)	61,9
Grau de ocupação (%)	73,08
Grau de sobreposição (%)	15,38

Fonte: Alexandra Katiuska Ramos Diaz, 2018

Já para o caso do coagrupamento a avaliação não é tão bem estabelecida, no entanto existem avaliações que dizem a respeito do resultado de agrupamento que pode ser derivado do processo de coagrupamento. E essas avaliações dão um indício da qualidade do coagrupamento.

Halkidi, Batistakis e Vazirgiannis (2002a) e Halkidi, Batistakis e Vazirgiannis (2002b) apresentam três abordagens para validar a qualidade de um grupo. O primeiro é baseado em critérios externos. Isto implica avaliar os resultados produzidos por um algoritmo baseado em uma organização dos dados possível e conhecida previamente (*ground truth* ou verdade fundamental). A segunda abordagem é baseada em critérios internos. Neste caso, os resultados de agrupamento são avaliados somente com relação aos próprios dados do conjunto de dados, sem ajuda de informações sobre o *ground truth* sob a organização dos dados. Quando esses critérios internos são usados, o que se pretende fazer de fato é responder a uma pergunta sobre se a solução encontrada é uma boa solução. A terceira abordagem de validade de agrupamento é baseada em critérios relativos. Aqui, a ideia básica é a avaliação de uma estrutura de agrupamento comparando-a com outros esquemas de agrupamento, resultantes do mesmo algoritmo, mas com valores de parâmetros de entrada diferentes.

Dentre as diferentes métricas de validação interna provenientes da área de agrupamento (HALKIDI; BATISTAKIS; VAZIRGIANNIS, 2001; DESGRAUPES, 2013), foram escolhidas para uso deste trabalho as seguintes: a similaridade média intra-grupos, similaridade média dos centroides inter-grupos e o índice Davies Bouldin. A escolha dessas métricas é decorrente do estudo da literatura descrita na seção 2.5. Na sequência deste texto cada uma das medidas é discutida:

Similaridade média intra-grupos: Do inglês *intra-cluster average similarity*, mede a similaridade média entre os objetos que pertencem ao mesmo grupo. Buono e Pio (2015) definem a similaridade média de um determinado grupo G_z pela equação 17:

$$IntraSim(G_z) = \frac{\sum_{i=1}^{p_z} \sum_{j=i+1}^{p_z} sim(\mathbf{x}_i, \mathbf{x}_j)}{0.5 \times p_z \times (p_z - 1)}, \quad (17)$$

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 $\text{sim}(\mathbf{x}_i, \mathbf{x}_j)$ é a função de similaridade¹⁴. A similaridade média intra-grupos é definida pela equação 18:

$$\text{IntraSim} = \frac{\sum_{i=1}^k \text{IntraSim}(G_i) \times p_i}{n}, \quad (18)$$

em que n é o número total de objetos e k é o número de grupos de objetos.

Similaridade média dos centroides inter-grupos: Do inglês *inter-cluster centroids average similarity*, mede a similaridade média entre objetos pertencentes a diferentes grupos. Buono e Pio (2015) calcula a similaridade entre os centroides dos grupos, em vez de calcular a média entre todos os objetos. Esta medida é definida pela equação 19:

$$\text{InterSim} = \frac{\sum_{i=1}^k \sum_{j=i+1}^k \text{sim}(\text{centroid}_i, \text{centroid}_j)}{0.5 \times k \times (k - 1)}, \quad (19)$$

em que centroid_i é o centroide do i – ésimo grupo, k é o número de grupos de objetos, e $\text{sim}(\text{centroid}_i, \text{centroid}_j)$ é a função de similaridade.

Índice Davies Bouldin (DB): É uma medida que se baseia na ideia de identificar em uma partição, grupos que sejam compactos e que exibam uma boa separabilidade entre si. Para isso é necessário definir uma medida geral da separação de grupos, baseada no cálculo da similaridade média de cada um dos grupos com o grupo mais próximo a ele (DAVIES; BOULDIN, 1979). O índice DB foi definido pelos autores Davies e Bouldin (1979) pela equação 20:

$$DB = \frac{1}{k} \sum_{\substack{i=1 \\ j=1 \dots k \\ i \neq j}}^k \max \left(\frac{\text{disp}(G_i) + \text{disp}(G_j)}{\text{dist}(G_i, G_j)} \right), \quad (20)$$

em que k é o número de grupos, $\text{disp}(G_i)$ é uma medida da dispersão do grupo G_i , $\text{disp}(G_j)$ é a medida da dispersão do grupo G_j e $\text{dist}(G_i, G_j)$ é a distância entre os centroides dos grupos G_i e G_j . A medida de dispersão é dada pela distância média entre os objetos pertencentes a um grupo. Na versão mais simples e adotada neste trabalho, a distância euclidiana é utilizada para os cálculos de $\text{dist}(G_i, G_j)$ e das medidas de dispersão. Valores pequenos para o índice DB indicam grupos compactos, cujos centros estão bem separados uns dos outros.

¹⁴ Neste trabalho, a função de similaridade foi implementada usando a distância euclidiana (SOLER et al., 2013).

No decorrer deste trabalho a medida de similaridade média intra-grupos e a medida de similaridade média dos centroides 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.

2.4 Análise de textos

Para aplicar algoritmos de análise de dados sobre dados textuais é comum modelá-los usando o modelo de espaço vetorial (SALTON; BUCKLEY, 1988; FELDMAN; SANGER, 2007; TURNEY; PANTEL, 2010; AMINE; ELBERRICHI; SIMONET, 2010; MARCACINI; REZENDE, 2010; MI; GAO; TAN, 2015). A ideia básica para construção de uma representação baseada em modelo de espaço vetorial é: extrair palavras que contenham conteúdo exclusivo do conjunto de documentos sob análise, tratando essas palavras como características¹⁵; representar cada documento com um vetor de frequências ponderadas associando documentos e palavras em um espaço de características.

O modelo de espaço vetorial de um conjunto de dados textuais considera que cada documento é representado por um vetor, cujas dimensões são as palavras presentes na coleção de documentos, a cada dimensão é associado um valor numérico que representa a frequência de ocorrência da cada palavra em cada documento (AMINE; ELBERRICHI; SIMONET, 2010). O processo de atribuir valores numéricos às coordenadas dos vetores é conhecido como atribuição de pesos, dando maior importância aos termos que são relevantes. Entre as possibilidades de atribuição de pesos existentes na literatura, as mais populares são (FELDMAN; SANGER, 2007):

- **Binária:** que usa a presença (1) ou ausência (0) de uma palavra no documento para criar um vetor binário;
- **Frequência do termo (TF):** do inglês *term frequency*, definida como o número de ocorrências de uma determinada palavra em um documento (SALTON; BUCKLEY, 1988).

¹⁵ Etapa de pré-processamento de texto, descrita no trabalho de DIAZ et al. (2018)

- **TF-IDF:** do inglês *term frequency – inverse document frequency*, avalia a importância de uma palavra para o documento em que ela ocorre e em relação a todos os documentos presentes na coleção sob análise (LESKOVEC; RAJARAMAN; ULLMAN, 2014). A atribuição de pesos TF-IDF é calculada pela equação 21:

$$tf-idf(t_i, d_j) = tf(t_i, d_j) \times idf(t_i, D), \quad (21)$$

em que $tf(t_i, d_j)$ é a ocorrência da palavra t_i no documento d_j e $idf(t_i, D)$ considera a ocorrência da palavra t_i sobre toda a coleção de documentos D . Adicionalmente a idf da palavra t_i é calculada pela equação 22:

$$idf(t_i, D) = \log \left(\frac{n}{n_i} \right), \quad (22)$$

em que n é o número total de documentos em D e n_i é o número de documentos em D que contém a palavra i .

Segundo Dhillon e Modha (2001) e Aggarwal e Zhai (2012), a característica distintiva da representação de texto é que a dimensionalidade da representação de textos é alta. Já que existe um grande número de palavras em uma língua, mesmo em conjuntos de documentos de tamanho moderado, a representação gera vetores de documentos com alta dimensionalidade. Além disso, a maioria dos documentos contém muito menos palavras (1 – 5% ou menos) do que o número total de palavras da coleção completa, portanto os vetores de documentos são esparsos.

A representação esparsa e de alta dimensionalidade de um conjunto de dados textuais requer o desenho ou a proposição de algoritmos específicos para lidar com a construção da representação e processamento dos dados. Este tópico tem sido bastante estudado na literatura de recuperação de informação, na qual muitas técnicas têm sido propostas para otimizar a representação de documentos e melhorar a precisão da correspondência de um documento com sua representação no modelo de espaço vetorial (SALTON; BUCKLEY, 1988; BAEZA-YATES; RIBEIRO-NETO, 2011). Dentro destas técnicas estão a seleção de atributos, a indexação semântica latente (do inglês *Latent Semantic Indexing* - LSI), a análise semântica latente probabilística (do inglês *Probabilistic Latent Semantic Analysis* - PLSA), entre outras (AGGARWAL; ZHAI, 2012).

A discussão apresentada nesta seção motiva o desenvolvimento do presente trabalho, uma vez que o objetivo deste trabalho é analisar o comportamento dos algoritmos básicos de biagrupamento e coagrupamento no contexto de alta dimensionalidade e alta esparsidade.

2.5 Estado da arte de biagrupamento e coagrupamento em dados textuais

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 (NADIF et al., 2013). Para propósitos desta pesquisa foi realizada uma revisão sistemática (RS) para construir um panorama sobre as pesquisas mais recentes na área. Nesta seção, são apresentados os resultados produzidos pela realização da RS de acordo com o protocolo apresentado no apêndice A.

As questões de pesquisa elaboradas foram: quais são os conjuntos de dados utilizados nestes estudos? (**Q1**); quais são os algoritmos de biagrupamento e coagrupamento utilizados em dados textuais? (**Q2**); como estes algoritmos de biagrupamento e coagrupamento estão caracterizados? (**Q3**); quais são os métodos de avaliação utilizados para estes algoritmos? (**Q4**) e quais são as aplicações na área de mineração de texto que podem utilizar o biagrupamento e o coagrupamento? (**Q5**).

Os resultados são mostrados considerando a análise geral da área em relação ao problema de biagrupamento e coagrupamento e a análise dos algoritmos de biagrupamento e coagrupamento apresentados por estes estudos, mostrando os resultados estruturados de maneira a responder às questões de pesquisa levantadas no protocolo do apêndice A.

2.5.1 Uma análise geral da área em relação ao problema de biagrupamento e coagrupamento

Em mineração de textos as tarefa de biagrupamento e coagrupamento são utilizadas em uma variedade de aplicações. Devido a isso, na literatura são encontrados diversos algoritmos orientados a resolver problemas específicos, o que é bom para determinados problemas. A partir da condução do protocolo, 340 estudos foram recuperados dos quais 80 foram selecionados depois de aplicar os critérios de inclusão e exclusão, e estes são considerados nesta seção. Na tabela 3 é mostrada a lista ordenada desses estudos, junto a seus anos de publicação e identificadores a eles atribuídos de acordo com o tipo de veículo no qual cada estudo foi publicado.

Tabela 3 – Lista de estudos selecionados

Ano	Conferência (ID/Ref.)	Periódicos (ID/Ref.)
2018		J03 Xuan et al. (2018)
2018		J28 Huang, Xu e Lv (2018)
2017		J29 Costa e Ortale (2017)
2017		J30 Mei et al. (2017)
2017	C01 Rho e Pensa (2017)	J01 Wang e Huang (2017)
2017		J02 Ailem, Role e Nadif (2017)
2016	C02 Honda et al. (2016)	J04 França (2016)
2016	C03 Allab, Labiod e Nadif (2016)	J05 Govaert e Nadif (2016)
2016	C04 Salah, Rogovschi e Nadif (2016)	J06 Ailem, Role e Nadif (2016)
2016		J07 Laclau e Nadif (2016)
2016		J08 Priam, Nadif e Govaert (2016)
2016		J09 Mei et al. (2017)
2015	C05 Ailem, Role e Nadif (2015)	J10 Bao et al. (2015)
2015	C06 Cao, Huang e Zhu (2015)	J11 Buono e Pio (2015)
2015		J12 Domeniconi e Laskey (2015)
2014	C07 Li e Wu (2014)	
2014	C08 Mei et al. (2014)	
2013	C09 Garg, Chaudhari e Narang (2013)	J13 Ma, Zhao e Shi (2013)
2013	C10 Nourashrafeddin, Milios e Arnold (2013)	J14 Song et al. (2013)
2013		J15 Yan, Chen e Tjhi (2013a)
2013		J16 Liu, Chen e Lu (2013)
2013		J17 Yan, Chen e Tjhi (2013b)
2012	C11 Cheng et al. (2012)	J18 Yang e Cai (2012)
2012	C12 Franca (2012)	J19 Banu e Inbarani (2012)
2012	C13 Huang et al. (2012)	
2011	C14 Wu, Lai e Wang (2011)	
2011	C15 Yan e Chen (2011)	
2011	C16 Gu, Ding e Han (2011)	
2011	C17 Xia, Zhang e Yao (2011)	
2010	C18 Shi, Fan e Philip (2010)	J20 Yoo e Choi (2010)
2010	C19 Hussain e Bisson (2010)	J21 Castro et al. (2010)
2010	C20 Ma et al. (2010)	J22 Chen, Wang e Dong (2010)
2010	C21 Li, Wu e Lu (2010)	J23 Bichot (2010)
2010	C22 Song et al. (2010)	
2010	C23 Greene e Cunningham (2010)	
2009	C24 Park et al. (2009)	J24 Tjhi e Chen (2009)
2009	C25 Ye et al. (2009)	
2009	C26 Chen, Wang e Dong (2009)	
2009	C27 Gu e Zhou (2009)	
2009	C28 Yoo e Choi (2009)	
2008	C29 Zhang e Xing (2008)	J25 Tjhi e Chen (2008)
2008	C30 Bisson e Hussain (2008)	
2008	C31 Ando e Suzuki (2008)	
2008	C32 Giannakidou et al. (2008)	
2007	C33 Dai et al. (2007)	J26 Tjhi e Chen (2007c)
2007	C34 Tjhi e Chen (2007b)	
2007	C35 Tjhi e Chen (2007a)	
2007	C36 Castro et al. (2007b)	
2007	C37 Mimaroglu e Uehara (2007)	
2006	C38 Rege, Dong e Fotouhi (2006)	J27 Tjhi e Chen (2006c)
2006	C39 Shafiei e Milios (2006)	
2006	C40 Tjhi e Chen (2006a)	
2006	C41 Hu et al. (2006)	
2006	C42 Tjhi e Chen (2006b)	
2005	C43 Long, Zhang e Yu (2005)	
2005	C44 Greene e Cunningham (2005)	
2005	C45 Tjhi e Chen (2005)	
2003	C46 Dhillon, Mallela e Modha (2003)	
2003	C47 Mandhani, Joshi e Kummamuru (2003)	
2003	C48 Kummamuru, Dhawale e Krishnapuram (2003)	
2002	C49 Frigui e Nasraoui (2002)	
2001	C50 Dhillon (2001)	

Fonte: Alexandra Katiuska Ramos Diaz, 2018

Com respeito ao tipo de veículo onde foram publicados cada estudo, foram encontrados 50 estudos primários publicados em conferências, representando 62,5% do total, e 30 estudos primários publicados em periódicos, representando 37,5% do total de estudos.

A tabela 4 mostra a distribuição temporal dos estudos primários. Vale ressaltar que 61,25% do total de estudos foram publicados nos últimos 10 anos, indicando um crescimento da área nos últimos anos. A figura 23 apresenta um gráfico sobre o comportamento deste crescimento, considerando o total de publicações por ano, independentemente do tipo de publicação. Ainda nesta figura é possível perceber que, nos anos de 2010 e 2016 ocorreram os picos de publicação dessa área de acordo com o que foi levantado nesta pesquisa.

Tabela 4 – Resumo dos estudos primários por tipo de publicação e por ano de publicação.

	2001	2002	2003	2005	2006	2007	2008	2009	2010
Conferência	1	1	3	3	5	5	4	5	6
Periódico	-	-	-	-	1	1	1	1	4
Total	1	1	3	3	6	6	5	6	10

	2011	2012	2013	2014	2015	2016	2017	2018
Conferência	4	3	2	2	2	3	1	-
Periódico	-	2	5	-	3	6	4	2
Total	4	5	7	2	5	9	5	2

Fonte: Alexandra Katiуска Ramos Diaz, 2018

Figura 23 – Resumo dos estudos primários por ano de publicação.

Fonte: Alexandra Katiуска Ramos Diaz, 2018

2.5.2 Análise dos algoritmos de biagrupamento e coagrupamento

Para um melhor entendimento da área e para fins de responder às questões de pesquisa, os estudos analisados são classificados de acordo com: os conjuntos de dados neles utilizados, em resposta à questão **Q1**; a modalidade de aprendizado e segundo a natureza dos algoritmos propostos, ambas classificações respondem às questões **Q2** e **Q3**; as medidas de avaliação neles utilizadas, em resposta à questão **Q4**; e com respeito às suas aplicações, em resposta à questão **Q5**.

Classificação dos estudos de acordo com os conjuntos de dados neles utilizados

Nos estudos selecionados, diferentes conjuntos de dados são utilizados para realização de experimentos e análise. Na tabela 5 são mostrados os conjuntos de dados utilizados em cada um dos estudos.

Dentre os conjuntos de dados, os mais utilizados são: o conjunto de dados *20-Newsgroups*, que consiste em aproximadamente 20.000 documentos contendo discussões sobre notícias extraídas de diversos fóruns, codificados em *ASCII plain text* e organizados de maneira quase uniforme em 20 grupos de notícias cobrindo basicamente 6 tópicos distintos (LANG, 1995); a coleção *CLASSIC3*¹⁶, composta das sub-coleções MEDLINE (1033 resumos de revistas médicas), CISI (1460 resumos de documentos de recuperação de informação) e CRANFIELD (1400 resumos de sistemas aerodinâmicos); a coleção de *Reuters*¹⁷, que consiste em uma grande coleção de matérias da Reuters News para uso em pesquisa e desenvolvimento de processamento de linguagem natural, recuperação de informações e sistemas de aprendizado de máquina, sendo a coleção *Reuters-21578* a mais usada; a coleção que foi nomeada como *Sinka & Corne*, pelo nome de seus autores e que consiste em uma coleção de documentos da web com aproximadamente 11.000 documentos organizados em 11 categorias (SINKA; CORNE, 2002); o conjunto de dados *WebACE*, que contém 2340 documentos de artigos de notícias de 20 tópicos diferentes publicados em outubro de 1997, coletados no projeto *WebACE* (HAN et al., 1998).

¹⁶ <http://www.dataminingresearch.com/index.php/2010/09/classic3-classic4-datasets>

¹⁷ <https://trec.nist.gov/data/reuters/reuters.html>

Tabela 5 – Classificação dos estudos de acordo com os conjuntos de dados neles utilizados

Conjunto de dados	ID	Total
<i>20-News</i>	C01, J04, J05, C03, J06, J07, J08, C04, J09, C05, J10, J11, C07, C08, C09, C10, J14, J15, J17, C11, C12, C14, C15, C16, C18, C22, J21, J22, C25, C26, C27, C28, J24, C29, C30, C31, J25, C33, J26, C40, C42, C43, C44, C46, J30, C47, C48	47
<i>CLASSIC3</i>	J02, J04, J05, C03, J06, J08, C05, C07, C09, C10, C14, J16, J21, J22, J23, J24, C21, C24, C26, C30, J25, J26, C38, C40, J27, C42, C43, C44, C46, C47, C48, C50	32
<i>Reuters</i>	C01, J02, J05, C03, J06, C05, J11, J15, C07, C09, C13, C15, C19, C23, J20, J22, J24, C26, C29, C31, C33, C37, C38, C41, C44, C50	26
<i>Sinka & Corne</i>	J07, J10, J11, J15, J17, J24, J25, J26, J27, C02, C05, C15, C16, C27, C35, C40, C42, C45	18
<i>WebACE</i>	C03, J06, C04, J09, C05, C07, C08, J11, J22, C26, C27, C41, C21, C38, J28, J30	16
<i>CSTR</i>	J02, J05, C03, J06, J07, C05, J11, C16, J20, C27, C28	11
<i>CLUTO toolkit</i>	J02, J05, C03, C04, J06, C05, J07, C06, J11, C07, C10	11
<i>TREC</i>	J09, C08, J17, C21, C41, C44, C38, J28, J30	9
<i>Yahoo_ K1</i>	J02, J15, J17, C14, J23, J25, C47, C48, C50	9
<i>CLASSIC4</i>	J02, J05, C03, C04, J06, J07, C05, C10	8
<i>TDT2</i>	J02, J05, J06, C04, C16, J28	6
<i>MovieLens</i>	J04, J12, C09, J21	4
<i>OHSUMED</i>	C21, J22, C26, C38	4
<i>Citeseer</i>	J03, J13, C20	3
<i>DUC</i>	J18, C17, C34	3
<i>CLASSIC</i>	C44, J28	2
<i>CORA</i>	C02, J03	2
<i>DBLP</i>	J13, C20	2
<i>Jester</i>	J12, C09	2
<i>NIPS</i>	C39, J28	2
<i>SRAA</i>	J14, C33	2
<i>Syskill & Weibert (SW)</i>	J26, C45	2
<i>URCS technical reports</i>	J13, C20	2
<i>Wikipedia XML</i>	C39, J29	2
<i>BASEHOCK</i>	J01	1
<i>BBC</i>	C44	1
<i>Brazilian newspaper Folha</i>	C36	1
<i>CLASSIC2</i>	C44	1
<i>CNAE</i>	J01	1
<i>Enron</i>	J09	1
<i>MSNBC</i>	J19	1
Dados sintéticos	C02, C04, C25, C49, J03, J07	7
Dados próprios	C06, J29, C49	3

Fonte: Alexandra Katiuska Ramos Diaz, 2018

Além dos conjuntos de dados utilizados na literatura, foram encontrados estudos que utilizam conjuntos de dados sintéticos (6 estudos) e conjuntos de dados de textos próprios (3 estudos), criados especificamente para testar a qualidade dos métodos propostos.

Classificação dos estudos de acordo com a modalidade de aprendizado utilizada

Os estudos foram classificados de acordo com a modalidade de aprendizado de máquina utilizada. Esta classificação é mostrada na tabela 6.

Os estudos listados na tabela 6 foram classificados estudos que aplicam aprendizagem semi-supervisionada e que aplicam aprendizagem não-supervisionada. Os estudos que propõem algoritmos de coagrupamento utilizando uma aprendizagem semi-supervisionada utilizam algum conhecimento anterior, como rótulos de classe ou restrições de pares de

Tabela 6 – Classificação dos estudos de acordo com a modalidade de aprendizado neles utilizada

Modalidade de aprendizado de máquina	ID	Total
Aprendizagem semi-supervisionada	C03, C06, J13, J14, J15, J17, J18, C13, C15, C18, C20, C22, J22, C26, C27, C33	16
Aprendizagem não-supervisionada	J01, J02, J03, C01, J04, J05, C02, J06, J07, J08, C04, J09, C05, J10, J11, J12, C07, C08, C09, C10, J16, C11, C12, C14, C16, C17, J19, C21, J20, J21, J23, C23, C24, C25, C28, J24, C30, C31, C32, J25, C34, C35, J26, C36, C37, C38, C39, C40, C41, J27, C42, C43, C44, C45, C46, C47, C48, C49, C50, J28, J29, J30	62

Fonte: Alexandra Katiuska Ramos Diaz, 2018

objetos de dados, para ajudar no coagrupamento (MA; ZHAO; SHI, 2013). Os estudos que propõem algoritmos de coagrupamento utilizando uma aprendizagem não-supervisionada não precisam da classe associada aos exemplos, tentando determinar se alguns deles podem ser coagrupados de alguma maneira.

Adicionalmente foram encontrados dois estudos fazendo alusão à aprendizagem supervisionada. Estes estudos são os trabalhos de Hussain e Bisson (2010) e de Zhang e Xing (2008). O trabalho de Hussain e Bisson (2010) propõe métodos de classificação inspirados no coagrupamento usando aprendizagem supervisionada. O trabalho de Zhang e Xing (2008) propõe o método de coagrupar conjuntos de conhecimento (informação dos documentos) e documentos de teste. Este método permite compartilhar a informação dos grupos de palavras entre ambos coagrupamentos, de maneira a encontrar grupos de documentos a partir da informação dos grupos de rótulos.

Classificação dos estudos de acordo com sua natureza

Segundo a natureza, os estudos foram classificados de acordo com as estratégias implementadas nos algoritmos propostos. Esta classificação é mostrada na tabela 7.

Tabela 7 – Classificação dos estudos de acordo com sua natureza

Natureza	ID	Total
Lógica Fuzzy	JJ09, C08, C09, J15, J16, J17, C15, J19, J07, J24, C34, C35, J25, J26, C40, J27, C42, C45, C48, C49, J30	21
Probabilístico	JJ02, J05, J08, J10, C06, J12, J14, C11, C14, C04, J18, C17, C22, C25, C29, C31, C33, C39, C46	19
Fatoração de Matrizes	JJ01, J03, C03, J11, C07, J13, C16, C20, C21, J20, J22, C26, C27, C28, J28, J29, C43	17
Modelagem com estruturas de grafos	J06, C05, C16, C18, J23, C23, C24, C32, C38, C41, C44, C50	12
Heurístico	J04, C02, C10, C12, J15, J21, J25, C36, C37, C47	10
Medida de co-similaridade	C19, C30	2

Fonte: Alexandra Katiuska Ramos Diaz, 2018

Os estudos listados na tabela 7 foram divididos neste trabalho em seis tipos:

- estudos que propõem algoritmos baseados em métodos que usam a lógica *fuzzy*. Nesta categoria um dos estudos mais representativos é o trabalho de Kummamuru, Dhawale e Krishnapuram (2003). Os autores modificam o algoritmo *Fuzzy Clustering for Categorical Multivariate Data* – FCCM (OH; HONDA; ICHIHASHI, 2001), maximizando a co-ocorrência de documentos e palavras (palavras-chave) nos grupos. Esta modificação permite a utilização do algoritmo no coagrupamento de conjuntos de dados de textos grandes, que é uma desvantagem do algoritmo original;
- estudos que propõem algoritmos baseados em métodos probabilísticos. Nesta categoria, os trabalhos de Dhillon, Mallela e Modha (2003) e Domeniconi e Laskey (2015) são os mais conhecidos. Dhillon, Mallela e Modha (2003) consideram a representação de documentos e palavras como uma tabela de contingência com distribuição de probabilidades conjuntas entre duas variáveis aleatórias discretas, tomando valores sobre as linhas e colunas. Ainda os autores consideram que o coagrupamento ótimo é aquele que leva a maior informação mútua entre as variáveis aleatórias agrupadas, minimizando a perda na informação mútua entre as variáveis aleatórias originais e as variáveis aleatórias agrupadas. O trabalho de Domeniconi e Laskey (2015), consideram os modelos bayesianos para o coagrupamento, desenvolvendo um modelo para o processo de geração de dados. Este modelo é invertido através de inferência bayesiana para inferir a associação de grupos, aprender características dos grupos e preencher observações ausentes;
- estudos que propõem algoritmos baseados em métodos de fatoração de matrizes, que modelam o problema de coagrupamento como uma matriz de dados diádicos. Nesta categoria o trabalho seminal foi o trabalho de Long, Zhang e Yu (2005);
- estudos que propõem algoritmos baseados em modelagem com estruturas de grafos, que modelam a coleção de documentos como um grafo bipartido de documentos e palavras. O trabalho mais representativo é apresentado por Dhillon (2001);
- estudos que apresentam algoritmos baseados em métodos heurísticos. Nesta categoria está o trabalho de Castro et al. (2007b), que consideram a geração de bigrupos como um problema de otimização combinatória multimodal, obtidos por meio do algoritmo de biagrupamento de inspiração imunológica (BIC-aiNet). Ainda no trabalho é mostrada a utilidade do algoritmo em conjunto de dados textuais. Adicionalmente,

nesta categoria está o trabalho de Cheng e Church (2000), mas não foi considerado nesta análise, por se concentrar na área de expressão gênica;

- estudos que apresentam algoritmos baseados na medida de co-similaridade. Nesta categoria está o trabalho de Bisson e Hussain (2008), que apresentam a medida de co-similaridade denominada como ‘*X-Sim*’. Esta medida está baseada na ideia de gerar simultaneamente as matrizes de similaridades entre linhas e colunas, sendo cada uma delas construída iterativamente na base da outra.

Classificação dos estudos de acordo com as medidas de avaliação neles utilizadas

Em relação aos métodos de avaliação foi constatado que os estudos avaliam seus resultados principalmente usando medidas de avaliação de agrupamento, e foi por este motivo que as medidas de avaliação expostas na seção 2.3 foram escolhidas para serem descritas. A tabela 8 mostra as medidas utilizadas nos estudos, sendo as medidas externas as favoritas na área. A seguir são descritas as medidas mais utilizadas em cada classificação.

Entre as medidas externas, as mais utilizadas são: a informação mútua normalizada (do inglês *Normalized Mutual Information* - NMI), que fornece uma indicação das informações compartilhadas entre os rótulos de grupos atribuídos ao resultado do agrupamento produzido pelo algoritmo e os rótulos verdadeiros já existentes representando conhecimento a priori (COVER; THOMAS, 2012); a acurácia (do inglês *Accuracy* - ACC), que descobre a relação um-a-um entre os grupos reais, definidos como a solução esperada para o problema, e o agrupamento gerado por um algoritmo (YAN; CHEN; TJHI, 2013b); a pureza (do inglês *Purity*), que avalia a coerência de um grupo, isto é, o grau em que um grupo contém documentos de uma única categoria (HUANG, 2008); a medida de precisão e revocação (do inglês *Precision and Recall*), em que a precisão (também chamada de preditividade positiva) é a fração de documentos recuperados que são relevantes para a consulta e a revocação (também conhecida como sensibilidade) é a fração dos documentos relevantes para a consulta que são recuperados com êxito (TJHI; CHEN, 2007c); a medida de *f-measure*, que combina a precisão e revocação para refletir a qualidade geral dos grupos resultantes (YAN; CHEN; TJHI, 2013b); e o índice de Rand ajustado (do inglês *Adjusted Rand Index* - ARI), que mede a qualidade entre duas partições (HUBERT; ARABIE, 1985). Os valores da informação mútua normalizada, acurácia, pureza, o índice rand ajustado

Tabela 8 – Classificação dos estudos de acordo com as medidas de avaliação neles utilizadas

Medidas de avaliação	ID	Total	Medida
Informação Mútua Normalizada	J01, J02, J04, J05, C03, J06, J07, C04, J09, C05, J10, C06, J11, J13, C10, J14, J15, J16, C13, C16, C18, C20, C22, J20, C23, C27, C30, C41, C44, J30	30	Externa
Acurácia	J01, J02, J05, C03, J06, J07, C05, J10, J11, J13, J15, J17, C15, C16, C19, C20, J20, J22, C23, C24, C26, C27, C28, C43, C48, J28	26	
Pureza	J04, C02, C06, C07, J16, C18, C21, J23, J24, J25, C35, J26, J27, C42, C45, C47, C50, J28, J29	19	
Precisão e Revocação	C09, J17, J24, C31, C32, J25, C35, J26, C39, C40, J27, C41, C42, C45, C46	15	
<i>F-measure</i>	J03, J09, C09, C10, C13, J23, C29, C32, C39, C41, J17, J30	12	
Índice rand ajustado	J02, C01, C03, J07, C04, J09, C06, C08, J23, J30	10	
<i>Micro-average precision</i>	C11, C12, C14, C17, C25, C30, C43, C46	8	
<i>Confusion matrix</i>	C25, C38, C41, J27, C46, C47, C49	7	
Entropia	C07, J16, C21, C41, C50	5	
<i>Error rate</i>	J08, C29, C33, C41	4	
<i>Perplexity</i>	J12, C39	2	
Verificação dos rótulos dos textos de cada bicluster	C36, J21	2	
Índice de Jaccard	J03	1	
Folkes & Mallows	J03	1	
<i>Pointwise Mutual Information</i> (PMI)	J04	1	
<i>Root Mean Square</i> (RMSE)	J19	1	
<i>Mean Absolute Error</i> (MAE)	J19	1	
ROUGE	J18, C17, C34	3	
Índice Davies–Bouldin	J08	1	Interna
<i>Intra-cluster average similarity</i>	J11	1	
<i>Inter-cluster centroids average similarity</i>	J11	1	
Resíduo quadrático médio	J21, C36	2	
Volume, sobreposição, ocupação, cobertura	J21	1	
Tempo de execução	C07, J15, C18, C37	4	
Complexidade do tempo	J15	1	
Uso da memória	C07	1	
Inspeção visual	C36	1	

Fonte: Alexandra Katiuska Ramos Diaz, 2018

variam de 0 a 1, sendo que um valor próximo a 1 significa um bom resultado em termos de qualidade de agrupamento.

No conjunto de medidas internas estão aqueles próprias para validação de agrupamento, tais como: o índice de Davies Bouldin, a medida de similaridade média intra-grupos e a medida de similaridade média dos centroides inter-grupos descritos na seção 2.3.

Além dessas medidas alguns estudos aplicam medidas próprias de biagrupamento, tais como: o resíduo quadrático médio, o volume, a cobertura do conjunto de dados, o grau de sobreposição, o grau de ocupação, todas estão descritas na seção 2.3.

Adicionalmente, alguns estudos medem a qualidade dos resultados do biagrupamento e coagrupamento em termos de tempo de execução, complexidade, uso da memória e

usando uma inspeção visual, que permite visualizar graficamente a qualidade dos bigrups gerados destacando a compactação dos dados (CASTRO et al., 2007b).

Classificação dos estudos de acordo com suas aplicações

Para analisar as aplicações, os estudos foram classificados de acordo com as tarefas utilizadas em mineração de texto. Esta classificação é mostrada na tabela 9.

Tabela 9 – Classificação dos estudos de acordo com as tarefas utilizadas em mineração de texto

Tarefas		ID	Total
Agrupamento	Texto	J01, J02, J03, C01, J04, J05, C02, C03, J06, J07, J08, C04, C05, J10, C06, J11, J12, C07, J13, C10, J14, J16, C11, C12, C13, C14, C16, C18, C20, C21, J19, C22, J20, J21, J22, J23, C23, C24, C25, C27, C28, C29, C30, C31, C32, C33, J26, C36, C37, C38, C39, C40, C41, J27, C43, C44, C45, C46, C47, C48, C50, J28, J29, J30	64
	Dados de expressão gênica	J01, J21, J22	3
	Imagens	J01, J22, C27	3
Classificação	Texto	J04, J09, C08, C09, J15, J17, C15, C19, C26, J24, J25, C33, C35, C42, C49	15
Sumarização	Texto	J18, C17, C34	3

Fonte: Alexandra Katiuska Ramos Diaz, 2018

Nos estudos listados na tabela 9 foram encontrados três trabalhos que utilizam a tarefa sumarização, 15 trabalhos que classificam ou categorizam documentos e 64 trabalhos focados na tarefa de agrupamento. Dentre os estudos que utilizam a tarefa de agrupamento estão os trabalhos que agrupam texto (resumos médicos, repórteres de notícias, documentos de publicações científicas, etc.) e os trabalhos que além de agrupar texto mostram a efetividade de seus métodos propostos em conjuntos de expressão gênica (três estudos) e imagens (três estudos).

Além da classificação da tabela 9, foram encontrados alguns estudos que aplicam coagrupamento para aplicações específicas. Estes estudos são mostrados na tabela 10.

Tabela 10 – Classificação dos estudos de acordo com suas aplicações

Aplicação	ID	Total
Expansão de consulta	J21	1
Modelagem de tópicos	J04, C31	2
Sistemas de recomendação	J12, C09, J04	3
Mineração da web	J19, C23, C32, J29	4

Fonte: Alexandra Katiuska Ramos Diaz, 2018

De acordo com o conteúdo da tabela 10, percebe-se que os estudos podem ser divididos em quatro tipos:

- Expansão de consulta (do inglês *query expansion*): coagrupa documentos e termos com a finalidade de reformular a consulta inicial enviada pelo usuário. Adicionando termos representativos e relacionados, provenientes do cogrupo de documentos ao qual pertencem os termos da consulta inicial (CASTRO et al., 2010).
- Modelagem de tópicos (do inglês *topic modelling*): processo de encontrar os termos de um documento que mais descrevem seus tópicos (descoberta de estruturas semânticas ocultas em um corpo de texto). Normalmente, um documento diz respeito a vários tópicos em diferentes proporções. É por esta razão que as análises do coagrupamento tornam-se úteis, permitindo descobrir os possíveis tópicos de cada grupo de documento (FRANÇA, 2016).
- Sistemas de recomendação: processo de agrupar clientes e produtos simultaneamente. Nestas aplicações, os clientes do mesmo grupo de clientes apresentam padrões semelhantes de compra no mesmo grupo de produtos (DOMENICONI; LASKEY, 2015).
- Mineração da web: além dos sistemas de recomendação existem outras aplicações na mineração da web que foram classificadas com o nome geral de “mineração da web”. Dentre estas aplicações estão a análise dos logs do servidor da web, a análise da formação e evolução de padrões nos dados ao longo do tempo, a captura das relações sociais entre os membros de uma comunidade e a descoberta de padrões em volumes de documentos em formato XML. Estas aplicações estão descritas a seguir.
 - Na análise dos logs do servidor da web, o coagrupamento descobre grupos de usuários que exibem classificações altamente correlacionadas em grupos de páginas (BANU; INBARANI, 2012).
 - Na análise da formação e evolução de padrões nos dados ao longo do tempo, o coagrupamento agrupa objetos (*bookmarks*) e recursos (*tags*) para identificar grupos ao longo do tempo. Algumas das aplicações utilizadas são: rastrear comunidades em desenvolvimento de usuários de redes sociais, identificar círculos de usuários com interesses compartilhados em sites de mídia social ou mapear o desenvolvimento de notícias individuais ao longo do tempo (GREENE; CUNNINGHAM, 2010).

- Na captura das relações sociais entre os membros de uma comunidade, o coagrupamento agrupa simultaneamente recursos e tags de dados sociais (GI-ANNAKIDOU et al., 2008).
- Na descoberta de padrões em volumes (geralmente muito grandes) de documentos em formato XML, o coagrupamento é útil em aplicações como processamento de consultas, integração e indexação de dados, recuperação e extração de informações, filtragem de documentos e navegação (COSTA; ORTALE, 2017).

A partir da análise desses estudos foi observado que os diferentes algoritmos de coagrupamento são propostos baseados nas técnicas utilizadas no agrupamento, tais como a fatoração de matrizes, a lógica *fuzzy*, os modelos probabilísticos, entre outros. Estudos que explicitamente citem algoritmos de biagrupamento são em número reduzido, fazendo um destaque aos trabalhos de Mimaroglu e Uehara (2007), Castro et al. (2007b), Banu e Inbarani (2012), Castro et al. (2010).

Os estudos analisados partem da premissa de que o biagrupamento e coagrupamento podem trazer resultados melhores comparado com os resultados do agrupamento em determinados contextos. Assim estes estudos propõem métodos que são validados usando conjuntos de dados e métricas de agrupamento e biagrupamento. Além disso, estes métodos podem ser aplicados em diferentes áreas da mineração de dados, mostrando suas vantagens e desvantagens na sua aplicação.

2.6 Considerações finais

Neste capítulo foram apresentadas as definições referentes à área de biagrupamento e coagrupamento, assim como foram estudados em detalhe um algoritmo básico para cada uma dessas tarefas. Estes algoritmos possuem versões aprimoradas que visam melhorar alguns de seus aspectos.

O exemplo apresentado nas seções 2.2.1 e 2.2.2 mostra a variabilidade dos bigrupos obtidos pelo biagrupamento, aplicando o algoritmo de Cheng e Church e os cogrupos obtidos pelo coagrupamento, aplicando o algoritmo de NBVD, em um mesmo conjunto de dados. Ambos algoritmos apresentam características específicas que os diferenciam e que lhes atribuem vantagens e desvantagens.

Quanto ao algoritmo de Cheng e Church destaca-se que ele apresenta uma natureza de encontrar um bigrupo por vez. Além disso, é essencialmente um método guloso, e como método guloso, ele faz uma escolha ótima para as condições locais, na esperança de que essa escolha leve a uma solução ótima para a situação global (CORMEN et al., 2009). Outra desvantagem que apresenta é a adição do ruído na busca dos próximos bigrupos, que pode alterar negativamente a descoberta dos próximos bigrupos.

Com respeito ao algoritmo de NBVD, destaca-se que ele faz as descobertas dos cogrupos simultaneamente, porém requer um pós-processamento nas matrizes U , S e V para a obtenção dos cogrupos resultantes. Além disso, o algoritmo de NBVD não é capaz de garantir a convergência para um mínimo global do problema de minimização, sendo possível encontrar diversas soluções em execuções diferentes do algoritmo, devido à inicialização de forma aleatória.

Nas próximas seções serão estudados e analisados o comportamento destes algoritmos em conjuntos de dados sintéticos e reais.

3 Método para experimentação

Esse trabalho tem como objetivo principal apresentar uma análise do comportamento do algoritmo para biagrupamento Cheng e Church e do algoritmo de coagrupamento NBVD, aplicado ao contexto de dados textuais. O presente capítulo apresenta o método de experimentação usado para construção do ambiente de experimentação, incluindo: a descrição dos conjuntos de dados sintéticos (Seção 3.1); a descrição do conjunto de dados reais e o pré-processamento aplicado sobre eles (Seção 3.2); uma visão geral dos aspectos de estudo explorados na experimentação e das medidas de avaliação aplicadas (Seção 3.3), o método estabelecido para experimentação com o algoritmo de Cheng e Church (Seção 3.4); o método estabelecido para experimentação com o algoritmo de NBVD (Seção 3.5); as configurações das experimentações para ambos algoritmos (Seção 3.6) e as considerações finais do capítulo (Seção 3.7).

3.1 Conjuntos de dados sintéticos

Os conjuntos de dados sintéticos são baseados nas diferentes estruturas de bigrupos ¹, propostas no trabalho de Madeira e Oliveira (2004) e descritas na seção 2.1.1.

Foram escolhidas as nove estruturas apresentadas pelos autores e mostradas graficamente na figura 24, com o objetivo de realizar um estudo amplo do comportamento dos algoritmos de Cheng e Church e de NBVD, criando um ambiente de experimentação controlado com diferentes níveis de esparsidade. A descrição destes conjuntos de dados estão baseadas no texto de Brunialti (2016). Considere que cada um dos quadrados maiores representa um conjunto de dados sintéticos, cujo tamanho real está representado pela altura (número de dados) e largura (número de atributos) e cada quadrado ou retângulo dentro de cada conjunto de dados, representados com diferentes tons da cor azul, representam um bigrupo da matriz original.

Os conjuntos de dados possuem 150 dados (linhas) e 150 atributos (colunas), sendo que as diferentes tonalidades da cor azul revelam a similaridade entre os dados de cada subconjunto de dados, e a intensidade da cor é proporcional à intensidade dos valores associados a cada atributo em cada dado, ou seja valores maiores são representados por tonalidades mais escuras e vice-versa. Todos estes conjuntos de dados são matrizes de

¹ Neste trabalho os conjunto de dados sintéticos são também considerados sob a visão de coagrupamento.

Figura 24 – Conjunto de dados sintéticos baseados nas diferentes estruturas de bigrupos

Fonte: Alexandra Katiuska Ramos Diaz, 2018

valores reais positivos gerados artificialmente, preenchidos em primeira instância com valores aleatórios uniformes no intervalo $[0, 1]$. Para a criação dos bigrupos, os tamanhos de número de linhas e colunas foram definidos de acordo com cada estrutura desejada. Para instanciar cada bigrupo, um conjunto de valores foi criado e distribuído pelas células do bigrupo da seguinte forma:

- um valor central $c \in O$, sendo $O = \{20, 40, 60, 80, 100, 120, 140, 160, 180\}$, foi aleatoriamente escolhido e retirado de O com o objetivo de assegurar que não houvesse bigrupos com o mesmo c em uma mesma matriz;
- o conjunto de valores usados para instanciar as células do bigrupo foi estabelecido mediante a adição da equação 23:

$$a_{ij} = c + \mathcal{U}(0, 1), \quad (23)$$

em que, considerando que um bigrupo é uma submatriz da matriz original de dados $\mathcal{A}$, a_{ij} pode ser considerado como o elemento que pertence à linha i e à coluna j do bigrupo da matriz $\mathcal{A}$; sendo i e j os índices das linhas e colunas escolhidos para compor esse bigrupo. Então os bigrupos serão preenchidos com a adição c e a distribuição uniforme $\mathcal{U}(0, 1)$.

Segundo a visão da tarefa de coagrupamento, mais cogrupos podem ser observados nos conjuntos de dados sintéticos, além daqueles que são de interesse nas análises neste trabalho. Esse cogrupos a mais estão coloridos em tons de azul claro. Todos os cogrupos podem ser visualizados na figura 25.

Figura 25 – Visão de coagrupamento nos conjuntos de dados sintéticos

Fonte: Alexandra Katiuska Ramos Diaz, 2018

Sob essa visão tem-se as seguintes organizações de cogrupos, seguindo a disposição de conjuntos usada na figura 25:

- *a*: um cogrupos de interesse principal e dois cogrupos não destacados na figura;
- *b*: três cogrupos de interesse principal e três cogrupos não destacados na figura;
- *c*: nove cogrupos de interesse principal;
- *d*: três cogrupos de interesse principal e três cogrupos não destacados na figura;
- *e*: três cogrupos de interesse principal e três cogrupos não destacados na figura;
- *f*: quatro cogrupos de interesse principal e um cogrupos não destacados na figura;
- *g*: quatro cogrupos de interesse principal e um cogrupos não destacados na figura;
- *h*: cinco cogrupos de interesse principal e quatro cogrupos não destacados na figura;
- *i*: quatro cogrupos de interesse principal e cinco cogrupos não destacados na figura.

3.2 Conjunto de dados reais - corpus EBC

O conjunto de dados reais utilizado neste trabalho foi o corpus EBC, elaborado no trabalho de Pagnossim (2018). Nas próximas subseções são detalhadas a descrição deste corpus e o pré-processamento realizado sobre ele.

3.2.1 Descrição do conjunto de dados

O corpus EBC é composto de uma coleção de textos extraídos do portal de notícias da Empresa Brasil de Comunicação (EBC)². Este corpus apresenta 1157 notícias obtidas no período de abril de 2017 até setembro de 2017, as quais estão organizadas em dez canais, conforme mostrado na tabela 11.

Tabela 11 – Organização do corpus EBC

Id	Canal da notícia	Total de notícias
1	Cidadania	96
2	Cultura	124
3	Educação	257
4	Esportes	200
5	Infantil	14
6	Política	86
7	Tecnologia	125
8	Economia	41
9	Internacional	89
10	Geral	125

Fonte: Adaptado de Pagnossim (2018)

O formato das notícias é de arquivo de texto, facilitando sua manipulação nos procedimentos de pré-processamento de texto. Um exemplo do conteúdo de uma notícia é o mostrado na figura 26.

3.2.2 Pré-processamento

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 (DIAZ et al., 2018).

A implementação do pré-processamento foi feita na linguagem R (R Development Core Team, 2008), e os procedimentos foram baseados no relatório técnico dos autores DIAZ et al. (2018). No referido relatório são apresentados os detalhes técnicos destes procedimentos. Os principais procedimentos aplicados foram:

- A preparação dos dados: conversão dos textos em letras minúsculas, a remoção de números, a remoção de pontuações e remoção do caracteres de Twitter (@, #).

² <http://www.ebc.com.br>

Figura 26 – Exemplo do conteúdo de uma notícia do corpus EBC

```

<channelcode>8</channelcode>

<newstitle>Temer comemora saques do FGTS e diz que preocupação é recuperar empregos</
newstitle>

<newsauthor>Marcelo Brandão – Repórter da Agência Brasil</newsauthor>

<newssubtitle></newssubtitle>

<newstext>O presidente Michel Temer disse, em vídeo divulgado hoje (1º), que a sua principal
preocupação é reduzir o desemprego no país, recuperando “os empregos perdidos”. O vídeo foi
divulgado na sua conta de Twitter no final da tarde. “Tenho como preocupação central o resgate
dos empregos perdidos e vou lutar por esses empregos até o último dia do meu governo”, disse.

O vídeo foi feito um dia após o término do prazo para o saque das contas inativas do Fundo de
Garantia do Tempo de Serviço (FGTS). O presidente comemorou os resultados da iniciativa do
governo de liberar o dinheiro para o trabalhador.

“Foi um sucesso. Aproximadamente R$ 43 bilhões repassados aos contribuintes, mais de 25
milhões de brasileiros beneficiados. Uma ação criativa e que deu um extraordinário resultado”,
disse. “Entrava governo, saía governo e ninguém nunca pensou nisso. Nunca ninguém fez nada.
Ficava aquele dinheirão parado, um dinheiro que era do trabalhador brasileiro”, completou.

O presidente ressaltou ainda que “não fez mais que sua obrigação” ao liberar o dinheiro das
contas inativas. “Pensei na necessidade que cada um desses 25 milhões de trabalhadores e
trabalhadores tinha em saldar suas dívidas, comprar algum bem, visitar um familiar ou poupar
onde e como quisesse. Não fiz mais do que a minha obrigação. E é isso que me motiva a pensar
em novas medidas para facilitar a sua vida, como essa por exemplo”.

O prazo para saque das contas inativas terminou ontem (31). A partir de hoje (1º), será
preciso comprovar, nas agências da Caixa Econômica Federal, doença impeditiva à locomoção ou
reclusão em regime fechado que impediu o saque no prazo.</newstext>

```

Fonte: Alexandra Katiuska Ramos Diaz, 2018

- A tokenização: extração de unidades mínimas de texto (*token*). No trabalho foi utilizado o espaço em branco como o caractere delimitador na separação dos *tokens*.
- A remoção de *stopwords*: consistiu em remover *tokens* que não possuem valor para o contexto, sendo úteis apenas para compreensão geral do texto. Neste trabalho, além de usar a lista de *stopwords* da linguagem R, foram adicionadas algumas palavras comuns no contexto do corpus EBC (*ebc*, *rádio*, *agência*, *brasil*) e palavras utilizadas na estrutura de armazenamento (*channelcode*, *newssubtitle*, *newstext*, *newstitle*, *newsauthor*).
- Para a representação vetorial dos textos foram usados três esquemas: a representação binária, a representação TF e a representação TF-IDF. Nas representações TF e TF-IDF foi usada a normalização norma-2, ou seja, a norma das colunas é igual a 1. Em cada uma das representações foram consideradas: as palavras com ocorrência em mais de dois documentos, pois as palavras que aparecem em apenas um documento não possuem informações no agrupamento de múltiplos documentos (YOO; CHOI, 2010); e as *s* melhores pontuações em termos de TF-IDF, em que $s = 5, 10, 15, 20$.

Sendo assim, nos experimentos foram utilizadas três formas de representação de documentos combinadas com cinco formas de seleção das características, resultando em um total de quinze configurações diferentes. Estas configurações estão relatadas na tabela 12 junto com um identificador equivalente a cada uma delas. A tabela ainda traz informação sobre o número de documentos e palavras totais em cada matriz de dados.

Tabela 12 – Conjunto de dados reais

Id	Representação	# Documentos	# Palavras
Mat. 1	Binária	1157	12422
Mat. 2		1157	3324
Mat. 3		1157	5943
Mat. 4		1157	8379
Mat. 5		1157	10628
Mat. 6	TF	1157	12422
Mat. 7		1157	3324
Mat. 8		1157	5943
Mat. 9		1157	8379
Mat. 10		1157	10628
Mat. 11	TF-IDF	1157	12422
Mat. 12		1157	3324
Mat. 13		1157	5943
Mat. 14		1157	8379
Mat. 15		1157	10628

Fonte: Alexandra Katiuska Ramos Diaz, 2018

3.3 Estratégia de estudo geral

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.

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 24(i)). 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 22500, um conjunto de dados esparsos com um nível de esparsidade de 10 significa que 10% do número total de células são valoradas com zero, ou seja, para o exemplo, os valores de 2250 células são substituídos por zeros de forma aleatória. Os níveis de esparsidade considerados neste trabalho variaram desde 0 até 98, com passos de tamanho de 10. Os

níveis de esparsidade considerados são: 0, 10, 20, 30, 40, 50, 60, 70, 80, 90, 98. Na figura 27, versões esparsas do conjunto de dados da estrutura i são apresentadas.

Figura 27 – Efeito da inserção de esparsidade no conjunto de dados da estrutura i , com níveis de esparsidade de 0, 30, 50, 70 e 90

Fonte: Alexandra Katiuska Ramos Diaz, 2018

O conjunto de dados reais, independente da forma de representação de dados usada, já apresenta alta esparsidade. O cálculo da porcentagem da esparsidade foi realizado seguindo a equação 24, a qual foi baseada no trabalho de Brunialti (2016).

$$Esparsidade = 100 - \frac{100 \times w}{n \times m} \quad (24)$$

em que w é o número total de palavras, n é o número de documentos e m o número total de palavras existentes nos documentos. Adicionalmente, $n \times m$ é equivalente ao tamanho da matriz de dados.

A porcentagem de esparsidade depende da ocorrência das palavras consideradas nas representações (w). Na tabela 13, são mostrados os níveis de esparsidade de cada matriz de dados pré-processadas.

Tabela 13 – Níveis de esparsidade das matrizes de dados extraídas do corpus EBC

Id	# de Documentos	# de palavras únicas	# Total de palavras	Esparsidade (%)
Mat. 1, Mat. 6, Mat. 11	1157	12422	149538	98,96
Mat. 2, Mat. 7, Mat. 12	1157	3324	48842	98,73
Mat. 3, Mat. 8, Mat. 13	1157	5943	78412	98,86
Mat. 4, Mat. 9, Mat. 14	1157	8379	100426	98,96
Mat. 5, Mat. 10, Mat. 15	1157	10628	115852	99,06

Fonte: Alexandra Katiuska Ramos Diaz, 2018

Para avaliação dos resultados dos algoritmos de Cheng e Church e do NBVD, quatro estratégias foram usadas: a aplicação das medidas de biagrupamento (média dos volumes, média dos RQMs, porcentagem de cobertura, porcentagem do grau de ocupação e porcentagem do grau de sobreposição); a aplicação das medidas internas de agrupamento (distância inter-grupos, distância intra-grupos e índice DB) calculados sobre a projeção dos bigrupos/cogrupos nas linhas dos conjuntos de dados; inspeção visual sobre representações

gráficas de matrizes reconstruídas em relação aos conjuntos de dados sintéticos; e avaliações qualitativas quando aplicados no corpus EBC.

As avaliações qualitativas estão baseadas na aplicabilidade dos algoritmos de Cheng e Church e do NBVD no domínio de mineração de texto. Esta avaliação é importante para determinar se os algoritmos são capazes de encontrar tópicos (grupos de palavras) e possivelmente explicar os grupos de notícias descobertos.

3.4 Método para experimentação: Cheng e Church

O algoritmo de Cheng e Church recebe três parâmetros como entrada, além do conjunto de dados para análise: 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 de estudo do presente trabalho, também é importante levar em consideração diferentes níveis de esparsidade que podem estar presentes nos dados. A estratégia de estudo adotada segue o disposto no esquema gráfico da figura 28, o qual mostra o fluxo de ações tomadas.

O algoritmo de Cheng e Church 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 EBC. As experimentações nos conjuntos de dados sintéticos nas suas versões originais estão relatados no *Exp.#1* e *Exp.#3*. As experimentações nos conjuntos de dados sintéticos nas versões com os diferentes níveis de esparsidade estão relatados no *Exp.#2* e *Exp.#4*. As experimentações nas matrizes de dados extraídas do corpus EBC estão relatadas no *Exp.#5* e *Exp.#6*. Estas experimentações estão descritas da seguinte forma:

- *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

Figura 28 – Esquema gráfico de estudo aplicado ao algoritmo Cheng e Church

Fonte: Alexandra Katiuska Ramos Diaz, 2018

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 EBC.
- *Exp. #6* - estudar o comportamento do algoritmo aplicado ao corpus EBC, 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 conhecidos dos corpus, respectivamente.

Para avaliação dos resultados de biagrupamento nos conjuntos de dados sintéticos foram consideradas as medidas de biagrupamento, 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 nas matrizes extraídas do corpus EBC 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.

3.5 Método para experimentação: NBVD

O algoritmo de NBVD recebe como parâmetros de entrada o conjunto de dados para análise, os valores considerados como critérios de parada (número máximo de iterações e a diferença do erro de reconstrução em duas iterações consecutivas), e os parâmetros k e l . A estratégia de estudo adotada está detalhada na figura 29.

Figura 29 – Esquema gráfico de estudo aplicado ao algoritmo NBVD

Fonte: Alexandra Katiuska Ramos Diaz, 2018

O algoritmo de NBVD foi aplicado nos conjuntos de dados sintéticos referentes às estruturas de bigrupos $(a, b, c, d, e, f, g, h, e, i)$ e no corpus EBC. 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 diferentes níveis de esparsidade). 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 EBC estão relatadas no *Exp. #3*. Estas experimentações são descritas da seguinte forma:

- *Exp. #1* - estudar os parâmetros k e l , e seu impacto nos resultados obtidos quando é aplicado nos conjuntos de dados sintéticos sem esparsidade.
- *Exp. #2* - estudar o comportamento do algoritmo conforme os níveis de esparsidade aumentam, quando aplicados aos conjuntos de dados sintéticos com diferentes esparsidades. Nesta experimentação foram considerados os valores para k e l que retornaram resultados satisfatórios do *Exp. #1*.

- *Exp. #3* - estudar o comportamento do algoritmo aplicado ao corpus EBC, 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 as medidas de biagrupamento, medidas internas de agrupamento na projeção nas linhas dos cogrupos, analisando o erro de reconstrução e realizando uma análise visual. 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 da mesma forma que o problema proposto pelo algoritmo 2. 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 EBC 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.

3.6 Configurações para experimentação

Tanto o algoritmo de Cheng e Church como o algoritmo de NBVD foram implementados usando a linguagem Python³. As experimentações dos algoritmos nos conjuntos de dados sintéticos foram executados em uma máquina de uso pessoal, com processador 1.8 GHz Intel Core i5, com 8 GB de memória RAM, e sistema operacional MacOS Sierra.

Com respeito às experimentações no conjunto de dados reais, eles foram executadas em dois servidores do Laboratório de Pesquisa em Inteligência Artificial (GrIA), sendo suas especificações:

- servidor 1 com processador de 2.0 GHz Intel Xeon(R), memória RAM de 7,8 GB e sistema operacional Ubuntu 16,04 LTS;
- servidor 2 com processador de 2.53GHz Intel Xeon(R), memória RAM de 31,4 GB e sistema operacional Ubuntu 16,04 LTS.

Para execução dos experimentos com o algoritmo de NBVD os valores dos critérios de parada foram de 1000 para o número máximo de iterações e 1×10^{-4} para a diferença

³ <https://www.python.org/>

do erro de reconstrução em duas iterações consecutivas (ζ). Os parâmetros k e l foram escolhidos baseados nas análises relatadas na seção 5.1. Da mesma forma, a configuração dos parâmetros do algoritmo de Cheng e Church estão detalhados na seção 4.1.

3.7 Considerações finais

Este capítulo teve como finalidade apresentar de forma detalhada como foram realizados os estudos dos algoritmos considerando: características dos conjuntos de dados sintéticos e reais; estratégia dos estudos em ambientes sem e com esparsidade; medidas de avaliação utilizadas. Além disso foram mostrados esquemas gráficos indicando os fluxos de ações executadas nas experimentações para os algoritmos estudados. Finalmente, foram descritas as configurações técnicas dos experimentos.

4 Resultados e Análises do algoritmo de Cheng e Church

Neste capítulo são apresentados os resultados obtidos a partir da realização dos experimentos com o algoritmo de Cheng e Church, considerando os conjuntos de dados sintéticos e o corpus EBC. Juntamente com a apresentação dos resultados, são tecidas análises. O capítulo é finalizado com um resumo das conclusões obtidas a partir das análises dos resultados.

A organização do capítulo segue a seguinte estrutura: na seção 4.1 são apresentadas as análises do comportamento do algoritmo em diferentes níveis de esparsidade presentes no conjuntos de dados sintéticos, na seção 4.2 são apresentados os resultados da aplicação do algoritmo sobre o corpus EBC junto com as análises obtidas, e na seção 4.3 são apresentadas as considerações finais sobre algoritmo.

4.1 Estudo do algoritmo de Cheng e Church

O estudo do algoritmo tem como objetivo analisar o seu comportamento em um ambiente sem esparsidade, aplicando-o ao conjunto de dados sintéticos em sua forma original e em um ambiente com esparsidade artificial, ou seja, aplicando-o aos conjuntos de dados sintéticos com a inserção de diferentes níveis de esparsidade.

4.1.1 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 como:

- Margem para o RQM de linhas e colunas no bigrupo (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 no procedimento de Remoção-Simples, uma vez que dificulta a retirada de linhas ou colunas. Além disso, a redução do RQM do bigrupo permite mais iterações. Por outro lado, um valor pequeno para α implica na remoção de muitas linhas ou colunas no procedimento de Remoção-

Multiplos-Nós. Assim, é interessante manter α tão alto quanto for possível, porém sem deixar a execução do algoritmo ineficiente. Os proponentes do algoritmo usam $\alpha = 1, 2$ (CHENG; CHURCH, 2000). 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 bigrupo (parâmetro δ): uma heurística sugerida por Cheng e Church (2000) foi aplicada para a determinação do espaço de busca para o δ . Nessa heurística, o valor do RQM mínimo entre os RQMs de cada bigrupo 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. 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 30, e a tabela 14 com o espaço de busca de δ para esse caso.

Figura 30 – Bigrupos com linhas exclusivas (estrutura d)

Fonte: Alexandra Katiuska Ramos Diaz, 2018

Tabela 14 – Espaço de busca para o parâmetro δ considerando o conjunto de dados com a estrutura d , ilustrado na figura 30

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

Fonte: Alexandra Katiuska Ramos Diaz, 2018

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 15, 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 .

Tabela 15 – 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

Fonte: Alexandra Katiuska Ramos Diaz, 2018

- Número de bigrupos (parâmetro K): este parâmetro assume valores correspondentes ao número de bigrupos conhecido para o conjunto de dados no *Exp. #1* e, para variação no *Exp. #3*, o parâmetro assume valores que variaram de 1 até no máximo o número de cogrupos 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 16.

Tabela 16 – Espaço de busca para o parâmetro K

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

Fonte: Alexandra Katiuska Ramos Diaz, 2018

4.1.2 *Exp. #1*: estudo do parâmetro α

O estudo do parâmetro α está baseado nos objetivos do algoritmo de Cheng e Church, descritos na seção 2.2.1. No presente estudo foi analisada a implicação dos diferentes valores do parâmetro α no tempo de execução do algoritmo, o volume, e no RQM dos bigrupos retornados em cada conjunto de dados sintéticos. A variação do tempo

de execução do algoritmo de Cheng e Church é mostrada na figura 31 em relação aos valores do parâmetro α , aplicados a todos os conjuntos de dados sintéticos, considerando δ_1 (Figura 31a), δ_2 (Figura 31b) e δ_3 (Figura 31c). Os valores de δ dependem de cada conjunto de dados, como é mostrado na tabela 15, e cada linha de tendência mostra o comportamento do algoritmo em cada conjunto de dados das estruturas a , b , c , d , e , f , g , h , e i .

Figura 31 – Variação do tempo de execução do algoritmo de Cheng e Church em relação ao parâmetro α para cada conjunto de dados sintéticos, para δ_1 , δ_2 e δ_3 .

Fonte: Alexandra Katuska Ramos Diaz, 2018

Segundo a natureza do algoritmo de encontrar um bigrupo por vez, é de se esperar que conjuntos de dados que apresentem maior número de bigrupos levem a um maior tempo de execução do algoritmo. Segundo essa lógica, comprovada nas experimentações, os tempos de execução do algoritmo sobre os conjuntos de dados referentes a todas as estruturas de bigrupos seguem, em geral, a seguinte ordem: $Tempo(a) < Tempo(b, d, e) < Tempo(f, g, i) < Tempo(h) < Tempo(c)$. Na figura 31 é observado que, os tempos de execução do algoritmo tendem a crescer com valores maiores do parâmetro

α , um comportamento parecido em todos os conjuntos de dados sintéticos. No caso do conjunto de dados da estrutura c o tempo de execução cresce rapidamente estabilizando a partir de $\alpha = 1,5$.

Uma visão detalhada da figura 31, mostrada na figura 32b, ilustra a implicação dos diferentes valores de α no tempo de execução do algoritmo de Cheng e Church, para o conjunto de dados sintético da estrutura d (linhas exclusivas, 32a). As três linhas de tendência mostram o comportamento do algoritmo considerando os três valores assumidos para o parâmetro δ .

Figura 32 – Variação do tempo de execução do algoritmo de Cheng e Church em relação ao parâmetro α , no conjunto de dados da estrutura d , para $\delta_1 = 0,007$, $\delta_2 = 0,077$ e $\delta_3 = 0,772$

Fonte: Alexandra Katiuska Ramos Diaz, 2018

Os resultados apresentados nas figuras 31 e 32 confirmam o exposto em Cheng e Church (2000), sobre o aumento de tempo de execução do algoritmo conforme o aumento do valor de α . Valores altos para o parâmetro α permitem menor número 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. Embora o consumo de tempo para o conjunto de dados da estrutura c seja maior que o consumo de tempo das outras estruturas, todos os conjuntos de dados apresentam comportamento semelhante diante da variação do parâmetro α .

Nessas execuções foi observado que diferentes valores do parâmetro α não influenciam significativamente nos volumes e nos RQMs dos bigrupos resultantes. Nos gráficos da figura 33, são mostradas a variação da média dos volumes (Figura 33a) e a média dos RQMs (Figura 33b) dos bigrupos, considerando variação em α .

Figura 33 – Variação da média do volume e da média do RQM dos bigrupos em relação do parâmetro α após a aplicação do algoritmo de Cheng e Church, para cada conjunto de dados sintéticos

De acordo com o observado na figura 33, valores pequenos para α levam à descoberta de bigrupos com volumes pequenos e com maior homogeneidade entre seus valores (RQMs baixos). Isso decorre do fato de que com valores pequenos de α , o procedimento de remoção de múltiplas linhas e colunas (Algoritmo 2) tende a eliminar maior número de linhas e colunas que satisfazem as condições das linhas ($H(\mathbf{i}, J) > \alpha \times H(I, J)$) e das colunas ($H(I, \mathbf{j}) > \alpha \times H(I, J)$). Contudo, a partir de valores de α em 1, 2, os volumes dos bigrupos e os RQMs não são mais significativamente influenciados. Quando há variações nessas medidas, tais variações são mínimas, tornando os volumes e os RQMs dos bigrupos estáveis em relação ao parâmetro α .

Para análise do melhor valor para o parâmetro α , também foram consideradas as medidas internas de avaliação de agrupamento (considerando a projeção dos bigrupos nas linhas) e as medidas de biagrupamento. Em ambas medidas são consideradas as variações de δ . Essas medidas são listadas na tabela 17 para o caso do conjunto de dados da estrutura d .

Os resultados para essa instância específica mostram que pequenos valores de α tendem a produzir resultados de agrupamento melhores, embora haja instabilidade nos resultados obtidos para medidas que usam distância entre grupos. Seguindo apenas essa análise, valores pequenos de α são mais vantajosos, uma vez que produzem, mais frequentemente, os melhores resultados de avaliação. Além disso, o tempo de execução do algoritmo é também minimizado.

Tabela 17 – Avaliação quantitativa dos biagrupamentos obtidos sobre o conjunto de dados da estrutura de bigrupos d , do ponto de vista do parâmetro α , para $\delta_1 = 0,007$, $\delta_2 = 0,077$ e $\delta_3 = 0,772$ e α variando em $\{1, 1,2, 1,5, 2, 2,5, 3\}$

α	δ	Dist. inter	Dist. intra	Índ. DB	Vol.	DP	RQM	Cob. (%)	Ocup. (%)	Sob. (%)	Tempo (seg)
1	δ_1	2,90	21,30	2,32	37,33	4,98	0,002	0,50	100	0	0,62
	δ_2	0,84	146,11	11,42	1979,33	1193,15	0,06	21,58	100	22,25	0,51
	δ_3	2,41	165,00	2,41	3733,33	3003,42	0,05	33,33	100	49,33	0,51
1, 2	δ_1	1,65	29,42	3,87	41,66	5,31	0,006	0,55	100	0	8,50
	δ_2	1114,07	238,40	0,008	5477,33	1334,57	0,07	73,03	100	0	1,60
	δ_3	1114,08	248,88	0,007	6333,33	817,85	0,07	84,44	100	0	1,04
1, 5	δ_1	274,15	2748,27	6,31	47	3,56	0,007	0,62	100	0	11,72
	δ_2	347,23	62496,20	4,3075	4595,66	1340,44	0,07	61,27	100	0	7,31
	δ_3	214,19	78846,28	12,38	4773,33	1950,75	0,07	63,64	100	0	12,75
2	δ_1	228,61	2646,70	12,60	48,66	0,47	0,006	0,64	100	0	36,57
	δ_2	288,05	66926,80	4,9015	4840	1564,67	0,07	64,53	100	0	31,23
	δ_3	270,72	76537,95	5,40	5500	1870,83	0,07	73,33	100	0	28,42
2, 5	δ_1	338,45	3340,92	3,57	46,33	3,09	0,007	0,61	100	0	36,82
	δ_2	288,05	66926,80	4,9015	4840	1564,67	0,07	64,53	100	0	31,31
	δ_3	270,72	76537,95	5,40	5500	1870,83	0,07	73,33	100	0	30,10
3	δ_1	228,61	2646,70	12,60	48,66	0,47	0,006	0,64	100	0	37,46
	δ_2	157,38	80591,11	19,25	4716,66	1736,06	0,07	62,88	100	0	31,60
	δ_3	270,72	76537,95	5,40	5500	1870,83	0,07	73,33	100	0	30,24
		1114,08	248,88	0,007	3166,66	471,40	0,08	42,22	100	0	

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 ocupação (Ocup. %), porcentagem do grau de sobreposição (Sob. %).

Fonte: Alexandra Katiuska Ramos Diaz, 2018

Uma forma alternativa para realização da análise dos resultados é comparar as medidas obtidas às medidas esperadas, ou seja, dado que o conjunto de dados e sua organização em grupos e bigrupos é conhecida, as medidas ideais de qualidade podem ser calculadas. Essas medidas são apresentadas na última linha da tabela 17 para essa instância de teste. Comparando as medidas ideais às obtidas a partir da variação dos parâmetros, observa-se o melhor desempenho ocorrendo para α menores, mais especificamente para os valores 1 e 1,2. Nota-se ainda que para o valor de $\alpha = 1,2$, grupos com medidas de qualidades ideais e RQM também ideais ou muito próximos do ideal são obtidos. Entretanto, o volume e a cobertura dos bigrupos apresentam valores maiores que os ideais. Essa constatação mostra que o algoritmo está trazendo mais células para o bigrupo do que é de fato necessário para caracterizá-los. Mas fazendo uma análise no desvio padrão pode-se observar que mesmo trazendo células a mais, existe uma distribuição equilibrada entre os números de células dos três bigrupos resultantes. Por outro lado, o resultado obtido com $\alpha = 1$ e δ_3 é também melhor em termos de volume e cobertura, porém, o desvio padrão mostra que a inserção de células não é adequada, já que existe uma variabilidade maior nos tamanhos dos três bigrupos resultantes. Na figura 34 são representados os três bigrupos obtidos pelo algoritmo para $\alpha = 1$, e $\delta_3 = 0,772$, mostrando suas diferenças de

tamanhos; as diferentes tonalidades da cor azul representam as semelhanças dos valores dentro dos bigrupos.

Figura 34 – Representação dos bigrupos resultantes do algoritmo de Cheng e Church usando o conjunto de dados da estrutura d , para $\alpha = 1$, e $\delta_3 = 0,772$

Fonte: Alexandra Katiuska Ramos Diaz, 2018

O comportamento do parâmetro α em todos os conjuntos de dados sintéticos e seguindo o estudo relatado no conjunto de dados da estrutura d pode ser observado na tabela 18. Nesta tabela são mostrados os valores de α que levam aos melhores resultados de avaliação quantitativa, comparados com os resultados desejados de cada conjunto de dados das estruturas de bigrupos.

Tabela 18 – Valores para o parâmetro α que levam aos melhores resultados de avaliação quantitativa comparados aos resultados desejados, para cada estrutura de bigrupos.

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

* considerando só a combinação com δ_1

Fonte: Alexandra Katiuska Ramos Diaz, 2018

Na análise geral e no que diz respeito às medidas de volume e cobertura, o comportamento do algoritmo nos conjuntos de dados das estruturas b , c e e foi similar ao do conjunto de dados da estrutura d . Embora com um $\alpha = 1$ o algoritmo apresenta resultados bem próximos aos resultados esperados, os volumes dos bigrupos obtidos apresentam uma maior variabilidade (desvio padrão alto), retornando melhores resultados considerando $\alpha = 1, 2$.

Finalmente, 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, mas apresenta variabilidade nos valores dos parâmetros de α quando se trata de atender diferentes objetivos de qualidade. Pode-se deduzir que valores de α iguais a 1, 1,2 e 1,5 apresentam os resultados mais promissores.

4.1.3 Exp. #1: estudo do parâmetro δ

Os valores de estudo para o parâmetro δ dependem de cada conjunto de dados sob análise, sendo que três valores estão sendo testados neste trabalho: $\delta_1 \leq \frac{\min(\text{RQM})}{10}$, $\delta_2 = \min(\text{RQM})$ e $\delta_3 \geq 10 * \min(\text{RQM})$.

O gráfico da figura 35 mostra a variação dos tempos de execução totais considerando as variações em δ , para cada conjunto de dados das estruturas de bigrupos.

Figura 35 – Variação dos tempos de execução totais em base a δ , do algoritmo de Cheng e Church, em cada conjunto de dados

Fonte: Alexandra Katiuska Ramos Diaz, 2018

Como o parâmetro δ é usado como um limite superior para a minimização do RQM, seria natural que quanto menor seu valor, mais iterações fossem necessárias no processo de minimização, como foi o caso dos conjuntos de dados das estruturas a , b , c , d , e , g e i . Comportamento diferente foi observado com o conjunto de dados das estrutura f , e h , que apresentam um maior tempo de execução com um δ_2 do que usando δ_1 .

O gráfico da figura 36, mostra uma visão detalhada das informações presentes na figura 35. Nele é possível observar a implicação dos diferentes valores de δ no tempo de execução do algoritmo, para o conjunto de dados da estrutura d (linhas exclusivas). Cada

bolha representa a relação entre o uso de um valor específico para δ e para α , e o tamanho das bolhas é proporcional ao tempo de execução do algoritmo mediante os valores destes parâmetros.

Figura 36 – Variação do tempo de execução do algoritmo de Cheng e Church em relação ao parâmetro δ no conjunto de dados da estrutura d , para $\delta = \{0,007, 0,077, 0,772\}$ e α variando em $\{1, 1,2, 1,5, 2, 2,5, 3\}$

Fonte: Alexandra Katiuska Ramos Diaz, 2018

Isolando o comportamento da variável δ , na instância de execução usada para ilustrar esta análise na figura 36, nota-se que de fato o menor valor testado (0,007) leva à necessidade de execução de mais iterações nos laços do algoritmo. Contudo, o mesmo comportamento não é observado para os valores 0,077 e 0,772 quando a análise considera o valor de $\alpha = 1,5$, já que o uso de $\delta_2 = 0,077$ exige menos tempo de execução do algoritmo do que $\delta_3 = 0,772$. Um comportamento semelhante à estrutura d , considerando o valor de $\alpha = 1,5$, é observado nos conjuntos de dados das estruturas b , g , e i ; nas demais estruturas (estruturas a , c , e , f e h) o algoritmo mostra tempos similares de execução usando δ_2 e δ_3 .

Para a análise do parâmetro δ em relação aos volumes e os RQMs dos bigrupos resultantes, observe os gráficos das figuras 37 e 38, que mostram a variação da média das médias dos volumes (Figura 37a) e da média das médias dos RQMs (Figura 38a) dos bigrupos resultantes após a aplicação do algoritmo em cada conjunto de dados. Uma visão detalhada é apresentada para a estrutura d mostrando a variação da média dos volumes (figura 37b) e a média dos RQMs (figura 38b) dos bigrupos retornados, em relação do parâmetro δ . Para ambas as figuras, os valores de δ são $\{0,007, 0,077, 0,772\}$, com α variando em $\{1, 1,2, 1,5, 2, 2,5, 3\}$, e o tamanho das bolhas é proporcional à média dos volumes, ou à média dos RQMs dos bigrupos retornados pelo algoritmo mediante a combinação dos valores destes parâmetros respectivamente.

Figura 37 – Variação da média das médias dos volumes dos bigrupos resultantes, em base a δ , após a aplicação do algoritmo de Cheng e Church, em cada conjunto de dados (a), e a variação da média dos volumes em relação a δ usando o conjunto de dados da estrutura d (b).

Fonte: Alexandra Katiuska Ramos Diaz, 2018

Figura 38 – Variação da média das médias dos RQMs dos bigrupos resultantes, em base a δ , após a aplicação do algoritmo de Cheng e Church, em cada conjunto de dados (a), e a variação da média dos RQMs em relação a δ usando o conjunto de dados da estrutura d (b).

Fonte: Alexandra Katiuska Ramos Diaz, 2018

Sendo o parâmetro δ um limiar para a minimização do RQM, é de se esperar, que usando valores de δ pequenos (δ_1) os volumes e os RQMs sejam menores do que usando δ_2 e δ_3 . Com relação ao comportamento do algoritmo usando δ_2 e δ_3 , a média dos volumes e dos RQMs são semelhantes, com exceção da estrutura c que apresenta RQM maior usando um δ_3 . Um comportamento similar presente em todas as estruturas é mostrado na estrutura d das figuras 37b e 38b, em que as médias dos volumes e dos RQMs dos bigrupos retornados usando $\alpha = 1, 2$ são maiores do que usando os demais valores de α para δ_2 e δ_3 .

Reorganizando na tabela 19 os dados da tabela 17, para análise do ponto de vista de valores ideais para δ e dos valores esperados para as medidas de acordo com a verdade fundamental do conjunto, observa-se que os melhores resultados de fato aparecem para δ_2 e δ_3 . Esse comportamento já podia ser esperado pois alcançar um RQM muito menor que o RQM que já se sabe ser o real (caso do δ_1) tenderia à construção de bigrupos pequenos. Também é possível observar ao usar um RQM alto (caso do δ_3) permitiu-se a construção de bigrupos maiores, uma vez que a remoção de linhas e colunas se deu com menos intensidade.

Tabela 19 – Avaliação quantitativa dos biagrupamentos obtidos sobre o conjunto de dados da estrutura de bigrupos d , ponto de vista do parâmetro δ , para $\delta_1 = 0,007$, $\delta_2 = 0,077$ e $\delta_3 = 0,772$ e α variando em $\{1, 1,2, 1,5, 2, 2,5, 3\}$

δ	α	Dist. inter	Dist. intra	Índ. DB	Vol.	DP	RQM	Cob. (%)	Ocup. (%)	Sob. (%)	Tempo (seg)
δ_1	1	2,90	21,30	2,32	37,33	4,98	0,002	0,50	100	0	0,62
	1,2	1,65	29,42	3,87	41,66	5,31	0,006	0,55	100	0	8,50
	1,5	274,15	2748,27	6,31	47	3,55	0,007	0,62	100	0	11,72
	2	228,61	2646,70	12,60	48,66	0,47	0,006	0,64	100	0	36,57
	2,5	338,45	3340,92	3,57	46,33	3,09	0,007	0,61	100	0	36,82
	3	228,61	2646,70	12,60	48,66	0,47	0,006	0,64	100	0	37,46
δ_2	1	0,84	146,11	11,42	1979,33	1193,14	0,06	21,58	100	22,25	0,51
	1,2	1114,07	238,40	0,008	5477,33	1334,57	0,07	73,03	100	0	1,60
	1,5	347,23	62496,20	4,3075	4595,66	1340,45	0,07	61,27	100	0	7,31
	2	288,05	66926,80	4,9015	4840	1564,68	0,07	64,53	100	0	31,23
	2,5	288,05	66926,80	4,9015	4840	1564,68	0,07	64,53	100	0	31,31
	3	157,38	80591,11	19,25	4716,66	1736,06	0,07	62,88	100	0	31,60
δ_3	1	2,41	165,00	2,41	3733,33	3003,42	0,05	33,33	100	49,33	0,51
	1,2	1114,08	248,88	0,007	6333,33	817,86	0,07	84,44	100	0	1,04
	1,5	214,19	78846,28	12,38	4773,33	1950,75	0,07	63,64	100	0	12,75
	2	270,72	76537,95	5,40	5500	1870,83	0,07	73,33	100	0	28,42
	2,5	270,72	76537,95	5,40	5500	1870,83	0,07	73,33	100	0	30,10
	3	270,72	76537,95	5,40	5500	1870,83	0,07	73,33	100	0	30,24
		1114,08	248,88	0,007	3166,66	471,40	0,08	42,22	100	0	

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 ocupação (Ocup. %), porcentagem do grau de sobreposição (Sob. %).

Fonte: Alexandra Kátiuska Ramos Diaz, 2018

Observando o comportamento do parâmetro δ em todos os conjuntos de dados, e seguindo o estudo relatado no conjunto de dados da estrutura d , na tabela 20 são mostrados os valores de δ que levam aos melhores resultados de avaliação quantitativa comparados com os resultados desejados para cada conjunto de dados.

Finalmente, a análise do parâmetro δ mostra 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. Pode-se deduzir que usando 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

Tabela 20 – Valores para o parâmetro δ que levam aos melhores resultados de avaliação quantitativa comparativa com os resultados desejados, para cada estrutura de bigrupos

Estrutura	Dist. inter	Dist. intra	Índ. DB	Vol.	RQM	Cob.	Ocup.	Sob.
a	-	δ_2, δ_3	-	δ_2, δ_3	δ_2, δ_3	δ_2, δ_3	qualquer	qualquer
b	δ_1	δ_1	δ_1	δ_3	δ_3	δ_3	qualquer	δ_1
c	δ_3	δ_1	δ_1	δ_3	δ_2	δ_3	qualquer	δ_1^*
d	δ_2, δ_3	δ_2, δ_3	δ_2, δ_3	δ_3	δ_2, δ_3	δ_3	qualquer	qualquer*
e	δ_3	δ_2	δ_3	δ_3	δ_2, δ_3	δ_3	qualquer	δ_1, δ_3
f	δ_3	δ_3	δ_2	δ_2	δ_2, δ_3	δ_2	qualquer	δ_1
g	δ_3	δ_3	δ_2	δ_2	δ_2, δ_3	δ_2	qualquer	δ_1
h	δ_3	δ_2	δ_2, δ_3	δ_2	δ_2, δ_3	δ_2	qualquer	δ_3
i	δ_3	δ_2	δ_3	δ_3	δ_2, δ_3	δ_3	qualquer	δ_3

* desconsiderando a combinação com $\alpha = 1$.

Fonte: Alexandra Katuska Ramos Diaz, 2018

δ_3 trazem resultados bons em termos de volume, RQMs e tempos de execução minimizados, parecendo portanto, ser boas escolhas.

Como considerações finais para a análise do *Exp. #1* tem-se que sob as condições originais dos conjuntos de dados sintéticos, a escolha ideal para os parâmetros é:

- $\alpha = 1, 1.2, 1.5$;
- $\delta = \min(\text{RQM})$ ou $> \min(\text{RQM})$;

Valores pequenos para a margem do RQM de linhas e colunas em um bigrupo (α) combinados a limiares próximos ao RQM mínimo esperado para os bigrupos do conjunto de dados, levam a execuções mais eficientes em tempo de execução, ao alcance de boas avaliações para validação interna de agrupamento construído a partir de projeções de bigrupos, e também a medidas mais próximas ao ideal para as avaliações de bigrupos.

4.1.4 *Exp. #2*: estudo do parâmetro α em conjunto de dados com níveis de esparsidade

O estudo do parâmetro α foi realizado sobre todos os conjuntos de dados sintéticos, alterados para apresentarem diferentes níveis de esparsidade. A escolha do parâmetro α foi baseada na análise da implicação dos valores deste parâmetro no tempo de execução do algoritmo nos conjuntos de dados nas suas versões esparsas.

De maneira geral, o comportamento do algoritmo quanto à variação do tempo de execução em relação à variação do parâmetro α é ilustrada na figura 39. A ilustração conta com a média dos tempos de execução do algoritmo sobre todos os níveis de esparsidade.

Figura 39 – Variação da média do tempo de execução sobre os diferentes níveis de esparsidade em relação ao parâmetro α , após a aplicação do algoritmo de Cheng e Church nos conjuntos de dados sintéticos, com δ_1 , δ_2 e δ_3

Fonte: Alexandra Katiuska Ramos Diaz, 2018

Na figura 39 é observado que, em todos os conjuntos de dados, tempos de execução menores ocorrem usando valores pequenos de α . Além disso, observa-se um aumento nos valores do tempo para valores $\alpha = 1, 2$ até $\alpha = 2$ em todos os conjuntos de dados.

Ainda na figura 39, a estrutura a apresenta um comportamento diferente e variado com respeito às demais estruturas, pois para δ_1 o comportamento do algoritmo é similar ao conjunto de dados da estrutura e . Entretanto, com valores de δ_2 e δ_3 , o algoritmo se torna insensível aos valores de α .

Uma visão detalhada da figura 39, mostrada nos gráficos da figura 40, ilustra a implicação dos diferentes valores de α no tempo de execução do algoritmo para o conjunto de dados da estrutura d . As linhas de tendência mostram o comportamento do algoritmo em cada nível de esparsidade.

Nos gráficos da figura 40 observa-se uma variabilidade do comportamento do algoritmo entre $\alpha = 1, 2$ até $\alpha = 2, 5$, seguido com uma estabilidade com tempos elevados

Figura 40 – Variação do tempo de execução do algoritmo de Cheng e Church em relação ao parâmetro α , no conjunto de dados da estrutura d com diferentes níveis de esparsidade, para valores de $\alpha \{1, 1,2, 1,5, 2, 2,5, 3\}$, $\delta_1 = 0,007$ (a), $\delta_2 = 0,077$ (b), $\delta_3 = 0,772$ (c) e níveis de esparsidade de $\{0, 10, 20, 30, 40, 50, 60, 70, 80, 90, 98\}$

Fonte: Alexandra Katiuska Ramos Diaz, 2018

usando valores de alpha maiores ou iguais a 2,5. Além disso é observado que o parâmetro δ não influencia significativamente no tempo de execução do algoritmo. O comportamento descrito no conjunto de dados da estrutura d foi similar para os demais conjuntos de dados.

Para análise da influência da esparsidade nos valores do parâmetro α , são mostrados os gráficos da figura 41. Esta figura ilustra a comparação da implicação dos valores de α no tempo de execução do algoritmo em um ambiente com esparsidade baixa e em um ambiente com esparsidade alta aplicado a todos os conjuntos de dados sintéticos. Nesta figura os gráficos da esquerda são as experimentações com esparsidade baixa e os gráficos da direita são as experimentações com esparsidade alta. As linhas de tendência representam cada conjunto de dados e são considerados os três valores para o parâmetro δ .

Figura 41 – Variação do tempo de execução do algoritmo de Cheng e Church em relação ao parâmetro α , em cada conjunto de dados sintéticos, com níveis de esparsidade baixa e alta, usando δ_1 , δ_2 e δ_3

Fonte: Alexandra Katiuska Ramos Diaz, 2018

Na figura 41 é observada uma influência da esparsidade no tempo de execução do algoritmo no intervalo de valores de α igual a 1,2 até 2,5, de forma geral para todos os conjuntos de dados. Neste intervalo existe uma variabilidade do comportamento do algoritmo em um ambiente de esparsidade baixo e em um ambiente de esparsidade alto. Esta variabilidade deve-se ao fato de que em algumas estruturas os tempos de execução

começam a subir a partir de um $\alpha = 1,2$ com níveis de esparsidade alto. Também foi observada uma baixa variabilidade entre os diferentes parâmetro de δ .

Ainda na figura 41, a estrutura *a* apresenta um comportamento diferente, em que o algoritmo passa a ser insensível à variação dos valores de α tanto em um ambiente de baixa como de alta esparsidade.

Finalmente, no algoritmo de Cheng e Church o parâmetro α age como um parâmetro de qualidade, tornando-se custoso para valores maiores de α . Das experimentações feitas nos conjuntos de dados com diferentes níveis de esparsidade, foi observado uma maior variabilidade do tempo de execução usando valores de α entre 1,2 e 2,5. Os valores de $\alpha = 1$ e $\alpha = 1,2$ não foram influenciados pela esparsidade alta, com exceção do comportamento na estrutura *c*.

Por fim, de acordo com as análises do *Exp.#1*, no que diz respeito às medidas de volume e cobertura, valores de $\alpha = 1$ retornaram bigrupos com maior variabilidade de seus volumes. Seguindo essas análises, uma boa escolha para o parâmetro α seria 1,2, pois apresenta resultados promissores.

4.1.5 *Exp.#2*: estudo do parâmetro δ em conjunto de dados com níveis de esparsidade

A análise do parâmetro δ foi realizada sobre os conjuntos de dados sintéticos, modificados para que apresentassem diferentes níveis de esparsidade. A escolha do parâmetro δ foi baseada na análise da implicação dos diferentes valores deste no tempo de execução, nos volumes e nos RQMs dos bigrupos retornados pelo algoritmo aplicado nos conjuntos de dados nas suas versões esparsas.

A figura 42 ilustra o comportamento do algoritmo de maneira geral quanto à variação do tempo de execução, o volume e o RQM em relação à variação do parâmetro δ . Esta figura conta com a média dos tempos de execução do algoritmo (Figura 42a), os volumes (Figura 42b) e os RQMs (Figura 42c) dos bigrupos resultantes sobre todos os níveis de esparsidade. Nestas análises foram considerados valores de α igual a 1,2 e o número de bigrupos já conhecido em cada conjunto de dados sintéticos.

Na figura 42a, é observado que, em geral, os valores do parâmetro δ não afetam de maneira significativa o tempo de execução do algoritmo. Custos ligeiramente menores, em termos de tempos de execução, são apresentados com δ_3 nas estruturas *a*, *d*, *e*, *g* e *i*.

Figura 42 – Variação da média dos tempos de execução (a), das médias dos RQM (b) e das médias dos volumes (c) sobre os diferentes níveis de esparsidade em relação ao parâmetro δ , após a aplicação do algoritmo de Cheng e Church nos conjuntos de dados sintéticos, com $\alpha = 1, 2$

Fonte: Alexandra Katiuska Ramos Diaz, 2018

Casos excepcionais a este comportamento são as estruturas f e c que apresentam custos ligeiramente maiores do que δ_2 . Ainda na figura, os tempos de execução do algoritmo nas estruturas b e h são iguais usando δ_2 e δ_3 .

Nas figuras 42b e 42c é observado o mesmo comportamento do *Exp.#1* no que diz a δ_1 , retornando bigrupos com volumes e RQMs pequenos. Já usando δ_3 o algoritmo retorna bigrupos com volumes maiores combinados com RQMs ligeiramente maiores em comparação ao δ_2 .

Para análise da influência da esparsidade nos valores do parâmetro δ , é mostrada a figura 43. Esta figura ilustra a comparação da implicação dos valores de δ no tempo de execução, o RQM e o volume em ambientes de baixa e alta esparsidade.

Em um ambiente de baixa esparsidade, como nas figuras 43a, 43c, e 43e, observa-se que usando δ_3 , resultados melhores são obtidos em comparação com aqueles obtidos com

Figura 43 – Variação dos tempos de execução, dos RQM e dos volumes do algoritmo de Cheng e Church em relação ao parâmetro δ , em cada conjunto de dados sintéticos, com níveis de esparsidade baixa e alta, usando $\alpha = 1, 2$

(a) Tempo de execução com esparsidade baixa

(b) Tempo de execução com esparsidade alta

(c) RQM dos bigrupos com esparsidade baixa

(d) RQM dos bigrupos com esparsidade alta

(e) Volume dos bigrupos com esparsidade baixa

(f) Volume dos bigrupos com esparsidade alta

Fonte: Alexandra Katiuska Ramos Diaz, 2018

δ_2 . Embora o uso de δ_3 leve a valores de RQMs maiores do que δ_2 , combinados com o fato de este retornar bigrupos com volumes maiores e tempos de execução menores, estes resultados são considerados bons de acordo com os objetivos do algoritmo.

Por outro lado, em um ambiente de alta esparsidade (nível de esparsidade de 98), como foi analisado anteriormente, as variações do tempo de execução, do volume e

do RQM não apresentam variações significativas em relação ao parâmetro δ , mantendo um comportamento constante. Isto induz a considerar como esparsidade alta um valor menor que 98, a fim de distinguir alguns comportamentos. Já nas figuras 43b, 43d, e 43f é observada uma estabilização quanto a estas três medidas usando δ_2 e δ_3 , mas ainda, melhores volumes são retornados usando δ_3 .

De acordo com as análises e as experimentações feitas, não parece ser uma boa escolha usar valores de δ pequenos (δ_1), já que estes tendem a retornar volumes baixos. Por outro lado, para δ_2 e δ_3 , apesar de que as medidas retornam valores levemente similares entre si, os resultados retornados usando δ_3 tendem a ser melhores do que usando δ_2 .

Finalmente, 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, parece ser uma boa escolha usar um limiar acima ao RQM mínimo para o parâmetro δ (δ_3).

4.1.6 *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 , detalhados na tabela 16. Nestas experimentações foram considerados $\alpha = 1, 2$ e δ_3 , escolhidos nos estudos dos itens anteriores. Esta análise está baseada nos objetivos do algoritmo de Cheng e Church e, por consequência estão sendo analisados os volumes, os RQMs dos bigrupos retornados e o tempo de execução do algoritmo. Além disso, serão analisadas as medidas da porcentagem da cobertura e a porcentagem do grau de sobreposição, já que estão relacionadas com o volume dos bigrupos. A variação destas medidas são mostradas nos gráficos da figura 44 nos conjuntos de dados das estruturas a , b , e i . Na figura os valores ideais destas estruturas são destacados por um círculo. Deve-se notar que todos os conjuntos de dados apresentam um comportamento semelhante às estruturas mostradas aqui.

Na figura 44, sendo a natureza do algoritmo 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 são as estruturas b e i , em que valores altos para

Figura 44 – Variação do tempo de execução, dos volumes, dos RQMs dos bigrupos, a porcentagem de cobertura e a porcentagem do grau de sobreposição em relação ao número de bigrupos (K) após a aplicação do algoritmo de Cheng e Church nos conjuntos de dados das estruturas a , b , e i na suas formas originais

Fonte: Alexandra Katiuska Ramos Diaz, 2018

o parâmetro K , como $K = 6$ na estrutura b e $K = 9$ na estrutura i , levam a tempos de execução altos comparados aos tempos usando seus valores ideais.

Com respeito ao RQM e o volume, se o número de bigrupos a procurar é maior do que o ideal, o algoritmo retorna bigrupos com RQMs levemente menores combinados com volumes pequenos. Além disso, é observado um aumento na porcentagem de cobertura e na porcentagem do grau de sobreposição devido a ser superior o número de elementos a ser considerados nos bigrupos (cobertura alta), e estes bigrupos apresentam uma maior semelhança entre si (grau de sobreposição alto).

Por outro lado, se o parâmetro K é menor do que o ideal, os bigrupos resultantes 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, existe um maior número de elementos que não pertencem a nenhum bigrupo (cobertura baixa), além disso, estes bigrupos têm baixas semelhanças entre si (sobreposição baixa).

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

A análise do comportamento do algoritmo foi realizada sobre os conjuntos de dados com diferentes níveis de esparsidade. Os bigrupos resultantes do algoritmo foram avaliados com as medidas de biagrupamento e as medidas internas de agrupamento, usando valores de α , δ e K fixos. Nestas experimentações foram usados $\alpha = 1, 2$ e δ_3 , valores escolhidos das análises do *Exp. #2*. Para o parâmetro K foram considerados os números de bigrupos já conhecidos em cada conjunto de dados sintéticos.

Nos gráficos da figura 45 são mostradas as variações das medidas internas de agrupamento (distância inter-grupos, distância intra-grupos, e o índice DB) em relação ao incremento de esparsidade nos conjuntos de dados sintéticos. Estas medidas foram calculadas nos grupos das linhas, formados pelas projeções nas linhas dos bigrupos resultantes do algoritmo.

A distância inter-grupos mede a distância média dos centroides dos diferentes grupos, valores maiores desta medida indicam maior grau de separação entre os centroides dos grupos das linhas. Na figura 45a é observada uma tendência decrescente em todos os conjuntos de dados em relação à esparsidade. Também é observado que em média a partir dos níveis de esparsidade de 70 até 98 os centroides dos grupos das linhas retornados pelo algoritmo tendem a ser similares entre si, ou seja os bigrupos retornados são similares uns a outros (valores da distância inter-grupos baixos). Este comportamento é similar em todos os conjuntos de dados.

Com respeito à distância intra-grupos, valores menores desta medida indicam maior grau de similaridade entre os objetos de um mesmo grupo. Na figura 45b é mostrado um comportamento quase estável dentro de um intervalo em todos os conjuntos de dados, com exceção dos conjuntos de dados das estruturas f e b . Também é percebido que a partir de um nível de esparsidade, em geral de 70, os elementos dos grupos das linhas apresentam maior semelhança entre si (valores da distância intra-grupos baixos).

Figura 45 – Variação das medidas internas de agrupamento em relação ao incremento dos níveis de esparsidade, em cada conjunto de dados sintéticos após a aplicação do algoritmo de Cheng e Church, usando $\alpha = 1, 2$ e δ_3

Fonte: Alexandra Katiuska Ramos Diaz, 2018

A medida do índice DB combina a relação do grau de compactação dos grupos (distância intra-grupos) e o grau de separação destes grupos (distância inter-grupos). Na figura 45c, o índice DB apresenta um comportamento variado para os conjuntos de dados. Nos conjuntos de dados das estruturas *a*, *b*, *d*, *e*, *h* e *i* sem esparsidade, os grupos das linhas são bons (valores baixo do índice DB), mas conforme os níveis de esparsidade aumentam estes se tornam ruins (valores altos do índice DB). Comportamento diferente acontece nos conjuntos de dados das estruturas *c*, *f* e *g*, sem esparsidade o algoritmo apresenta problemas ao agrupar estes conjuntos retornando valores altos do índice DB. Este comportamento tende a melhorar com o aumento da esparsidade, mas piora com altos níveis de esparsidade.

O fato do índice DB apresentar valores altos nos níveis de esparsidade alta, indica grupos semelhantes entre si. Este comportamento é explicado pela proximidade dos centroides dos grupos (distância inter-grupos).

As análises das medidas do volume e os RQMs são mostradas nos gráficos da figura 46. Estas figuras ilustram as médias dos volumes e dos RQMs dos bigrupos retornados pelo algoritmo aplicados a conjuntos com diferentes níveis de esparsidade.

Figura 46 – Variação da média dos volumes e dos RQMs em relação aos incremento dos níveis de esparsidade, em cada conjunto de dados sintéticos, após a aplicação do algoritmo de Cheng e Church, usando $\alpha = 1, 2$ e δ_3

Fonte: Alexandra Katiuska Ramos Diaz, 2018

Na figura 46a é observada a vulnerabilidade dos bigrupos retornados frente ao incremento da esparsidade. Em geral, os volumes dos bigrupos diminuem em níveis de esparsidade de 10, apresentando depois leves variações conforme os níveis de esparsidade aumentam. Este comportamento continua até níveis de esparsidade de 60 ou 70 em que os volumes dos bigrupos tendem a aumentar, alcançando valores elevados com níveis de esparsidade altos.

O comportamento dos RQMs dos bigrupos é observado na figura 46b). Estes valores aumentam até um nível de esparsidade de 40 (em algumas estruturas é até 20 de esparsidade), para depois diminuir chegando a valores próximos de 0 com esparsidades altas.

As medidas do volume e o RQM dos bigrupos (Figuras 46a e 46b) mostram que em determinado nível de esparsidade o algoritmo tende a considerar os elementos nulos no biagrupamento, levando à obtenção de bigrupos com altos volumes e ao mesmo tempo baixos RQMs.

As análises das medidas da porcentagem da cobertura, do grau de sobreposição e do grau de ocupação em relação ao incremento dos níveis de esparsidade são mostradas nos gráficos da figura 47.

A porcentagem da cobertura (Figura 47a) apresenta relação com o volume dos bigrupos. Já que se o algoritmo tende a retornar bigrupos com volumes altos é de se esperar

Figura 47 – Variação da porcentagem de cobertura, do grau de sobreposição e 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 de Cheng e Church, usando $\alpha = 1, 2$ e δ_3

que os elementos destes cubram a maior parte do conjunto de dados. Adicionalmente, se os bigrupos tendem a ser maiores em volume, a porcentagem de elementos sobrepostos entre eles tende a ser maior (Figura 47b).

Em relação ao grau de ocupação, a figura 47c confirma o comportamento relatado nas medidas anteriores. Esta figura mostra uma tendência decrescente nas porcentagens dos elementos não nulos nos bigrupos resultantes. Ainda a figura ilustra que, conforme os níveis de esparsidade aumentam (aumento de elementos nulos nos conjuntos de dados) o algoritmo tende a considerar os elementos nulos como importantes para o biagrupamento e a desprezar a informação útil. Este comportamento leva a retornar bigrupos formados principalmente com elementos nulos.

Com base nas análises das tendências das medidas internas e das medidas de biagrupamento em relação ao incremento da esparsidade, pode se 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 (2004), bigrupos 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 principalmente os elementos nulos no biagrupamento. Este comportamento é mais pronunciado a partir dos níveis de esparsidade de 60 ou 70, em que o algoritmo tende a biagrupar em torno aos elementos nulos dos conjuntos de dados e a desconsiderar a informação relevante no biagrupamento. Por consequência, os bigrupos retornados apresentam maiores volumes e valores de RQMs próximos a zero (maior quantidade de elementos nulos). Adicionalmente, estes bigrupos são semelhantes entre si, com maiores sobreposições entre eles.

4.2 Algoritmo de Cheng e Church aplicado no corpus EBC

O algoritmo de Cheng e Church foi aplicado no corpus EBC com a finalidade de estudar seu comportamento em um ambiente de alta esparsidade e analisar os resultados retornados por este. A representação do corpus EBC varia dependendo de suas configurações, todas estas detalhadas na tabela 12, junto com seus identificadores nomeados de matrizes desde Mat. 1 até Mat. 15.

Para a aplicação do algoritmo, foi visto nas seções anteriores que são necessários três parâmetros de entrada, os quais serão escolhidos em base a algumas experimentações. A seguir são apresentadas as experimentações (*Exp.#5*) para a determinação destes parâmetros, e as análises quantitativas e qualitativas a partir da aplicação do algoritmo de Cheng e Church no corpus EBC (*Exp.#6*).

4.2.1 *Exp.#5*: determinação dos parâmetros α , δ e K

Segundo a análise feita nas experimentações nos conjunto de dados sintéticos, o valor de α utilizado foi de 1, 2, dado que ele retorna resultados bons quando é aplicado à ambientes esparsos. Com respeito ao parâmetro K , foram escolhidos valores em torno do número de canais já conhecido do corpus EBC, sendo estes 3, 5, 7 e 10.

Para a determinação do valor do parâmetro δ , foram obtidas aleatoriamente 1000 submatrizes com cada combinação de tamanho, a partir de uma matriz escolhida aleatoriamente das 15 matrizes produzidas do corpus EBC. A partir das submatrizes obtidas foram

calculados seus RQMs e mostrados na tabela 21. Nesta tabela são mostrados os diferentes números de linhas e colunas das submatrizes obtidas, além das pontuações mais baixas, as pontuações mais altas, e a moda das pontuações dos valores dos RQMs calculados nestas submatrizes.

Tabela 21 – RQMs estimados de 1000 submatrizes geradas aleatoriamente para cada combinação de tamanho

Linhas	Colunas	RQM		
		Mínimo	Máximo	Moda
100	10	0,00021	0,00256	0,00063
	25	0,00035	0,00139	0,00073
	50	0,00046	0,00128	0,00074
	100	0,00052	0,00125	0,00072
133	10	0,00026	0,00186	0,00056
	25	0,00035	0,00156	0,00071
	50	0,00046	0,00128	0,00081
	100	0,00053	0,00109	0,00078
171	10	0,00027	0,00155	0,0006
	25	0,00044	0,00128	0,00128
	50	0,0005	0,00119	0,0008
	100	0,00055	0,00116	0,00078
240	10	0,00037	0,0019	0,00063
	25	0,0005	0,00121	0,00072
	50	0,00057	0,00111	0,00074
	100	0,00058	0,00106	0,00076
400	10	0,00043	0,00142	0,00069
	25	0,00051	0,00105	0,00075
	50	0,00059	0,00106	0,00076
	100	0,00062	0,00099	0,00077

Fonte: Alexandra Katiuska Ramos Diaz, 2018

A partir da tabela 21, é possível observar o intervalo de valores pelos quais o RQM flutua no conjunto de dados reais, sendo este de 0,00021 até 0,00256. Dado que os valores do RQM apresentam valores pequenos nas matrizes produzidas do corpus EBC, foi necessário fazer experimentações com diferentes valores de δ , em uma matriz obtida aleatoriamente das matrizes do corpus EBC. Estas experimentações são mostradas na tabela 22, em que os bigrupos retornados pelo algoritmo foram avaliados considerando as medidas de biagrupamento. Nesta tabela os valores de δ considerados foram entorno aos valores mínimos do RQM (0, 0,001, 0,1, 0,5, 1), com $\alpha = 1, 2$ e com $K = \{3, 5, 7, 10\}$.

A tabela 22 apresenta as experimentações realizadas em uma matriz do corpus EBC. Nesta tabela pode-se observar que, valores muito pequenos como $\delta = 0$, retorna bigrupos triviais, em que todos seus elementos são zeros, ou seja, o algoritmo acha bigrupos com zeros constantes. Adicionalmente, valores altos para o parâmetro δ como 0,1, 0,5 e 1, o algoritmo tende a retornar bigrupos com volumes grandes, ou seja, sendo grande o limiar todos os elementos do conjunto de dados são colocados nos bigrupos. Similarmente, com um $\delta = 0.001$ o algoritmo retorna bigrupos com volumes grandes, pois aceita quase todos

Tabela 22 – Medidas de biagrupamento com $K = \{3, 5, 7, 10\}$, $\delta = \{0, 0, 001, 0, 1, 0, 5, 1\}$, usando $\alpha = 1, 2$

K	δ	Vol.	RQM	Cob. (%)	Ocup. (%)	Sob. (%)	Tempo (seg)
3	0	56825,3	0	14,17	0	3,74	128,54
	0,001	390798	0,0003	100	6,57	1,08	8115,74
	0,1	1159839	0,0008	100	6,57	100	18,68
	0,5	1159839	0,0008	100	6,57	100	18,81
	1	1159839	0,0008	100	6,57	100	18,85
5	0	45185,8	0	17,46	0	11,59	149,59
	0,001	238710	0,0002	100	6,57	2,91	16202,76
	0,1	1159839	0,0008	100	6,57	100	32,31
	0,5	1159839	0,0008	100	6,57	100	32,97
	1	1159839	0,0008	100	6,57	100	231,52
7	0	40901,7	0	20,57	0	20,01	178,45
	0,001	182258,3	0,0001	100	6,57	19,99	24587,79
	0,1	1159839	0,0008	100	6,57	100	45,26
	0,5	1159839	0,0008	100	6,57	100	44,82
	1	1159839	0,0008	100	6,57	100	206,19
10	0	36513,1	0	22,72	0	24,73	206,19
	0,001	147777,4	0,0001	100	6,57	14,67	32786,32
	0,1	1159839	0,0008	100	6,57	100	58,28
	0,5	1159839	0,0008	100	6,57	100	57,95
	1	1159839	0,0008	100	6,57	100	58,44

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

Fonte: Alexandra Katiuska Ramos Diaz, 2018

os elementos dentro dos bigrupos. Seguindo este raciocínio e a partir do RQMs calculados na tabela 21, parece realmente coerente fazer as experimentações com valores de $\delta < 0.001$, como $\delta = 0,0008$.

4.2.2 Exp. #6: Avaliações quantitativas

Para avaliar o algoritmo de Cheng e Church de forma quantitativa foram realizados experimentos considerando a tarefa de biagrupamento de documentos. Estes resultados foram avaliados e analisados usando as medidas de biagrupamento e as medidas internas de agrupamento calculadas nos grupos formados pela projeção das linhas dos bigrupos resultantes. Ambas medidas foram descritas na seção 2.3.

Com o fim de mostrar uma melhor visualização, estas medidas são apresentadas usando mapas de calor e diagramas de caixas. Nos mapas de calor as cores são associadas de acordo com os valores de cada medida, com variações da cor amarela à cor vermelha. A cor amarela representa resultados bons e a cor vermelha piores resultados. Cada mapa de calor mostra os resultados obtidos em uma determinada medida, sendo que, no eixo x estão as diferentes entradas com identificadores de Mat. 1 até Mat. 15 e no eixo y os diferentes

valores usados para o parâmetro K . Além disso, as matrizes com identificadores de Mat. 1 até Mat. 5 apresentam representações binárias, as matrizes com identificadores de Mat. 6 até Mat. 10 apresentam representações TF e as matrizes com identificadores de Mat. 11 até Mat. 15 apresentam representações TF-IDF. Com respeito aos diagramas de caixa, eles mostram os valores de uma medida determinada (posicionada no eixo x), por cada representação (posicionada no eixo y). As representações usadas foram as representações binárias, TF e TF-IDF. A seguir são apresentados os resultados e análises de cada uma das medidas de avaliação.

Na figura 48 é apresentado o mapa de calor das porcentagens do grau de ocupação em relação aos diferentes números de bigrupos, aplicados às matrizes do corpus EBC.

Figura 48 – Mapa de calor da porcentagem do grau de ocupação dos resultados obtidos após a aplicação do algoritmo de Cheng e Church nas matrizes obtidas do corpus EBC, com $K = \{3, 5, 7, 10\}$, $\alpha = 1, 2$ e $\delta = 0,0008$

Fonte: Alexandra Katiuska Ramos Diaz, 2018

Segundo a figura 48, é observado que o algoritmo tende a retornar bigrupos com maior quantidade de zeros, levando a valores baixos nos graus de ocupação. Com respeito às matrizes com representações binárias, elas tendem a trazer valores de grau de ocupação próximos ou iguais a zeros. Este comportamento talvez seja pelo fato do algoritmo procurar bigrupos com RQMs baixos, priorizando o biagrupamento dos elementos nulos e eliminando elementos distintos a estes.

Ainda na figura 48, as representações TF e TF-IDF trouxeram resultados semelhantes entre cada par de matrizes. Por exemplo, os pares de matrizes: (Mat. 6, Mat. 11), (Mat. 7, Mat. 12), (Mat. 8, Mat. 13) e (Mat. 10, Mat. 15) retornaram bigrupos com porcentagem do grau de ocupação semelhantes entre si. Dos pares de matrizes, os primeiros elementos de

cada par tem representações TF e os segundos elementos de cada par têm representações TF-IDF. Dentre estes pares de matrizes, as matrizes com identificadores Mat. 7 e Mat. 12 obtiveram melhores resultados.

Os resultados da porcentagem do grau de sobreposição são apresentados no mapa de calor da figura 49.

Figura 49 – Mapa de calor da porcentagem do grau de sobreposição dos resultados obtidos após a aplicação do algoritmo de Cheng e Church nas matrizes obtidas do corpus EBC, com $K = \{3, 5, 7, 10\}$, $\alpha = 1, 2$ e $\delta = 0,0008$

Fonte: Alexandra Katiuska Ramos Diaz, 2018

De acordo com a figura 49, percebe-se que menores valores para o parâmetro K trazem menores porcentagens de sobreposições, obtendo os melhores resultados com valores de K igual a 3. Além disso, as matrizes com representações binárias trazem menores porcentagens de sobreposição do que as outras representações. Quanto às representações TF e TF-IDF apresentam o mesmo comportamento descrito no grau de ocupação, valores semelhantes entre si são retornados entre cada par de matrizes. Destas matrizes as que apresentam melhores resultados são a Mat. 6, Mat. 11 seguidas da Mat. 7 e a Mat. 12.

Os resultados do índice DB são mostrados no mapa de calor da figura 50.

Conforme a figura 50, é observado que, valores baixos do índice DB são retornados pelas representações binárias, seguidas das matrizes com identificadores Mat. 6, Mat. 11, Mat. 7 e Mat. 12. Nas representações TF e TF-IDF, ainda é observado o comportamento percebido nas medidas anteriores, no que diz respeito aos resultados próximos entre si, são retornados entre cada par de matrizes.

Figura 50 – Mapa de calor do índice DB dos resultados obtidos após a aplicação do algoritmo de Cheng e Church nas matrizes obtidas do corpus EBC, com $K = \{3, 5, 7, 10\}$, $\alpha = 1, 2$ e $\delta = 0,0008$

Fonte: Alexandra Katuska Ramos Diaz, 2018

As medidas da porcentagem de cobertura e os volumes dos bigrupos apresentam comportamentos semelhantes, o que leva a uma análise conjunta. Estas medidas são mostradas nos gráficos da figura 51 por meio de diagramas de caixas.

Figura 51 – Diagrama de caixa do volume (a) e da porcentagem de cobertura (b) dos resultados obtidos após a aplicação do algoritmo de Cheng e Church nas matrizes obtidas do corpus EBC, com $K = \{3, 5, 7, 10\}$, $\alpha = 1, 2$ e $\delta = 0,0008$

Fonte: Alexandra Katuska Ramos Diaz, 2018

Das figuras 51a e 51b é deduzido que, matrizes com representações TF e TF-IDF tendem a retornar bigrupos com volumes grandes, cobrindo em média 80% da matriz de dados. As matrizes com representações binárias tendem a retornar bigrupos com volumes menores, mas que ainda cobrem em média 53% da matriz de dados. Esse fato corrobora e explica a alta sobreposição presente nos bigrupos resultantes.

Os resultados das medidas de RQM dos bigrupos, distância intra-grupos e distância inter-grupos são apresentados nos diagramas de caixa da figura 52.

Figura 52 – Diagrama de caixa dos RQMs (a), distância intra-grupos (b) e distância inter-grupos (c) dos resultados obtidos após a aplicação do algoritmo de Cheng e Church nas matrizes obtidas do corpus EBC, com $K = \{3, 5, 7, 10\}$, $\alpha = 1, 2$ e $\delta = 0,0008$

Fonte: Alexandra Katiuska Ramos Diaz, 2018

O diagrama da caixa dos RQMs da figura 52a mostra certas peculiaridades. Por exemplo, a maioria das matrizes binárias (80%) retornam bigrupos com RQMs iguais a zero, distorcendo o cálculo da média e dificultando sua visualização. Nas matrizes com representações TF e TF-IDF, mesmo tendo valores de suas médias próximas umas das outras, nas matrizes com representação TF a maioria (85%) de suas matrizes retornam valores de RQM menores ou iguais a 0,0002. Adicionalmente, algumas das matrizes da representação TF (15%) retornam valores de RQM atípicos mais altos ou iguais a 0,0003, gerando valores de RQM em média de $0,00019 \pm 0,00008$. O contrário acontece nas matrizes com representação TF-IDF, em que a maioria das matrizes (95%) retornam valores de RQM maiores ou iguais a 0,0002, gerando um RQM em média de $0,00023 \pm 0,00007$. Visto de uma perspectiva de distribuição dos dados, as matrizes com representações TF-IDF

trouxeram melhores resultados, tendo a maior parte de seus valores distribuídos em torno à média.

Os diagramas de caixa das figuras 52b e 52c mostram os valores obtidos pelas distância intra-grupos e pela distância inter-grupos depois da aplicação do algoritmo nas matrizes obtidas do corpus EBC. Nestas figuras é observado que as representações TF e TF-IDF apresentam objetos similares entre seus grupos (valores da distância intra-grupos baixa), e ao mesmo tempo os centroides dos grupos são próximos (valores de distância inter-grupos baixa). O fato de não ter uma boa separação entre os grupos explica os valores altos no índice DB nas representações TF e TF-IDF. O contrário acontece nas representações binárias, embora estas apresentam uma maior separabilidade entre seus grupos (valores da distância inter-grupos alta) e menor grau de compactação dos objetos dentro de cada grupo (valores de distância intra-grupos alta), os valores do índice DB parecem ser bons.

Finalmente, segundo as análises das medidas, foi visto que embora nas representações binárias tragam melhores valores de índice DB, distância inter-grupos e porcentagens do grau de sobreposição baixos, o algoritmo apresenta dificuldades na procura dos bigrupos, direcionando a busca deles aos elementos nulos. Este fato leva à obtenção de bigrupos com porcentagens do grau de ocupação e valores dos RQMs extremamente baixos, sendo na maioria dos casos iguais a zeros.

Com respeito às matrizes com representações TF e TF-IDF, seus resultados obtidos são semelhantes entre cada par de matrizes. Estes pares são formados pelo igual número de palavras em cada uma delas. Segundo a tabela 12 os pares de matrizes são: o par (Mat. 6, Mat. 11), o par (Mat. 7, Mat. 12), o par (Mat. 8, Mat. 13), e o par (Mat. 10, Mat. 15); em que as matrizes primeiras de cada par apresentam representações TF e as segundas matrizes de cada par apresentam representações TF-IDF. Dos pares os que trouxeram melhores resultados foi o par (Mat. 7, Mat. 12).

Resultados diferentes foram observados nos gráficos dos RQMs em relação às representações TF e TF-IDF, obtendo melhores resultados as representações TF-IDF. Na tabela 23 são apresentadas as médias das medidas avaliadas para cada representação.

A escolha dos melhores resultados foi baseada a partir da criação de uma ordem de prioridade das medidas. Essa ordem de prioridade foi construída segundo o comportamento das medidas em ambientes esparsos. Assim, foram considerados: valores altos na porcentagens do grau de ocupação, valores baixos nas porcentagens dos grau de sobreposição,

Tabela 23 – Médias dos medidas por representação, obtidas depois da aplicação do algoritmo de Cheng e Church nas matrizes obtidas do corpus EBC, com $K = \{3, 5, 7, 10\}$, $\alpha = 1, 2$ e $\delta = 0,0008$

Medidas	Binária	TF	TF-IDF
Ocupação (%)	$0,008 \pm 0,01$	$0,9 \pm 0,15$	$0,94 \pm 0,14$
Sobreposição (%)	$35,1 \pm 17,6$	$48,5 \pm 14,5$	$49,1 \pm 13,46$
Cobertura (%)	$52,69 \pm 9,01$	$78,35 \pm 8,05$	$79,46 \pm 7,45$
Volume	$1,1e+06 \pm 4,4e+05$	$2,1e+06 \pm 9,7e+05$	$2,2e+06 \pm 1e+03$
RQM	$2e-05 \pm 5e-05$	$1,9e-04 \pm 8e-05$	$2,3e-04 \pm 7e-05$
Índice DB	$18,86 \pm 4,1$	$45,82 \pm 6,2$	$46,5 \pm 6,2$
Dist. inter-grupos	$1,79 \pm 0,26$	$0,26 \pm 0,24$	$0,27 \pm 0,28$
Dist. intra-grupos	$3619,4 \pm 736,3$	$1535,9 \pm 444,8$	$1534,7 \pm 466,7$

Fonte: Alexandra Katiuska Ramos Diaz, 2018

valores baixos no índice DB, RQM¹ e distância intra-grupos, e valores altos na distância inter-grupos, volumes e na porcentagem de cobertura.

Assim, segundo a tabela 23 e as análises feitas em cada medidas a matriz que obteve melhores resultados foi a matriz com identificador Mat. 12, a qual apresenta 98,73% de esparsidade. Além disso, analisando os diferentes valores de K e considerando que porcentagens de sobreposição altos não ajudam nas análises qualitativas, foi escolhido o valor para o parâmetro K igual a 3. Os valores das medidas obtidos por estas configurações estão detalhados na tabela 24, sendo seus bigrupos resultantes usados para as análises qualitativas mostradas na próxima seção.

Tabela 24 – Melhores resultados obtidas após a aplicação do algoritmo de Cheng e Church na Mat. 12 do corpus EBC, com $K = 3$, $\alpha = 1, 2$ e $\delta = 0,0008$

Medidas	$K = 3$
Ocupação (%)	1,09
Sobreposição (%)	31,7
Cobertura (%)	65,2
Volume	1101212
RQM	0,0002
Índice DB	46,13
Dist. intra-grupos	1258,55
Dist. inter-grupos	0,11

Fonte: Alexandra Katiuska Ramos Diaz, 2018

¹ Dado que os valores dos RQMs são baixos na maioria dos resultados, foi levada em conta a distribuição dos dados.

4.2.3 Exp. #6: Avaliações qualitativas

Para desenvolver esta análise foi escolhida a configuração que obteve melhores resultados considerando as análises da seção anterior.

O algoritmo de Cheng e Church é capaz de encontrar bigrupos, estes bigrupos estão compostos por grupos de notícias e grupos de palavras, agrupados dependendo de suas similaridades parciais. Uma forma de visualização destes bigrupos é a nuvem de palavras, a qual permite a interpretação das palavras pertencentes a cada bigrupo. Na figura 53 é apresentada esta visualização a partir de nuvens de top-30 palavras, na qual o tamanho das palavras é definido pela ocorrência destas em cada bigrupo.

Figura 53 – Visualização em nuvem de palavras das top-30 palavras para cada bigrupo gerado pelo algoritmo de Cheng e Church, $K = 3$, $\alpha = 1, 2$ e $\delta = 0,0008$

Fonte: Alexandra Katiuska Ramos Diaz, 2018

Da figura 53, por meio da inspeção visual das palavras pertencentes a cada bigrupo, foi possível atribuir rótulos para cada um destes. Desta forma, são visualizados três tópicos diferentes: “educação” (bigrupo 1 - B# 1), “política” (bigrupo 2 - B# 2) e “futebol” (bigrupo 3 - B# 3).

Os bigrupos obtidos podem ser analisados a partir dos grupos de notícias e os grupos de palavras que formam este. Os grupos de notícias e os grupos de palavra são obtidos pelas projeções nas linhas e colunas destes bigrupos, respectivamente.

Assim, nas projeções das linhas podem ser obtidos três grupos de notícias, as quais são: o grupo de notícias 1 (GN# 1) formado pela projeção nas linhas do B# 1, o grupo de notícias 2 (GN# 2) formado pela projeção nas linhas do B# 2, e a projeção nas linhas do B# 3 formam o grupo de notícias 3 (GN# 3). Estes grupos de notícias podem ser comparados com os canais conhecidos do corpus EBC, com o intuito de ajudar no

entendimento de como os grupos de notícias foram agrupados. Na figura 54 é mostrado esta análise.

Figura 54 – Comparação dos grupos de notícias com os canais conhecidos, após a aplicação do algoritmo de Cheng e Church, com $K = 3$, $\alpha = 1,2$ e $\delta = 0.0008$

Fonte: Alexandra Katiuska Ramos Diaz, 2018

Ainda na figura 54, é observado que os três grupos de notícias formados a partir dos bigrupos apresentam notícias pertencentes à maioria dos canais, com similares porcentagens de cada um. Por exemplo:

- o GN# 1 composto de 755 notícias, as quais pertencem principalmente aos canais de educação (28,02%), cultura (11,93%), política (11,8%) e tecnologia (10,8%);
- o GN# 2 composto de 562 notícias, as quais pertencem principalmente aos canais de educação (17,26%), esportes (14,8%), geral (14,21%) e tecnologia (12,2%);
- o GN# 3 composto de 494 notícias, as quais pertencem principalmente aos canais de educação (28,22%) e política (24,05%).

Nos grupos de notícias foram encontradas sobreposições, o que significa que existem notícias pertencentes a mais de um grupo de notícias, sendo estas agrupadas dependendo das caracterizações das palavras que ocorrem nelas. Nestes grupos de notícias foram encontradas grandes porcentagens de sobreposições. Por exemplo, existem 160 notícias pertencentes ao GN# 1 e ao mesmo tempo ao GN# 2, 494 notícias pertencentes ao GN# 1 e ao GN# 3, 75 notícias pertencentes ao GN# 2 e ao GN# 3 e 75 notícias pertencentes aos três grupos de notícias (GN# 1, GN# 2 e GN# 3). Cabe ressaltar que todas as notícias pertencentes ao GN# 3 pertencem ao mesmo tempo ao GN# 1.

Figura 56 – Conteúdo da notícia 11022, pertencente ao GN# 1 e ao GN# 3, após a aplicação do algoritmo de Cheng e Church

Fonte: Alexandra Katiuska Ramos Diaz, 2018

foram encontradas notícias sobre educação e futebol que justifiquem a sobreposição e o agrupamento destas.

Da mesma forma das projeções nas linhas, os bigrupos resultantes podem ser analisados pela formação dos grupos de palavras. Assim, nas projeções nas colunas são obtidos três grupos de palavras. Estes grupos de palavras são: o grupo de palavras 1 (GP# 1), formado pela projeção do B# 1 nas colunas; o grupo de palavras 2 (GP# 2), formado pela projeção do B# 2 nas colunas e o grupo de palavras 3 (GP# 3), formado pela projeção do B# 3 nas colunas.

Nestes grupos de palavras é observado que existem sobreposições (Figura 53), ou seja existem palavras pertencentes a mais de um grupo de palavras. Por exemplo, na figura 53 as palavras "jogo", "vasco", "januário" pertencem ao mesmo tempo ao GP# 2 (obtido pelo B# 2) e ao GP# 3 (obtido pelo B# 3). Este fato indica que existem notícias

tanto no B# 2 quanto no B# 3, nas quais as palavras “jogo”, “vasco” e “janeiro” ocorrem. Exemplos destas notícias são as notícias 11034 e 11509 pertencentes aos B# 2 e B# 3 respectivamente, nas quais parte de seus conteúdos são mostrados na figura 57 com destaque nas palavras “jogo”, “vasco”, “janeiro”.

Figura 57 – Ocorrência da palavra “jogo”, “vasco”, “janeiro” nas notícias 11034 e 11509, após a aplicação do algoritmo de Cheng e Church

Fonte: Alexandra Katiуска Ramos Diaz, 2018

Além das sobreposições de linhas e colunas foi observado que existem sobreposições nos bigrupos, ou seja, o par (notícia, palavra que ocorre nela), podem pertencer a mais de um bigrupo. Um exemplo destas sobreposições é a notícia 4425, em que a palavra “ministério”, que ocorre na notícia, permite o biagrupamento dela tanto no B# 1 quanto o B# 3.

A fim de visualizar a notícia 4425, é mostrada parte de seu conteúdo na figura 58, com destaque da palavra “ministério” na cor verde. Também nesta figura são destacadas as possíveis palavras pertencentes ao B# 1 na cor amarela e as possíveis palavras pertencentes ao B# 2 na cor azul.

Para fins de visualização do efeito produzido pela alta sobreposição nos bigrupos, são mostradas as nuvens das top-30 palavras dos bigrupos resultantes após a aplicação do algoritmo com valores de $K = 7$. Esta visualização é mostrada na figura 59.

Da figura 59 pode-se observar que, a partir do terceiro bigrupo (B# 3) o algoritmo tende a retornar bigrupos semelhantes a este sem maior diferenciação de tópicos. Este fato

Figura 58 – Exemplo de sobreposição de bigrupos na notícia 4425, após a aplicação do algoritmo de Cheng e Church

Fonte: Alexandra Katiuska Ramos Diaz, 2018

torna difícil a rotulação dos bigrupos. Além disso, foram observadas maiores porcentagens de sobreposição tanto nas projeções de linhas e colunas como as sobreposições nos bigrupos. A porcentagem do grau de sobreposição apresentado por esta configuração é de 56,98%.

Ainda na figura 59, são visualizados três tópicos diferentes: o bigrupos 1 (B# 1), referido à “educação”, o bigrupos 2 (B# 2) referido à “política” e os bigrupos 3, 4, 5, 6 e 7 (B# 3, B# 4, B# 5, B# 6 e B# 7) são referidos ao “futebol”. Além disso, é observado que os três bigrupos obtidos usando $K = 3$ são similares aos bigrupos obtidos usando $K = 7$. Por exemplo, o B# 1 usando $K = 3$ é semelhante ao B# 1 usando $K = 7$, já que ambos apresentam tópicos referidos à “educação”. De forma similar, o B# 2 usando $K = 3$ é semelhante ao B# 2 usando $K = 7$, ambos com tópicos referentes à “política”. Similarmente, acontece com o B# 3 usando $K = 3$ é semelhante aos bigrupos B# 3, B# 4, B# 5, B# 6 e B# 7 usando $K = 7$, ambos com tópicos referidos ao “futebol” mas com diferentes valores na ocorrência de palavras destes.

Finalmente, segundo as experimentações foi percebido que valores maiores para o parâmetro K (maiores que 3) retornam bigrupos com altas porcentagens de sobreposição, sendo estes semelhantes entre si.

Figura 59 – Visualização em nuvem de palavras das top-30 palavras para cada bigrupo gerados pelo algoritmo de Cheng e Church, com $K = 7$, $\alpha = 1,2$ e $\delta = 0.0008$

Fonte: Alexandra Katiuska Ramos Diaz, 2018

Com respeito à comparação com os canais do corpus, nas experimentações foram visualizados só três tópicos dos dez conhecidos, sendo esses: educação, política e esportes (especificamente futebol). Isto é provavelmente causado pela grande quantidade de notícias pertinentes a estes canais.

Além das experimentações mostradas, foram feitas experimentações no corpus trocando os valores nulos do conjunto de dados por números aleatórios uniformes de 0 a 1, com o objetivo de que o algoritmo desconsidere a formação de bigrupos com elementos nulos. Mas a inserção destes valores aleatórios alteraram negativamente a descoberta dos bigrupos, trazendo bigrupos triviais (todos os elementos pertencentes aos bigrupos foram compostos pelos elementos nulos). Estes resultados não foram inseridos ou discutidos porque não eram melhores do que os já apresentados.

4.3 Considerações finais

Este capítulo apresentou os estudos e análise do algoritmo de Cheng e Church aplicado em conjuntos de dados sintéticos com e sem esparsidade, e no conjunto de dados real. As experimentações tiveram a finalidade de analisar o comportamento do

algoritmo nestes ambientes a partir das medidas de biagrupamento e das medidas internas de agrupamento.

Para ambos estudos foram necessários experimentações iniciais, a fim de escolher valores adequados para os parâmetros do algoritmo que retornassem resultados promissores em ambientes esparsos.

Do ponto de vista das experimentações nos conjuntos de dados sintéticos, foi observado que o algoritmo não apresenta uma boa adequabilidade aos níveis de esparsidade altos. Este comportamento faz que o algoritmo tenda a redirecionar a busca dos bigrupos em torno dos elementos nulos, retornando bigrupos com alta esparsidade e sobrepostos entre si.

Do ponto de vista das experimentações no corpus EBC, a partir das avaliações quantitativas é deduzido que o algoritmo mostrou dificuldades na procura dos bigrupos nas representações binárias, retornando bigrupos compostos com elementos nulos. Esse fato, no sentido do algoritmo agrupar elementos coerentes menores que um limiar, são considerados bons resultados, mas no contexto dos dados textuais, bigrupos compostos na sua totalidade com elementos nulos não trazem significados relevantes para as análises. Sobre as representações TF e TF-IDF, o algoritmo apresentou similares comportamentos em elas, mas ainda o algoritmo tende a procurar os bigrupos em torno aos elementos nulos, retornando bigrupos com esparsidade alta.

As avaliações qualitativas foram feitas a partir da configuração que obteve melhores resultados nas medidas avaliadas. Esta análise mostrou que, embora o algoritmo consiga diferenciar alguns tópicos e biagrupar as notícias pelas suas similaridades parciais. Este é afetado pela esparsidade retornando bigrupos com tópicos semelhantes uns a outros e com altas porcentagens de sobreposição. Esse fato faz com que o algoritmo não seja ideal para o biagrupamento em dados textuais e com alta esparsidade.

5 Resultados e Análises do algoritmo de NBVD

Neste capítulo são apresentados os resultados junto com as análises obtidas das experimentações do algoritmo de NBVD aplicado aos conjuntos de dados sintéticos e ao corpus EBC. O algoritmo foi estudado em dois ambientes: um ambiente sem esparsidade e em um ambiente com esparsidade. No estudo em um ambiente sem esparsidade, o algoritmo foi aplicado nos conjuntos de dados sintéticos na suas formas originais. No estudo em um ambiente com esparsidade, o algoritmo foi aplicado também aos conjuntos de dados sintéticos, mas estes foram alterados com inserções de diferentes níveis de esparsidade (descritos na seção 3.3), e aplicado também em um conjunto de dados reais (corpus EBC).

Na seção 5.1 é apresentado o estudo e análise do algoritmo aplicado nos conjuntos de dados sintéticos, analisando seu comportamento sem e com esparsidade, na seção 5.2 é apresentado a aplicação do algoritmo ao corpus EBC, e finalmente são apresentadas as considerações finais destes estudos na seção 5.3.

5.1 Estudo do algoritmo de NBVD

O estudo do algoritmo de NBVD nos conjuntos de dados sintéticos está baseado em duas análises: na análise do impacto da eleição dos parâmetros k e l aplicados nos conjuntos de dados em suas formas originais; e a análise do comportamento do algoritmo em um ambiente com esparsidade, quando é aplicado nos conjuntos de dados com diferentes níveis de esparsidade. Estas análises estão baseadas nos cogrupos retornados de interesse principal dos conjuntos de dados sintéticos (descritos na seção 3.1).

5.1.1 Espaço de busca para os parâmetros do algoritmo

Para as análises do *Exp. #1*, foram escolhidos valores para os parâmetros de k e l a partir do conhecimento a priori dos cogrupos de interesse principal e segundo o número máximo de grupos de linhas e colunas formados nos conjuntos de dados sintéticos. Esta escolha teve a finalidade de obter um espaço de busca adequado para a análise. A seguir é apresentada a estratégia utilizada para a obtenção do número máximo de grupos de linhas e de colunas formados dos conjuntos de dados sintéticos.

Dado um determinado número de cogrupos, conhecidos em cada estrutura e de interesse principal, o número máximo de grupos de linhas (k) necessárias para formar estes cogrupos é dado pela equação 25:

$$\# \text{ max. de grupos} = 2(P) + 1, \quad (25)$$

em que P é o número de cogrupos conhecidos em cada estrutura e de interesse principal.

De maneira gráfica a figura 60 apresenta três exemplos básicos. Nesta figura os cogrupos de interesse principal estão destacados da cor azul e os números máximos de grupos das linhas formados são mostradas de cor rosa no lado esquerdo. Ainda na figura destaca-se no lado direito os possíveis grupos de linhas a serem formados.

Figura 60 – Representação do número máximo de grupos de linhas

Fonte: Alexandra Katiuska Ramos Diaz, 2018

Na figura 60a é observado que, o número máximo de grupos de linhas necessários para formar o cogrupos de interesse principal é três. Mas estes grupos podem ser menores, por exemplo dois, se são consideradas o primeiro e o terceiro grupo em um (linhas da cor verde do lado direito da figura). Similarmente, na figura 60b o número máximo de grupos de linhas é cinco, mas poderiam formar-se três grupos de linhas, considerando o primeiro, o terceiro e o quinto grupo em um (linhas da cor verde do lado direito da figura). Na figura 60c, como máximo são cinco os grupos de linhas formados, mas o número de grupos poderia ser menor. Por exemplo, o número de grupos de linhas pode ser quatro se são combinados o primeiro e o quinto grupo (linhas da cor verde do lado direito da figura), ou três combinando também o segundo e o terceiro grupo de forma a considerar o cogrupos colorido de azul claro em sua totalidade (linhas da cor vermelha do lado direito da figura).

Com base no estudo dos grupos de linhas, são considerados os grupos de colunas (l), sendo mostrado na tabela 25 o espaço de busca para ambos parâmetros. Este espaço

de busca assumiu valores que variaram de 2 até o máximo número de grupos de linhas ou colunas necessárias para formar os cogrupos de interesse (calculados pela equação 25), com passos de tamanho 1 ou 3 a depender do conjunto de dados.

Tabela 25 – Espaço de busca para os parâmetro k e l , em cada conjunto de dados das estruturas

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

Fonte: Alexandra Katiuska Ramos Diaz, 2018

Para as experimentações do algoritmo nos conjuntos de dados com diferentes níveis de esparsidade, foram escolhidos os mínimos valores possíveis para os parâmetros de k e l que garantem execuções boas em termos de erro de reconstrução. Esta escolha foi baseada nas análises do *Exp.#1* e resumidos na tabela 28.

5.1.2 *Exp.#1*: estudo dos parâmetros k e l

Para esta análise foram utilizadas diferentes combinações de valores para k e l em cada conjunto de dados, as quais são detalhadas na tabela 25. Para medir o impacto da eleição destes parâmetros, foi usado o erro de reconstrução e uma análise visual da reconstrução retornada pelo algoritmo. Ambas medidas foram calculadas em cada combinação de valores de k e l , e aplicadas a cada conjunto de dados sintéticos.

Nos gráficos da figura 61 são mostrados os diferentes valores para os parâmetros de k e l aplicados nos conjuntos de dados da estrutura d (Figura 61a) e a estrutura i (Figura 61b), juntamente com os erros de construção obtidos. Nestes gráficos, o tamanho das bolhas é proporcional ao erro de reconstrução.

De acordo com a figura 61, percebe-se que é necessário estudar uma combinação de valores para k e l que permita ao algoritmo de NBVD realizar operações matriciais com

Figura 61 – Variação do erro de reconstrução do algoritmo de NBVD em relação aos parâmetros k e l nos conjuntos de dados das estruturas $d(a)$ e i (b)

Fonte: Alexandra Katiuska Ramos Diaz, 2018

sucesso e consequentemente obter resultados satisfatórios, sendo que valores pequenos nos parâmetros k e l produzem resultados piores (maiores erros de reconstrução).

Comportamentos similares aos conjuntos de dados das estruturas d e i foram encontrados nos demais conjuntos de dados. Estes comportamentos são apresentados na tabela 26. Nesta tabela é mostrado a partir de quais valores para os parâmetros k e l o algoritmo tende a trazer erros de reconstrução menores em cada conjunto de dados.

Tabela 26 – Combinação de valores dos parâmetros de k e l que levam aos melhores resultados de erro de reconstrução para cada conjunto de dados sintético

Conjuntos de dados	k	l
a	2	2
b	≥ 3	≥ 3
c	≥ 3	≥ 3
d	≥ 3	≥ 3
e	≥ 3	≥ 3
f	2	≥ 4
g	≥ 3	≥ 3
h	≥ 6	≥ 5
i	≥ 4	≥ 4

Fonte: Alexandra Katiuska Ramos Diaz, 2018

Além do erro de reconstrução, foram analisadas de maneira visual as representações gráficas das matrizes reconstruídas. Esta análise teve como objetivo determinar a partir de quais valores para os parâmetros de k e l é possível visualizar uma boa reconstrução. Os resultados desta análise são apresentados na tabela 27, mostrando a comparação dos conjuntos originais e as matrizes reconstruídas, assim como a partir de quais valores para k e l o algoritmo trouxe boas reconstruções.

Tabela 27 – Representação visual da reconstrução dos conjunto de dados sintéticos, após a aplicação do algoritmo de NBVD, com diferentes valores de k e l

Conj. original	Reconstrução	k e l	Conj. original	Reconstrução	k e l		
Estrutura a :			$k \geq 2$ $l \geq 2$	Estrutura b :			$k \geq 3$ $l \geq 3$
Estrutura c :			$k \geq 3$ $l \geq 3$	Estrutura d :			$k \geq 3$ $l \geq 3$
Estrutura e :			$k \geq 3$ $l \geq 3$	Estrutura f :			$k \geq 2$ $l \geq 2$
Estrutura g :			$k \geq 3$ $l \geq 3$	Estrutura h :			$k \geq 5$ $l \geq 5$
Estrutura i :			$k \geq 4$ $l \geq 4$				

Fonte: Alexandra Katiuska Ramos Diaz, 2018

De acordo com a tabela 27, embora as reconstruções retornadas pelo algoritmo sejam boas para os conjuntos de dados das estruturas c , f , g , h e i , o algoritmo apresentou dificuldades para produzir a separação dos cogrupos adjacentes, gerando sobreposições entre eles.

Finalmente, as análises nos conjuntos de dados sintéticos mostra que o algoritmo trouxe resultados melhores (erros de reconstrução menores e boas reconstruções visuais) quando a combinação dos valores de k e l tende a crescer. Seguindo esse raciocínio, valores suficientemente altos para os parâmetros k e l poderiam ser bons, mas sendo o algoritmo baseado em cálculos matriciais, os quais tendem a apresentar alta complexidade computacional com valores altos de k e l e em conjuntos de dados grandes, valores altos para esses parâmetros não parecem ser boas escolhas. Portanto, é preferível escolher valores de k e l que sejam o suficientemente altos para produzir resultados satisfatórios, e ao mesmo tempo suficientemente baixos para minimizar o custo computacional.

A partir das análises referentes ao erro de reconstrução e à análise visual das matrizes de reconstrução, foram escolhidos valores dos parâmetros k e l para cada conjunto de dados sintéticos. Os valores considerados foram os menores valores possíveis para k e l que permitiram um erro de reconstrução pequeno. Estes valores são apresentados na tabela 28 e foram utilizados nas experimentações do algoritmo em conjuntos de dados com níveis de esparsidade.

Tabela 28 – Valores dos parâmetros de k e l que obtiveram bons resultados do algoritmo de NBVD nos conjuntos de dados sintéticos

	Conjuntos de dados								
	a	b	c	d	e	f	g	h	i
k	2	3	3	3	3	2	3	6	4
l	2	3	3	3	3	4	3	5	4

Fonte: Alexandra Katiuska Ramos Diaz, 2018

5.1.3 *Exp. #2*: estudo do comportamento do algoritmo nos conjuntos de dados com níveis de esparsidade

Neste estudo foram analisadas as medidas internas de agrupamento calculadas nos grupos das linhas, as medidas de biagrupamento e o erro de reconstrução do algoritmo de NBVD em um ambiente com esparsidade. Para estas análises foram considerados os valores de k e l apresentados na tabela 28 e os conjuntos de dados com diferentes níveis de esparsidade.

As análises das medidas da distância inter-grupo, distância intra-grupos e o índice DB são apresentadas na figura 62. Estas figuras mostram as variações das medidas em relação ao incremento dos níveis de esparsidade em cada conjunto de dados sintéticos.

Figura 62 – Variação das medidas internas de agrupamento em relação ao incremento dos níveis de esparsidade, em cada conjunto de dados sintéticos após a aplicação do algoritmo de NBVD

Fonte: Alexandra Katiuska Ramos Diaz, 2018

A medida da distância inter-grupos mede a separação dos centroides dos grupos. Valores altos nesta medida indicam separações melhores desses centroides. Na figura 62a é mostrada uma tendência decrescente conforme a esparsidade aumenta, isto implica que os centroides dos grupos das linhas tendem a estar mais próximos uns dos outros, seguindo o aumento da esparsidade. Uma diminuição pronunciada pode ser observada a partir do nível de esparsidade de 90 até 98, em que os centroides dos grupos retornados são muito semelhantes entre si.

A medida da distância intra-grupos mede a similaridade entre os objetos que pertencem aos mesmos grupos. Valores baixos desta medida indicam similaridades maiores entre os objetos. Na figura 62b é mostrado que, conforme a esparsidade aumenta o algoritmo tende a retornar grupos com valores cada vez mais diferentes entre si (inserção de elementos nulos nos grupos), ocasionando uma baixa similaridade entre seus elementos (incremento dos valores da distância intra-grupos). Este comportamento continua até alcançar seus valores mais altos em torno de níveis de esparsidade de 40 ou 50 (a depender

do conjunto de dados). A partir deste intervalo de esparsidade, os elementos dos grupos retornados começam a se tornar novamente parecidos, porém a similaridade é devido à grande quantidade de elementos nulos (os objetos são similares porque possuem muitos elementos nulos), causando uma diminuição dos valores da distância intra-grupos.

Os gráficos do índice DB (Figura 62c) mostram um leve incremento deste conforme os níveis de esparsidade aumentam. Embora exista uma maior similaridade dos elementos a seus grupos (diminuição dos valores das distâncias intra-grupos), estes grupos tendem a se assemelhar (diminuição dos valores das distâncias inter-grupos), afetando o índice DB. A partir dos níveis de esparsidade de 90 até 98 o índice DB piora, já que a separação entre os centroides dos grupos resultantes diminui.

Ainda na figura 62, com respeito ao conjunto de dados da estrutura a , o número mínimo de grupos de linhas e colunas para encontrar o cogruppo de interesse principal é 2 ($k = 2$ e $l = 2$), levando à formação de dois grupos de linhas semelhantes entre si, com centroides próximos. Este comportamento é ilustrado com valores diferentes de 0 na distância inter-grupos, e elevados valores no índice DB.

As análises das medidas dos volumes e os RQMs dos cogrupos resultantes são mostrados na figura 63. Nesta figura, é mostrado o comportamento do algoritmo frente aos níveis de esparsidade com respeito a estas medidas.

Figura 63 – Variação da média dos volumes e dos RQMs em relação aos incremento dos níveis de esparsidade, em cada conjunto de dados sintéticos, após a aplicação do algoritmo de NBVD

Fonte: Alexandra Katiuska Ramos Diaz, 2018

De acordo com a figura 63a é observado que os volumes dos cogrupos retornados em relação à esparsidade são estáveis, com uma tendência de queda em níveis de esparsidade altos. Este comportamento provavelmente ocorre pela diminuição de elementos não nulos no conjunto de dados. Adicionalmente, na figura 63b é percebido que os valores dos

RQMs dos cogrupos retornados apresentam um comportamento semelhante à distância intra-grupos (Figura 62b). Conforme os níveis de esparsidade aumentam, a coerência entre os elementos dos cogrupos tende a diminuir (aumento dos valores dos RQMs) até níveis de esparsidade de 40 ou 50. A partir deste intervalo, os cogrupos retornados apresentam novamente maior coerência entre seus elementos (diminuição dos valores dos RQMs).

As análises da porcentagem de cobertura, da porcentagem do grau de sobreposição e da porcentagem do grau de ocupação são mostradas na figura 64.

Figura 64 – Variação da porcentagem de cobertura, do grau de sobreposição e do grau de ocupação dos bigrupos em relação aos incremento dos níveis de esparsidade, em cada conjunto de dados sintéticos após a aplicação do algoritmo de NBVD

Fonte: Alexandra Katiuska Ramos Diaz, 2018

De acordo com a figura 64a percebe-se que as porcentagens da cobertura dos cogrupos não apresentam maior variação em relação à esparsidade. Este comportamento é similar ao volume dos cogrupos. Variações relevantes são produzidas com níveis de esparsidade altos (90 até 98), em que tanto o volume como a porcentagem de cobertura dos cogrupos diminuem pela diminuição de elementos não nulos nos conjuntos de dados. Em relação à porcentagem do grau de sobreposição, a figura 64b mostra um comportamento estável conforme a esparsidade aumenta (exceção para o conjunto de dados da estrutura

a). A partir do nível de esparsidade de 90 as sobreposição dos cogrupos aumentam consideravelmente.

A porcentagem do grau de ocupação dos cogrupos resultantes (Figura 64c) ajuda no entendimento das outras medidas. Esta medida apresenta uma tendência linear decrescente em relação à esparsidade, implicando uma maior existência de elementos nulos nos cogrupos resultantes. A existência de cogrupos com um maior número de elementos nulos leva a valores baixos na distância inter-grupos, na distância intra-grupos, e nos RQMs e valores altos no índice DB. Em níveis de esparsidade altos (90 até 98) o algoritmo tende a retornar cogrupos com volumes levemente menores e com altos graus de sobreposição entre eles.

Além das medidas internas de agrupamento e das medidas de biagrupamento, o algoritmo foi avaliado utilizando o erro de reconstrução. O comportamento dos valores desta medida é mostrado nos gráficos da figura 65. Adicionalmente é mostrado a soma residual de quadrados (RSS)¹ em relação à esparsidade.

Figura 65 – Variação do erro de reconstrução em relação aos incremento dos níveis de esparsidade, em cada conjunto de dados sintéticos, após a aplicação do algoritmo de NBVD (a), Variação do RSS dos elementos nulos e não nulos na estrutura d

Fonte: Alexandra Katiuska Ramos Diaz, 2018

Segundo a figura 65a, inicialmente a adição dos elementos nulos piora o erro de reconstrução, atingindo seus maiores valores em um nível de esparsidade aproximado de 50. A partir deste nível de esparsidade o algoritmo tende a se adaptar aos elementos nulos, dado que os conjuntos de dados apresentam uma maior quantidade deles, retornando melhores resultados em termos de erro de reconstrução. Para detalhar este comportamento

¹ Do inglês *Residual Sum of Squares*, uma medida da discrepância entre os dados e um modelo de estimativa. Um RSS pequeno indica um ajuste apertado do modelo aos dados. Esta medida é definida como: $\sum_{i=1}^n (y_i - f(x_i))^2$, em que y_i é o i -ésimo valor da variável a ser prevista e $f(x_i)$ é o valor previsto de y_i .

a figura 65b mostra o cálculo dos RSS gerados pelos elementos nulos (linha azul) e pelos elementos não nulos (linha vermelha) em relação à esparsidade na estrutura d . Nesta figura é notado que os elementos nulos em conjuntos de dados com pouca esparsidade (0 até 40) geram valores de RSS maiores do que os gerados pelos elementos não nulos. Este comportamento não se repete em conjuntos de dados com alta esparsidade (60 até 98), em que os valores dos RSS gerados pelos elementos nulos tendem a diminuir, pela adaptação do algoritmo a estes elementos, levando a menores valores destes comparados com os valores gerados pelos elementos não nulos. Adicionalmente, na figura é observado que no nível de esparsidade de 50, o algoritmo produz um maior RSS para ambos elementos, já que a quantidade de elementos nulos e não nulos é igual.

Finalmente, com base no comportamento das análises das figuras mostradas anteriormente, pode-se deduzir que o algoritmo de NBVD é adaptável ao incremento da esparsidade. A natureza do algoritmo lhe permite encontrar uma equivalência entre a importância dos elementos não nulos e os elementos nulos dentro dos conjuntos de dados esparsos (figura 65). Embora os cogrupos retornados sejam esparsos² (diminuição do grau de ocupação), estes ainda são agrupados com base nas informações relevantes dos conjuntos de dados e não em torno aos elementos nulos. Isto permite uma boa separação dos cogrupos e até mesmo preservar a qualidade da reconstrução em relação ao conjunto original (comportamento mostrado na figura 66), adicionalmente leva a um comportamento estável na média dos volumes e da porcentagem de cobertura.

Figura 66 – Reconstrução visual dos resultados do algoritmo de NBVD no conjunto de dados da estrutura d com diferentes os níveis de esparsidade de 30, 50, 70, 90 e 98

Fonte: Alexandra Katiuska Ramos Diaz, 2018

A partir de uma esparsidade de 90 até 98, o algoritmo diminui sua capacidade de separar efetivamente as informações, retornando cogrupos semelhantes entre si. Este comportamento traz valores altos na porcentagem do grau de sobreposição e o índice DB.

² Nos níveis de esparsidade de 0 até 90

5.2 Algoritmo de NBVD aplicado no corpus EBC

A fim de analisar o comportamento do algoritmo de NBVD no contexto da resolução da tarefa de coagrupamento em mineração de texto foi utilizado o corpus EBC. Este corpus está descrito na seção 3.2 juntamente com o pré-processamento realizado sobre ele.

A representação do corpus varia dependendo das configurações de pré-processamento utilizado, as quais estão descritas na tabela 12. Na tabela, cada matriz de dado pode ser usada como entrada para o algoritmo e essas são nomeadas com identificadores de Mat. 1 até Mat. 15. Estes identificadores são usados nas análises dos resultados da aplicação do algoritmo sobre os dados textuais.

Para a eleição adequada dos parâmetros k e l , de forma a produzir resultados satisfatórios, e segundo as análises da subseção 5.1.2 é recomendável selecionar valores baseados em progressões numéricas. Nas experimentações no corpus EBC, podem ser usadas duas heurísticas: o uso de progressões aritméticas crescentes (3, 5, 6, 7, 9, 11); ou a eleição baseada no número de classes conhecidas no corpus, com passos de tamanho de 2 e 3 para obter valores inferiores (3, 5, 7, 10). Neste trabalho foi utilizada a segunda heurística, sendo os valores de k e l iguais a 3, 5, 7 e 10, com todas suas combinações entre eles.

Nas próximas seções são apresentadas as análises quantitativas e qualitativas obtidas a partir da aplicação do algoritmo de NBVD sobre o corpus EBC.

5.2.1 *Exp. #3*: Avaliações quantitativas

Os resultados obtidos pelo algoritmo foram avaliados e analisados usando as medidas de biagrupamento, as medidas internas de agrupamento e o erro de reconstrução. Estas medidas são apresentadas usando mapas de calor e mostrando suas distribuições de valores em cada representação do corpus.

Para os mapas de calor as cores são associadas de acordo com os valores de cada medição, variando da cor amarela à cor vermelha. A cor amarela representa bons resultados, e a cor vermelha resultados piores. Os mapas mostram os resultados obtidos para uma determinada medida, com diferentes matrizes de dados de entradas (eixo x), usando

diferentes valores para o parâmetro l (eixo y) e fixando um determinado valor para o parâmetro k .

Na figura 67 é mostrado o mapa de calor do erro de reconstrução. De acordo com esta análise, observa-se que as matrizes que têm representações binárias (Mat. 1 até Mat. 5) geraram maior erro de reconstrução. Ainda na figura pode-se visualizar que matrizes com identificadores de Mat. 7 e Mat. 12, pertencentes às representações TF e TF-IDF respectivamente, obtiveram os melhores resultados. Este comportamento é semelhante para todos os valores de k (3, 5, 7, 10). Ambas matrizes de dados (Mat. 7 e Mat. 12) são caracterizadas por ter menor número de palavras no corpus e diferenciando-se entre elas pelo seu tipo de representação.

Figura 67 – Mapa de calor do erro de reconstrução dos resultados obtidos a partir da aplicação do algoritmo de NBVD nas matrizes obtidas do corpus EBC, com $k = 3$ e $l = \{3, 5, 7, 10\}$

Fonte: Alexandra Katiuska Ramos Diaz, 2018

A medida do grau de ocupação mostra a porcentagem de elementos não nulos nos cogrupos. Por causa da alta esparsidade inerente ao corpus, os cogrupos resultantes são também esparsos, com baixos graus de ocupação. Nas experimentações (Figura 68) foi observado que maiores valores do grau de ocupação foram retornados pelas representações binárias ($3, 99 \pm 0, 72$), seguidas da representação TF-IDF ($2, 9 \pm 0, 45$) e a representação TF ($2, 3 \pm 0, 46$). Ainda, na figura é observado que matrizes com menor número de palavras (ou seja, matrizes de dimensionalidade menor), como são as matrizes Mat. 7 e Mat. 12, apresentaram valores de grau de ocupação melhores.

Sobre a porcentagem do grau de sobreposição, as representações que apresentaram maior porcentagem foram as representações binárias (em média é de $36, 57 \pm 16, 43$), seguidas

Figura 68 – Mapa de calor da porcentagem do grau de ocupação dos resultados obtidos a partir da aplicação do algoritmo de NBVD nas matrizes obtidas do corpus EBC, com $k = 7$ e $l = \{3, 5, 7, 10\}$

Fonte: Alexandra Katiuska Ramos Diaz, 2018

das matrizes com representação TF ($33,06 \pm 16,11$) e as matrizes com representação TF-IDF ($31,06 \pm 16,23$). Diante da análise de todos os resultados, foi observado que os melhores resultados emergem quando os valores dos parâmetros k e l são iguais. Por exemplo, na figura 69, o mapa de calor mostra que, para valores de $k = 3$ e $l = 3$, as porcentagens de sobreposição são mais baixas (cor amarelo) do que nas demais variações. Também foi observado que o grau de sobreposição aumenta conforme os valores dos parâmetros de k e l aumentam.

Figura 69 – Mapa de calor da porcentagem do grau de sobreposição dos resultados obtidos a partir da aplicação do algoritmo de NBVD nas matrizes obtidas do corpus EBC, com $k = 3$ e $l = \{3, 5, 7, 10\}$

Fonte: Alexandra Katiuska Ramos Diaz, 2018

Similarmente, resultados melhores para o índice DB são alcançados quando os valores de k e l são iguais. De maneira geral, as matrizes com representações binárias apresentaram melhores resultados (em média de $13,82 \pm 6,74$) do que as representações TF e TF-IDF, diante deste índice. Estas últimas apresentaram resultados próximos, sendo estes em média de $24,28 \pm 6,62$ e $22,9 \pm 6,7$ respectivamente. Além disso, na figura 70, considerando as representações TF e TF-IDF, é observado um comportamento similar ao grau de ocupação, em que matrizes com menor número de palavras como são as matrizes Mat. 7 e Mat. 12, apresentaram melhores resultados.

Figura 70 – Mapa de calor do índice DB dos resultados obtidos a partir da aplicação do algoritmo de NBVD nas matrizes obtidas do corpus EBC, com $k = 7$ e $l = \{3, 5, 7, 10\}$

Fonte: Alexandra Katiuska Ramos Diaz, 2018

Os resultados das medidas do volume, da porcentagem da cobertura, RQM, distância inter-grupos e a distância intra-grupos seguem analisados por meio de diagramas de caixa. Nesses gráficos, o eixo y apresenta três representações utilizadas nas experimentações binário, TF e TF-IDF, e o eixo x mostra as distribuições dos valores das medidas.

Analisando os gráficos da figura 71 é observado que, embora a representação binária gere uma distribuição simétrica em seus valores de volume e porcentagem de cobertura, os volumes e cobertura retornados por ela foram menores do que as representações TF e TF-IDF. Os valores das médias e seus desvios padrão são relatados na tabela 29.

A partir dos gráficos da figura 72 conclui-se que, as representações TF e TF-IDF produziram valores de RQMs (apresentado em escala logarítmica para melhorar a visualização), distância inter-grupos e distância intra-grupos com menor variabilidade entre eles, sendo estes próximos entre si. Também pode ser observado que essas representações

Figura 71 – Diagrama de caixa do volume (a) e da porcentagem de cobertura (b) dos resultados obtidos após a aplicação do algoritmo de NBVD nas matrizes obtidas do corpus EBC

Fonte: Alexandra Katiuska Ramos Diaz, 2018

geraram resultados melhores nos valores de RQM e distância intra-grupos e piores nas distância inter-grupos com respeito à representação binária. Na tabela 29 são apresentadas as médias das medidas separadas por representação (binária, TF e TF-IDF).

Figura 72 – Diagrama de caixa do RQM (a), distância inter-grupos (b) e distância intra-grupos (c) dos resultados obtidos após a aplicação do algoritmo de NBVD nas matrizes obtidas do corpus EBC

Fonte: Alexandra Katiuska Ramos Diaz, 2018

Tabela 29 – Médias das medidas por representação, obtidas depois da aplicação do algoritmo de NBVD nas matrizes obtidas do corpus EBC

Medidas	Binária	TF	TF-IDF
Ocupação (%)	3,99 ± 0,72	2,3 ± 0,46	2,9 ± 0,45
Sobreposição (%)	36,57 ± 16,43	33,06 ± 16,11	31,06 ± 16,23
Cobertura (%)	25,47 ± 5,07	31,73 ± 4,83	33,29 ± 4,91
Volume	434,6e+03 ± 203,5e+03	560,6e+03 ± 270,1e+03	562,1e+03 ± 265,8e+03
RQM	0,04 ± 5e-03	0,002 ± 3e-04	0,002 ± 3e-04
Índice DB	13,82 ± 6,74	24,28 ± 6,62	22,9 ± 6,7
Dist. inter-grupos	2,34 ± 0,4	0,3 ± 0,005	0,3 ± 0,05
Dist. intra-grupos	4,17 ± 1,06	1,13 ± 344,8	1,13 ± 338,3
Erro de reconstrução	284,3 ± 53,2	89,9 ± 18,9	86,9 ± 18,9

Fonte: Alexandra Katiuska Ramos Diaz, 2018

A escolha dos melhores resultados para seguir com a análise qualitativa foi baseada em uma ordem de prioridade das medidas. Esta ordem foi construída segundo o comportamento das medidas em ambientes esparsos. A ordem considerada foi: menores valores do erro de reconstrução, altos graus da porcentagem da ocupação³, baixas porcentagens dos graus de sobreposição⁴, valores baixos do índice DB, dos RQM e da distância intra-grupos, valores altos nas distâncias inter-grupos e valores altos nas porcentagens de cobertura e volumes.

Embora exista uma variabilidade de resultados com respeito à distância inter-grupos, ao grau de ocupação e ao índice DB, as matrizes com representações TF-IDF permitiram produzir, em sua maioria, resultados melhores do que as outras representações. Na representação TF-IDF, a matriz Mat.12, com 98,73% de esparsidade, foi a que obteve os melhores resultados nas medidas avaliadas. Além disso, na análise dos dados foi observado que nas medidas do grau de sobreposição e do índice DB, os melhores resultados foram produzidos quando k e l são iguais. Na tabela 30 são relatados os melhores resultados obtidos com o uso da matriz de dados Mat.12.

Para a realização da análise qualitativa sobre os cogrupos obtidos nas configurações relatadas na tabela 30, foram escolhidos os resultados obtidos com os parâmetros k e l baixos ($= 3$) e altos ($= 7$ e $= 10$). Essas análises são apresentadas na próxima seção.

³ Dado que os cogrupos são esparsos, é de interesse analisar cogrupos que tenham a maior quantidade de elementos não nulos

⁴ Elevados porcentagens do grau de sobreposição dificultam a análise pela aparição de cogrupos semelhantes entre si

Tabela 30 – Melhores resultados obtidas depois da aplicação do algoritmo de NBVD nas matrizes obtidas do corpus EBC, considerando a configuração de pré-processamento que gera a Mat. 12, com $k = l = 3$, $k = l = 7$ e $k = l = 10$

Medidas	$k = 3, l = 3$	$k = 7, l = 7$	$k = 10, l = 10$
Ocupação (%)	3,98	3,77	3,40
Sobreposição (%)	6,04	12,63	27,31
Cobertura (%)	22,73	30,23	33,87
Volume	309980,6	189741,42	160547,7
RQM	0,002	0,002	0,002
Índice DB	15, 95	15,55	18,48
Dist. inter-grupos	0,23	0,27	0,25
Dist. intra-grupos	840,54	614,1	578,83
Erro de reconstrução	57,02	56,52	56,32

Fonte: Alexandra Katiuska Ramos Diaz, 2018

5.2.2 Exp. #3: Avaliações qualitativas

A análise apresentada nesta seção tem a finalidade de discutir a relação existente entre os grupos de palavras e os grupos de notícias encontrados pelo algoritmo de NBVD. Além disso, intenciona mostrar os resultados obtidos pelo algoritmo em termos semânticos quando trabalha com valores de k e l pequenos e grandes. Nesta análise foram consideradas as configurações relatadas na tabela 30.

O algoritmo de NBVD é capaz de encontrar grupos de notícias e grupos de palavras. Ademais, cada grupo de notícias está relacionado com os grupos de palavras por um fator (matriz S), de modo que cada grupo de palavras está relacionado com grupos de notícias. Para o caso de $k = l = 3$, a relação dos grupos de linhas com os grupos de palavras, estabelecida na matriz S da decomposição construída pelo NBVD, pode ser visualizada na tabela 31.

Tabela 31 – Matriz S gerada por após a aplicação do algoritmo de NBVD no corpus com $k = 3, l = 3$, considerando a matriz de dados Mat.12

	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

Fonte: Alexandra Katiuska Ramos Diaz, 2018

A partir da matriz S , é possível relacionar os grupos de notícias e os grupos de palavras que formam cada cogrupos. Na tabela 31 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 cogrupos 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 cogrupos (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 cogrupos (C# 3). Adicionalmente, por meio do processo de pós-processamento detalhado no capítulo 2, 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 das palavras pertencentes a cada cogrupos é a nuvem de palavras. Na nuvem, o tamanho das palavras é definido pela magnitude de sua ocorrência em cada cogrupos. Na figura 73 é mostrada a visualização a partir de nuvens de 30 palavras por cada cogrupos. Ainda na figura, por meio de inspeção visual das palavras pertencentes a cada cogrupos é possível rotular cada um destes cogrupos. Desta forma foram obtidos três cogrupos referentes aos tópicos de “Cidadania”, “Educação” e “Política”.

Figura 73 – Visualização em nuvem de palavras das top-30 palavras para cada cogrupos gerado pelo algoritmo NBVD

Fonte: Alexandra Katiuska Ramos Diaz, 2018

Uma análise possível para discutir os cogrupos formados é comparar a projeção dos cogrupos nas linhas (matriz U) com as classes conhecidas para o conjunto de dados. Ou seja, comparar os grupos de notícias pertencentes a cada cogrupos, com seus canais já conhecidos, possibilitando o entendimento da formação destes grupos de notícias. Na figura 74 é apresentada esta análise. Cada cor é associada a um canal, e a extensão da área associada a essa cor nas barras do gráfico equivale à porcentagem de notícias pertencentes ao canal, por cada grupo de notícias descoberto pelo algoritmo NBVD.

Figura 74 – Comparação dos grupos de notícias com os canais conhecidos

Ademais, na figura observa-se que o GN# 1, formado pela projeção do C# 1, apresenta 408 notícias, das quais a maioria pertence aos canais de cidadania (26,2%), geral (18,92%), cultura (18,1%) e tecnologia (16,5 %), isso implica que o GN # 1 é caracterizado por palavras relevantes sobre esses tópicos. Do mesmo modo o GN # 2, com 275 notícias, está caracterizado por palavras referentes ao canal de educação (68,9 %). E no GN # 3, das 283 notícias, 35,9 % pertencem ao canal de política, implicando sua caracterização.

No contexto de análise proveniente de coagrupamento, uma notícia pode pertencer a dois ou mais grupos de notícias, a depender dos grupos de palavras que caracterizam cada cogrupamento. A sobreposição de linhas (notícias) nos grupos pode ser explicada por meio da figura 75. Na figura, a notícia 1 pertence ao grupo de linhas da cor laranja por conter as palavras que caracteriza o grupo (palavra P1) e, ao mesmo tempo, pertence ao grupo de linhas da cor verde por conter as palavras que caracterizam esse grupo (palavra P2).

Figura 75 – Exemplo gráfico de sobreposição de linhas no corpus EBC

Fonte: Alexandra Katiuska Ramos Diaz, 2018

Nos resultados obtidos foram observadas diferentes sobreposições nos grupos de notícias: 65 notícias que pertencem ao GN# 1 e ao mesmo tempo ao GN# 2; 74 notícias

referentes ao GN# 1 e ao GN# 3; 75 notícias que estão no GN# 2 e no GN# 3; e 20 notícias que pertencem aos três grupos de notícias, totalizando 234 notícias pertencentes a mais de um grupo de notícias. Analisando um exemplo de sobreposição de notícias, a figura 76 mostra o conteúdo da notícia 10999 pertencente ao GN# 2 e ao GN# 3.

Figura 76 – Conteúdo da notícia 10999, pertencente ao GN# 2 e ao GN# 3

<newstitle>Justiça manda Instituto Federal de São Paulo adotar prova para seleção de alunos</newstitle>
<newssubtitle></newssubtitle>

<newstext>A Justiça Federal determinou que o Instituto Federal de Educação, Ciência e Tecnologia de São Paulo (IFSP) deixe de adotar o histórico escolar como critério de seleção de novos estudantes para o segundo semestre deste ano e que seja estabelecida a prestação de prova como parâmetro para a escolha dos alunos. De acordo com a decisão liminar, expedida na quarta-feira (5) passada, a instituição deve alterar o Edital 385 deste ano.

O instituto é alvo de uma ação civil pública do MPF por violar os princípios da isonomia e da razoabilidade ao prever a seleção dos candidatos com base na média global de notas do ensino fundamental, informou o Ministério Público Federal em São Paulo (MPFSP). “Sem uniformidade nos currículos e nas condições das escolas no país, o critério não garante que os estudantes com maior merecimento ingressem no instituto”, divulgou o órgão.

“Infelizmente não há vagas para todos. Por isso, e como os recursos que mantêm os cursos técnicos viabilizados pelo IFSP são públicos, deve ser realizada a seleção dos candidatos pelo critério que melhor e igualitariamente afira o mérito de cada um deles: prova”, argumentou o MPF nos pedidos acolhidos pela Justiça Federal. “É, a prova, o instrumento de avaliação que melhor contempla o tratamento isonômico para comparação dos conhecimentos adquiridos ao longo da vida escolar.”

De acordo com o MPF, embora a Lei de Diretrizes e Bases da Educação (Lei 9.394/96) não defina os critérios a serem adotados nos processos seletivos, é descabido o emprego de um parâmetro diferente do utilizado em outros concursos públicos dessa natureza, como o Exame Nacional do Ensino Médio (Enem) e vestibulares.

Nos últimos nove anos, o IFSP havia aplicado provas para a seleção dos estudantes. “O MPFSP rebateu as alegações do instituto de que a mudança do critério se deveu a dificuldades financeiras. Como nas edições anteriores, o pagamento de taxas de inscrição poderia subsidiar os custos da aplicação dos exames. Além disso, a Procuradoria ressaltou que a restrição orçamentária não deve ser pretexto para o desrespeito ao princípio da isonomia”, informou a instituição.

GN# 2

GN# 3

Fonte: Alexandra Katiuska Ramos Diaz, 2018

Na figura 76 é possível observar que, embora a notícia 10999 apresente palavras tais como “educação”, “alunos”, “semestre”, “estudantes”, etc. que a caracterizam no GN# 2 (maioria de notícias do canal de educação), ela apresenta também palavras tais como “federal”, “ministério” e “lei”, que permitem à notícia ser também agrupada no GN# 3 (maioria de notícias do canal de política).

Similarmente às sobreposições das projeções das linhas dos cogrupos, é possível analisar as sobreposições das projeções das colunas (palavras). Esta sobreposição pode ser explicada com a existência de uma ou mais palavras que pertencem a diferentes grupos de palavras. No exemplo da figura 77, a palavra P1 é agrupada no grupo de palavras da cor laranja por ocorrer na notícia 1 e ao mesmo tempo agrupada no grupo de palavras da cor verde por ocorrer na notícia 2.

Figura 77 – Exemplo gráfico de sobreposição de colunas no corpus EBC

Fonte: Alexandra Katiuska Ramos Diaz, 2018

Na tabela 32 são mostradas as top-30 palavras pertencentes a cada grupo de palavras da matriz S. Na tabela é possível observar que existem palavras que aparecem em diferentes grupos. Um exemplo é a palavra “social” que aparece 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 maior ocorrência é no GP# 3, (veja o tamanho da palavra no C# 1 visualizado na nuvem de palavras da figura 73). Um exemplo da ocorrência da palavra “social” é apresentado na figura 78 nas notícias 2425 e 6008.

Tabela 32 – Top-30 palavras para cada grupo de palavras

GP# 3 “Cidadania”		GP# 1 “Educação”		GP# 2 “Política”	
civil	sociedade	médio	privadas	prisão	econômica
pessoas	violência	estudante	resultado	aprovação	lei
mundo	direitos	social	fundo	supremo	votação
saúde	pesquisa	recursos	candidatos	comissão	câmara
situação	zona	inscrições	sistema	julgamento	social
acordo	municipal	enem	semestre	contas	governo
área	secretaria	salários	alunos	recursos	deputados
federal	número	universidade	exame	sessão	tribunal
gente	polícia	superior	ensino	acordo	congresso
população	importante	instituições	educação	juiz	república
ministério	comunidade	prazo	renda	ministro	pedido
vítimas	casos	curso	curiosos	federal	operação
humanos	uso	estudantil	programa	ministério	decisão
local	social	mec	vagas	proposta	eleições
mostra	segurança	estudantes	financiamento	valor	temer

Fonte: Alexandra Katiuska Ramos Diaz, 2018

A figura 78 ajuda na interpretação da função da palavra “social”. Ainda que seu significado não mude, ela adquire uma função especial em cada contexto. Na notícia 2425, pertencente ao C# 1, a palavra “social” se refere à inclusão do “nome social” nos registros civis para transexuais. Por outro lado, na notícia 6008 ela se refere ao pagamento de contas por “rede social” como o Facebook.

Figura 78 – Ocorrência da palavra “social” nas notícias 6008 e 2425

Notícia 2425

```
<newstitle>Entidades defendem no STF mudança de registro civil para transexuais</newstitle>
<newssubtitle></newssubtitle>
<newstext>Duas entidades que lutam pelos direitos dos transexuais defenderam hoje (20) no Supremo Tribunal Federal (STF) a possibilidade de alteração do nome no registro civil sem a realização de cirurgia de mudança de sexo. O plenário da Corte iniciou o julgamento de um recurso contra decisão da Justiça do Rio Grande do Sul, que negou autorização para que um cartório local aceitasse a inclusão do nome social como verdadeira identificação civil. Os magistrados entenderam que deve prevalecer o princípio da veracidade nos registros públicos.
Ao subirem à tribuna do STF, os advogados da Anis – Instituto de Bioética, Direitos Humanos e Gênero e da Associação Brasileira de Lésbicas, Gays, Bissexuais, Travestis e Transexuais (ABGLT) afirmaram que o requisito para alteração é inconstitucional.
De acordo com o advogado Leonardo Almeida Lage, representante da Anis, a alteração do registro é uma necessidade essencial para a vida dos transexuais e está de acordo com o princípio da proteção constitucional da dignidade humana.
"As consequências para a vida dessa pessoa são absolutamente nefastas, incluindo o isolamento social, o sentimento profundo de
```

Notícia 6008

```
<newstitle>Tecnologia pioneira permitirá transferir recursos e pagar contas por rede social</newstitle>
<newssubtitle></newssubtitle>
<newstext>Uma tecnologia pioneira desenvolvida no Brasil pretende dar agilidade e segurança a quem precisar transferir recursos para a conta bancária de terceiros. Dentro de uma semana, o Banco do Brasil (BB) lançará uma ferramenta de envio de dados bancários por meio de códigos QR (desenhos com informações que podem ser lidas por câmeras fotográficas) e de redes sociais, facilitando a divisão de contas em bares e restaurantes.
Algumas redes sociais, como o Facebook, permitem transferências de valores por meio da ferramenta de mensagens. No entanto, nos moldes desenvolvidos pelo Banco do Brasil, Mastroeni assegura que a tecnologia é pioneira no mundo.
"Lá, a operação parte de quem faz a transferência e só pode ser feita numa única rede social, que é o Facebook. Na nossa ferramenta, quem vai receber os recursos envia os dados criptografados para quem vai pagar por meio de qualquer rede social", esclareceu o diretor.</newstext>
```

SOCIAL

Fonte: Alexandra Katiuska Ramos Diaz, 2018

Além dos cogrupos obtidos para os valores de k e l igual a 3, foram analisados os cogrupos obtidos com valores de k e l igual a 7 e a 10, com o objetivo de mostrar as diferenças de resultados entre os cogrupos resultantes quando estes valores são maiores.

Nas tabelas 33 e 34 são apresentadas as matrizes S obtidas para $k = 7$, $l = 7$, e $k = 10$, $l = 10$ respectivamente. Nas referidas tabelas pode-se observar os fatores que relacionam os grupos de palavras e os grupos de notícias, com um destaque para os fatores que formam os cogrupos de interesse.

Tabela 33 – Matriz S gerada por após a aplicação do algoritmo de NBVD no corpus com $k = 7$, $l = 7$

	GP# 1	GP# 2	GP# 3	GP# 4	GP# 5	GP# 6	GP# 7
GN# 1	0,0	0,0	0,0	0,0	0,0	0,0	2,76
GN# 2	0,0	0,0	0,0	0,0	0,0	3,68	0,0
GN# 3	0,0	2,94	0,0	0,0	0,0	0,0	0,0
GN# 4	0,0	0,0	0,0	0,0	3,6	0,0	0,0
GN# 5	0,0	0,0	0,0	3,14	0,0	0,0	0,0
GN# 6	0,0	0,0	2,87	0,0	0,0	0,0	0,0
GN# 7	2,61	0,0	0,0	0,0	0,0	0,0	0,0

Fonte: Alexandra Katiuska Ramos Diaz, 2018

Na tabela 33 é observada a formação de 7 cogrupos de interesse para a análise. Sendo estes formados por: o GN# 1 e o GP# 7 formam o C# 1; o C# 2 esta formado pelo GN# 2 e o GP# 6; o C# 3 é formado pelo GN# 3 e o GP# 2; o C# 4 formado pelos GN# 4 e o GP# 5; o C# 5 formado pelos GN# 5 e os GP# 4; o C# 6 formado pelos GN# 6 e o GP# 3 e os GN# 7 junto com o GP# 1 formam o C# 7.

A figura 79 apresenta a visualização de nuvens das 30 palavras mais frequentes pertencentes a cada cogruppo, onde o tamanho das palavras é proporcional à ocorrência destas nos cogrupos. Uma rotulação foi realizada a partir da análise visual das palavras que os compõem. Desta forma, os tópicos encontrados nos cogrupos foram: “Futebol”, “Cidadania - segurança”, “Política - Poder legislativo”, “Cidadania - tecnologia”, “Educação”, “Cultura”, “Política - Poder judiciário”.

Figura 79 – Visualização em nuvem de palavras das top-30 palavras para cada cogruppo gerado pelo algoritmo NBVD, com $k = 7$ e $l = 7$

Fonte: Alexandra Katiuska Ramos Diaz, 2018

Cabe ressaltar que ambos cogrupos, C# 3 e C# 7, apresentam tópicos referentes à política, mas diferenciados em alguns aspectos. Por exemplo, o C# 3 apresenta palavras como “governo”, “câmara”, “congresso”, “proposta”, “lei”, etc. caracterizando-o com tópicos referidos ao poder legislativo. Do mesmo modo a C# 7 tem palavras como “tribunal”, “pedido”, “juiz”, “prisão”, “julgamento”, etc cujos tópicos são referentes ao poder judiciário.

A partir desta figura, podem ser visualizadas algumas diferenças entre os cogrupos obtidos com valores de k e l igual a 3 (Figura 73) e os cogrupos obtidos com valores de k e l igual a 7 (Figura 79). Estas diferenças são:

- para valores maiores de k e l ($k = 7, l = 7$), podem ser visualizados mais tópicos do que os três visualizados com k e l igual 3, estes tópicos são: “futebol” (C# 1), “cultura” (C# 6), e “Cidadania - tecnologia” (C# 4);
- é percebido que os cogrupos C# 3 e C# 7, obtidos com valores de k e l igual a 7 (Figura 79), formam o cogrupo C# 3 resultado dos valores de k e l igual a 3 (Figura 73).

A figura 80 mostra a comparação dos grupos de notícias, formados pela projeção dos cogrupos nas linhas, com os canais já conhecidos do corpus EBC. De maneira mais detalhada, o conteúdo da figura ilustra que:

- o GN# 1 (formado pela projeção nas linhas do C# 1) apresenta 211 notícias das quais 84% pertencem ao canal de esportes;
- o GN# 2, formado pela projeção nas linhas do C# 2, apresenta 251 notícias das quais 30,7% pertencem ao canal da cidadania e 27,8% ao canal de geral;
- o GN# 3, formado pela projeção nas linhas do C# 3, está formado por 251 notícias das quais 27,1% pertencem ao canal da política, 16,1% ao canal da educação e 14% aos canais da economia e internacional;
- o GN# 4 (projeção nas linhas do C# 4) apresenta 232 notícias com 33,1% no canal de tecnologia e um 25,1% no canal da cidadania;
- o GN# 5 (projeção nas linhas do C# 5) com 241 notícias das quais o 76,7% pertencem ao canal da educação;
- o GN# 6 (projeção nas linhas do C# 6) apresenta 244 notícias das quais 54,9% pertencem ao canal da cultura;
- e por último o GN# 7, formado pela projeção do C# 7, apresenta 191 notícias das quais 43,1% pertencem ao canal da política.

A comparação das notícias junto com os canais aos quais pertencem ajuda no entendimento de como foram agrupadas estas notícias para a formação dos cogrupos.

De forma semelhante aos resultados obtidos com valores de k e l igual a 7, a tabela 34 mostra a matriz S , resultante dos valores de k e l igual a 10. Nesta tabela é observado a formação de 11 cogrupos de interesse, tais como: o C# 1 é formado pelos GN# 1 e o GP# 4; o C# 2 está formado pelos GN# 2 e o GP# 5; ainda o GN# 2 junto com o GP 7 formam o C# 3; o C# 4 esta formado pelo GN# 3 e o GP# 9, o C# 5 esta formado pelo GN# 4 e o GP# 4; o C# 6 formado pelos GN# 5 e GP# 6; o C# 7 formado

Figura 80 – Comparação dos grupos de notícias com os canais conhecidos, com $k = 7$, $l = 7$

pelos GN# 3 e GP# 3; o C# 8 formado pelos GN# 7 e GP# 10; o C# 9 formado pelos GN# 8 e GP# 1; o C# 10 formado pelos GN# 9 e GP# 2 e finalmente os GN# 10 e GP# 8 formam o C# 11.

Tabela 34 – Matriz S gerada por após a aplicação do algoritmo de NBVD no corpus com $k = 10$, $l = 10$

	GP# 1	GP# 2	GP# 3	GP# 4	GP# 5	GP# 6	GP# 7	GP# 8	GP# 9	GP# 10
GN# 1	0,0	0,0	0,0	2,29	0,0	0,0	0,0	0,0	0,0	0,0
GN# 2	0,0	0,0	0,0	0,0	4,2	0,0	2,16	0,0	0,0	0,0
GN# 3	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	4,68	0,0
GN# 4	0,0	0,0	0,0	4,0	0,0	0,0	0,0	0,0	0,0	0,0
GN# 5	0,0	0,0	0,0	0,0	0,0	4,67	0,0	0,0	0,0	0,0
GN# 6	0,0	0,0	4,47	0,0	0,0	0,0	0,0	0,0	0,0	0,0
GN# 7	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	3,58
GN# 8	3,69	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0
GN# 9	0,0	3,92	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0
GN# 10	0,0	0,0	0,0	0,0	0,0	0,0	0,0	5,14	0,0	0,0

Fonte: Alexandra Katiuska Ramos Diaz, 2018

Cabe ressaltar que na matriz S da tabela 34, o GN# 2 é parte da formação de dois cogrupos (C# 2 e C# 3). Esse fato resulta em dois cogrupos formados pelas mesmas notícias, contudo, mediante caracterizações diferentes no que diz respeito ao grupo de palavras. Na figura 81 pode-se observar as nuvens das top-30 palavras pertencentes a cada cogrupo junto com sua rotulação, obtida a partir de uma análise visual das palavras. Nesta figura, embora os C# 2 e C# 3 apresentem palavras em comum relacionadas à “Cidadania”, o C# 2 possui palavras próprias que o caracterizam como “educação”, “aula”,

“fundamental”, “alunos”. Da mesma maneira, o C# 3 possui palavras próprias que ajudam na sua caracterização, como “municipal”, “crime”, “fiocruz”⁵.

Ainda na tabela da matriz S é observado que o GP# 4 é parte da formação dos cogrupos C# 1 e C# 5, levando à sobreposição de palavras entre estes. Embora ambos cogrupos apresentem palavras iguais relacionadas à “violência”, “mortos”, “segurança”, entre outros (Figura 81), o C# 5 evidencia palavras que ajudam a fazer uma pequena diferenciação entre eles, tais como: “acordo”, “governo”, “federal”, “ministério”.

Figura 81 – Visualização em nuvem de palavras das top-30 palavras para cada cogruppo gerado pelo algoritmo NBVD, com $k = 10$ e $l = 10$

Fonte: Alexandra Katiuska Ramos Diaz, 2018

⁵ Fundação pública Oswaldo Cruz, cujo objetivo é promover a saúde e o desenvolvimento social, assim com gerar e difundir conhecimento científico (<https://portal.fiocruz.br/>).

Nas figuras 79 e 81, podem ser visualizadas algumas diferenças entre os cogrupos obtidos com k e l igual 10 (Figura 81) e os cogrupos obtidos com k e l igual a 7 (Figura 79), sendo estas:

- podem ser visualizadas algumas diferenciações entre os tópicos, tais como referidos à cidadania em relação à educação e a cidadania em geral (C# 2 e C# 3); administração e o tópico de economia em relação a tecnologia;
- e provavelmente os cogrupos C# 1 e C# 5 obtidos dos valores de k e l igual a 10 formam o cogrupos C# 2 obtido com k e l igual a 7.

Figura 82 – Comparação dos grupos de notícias com os canais conhecidos, com $k = 10$ e $l = 10$

Fonte: Alexandra Katiushka Ramos Diaz, 2018

A figura 82 mostra a comparação dos grupos de notícias com os canais já conhecidos. De acordo com esta análise:

- o GN# 1, formado pela projeção nas linhas do C# 1, tem 211 notícias, distribuídas principalmente nos canais de geral (29,2%), internacional (26,6%) e cidadania (25,4%);
- o GN# 2, formado a partir da projeção nas linhas do C# 2 e o C# 3, apresenta 251 notícias distribuídas principalmente nos canais de cidadania (39,9%), geral (18,4%), educação (16,7%) e tecnologia (16,2%);
- o GN# 3, formado pela projeção nas linhas do C# 4, apresenta 232 notícias das quais o 56,7% pertencem ao canal da cultura.
- o GN# 4, formado pela projeção nas linhas do C# 5, apresenta 241 notícias distribuídas nos canais de geral (30,7%), cidadania (25,5%) e internacional (23,9%);

- o GN# 5, formado pela projeção nas linhas do C# 6, apresenta 244 notícias das quais o 79,3% pertencem ao canal da educação;
- o GN# 6, formado pela projeção nas linhas do C# 7, apresenta 191 notícias das quais o 42,8% pertencem ao canal da política e o 19,5% ao canal geral;
- o GN# 7, formado pela projeção nas linhas do C# 8, apresenta 187 notícias distribuídas principalmente pelos canais de política (30,6%), educação (18,1%) e internacional (16,5 %);
- o GN# 8, formado pela projeção nas linhas do C# 9, apresenta 225 notícias, das quais o 84% pertencem ao canal de esportes;
- o GN# 9, formado pela projeção nas linhas do C# 10, apresenta 291 notícias distribuídas principalmente nos canais de tecnologia (37,9%), educação (13,3%), economia e cidadania com um 12%;
- e finalmente o GN# 10, formado pela projeção nas linhas do C# 11, apresenta 232 notícias distribuídas principalmente nos canais de educação (27,3%) e cidadania (26,3%).

Sobre a questão de sobreposições, foi observado que além das sobreposições nas projeções nas linhas e nas projeções nas colunas, existem sobreposições nos cogrupos, ou seja uma notícia pode pertencer a um cogrupo por meio da caracterização de algumas palavras e ao mesmo tempo pertencer a outro cogrupo pela mesma caracterização dessas palavras. Um exemplo desta sobreposição é a notícia 10992 da figura 83, pertencente ao canal geral, que apresenta palavras como “escola” e “alunos” que a caracterizam ao cogrupo C# 2 e ao mesmo tempo a caracterizam ao cogrupo C# 6 (Figura 82). Ainda na figura do exemplo estão marcados da cor amarela as possíveis palavras pertencentes ao C# 2 e da cor azul as possíveis palavras pertencentes ao C# 6.

Finalmente, segundo as experimentações anteriores foi percebido que com valores maiores para os parâmetros k e l , o grau de sobreposição aumenta, assim mesmo as análises foram feitas com valores de k e l iguais baseados nas observações de que estes valores retornam menores graus de sobreposição.

Com respeito à comparação com os canais já conhecidos, maiores valores nos parâmetros de k e l permitem uma maior visualização e diferenciação de tópicos, mas ao custo de aumentar as sobreposições tanto de linhas como de colunas, levando à existência de cogrupos com tópicos semelhantes.

Figura 83 – Exemplo de sobreposição de cogrupos na notícia 10992, após a aplicação do algoritmo de NBVD

Além disso, foi observado que nenhum dos cogrupos de interesse resultantes obtiveram tópicos referidos ao canal infantil, e nenhum dos grupos formados pelas projeções nas linhas mostraram notícias pertencentes a este mesmo canal, provavelmente por conta da baixa quantidade de notícias existentes neste canal.

5.3 Considerações finais

Este capítulo apresentou as experimentações do algoritmo de NBVD sobre conjuntos de dados sintéticos sem e com esparsidade, e sobre conjunto de dados reais (aplicado ao corpus EBC), com a finalidade de analisar seu comportamento nestes ambientes.

Sobre as diferentes estruturas dos conjuntos de dados sintéticos que foram testadas, foi percebido que o algoritmo apresentou dificuldades na separação dos cogrupos adjacentes, como é no caso das estruturas *c*, *f*, *g*, *h* e *i*, embora suas reconstruções tenham sido boas, os cogrupos retornados por este apresentaram maiores graus de sobreposição.

Do ponto de vista dos dados sintéticos esparsos, foi observada uma adaptação do algoritmo frente ao incremento dos níveis de esparsidade. Embora os cogrupos resultantes sejam esparsos, dada a natureza do ambiente, estes cogrupos foram bons em termos de agrupamento. Além disso, foi observado que em níveis de esparsidade altos (90-98), o

algoritmo diminui sua capacidade de agrupar dados, retornando cogrupos semelhantes entre si, o que também leva a um incremento na porcentagem do grau de sobreposição.

Do ponto de vista dos experimentos no corpus EBC, foi observado que o algoritmo não apresentou bons resultados com representações binárias, retornando erros de reconstrução elevados. Os melhores resultados foram obtidos com representações TF-IDF. Nas análises qualitativas, o algoritmo mostrou ser capaz de gerar informação sobre os dados, e fornecer como cada grupo de notícias se organiza em termos de grupos de palavras. Além disso, o algoritmo apresentou uma boa capacidade de separação dos assuntos presentes em cada grupo de notícias, permitindo fazer rotulações sobre os tópicos referidos em cada cogruppo. Adicionalmente, foi observado que com valores altos de k e l , o algoritmo perde sua capacidade de separação de tópicos, retornando alguns cogrupos muito semelhantes entre si.

Sobre a sobreposição, o algoritmo permite que uma notícia possa pertencer a mais de um grupo de notícias (sobreposições nas projeções nas linhas), uma palavra possa pertencer a mais de um grupo de palavras (sobreposições nas projeções nas colunas), o par notícias e palavra possa pertencer a mais de um cogruppo (sobreposições nos cogrupos) e um mesmo conjunto de notícias pode estar associado por dois “motivos” diferentes, ou seja, estar relacionado a dois grupos de palavras diferentes. Essas constatações mostram que a estrutura de cogrupos presente em dados textuais é muito bem representada pela estrutura de bigrupos de formatos arbitrários, mostrada na figura 84.

Figura 84 – Estrutura de bigrupos de formato arbitrários: equivalente à estrutura de cogrupos encontrada no corpus EBC, usando representação TF-IDF e $k = l = 10$)

Fonte: Madeira e Oliveira (2004)

6 Conclusão

Este trabalho teve como objetivo apresentar uma análise do comportamento do algoritmo de biagrupamento Cheng e Church e do algoritmo de coagrupamento NBVD aplicados ao contexto de dados textuais. Analisando suas dificuldades, aspectos positivos e negativos da utilização destes em um ambiente de experimentação caracterizado pela alta dimensionalidade e esparsidade, como são os dados textuais.

A motivação de usar as tarefas de biagrupamento e coagrupamento foi a extração de informações relevantes obtidas através das análises destas tarefas, a qual pode ser interessante e útil em determinados aspectos, como no contexto de mineração de texto, sendo que nas análises os grupos de documentos e os grupos de palavras podem acabar revelando informações ocultas dos dados.

A partir de um estudo detalhado de ambas áreas (Capítulo 2), constatou-se que existem algumas lacunas que podem ser trabalhadas, tais como estudos de algoritmos de diferentes naturezas aplicados a um contexto específico. Assim, mediante critérios estabelecidos no capítulo 1, foram escolhidos dois algoritmos, um algoritmo pertencente à tarefa de biagrupamento e outro à tarefa de coagrupamento. Também, por meio da revisão sistemática detalhada no capítulo 2, foram escolhidos os métodos de avaliação comuns na área para avaliar os resultados produzidos por estes. Além disso, no capítulo 3 foi criada uma estrutura para as análises destes, considerando conjuntos de dados sintéticos e um conjunto de dados reais. Ambos conjuntos estão caracterizados pela alta dimensionalidade e esparsidade, objetivos desta análise.

A partir das experimentações, avaliações dos resultados e análises destes, foram respondidas as questões de pesquisa nos capítulos 4 e 5, mostrando quais foram suas dificuldades em ambientes de experimentação caracterizados pela alta esparsidade e dimensionalidade e analisando se estes algoritmos são capazes de gerar informações adicionais dos dados reais (análise qualitativa). Também, as análises foram delineadas na interpretação semântica das sobreposição existentes nas projeções nas linhas, colunas e sobreposição de bigrupos/cogrupos.

Assim as contribuições deste trabalho são:

- Aprofundamento das definições de biagrupamento e coagrupamento. Na literatura em muitos artigos estes termos são usados de forma intercambiável, assemelhando

suas definições. Abrindo a necessidade do entendimento claro de ambas definições assim como suas diferenças.

- Entendimento claro da influência que a alta esparsidade e a dimensionalidade exercem sobre cada um desses algoritmos que foram escolhidos para estudo. Ambos algoritmos são usados em diferentes aplicações, mas não existem estudos que mostrem suas dificuldades quando são sujeitos a ambientes de alta esparsidade e dimensionalidade. Entender suas dificuldades, vantagens e desvantagens pode levar a propor possíveis alterações nas características ou aspectos específicos destes, levando a alcançar melhorias neste contexto.
- Apresentação de uma forma de interpretação (pós-processamento) da obtenção das partições de linhas, colunas e as relações entre estas duas partições a partir das matrizes U , S e V no algoritmo NBVD. Este pós-processamento levou à geração de cogrupos que poderiam ter sobreposições nas projeções nas linhas e colunas entre cogrupos, sendo úteis para as análises.
- Estabelecimento de uma estrutura de análise, útil no estudo do comportamento dos algoritmos em ambientes sem esparsidade, com diferentes níveis de esparsidade e alta esparsidade. Esta estrutura permitiu, também, um estudo da parametrização dos valores de entrada de cada algoritmo.
- Apresentação de uma interpretação semântica dos resultados obtidos pelos algoritmos de biagrupamento e coagrupamento aplicados a dados textuais, assim como compreender de forma semântica as sobreposições nas projeções nas linhas e colunas e sobreposição dos bigrupos/cogrupos.

Assim, as experimentações mostraram melhores adaptações do algoritmo de coagrupamento NBVD ao problema da alta esparsidade e dimensionalidade em dados textuais, permitindo gerar informação dos dados e fornecer como cada grupo de notícia é organizado em relação a cada grupo de palavras. Adicionalmente, o algoritmo de NBVD apresentou uma boa capacidade de separação de temas, permitindo fazer rotulações dos tópicos referidos em cada cogrupo resultante.

Já o algoritmo de biagrupamento Cheng e Church apresentou problemas nos ambientes de alta esparsidade, mostrando sensibilidade e afetando sua capacidade de encontrar os bigrupos. Assim, embora os resultados obtidos sejam bons de acordo com os objetivos do algoritmo, no contexto dos dados textuais, suas análises apresentam baixa

relevância. De acordo com esta análise, os bigrupos obtidos ainda geram informações e conseguem explicar as organizações dos grupos de notícias e os grupos de palavras, mas observou-se que os resultados apresentam porcentagens de sobreposição elevadas, o que geram grande quantidade de tópicos sobrepostos, diminuindo sua capacidade de separação de tópicos e dificultando o processo de rotulação destes.

O fato de o algoritmo de Cheng e Church ser sensível à esparsidade se deve à sua natureza de encontrar bigrupos coerentes com RQMs baixos, considerando os elementos nulos como casos especiais dos valores coerentes. Ainda nestas experimentações foram testadas algumas variações a fim de direcionar a busca dos bigrupos aos elementos não nulos, mas estas inserções alteraram negativamente a descoberta dos bigrupos. Ficando como trabalhos futuros as adequações do algoritmo para dados textuais.

6.1 *Limitações do trabalho*

Este trabalho apresenta algumas limitações tais como:

- Nas experimentações, os algoritmos apresentados nem sempre são capazes de alcançar soluções ótimas. Por exemplo, o algoritmo de Cheng e Church é um método guloso, e como tal nem sempre produz soluções ótimas (CORMEN et al., 2009). Por outro lado, o algoritmo de NBVD não é capaz de garantir convergência para um mínimo global do problema de minimização, assim é possível encontrar diversas soluções em execuções diferentes do algoritmo devido à inicialização de forma aleatória.
- Sobre a utilização do corpus EBC, embora seja um corpus com as características desejadas para as análises, este é um conjunto pequeno voltado para um certo tipo de texto e de notícias. Provavelmente a utilização de conjuntos de dados com maior diversidade possa levar a conclusões diferentes das aqui apresentadas.
- Sobre a escolha específica dos algoritmos, o fato de usar uma instância específica do biagrupamento, como é o algoritmo de Cheng e Church não leva à conclusão de que a tarefa de biagrupamento não seja adequada no contexto dos dados textuais. Já que existem na literatura uma variedade de algoritmos que resolvem esta tarefa e que poderiam ser melhores neste contexto.

6.2 *Trabalhos futuros*

Durante a realização deste trabalho foram encontradas algumas dificuldades e conclusões que abrem espaço para aspectos a serem explorados. Fica em aberto a perspectiva de novas pesquisas e expansão dos conceitos introduzidos aqui, como a utilização destes algoritmos em outras áreas da mineração de dados, assim como a busca de outros algoritmos capazes de resolver a tarefa de biagrupamento trazidos à área de mineração de dados e testados nos dados textuais.

Sobre o algoritmo de Cheng e Church, as dificuldades encontradas nos dados textuais podem levar à proposição de melhorias e de novas técnicas de biagrupamento que visem melhorar suas dificuldades com a esparsidade.

Sobre o algoritmo de NBVD, a necessidade do pós-processamento para a obtenção das partições de linhas, colunas e a relação destas, é uma desvantagem dos métodos de fatoração de matrizes. Um estudo mais profundo pode levar a novas técnicas de obtenção destas partições que visem trazer resultados interessantes de serem analisados.

Sobre os métodos de avaliação de biagrupamento e coagrupamento, o fato de serem escassos e pouco comuns na literatura abrem espaços para estudos que visem encontrar novas técnicas focadas especificamente na validação destas tarefas.

Referências¹

- AGGARWAL, C. C.; ZHAI, C. A survey of text clustering algorithms. In: *Proceedings of the Mining text data*. [S.l.]: Springer, 2012. p. 77–128. Citado na página 78.
- 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 on Conference on Information and Knowledge Management*. [S.l.], 2015. p. 1807–1810. Citado na página 80.
- 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. Citado na página 80.
- 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. Citado 2 vezes nas páginas 27 e 80.
- 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*. [S.l.], 2016. p. 347–356. Citado na página 80.
- AMINE, A.; ELBERRICHI, Z.; SIMONET, M. Evaluation of text clustering methods using wordnet. *The International Arab Journal of Information Technology (IAJIT)*, Citeseer, v. 7, n. 4, p. 349–357, 2010. Citado 2 vezes nas páginas 24 e 77.
- ANAGNOSTOPOULOS, A.; DASGUPTA, A.; KUMAR, R. Approximation algorithms for co-clustering. In: ACM. *Proceedings of the twenty-seventh ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*. [S.l.], 2008. p. 201–210. Citado 3 vezes nas páginas 25, 33 e 46.
- ANDO, S.; SUZUKI, E. Unsupervised cross-domain learning by interaction information co-clustering. In: IEEE. *Proceedings of the 2008 Eighth IEEE International Conference on Data Mining (ICDM'08)*. [S.l.], 2008. p. 13–22. Citado na página 80.
- BAEZA-YATES, R.; RIBEIRO-NETO, B. *Modern Information Retrieval: The Concepts and Technology behind Search (ACM Press Books)*. [S.l.]: Addison-Wesley Professional Harlow, 2011. Citado na página 78.
- 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. Citado 2 vezes nas páginas 33 e 46.
- BANU, P. N.; INBARANI, H. Analysis of click stream patterns using soft biclustering approaches. In: *Proceedings of the Systems Approach Applications for Developments in Information Technology*. [S.l.]: IGI Global, 2012. p. 212–224. Citado 3 vezes nas páginas 80, 89 e 90.
- BAO, B.-K. et al. Joint local and global consistency on interdocument and interword relationships for co-clustering. *IEEE transactions on cybernetics*, IEEE, v. 45, n. 1, p. 15–28, 2015. Citado na página 80.

¹ De acordo com a Associação Brasileira de Normas Técnicas. NBR 6023.

BERKHIN, P. A survey of clustering data mining techniques. In: _____. *Proceedings of the Grouping Multidimensional Data: Recent Advances in Clustering*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2006. p. 25–71. ISBN 978-3-540-28349-2. Disponível em: http://dx.doi.org/10.1007/3-540-28349-8_2. Citado na página 25.

BHATTACHARYA, A.; CHOWDHURY, N.; DE, R. K. Comparative analysis of clustering and biclustering algorithms for grouping of genes: co-function and co-regulation. *Current Bioinformatics*, Bentham Science Publishers, v. 7, n. 1, p. 63–76, 2012. Citado na página 30.

BICHOT, C.-E. Co-clustering documents and words by minimizing the normalized cut objective function. *Journal of Mathematical Modelling and Algorithms*, Springer, v. 9, n. 2, p. 131–147, 2010. Citado na página 80.

BISSON, G.; HUSSAIN, F. Chi-sim: A new similarity measure for the co-clustering task. In: IEEE. *Proceedings of the 2008 Seventh International Conference on Machine Learning and Applications (ICMLA'08)*. [S.l.], 2008. p. 211–217. Citado 2 vezes nas páginas 80 e 86.

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. Citado 6 vezes nas páginas 66, 67, 68, 72, 92 e 98.

BRUNIALTI, L. F. et al. The binovnmf algorithm: Overlapping columns co-clustering based on non-negative matrix tri-factorization. In: IEEE. *Proceedings of the 2017 6th Brazilian Conference on Intelligent Systems (BRACIS)*. [S.l.], 2017. p. 330–335. Citado 2 vezes nas páginas 30 e 72.

BRYAN, K.; CUNNINGHAM, P.; BOLSHAKOVA, N. Biclustering of expression data using simulated annealing. In: IEEE. *Proceedings of the 2005 18th IEEE Symposium on Computer-Based Medical Systems*. [S.l.], 2005. p. 383–388. Citado na página 49.

BUDGEN, D.; BRERETON, P. Performing systematic literature reviews in software engineering. In: ACM. *Proceedings of the 28th international conference on Software engineering*. [S.l.], 2006. p. 1051–1052. Citado na página 194.

BULTEAU, L. et al. Co-clustering under the maximum norm. *Algorithms*, Multidisciplinary Digital Publishing Institute, v. 9, n. 1, p. 17, 2016. Citado na página 28.

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. Citado 5 vezes nas páginas 49, 50, 75, 76 e 80.

BUSYGIN, S.; PROKOPYEV, O.; PARDALOS, P. M. Biclustering in data mining. *Computers & Operations Research*, Elsevier, v. 35, n. 9, p. 2964–2987, 2008. Citado 2 vezes nas páginas 29 e 194.

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*. [S.l.], 2015. p. 79–89. Citado na página 80.

CASTRO, P. A. D. et al. Evaluating the performance of a biclustering algorithm applied to collaborative filtering-a comparative analysis. In: IEEE. *Proceedings of the 2007 7th International Conference on Hybrid Intelligent Systems (HIS 2007)*. [S.l.], 2007. p. 65–70. Citado na página 30.

CASTRO, P. A. de et al. Applying biclustering to text mining: an immune-inspired approach. In: *Proceedings of the Artificial Immune Systems*. [S.l.]: Springer, 2007. p. 83–94. Citado 4 vezes nas páginas 80, 85, 88 e 90.

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. Citado 5 vezes nas páginas 73, 74, 80, 89 e 90.

CHARRAD, M.; AHMED, M. B. Simultaneous clustering: A survey. In: SPRINGER. *Proceedings of the International Conference on Pattern Recognition and Machine Intelligence*. [S.l.], 2011. p. 370–375. Citado na página 33.

CHEN, Y.; WANG, L.; DONG, M. Semi-supervised document clustering with simultaneous text representation and categorization. In: SPRINGER. *Proceedings of the Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. [S.l.], 2009. p. 211–226. Citado na página 80.

CHEN, Y.; WANG, L.; DONG, M. Non-negative matrix factorization for semisupervised heterogeneous data coclustering. *IEEE Transactions on Knowledge and Data Engineering*, IEEE, v. 22, n. 10, p. 1459–1474, 2010. Citado na página 80.

CHENG, W. et al. Hierarchical co-clustering based on entropy splitting. In: ACM. *Proceedings of the 21st ACM international conference on Information and knowledge management*. [S.l.], 2012. p. 1472–1476. Citado na página 80.

CHENG, Y.; CHURCH, G. M. Biclustering of expression data. In: *Proceedings of the Intelligent Systems for Molecular Biology*. [S.l.: s.n.], 2000. v. 8, n. 2000, p. 93–103. Citado 11 vezes nas páginas 33, 49, 50, 55, 56, 57, 58, 73, 86, 106 e 109.

CORMEN, T. H. et al. *Introduction to algorithms*. [S.l.]: MIT press, 2009. Citado 2 vezes nas páginas 91 e 180.

COSTA, G.; ORTALE, R. Machine learning techniques for xml (co-)clustering by structure-constrained phrases. *Information Retrieval Journal*, v. 21, p. 24–55, 2017. Citado 2 vezes nas páginas 80 e 90.

COVER, T. M.; THOMAS, J. A. *Elements of information theory*. [S.l.]: John Wiley & Sons, 2012. Citado na página 86.

DAI, W. et al. Co-clustering based classification for out-of-domain documents. In: ACM. *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*. [S.l.], 2007. p. 210–219. Citado na página 80.

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. Citado na página 76.

- DEODHAR, M.; GHOSH, J. Scoal: A framework for simultaneous co-clustering and learning from complex data. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, ACM, v. 4, n. 3, p. 11, 2010. Citado na página 33.
- 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*. [S.l.], 2009. p. 241–248. Citado 2 vezes nas páginas 33 e 49.
- DESGRAUPES, B. Clustering indices. *University of Paris Ouest-Lab ModalX*, v. 1, p. 34, 2013. Citado na página 75.
- DHILLON, I. S. Co-clustering documents and words using bipartite spectral graph partitioning. In: ACM. *Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining*. [S.l.], 2001. p. 269–274. Citado 2 vezes nas páginas 80 e 85.
- 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. Citado 4 vezes nas páginas 47, 49, 80 e 85.
- DHILLON, I. S.; MODHA, D. S. Concept decompositions for large sparse text data using clustering. *Machine learning*, Springer, v. 42, n. 1, p. 143–175, 2001. Citado na página 78.
- DIAZ, A. K. R. et al. *Uma análise comparativa das ferramentas de pré-processamento de dados textuais: NLTK, PreText e R*. [S.l.], 2018. Citado 3 vezes nas páginas 31, 77 e 95.
- 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*. [S.l.], 2006. p. 126–135. Citado 2 vezes nas páginas 50 e 72.
- DOMENICONI, C.; LASKEY, K. Bayesian co-clustering. *Wiley Interdisciplinary Reviews: Computational Statistics*, Wiley Online Library, v. 7, n. 5, p. 347–356, 2015. Citado 3 vezes nas páginas 80, 85 e 89.
- FELDMAN, R.; SANGER, J. *The text mining handbook: advanced approaches in analyzing unstructured data*. [S.l.]: Cambridge university press, 2007. Citado 2 vezes nas páginas 24 e 77.
- 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 (ICMLA)*. [S.l.], 2012. v. 1, p. 464–467. Citado na página 80.
- FRANCA, F. O. de. *Bicusterização na Análise de Dados Incertos*. Tese (Doutorado) — Universidade Estadual de Campinas, 2010. Citado 3 vezes nas páginas 27, 30 e 50.
- FRANÇA, F. O. de. A hash-based co-clustering algorithm for categorical data. *Expert Systems with Applications*, Elsevier, v. 64, p. 24–35, 2016. Citado 3 vezes nas páginas 27, 80 e 89.
- FRIGUI, H.; NASRAOUI, O. Simultaneous categorization of text documents and identification of cluster-dependent keywords. In: IEEE. *Proceedings of the 2002 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE'02)*. [S.l.], 2002. v. 2, p. 1108–1113. Citado na página 80.

- GARG, V. K.; CHAUDHARI, S.; NARANG, A. Multi-regularization for fuzzy co-clustering. In: SPRINGER. *Proceedings of the International Conference on Neural Information Processing*. [S.l.], 2013. p. 67–75. Citado na página 80.
- GIANNAKIDOU, E. et al. Co-clustering tags and social data sources. In: IEEE. *Proceedings of the 2008 The Ninth International Conference on Web-Age Information Management (WAIM '08)*. [S.l.], 2008. p. 317–324. Citado 2 vezes nas páginas 80 e 90.
- GOVAERT, G.; NADIF, M. Mutual information, phi-squared and model-based co-clustering for contingency tables. *Advances in Data Analysis and Classification*, Springer, p. 1–34, 2016. Citado na página 80.
- GREENE, D.; CUNNINGHAM, P. Producing accurate interpretable clusters from high-dimensional data. In: SPRINGER. *Proceedings of the European Conference on Principles of Data Mining and Knowledge Discovery*. [S.l.], 2005. p. 486–494. Citado na página 80.
- GREENE, D.; CUNNINGHAM, P. Spectral co-clustering for dynamic bipartite graphs. In: SUN SITE CENTRAL EUROPE (CEUR). *Paper presented at the Workshop on Dynamic Networks and Knowledge Discovery (DyNAK 2010) at the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML PKDD 2010), Barcelona, September 24th 2010*. [S.l.], 2010. Citado 2 vezes nas páginas 80 e 89.
- GU, Q.; DING, C.; HAN, J. On trivial solution and scale transfer problems in graph regularized nmf. In: *Proceedings of the IJCAI Proceedings-International Joint Conference on Artificial Intelligence*. [S.l.: s.n.], 2011. v. 22, n. 1, p. 1288. Citado na página 80.
- GU, Q.; ZHOU, J. Co-clustering on manifolds. In: ACM. *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*. [S.l.], 2009. p. 359–368. Citado na página 80.
- 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. Citado na página 75.
- HALKIDI, M.; BATISTAKIS, Y.; VAZIRGIANNIS, M. Cluster validity methods: part i. *ACM Sigmod Record*, ACM, v. 31, n. 2, p. 40–45, 2002. Citado 3 vezes nas páginas 72, 73 e 75.
- HALKIDI, M.; BATISTAKIS, Y.; VAZIRGIANNIS, M. Clustering validity checking methods: part ii. *ACM Sigmod Record*, ACM, v. 31, n. 3, p. 19–27, 2002. Citado 3 vezes nas páginas 72, 73 e 75.
- HAN, E.-H. et al. Webace: a web agent for document categorization and exploration. In: ACM. *Proceedings of the second international conference on Autonomous agents*. [S.l.], 1998. p. 408–415. Citado na página 82.
- 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. Citado 3 vezes nas páginas 44, 45 e 73.

HOCHREITER, S. et al. Fabia: factor analysis for bicluster acquisition. *Bioinformatics*, Oxford University Press, v. 26, n. 12, p. 1520–1527, 2010. Citado na página 49.

HOFMANN, T. Probabilistic latent semantic analysis. In: MORGAN KAUFMANN PUBLISHERS INC. *Proceedings of the Fifteenth conference on Uncertainty in artificial intelligence*. [S.l.], 1999. p. 289–296. Citado na página 47.

HOFMANN, T.; PUZICHA, J.; JORDAN, M. I. Learning from dyadic data. In: *Proceedings of the Advances in neural information processing systems*. [S.l.: s.n.], 1999. p. 466–472. Citado na página 48.

HONDA, K. et al. Mmms-induced k-member co-clustering for k-anonymization of cooccurrence information. In: IEEE. *Proceedings of the 2016 International Joint Conference on Neural Networks (IJCNN)*. [S.l.], 2016. p. 2961–2966. Citado 2 vezes nas páginas 27 e 80.

HU, T. et al. Preserving patterns in bipartite graph partitioning. In: IEEE. *Proceedings of the 2006 18th IEEE International Conference on Tools with Artificial Intelligence (ICTAI'06)*. [S.l.], 2006. p. 489–496. Citado na página 80.

HUANG, A. Similarity measures for text document clustering. In: *Proceedings of the sixth new zealand computer science research student conference (NZCSRSC2008)*, Christchurch, New Zealand. [S.l.: s.n.], 2008. p. 49–56. Citado na página 86.

HUANG, F. et al. Semi-supervised hierarchical co-clustering. In: SPRINGER. *Proceedings of the International Conference on Rough Sets and Knowledge Technology*. [S.l.], 2012. p. 310–319. Citado na página 80.

HUANG, S.; XU, Z.; LV, J. Adaptive local structure learning for document co-clustering. *Knowledge-Based Systems*, Elsevier, v. 148, p. 74–84, 2018. Citado 2 vezes nas páginas 27 e 80.

HUBERT, L.; ARABIE, P. Comparing partitions. *Journal of classification*, Springer, v. 2, n. 1, p. 193–218, 1985. Citado na página 86.

HUSSAIN, S. F.; BISSON, G. Text categorization using word similarities based on higher order co-occurrences. In: SIAM. *Proceedings of the 2010 SIAM International Conference on Data Mining*. [S.l.], 2010. p. 1–12. Citado 2 vezes nas páginas 80 e 84.

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. Citado na página 46.

IHMELS, J.; BERGMANN, S.; BARKAI, N. Defining transcription modules using large-scale gene expression data. *Bioinformatics*, Oxford University Press, v. 20, n. 13, p. 1993–2003, 2004. Citado na página 49.

KUMMAMURU, K.; DHAWALE, A.; KRISHNAPURAM, R. Fuzzy co-clustering of documents and keywords. In: IEEE. *Proceedings of the 2003 12th IEEE International Conference on Fuzzy Systems (FUZZ'03)*. [S.l.], 2003. Citado 2 vezes nas páginas 80 e 85.

LACLAU, C.; NADIF, M. Hard and fuzzy diagonal co-clustering for document-term partitioning. *Neurocomputing*, Elsevier, v. 193, p. 133–147, 2016. Citado na página 80.

- LANG, K. Newsweeder: Learning to filter netnews. In: *Proceedings of the Twelfth International Conference on Machine Learning*. [S.l.: s.n.], 1995. p. 331–339. Citado na página 82.
- 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. Citado 2 vezes nas páginas 49 e 50.
- LEE, M. et al. Biclustering via sparse singular value decomposition. *Biometrics*, Wiley Online Library, v. 66, n. 4, p. 1087–1095, 2010. Citado na página 33.
- LESKOVEC, J.; RAJARAMAN, A.; ULLMAN, J. D. *Mining of massive datasets*. [S.l.]: Cambridge university press, 2014. Citado na página 78.
- LI, L. et al. A comparison and evaluation of five biclustering algorithms by quantifying goodness of biclusters for gene expression data. *BioData mining*, BioMed Central, v. 5, n. 1, p. 8, 2012. Citado na página 30.
- LI, Z.; WU, X. Single multiplicatively updated matrix factorization for co-clustering. In: ACM. *Proceedings of the 29th Annual ACM Symposium on Applied Computing*. [S.l.], 2014. p. 97–104. Citado na página 80.
- LI, Z.; WU, X.; LU, Z. Fast orthogonal nonnegative matrix tri-factorization for simultaneous clustering. In: SPRINGER. *Proceedings of the Pacific-Asia Conference on Knowledge Discovery and Data Mining*. [S.l.], 2010. p. 214–221. Citado na página 80.
- LIU, N.; CHEN, F.; LU, M. Spectral co-clustering documents and words using fuzzy k-harmonic means. *International Journal of Machine Learning and Cybernetics*, Springer, v. 4, n. 1, p. 75–83, 2013. Citado na página 80.
- LONG, B.; ZHANG, Z. M.; YU, P. S. Co-clustering by block value decomposition. In: ACM. *Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining*. [S.l.], 2005. p. 635–640. Citado 10 vezes nas páginas 33, 47, 49, 63, 64, 65, 66, 68, 80 e 85.
- 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. Citado 2 vezes nas páginas 80 e 84.
- MA, H. et al. Orthogonal nonnegative matrix tri-factorization for semi-supervised document co-clustering. In: SPRINGER. *Proceedings of the Pacific-Asia Conference on Knowledge Discovery and Data Mining*. [S.l.], 2010. p. 189–200. Citado na página 80.
- MADEIRA, S. C.; OLIVEIRA, A. L. Biclustering algorithms for biological data analysis: a survey. *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB)*, IEEE Computer Society Press, v. 1, n. 1, p. 24–45, 2004. Citado 21 vezes nas páginas 25, 28, 29, 31, 34, 36, 38, 39, 40, 41, 42, 43, 44, 45, 46, 50, 73, 92, 130, 177 e 194.
- MANDHANI, B.; JOSHI, S.; KUMMAMURU, K. A matrix density based algorithm to hierarchically co-cluster documents and words. In: ACM. *Proceedings of the 12th international conference on World Wide Web*. [S.l.], 2003. p. 511–518. Citado na página 80.

MARCACINI, R. M.; REZENDE, S. O. Incremental construction of topic hierarchies using hierarchical term clustering. In: *Proceedings of the 22nd International Conference on Software Engineering and Knowledge Engineering (SEKE 2010)*. [S.l.: s.n.], 2010. p. 553. Citado 2 vezes nas páginas 24 e 77.

MEI, J.-P. et al. Incremental fuzzy clustering for document categorization. In: IEEE. *Proceedings of the 2014 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*. [S.l.], 2014. p. 1518–1525. Citado na página 80.

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. Citado 2 vezes nas páginas 27 e 80.

MI, G.; GAO, Y.; TAN, Y. Apply stacked auto-encoder to spam detection. In: SPRINGER. *Proceedings of the International Conference in Swarm Intelligence*. [S.l.], 2015. p. 3–15. Citado 2 vezes nas páginas 24 e 77.

MIMAROGLU, S.; UEHARA, K. Bit sequences and biclustering of text documents. In: IEEE. *Proceedings of the 2007 Seventh IEEE International Conference on Data Mining Workshops (ICDM Workshops 2007)*. [S.l.], 2007. p. 51–56. Citado 2 vezes nas páginas 80 e 90.

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. Citado na página 33.

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. Citado 4 vezes nas páginas 38, 41, 73 e 194.

NADIF, M. et al. *Co-Clustering: Models, Algorithms and Applications*. [S.l.]: John Wiley & Sons, 2013. Citado 3 vezes nas páginas 27, 29 e 79.

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)*. [S.l.], 2013. p. 446–453. Citado na página 80.

OH, C.-H.; HONDA, K.; ICHIHASHI, H. Fuzzy clustering for categorical multivariate data. In: IEEE. *Proceedings of the Joint 9th IFSA World Congress and 20th NAFIPS International Conference*. [S.l.], 2001. v. 4, p. 2154–2159. Citado na página 85.

ORZECOWSKI, P.; BORYCZKO, K. Hybrid biclustering algorithms for data mining. In: SPRINGER. *Proceedings of the European Conference on the Applications of Evolutionary Computation*. [S.l.], 2016. p. 156–168. Citado na página 33.

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. Citado 3 vezes nas páginas 31, 94 e 95.

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 Eighth IEEE International Conference on Data Mining*. [S.l.], 2008. p. 512–521. Citado na página 35.

PARK, L. A. et al. Adapting spectral co-clustering to documents and terms using latent semantic analysis. In: SPRINGER. *Proceedings of the Australasian Joint Conference on Artificial Intelligence*. [S.l.], 2009. p. 301–311. Citado na página 80.

PEETERS, R. The maximum edge biclique problem is np-complete. *Discrete Applied Mathematics*, v. 131, n. 3, p. 651 – 654, 2003. ISSN 0166-218X. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0166218X03003330>. Citado na página 28.

PENSA, R. G.; BOULICAUT, J.-F. Constrained co-clustering of gene expression data. In: SIAM. *Proceedings of the 2008 SIAM International Conference on Data Mining*. [S.l.], 2008. p. 25–36. Citado na página 33.

PENSA, 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. Citado 3 vezes nas páginas 28, 34 e 48.

PENSA, R. G.; IENCO, D.; MEO, R. Hierarchical co-clustering: off-line and incremental approaches. *Data Mining and Knowledge Discovery*, Springer, p. 1–34, 2014. Citado na página 25.

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. Citado na página 29.

PRELIĆ, A. et al. A systematic comparison and evaluation of biclustering methods for gene expression data. *Bioinformatics*, Oxford University Press, v. 22, n. 9, p. 1122–1129, 2006. Citado 2 vezes nas páginas 30 e 49.

PRIAM, R.; NADIF, M.; GOVAERT, G. Generalized topographic block model. *Neurocomputing*, Elsevier, v. 173, p. 442–449, 2016. Citado na página 80.

QU, H. et al. An improved biclustering algorithm and its application to gene expression spectrum analysis. *Genomics, proteomics & bioinformatics*, Elsevier, v. 3, n. 3, p. 189–193, 2005. Citado na página 63.

R Development Core Team. *R: A Language and Environment for Statistical Computing*. Vienna, Austria, 2008. ISBN 3-900051-07-0. Disponível em: <http://www.R-project.org>. Citado na página 95.

REGE, M.; DONG, M.; FOTOUHI, F. Co-clustering documents and words using bipartite isoperimetric graph partitioning. In: IEEE. *Proceedings of the 2006 Sixth International Conference on Data Mining (ICDM'06)*. [S.l.], 2006. p. 532–541. Citado na página 80.

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*. [S.l.], 2017. p. 457–467. Citado na página 80.

- SALAH, A.; ROGOVSCHI, N.; NADIF, M. Stochastic co-clustering for document-term data. In: SIAM. *Proceedings of the 2016 SIAM International Conference on Data Mining*. [S.l.], 2016. p. 306–314. Citado 2 vezes nas páginas 27 e 80.
- SALTON, G.; BUCKLEY, C. Term-weighting approaches in automatic text retrieval. *Information processing & management*, Elsevier, v. 24, n. 5, p. 513–523, 1988. Citado 2 vezes nas páginas 77 e 78.
- SALTON, G.; WONG, A.; YANG, C. S. A vector space model for automatic indexing. *Commun. ACM*, ACM, New York, NY, USA, v. 18, n. 11, p. 613–620, nov. 1975. ISSN 0001-0782. Disponível em: <http://doi.acm.org/10.1145/361219.361220>. Citado na página 24.
- SHAFIEI, M. M.; MILIOS, E. E. Latent dirichlet co-clustering. In: IEEE. *Proceedings of the 2006 Sixth International Conference on Data Mining (ICDM'06)*. [S.l.], 2006. p. 542–551. Citado na página 80.
- SHI, X.; FAN, W.; PHILIP, S. Y. Efficient semi-supervised spectral co-clustering with constraints. In: IEEE. *Proceedings of the 2010 IEEE 10th International Conference on Data Mining (ICDM)*. [S.l.], 2010. p. 1043–1048. Citado na página 80.
- SILVA, L. A. da; PERES, S.; BOSCARIOLI, C. *Introdução À Mineração de Dados - Com Aplicação Em R*. São Paulo, Brasil: Elsevier Editora Ltda., 2016. Citado 3 vezes nas páginas 24, 72 e 73.
- SINKA, M. P.; CORNE, D. W. A large benchmark dataset for web document clustering. *Soft computing systems: design, management and applications*, IOS Press, Amsterdam, v. 87, p. 881–890, 2002. Citado na página 82.
- SOLER, J. et al. Data clustering and similarity. In: *Proceedings of the 26th International Florida Artificial Intelligence Research Society Conference (FLAIRS 2013)*. [S.l.: s.n.], 2013. Citado na página 76.
- SONG, Y. et al. Constrained coclustering for textual documents. In: *Proceedings of the Association for the Advancement of Artificial Intelligence (AAAI)*. [S.l.: s.n.], 2010. Citado na página 80.
- SONG, Y. et al. Constrained text coclustering with supervised and unsupervised constraints. *IEEE Transactions on Knowledge and Data Engineering*, IEEE, v. 25, n. 6, p. 1227–1239, 2013. Citado na página 80.
- TANAY, A.; SHARAN, R.; SHAMIR, R. Discovering statistically significant biclusters in gene expression data. *Bioinformatics*, Oxford University Press, v. 18, n. suppl_1, p. S136–S144, 2002. Citado na página 49.
- 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. Citado 3 vezes nas páginas 29, 50 e 194.
- TJHI, W.-C.; CHEN, L. Fuzzy co-clustering of web documents. In: IEEE. *Proceedings of the 4th International Conference on Cyberworlds (CW 2005)*. [S.l.], 2005. p. 7–pp. Citado na página 80.

- TJHI, W.-C.; CHEN, L. Flexible fuzzy co-clustering with feature-cluster weighting. In: IEEE. *Proceedings of the 2006 9th International Conference on Control, Automation, Robotics and Vision (ICARCV'06)*. [S.l.], 2006. p. 1–6. Citado na página 80.
- TJHI, W.-C.; CHEN, L. A new fuzzy co-clustering algorithm for categorization of datasets with overlapping clusters. In: SPRINGER. *Proceedings of the International Conference on Advanced Data Mining and Applications*. [S.l.], 2006. p. 328–339. Citado na página 80.
- TJHI, W.-C.; CHEN, L. A partitioning based algorithm to fuzzy co-cluster documents and words. *Pattern Recognition Letters*, Elsevier, v. 27, n. 3, p. 151–159, 2006. Citado na página 80.
- TJHI, W.-C.; CHEN, L. Dual fuzzy-possibilistic co-clustering for document categorization. In: IEEE. *Proceedings of the 2007 Seventh IEEE International Conference on Data Mining Workshops (ICDM Workshops 2007)*. [S.l.], 2007. p. 259–264. Citado na página 80.
- TJHI, W.-C.; CHEN, L. Experimental study on fuzzy word memberships for multi-document summarization. In: IEEE. *Proceedings of the 2007 6th International Conference on Information, Communications & Signal Processing*. [S.l.], 2007. p. 1–5. Citado na página 80.
- TJHI, W.-C.; CHEN, L. Possibilistic fuzzy co-clustering of large document collections. *Pattern Recognition*, Elsevier, v. 40, n. 12, p. 3452–3466, 2007. Citado 2 vezes nas páginas 80 e 86.
- TJHI, W.-C.; CHEN, L. A heuristic-based fuzzy co-clustering algorithm for categorization of high-dimensional data. *Fuzzy Sets and Systems*, Elsevier, v. 159, n. 4, p. 371–389, 2008. Citado na página 80.
- TJHI, W.-C.; CHEN, L. Dual fuzzy-possibilistic coclustering for categorization of documents. *IEEE Transactions on Fuzzy Systems*, IEEE, v. 17, n. 3, p. 532–543, 2009. Citado na página 80.
- TURNEY, P. D.; PANTEL, P. From frequency to meaning: Vector space models of semantics. *Journal of artificial intelligence research*, v. 37, p. 141–188, 2010. Citado 2 vezes nas páginas 24 e 77.
- WANG, H. et al. Fast nonnegative matrix tri-factorization for large-scale data co-clustering. In: *Proceedings of the IJCAI Proceedings-International Joint Conference on Artificial Intelligence*. [S.l.: s.n.], 2011. v. 22, n. 1, p. 1553. Citado 3 vezes nas páginas 30, 50 e 72.
- WANG, S.; HUANG, A. Penalized nonnegative matrix tri-factorization for co-clustering. *Expert Systems with Applications*, Elsevier, v. 78, p. 64–73, 2017. Citado na página 80.
- WU, J.-S.; LAI, J.-H.; WANG, C.-D. A novel co-clustering method with intra-similarities. In: IEEE. *Proceedings of the 2011 IEEE 11th International Conference on Data Mining Workshops*. [S.l.], 2011. p. 300–306. Citado na página 80.
- XIA, Y.; ZHANG, Y.; YAO, J. Co-clustering sentences and terms for multi-document summarization. In: SPRINGER. *Proceedings of the International Conference on Intelligent Text Processing and Computational Linguistics*. [S.l.], 2011. p. 339–352. Citado na página 80.

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. Citado 2 vezes nas páginas 27 e 80.

YAN, Y.; CHEN, L. Label-based semi-supervised fuzzy co-clustering for document categorization. In: IEEE. *Proceedings of the 2011 8th International Conference on Information, Communications and Signal Processing (ICICS)*. [S.l.], 2011. p. 1–5. Citado na página 80.

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. Citado na página 80.

YAN, Y.; CHEN, L.; TJHI, W.-C. Semi-supervised fuzzy co-clustering algorithm for document categorization. *Knowledge and information systems*, Springer, v. 34, n. 1, p. 55–74, 2013. Citado 2 vezes nas páginas 80 e 86.

YANG, J. et al. Enhanced biclustering on expression data. In: IEEE. *Proceedings of the 2003 Third IEEE Symposium on Bioinformatics and Bioengineering*. [S.l.], 2003. p. 321–327. Citado na página 49.

YANG, J. et al. An improved biclustering method for analyzing gene expression profiles. *International Journal on Artificial Intelligence Tools*, World Scientific, v. 14, n. 05, p. 771–789, 2005. Citado 2 vezes nas páginas 49 e 63.

YANG, J. et al. /spl delta/-clusters: capturing subspace correlation in a large data set. In: IEEE. *Proceedings of the 2002 18th International Conference on Data Engineering*. [S.l.], 2002. p. 517–528. Citado na página 49.

YANG, L.; CAI, X. Semi-supervised co-clustering for query-oriented theme-based summarization. *Research Journal of Applied Sciences, Engineering and Technology*, Maxwell Science Publishing, v. 4, n. 18, p. 3410–3414, 2012. Citado na página 80.

YE, Y. et al. Feature weighting information-theoretic co-clustering for document clustering. *2009 2nd International Conference on Computer Science and its Applications*, p. 1–6, 2009. Citado na página 80.

YOO, J.; CHOI, S. Probabilistic matrix tri-factorization. In: IEEE. *Proceedings of the 2009 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*. [S.l.], 2009. p. 1553–1556. Citado na página 80.

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. Citado 9 vezes nas páginas 30, 47, 49, 50, 65, 67, 68, 80 e 96.

ZHANG, C.; XING, D. Knowledge-supervised learning by co-clustering based approach. In: IEEE. *Proceedings of the 2008 Seventh International Conference on Machine Learning and Applications (ICMLA'08)*. [S.l.], 2008. p. 773–776. Citado 2 vezes nas páginas 80 e 84.

Apêndice A – Protocolo da revisão Sistemática

Segundo Budgen e Brereton (2006) uma revisão sistemática da literatura é um meio de identificar, avaliar e interpretar toda a pesquisa disponível relevante para uma determinada questão de pesquisa, área temática ou fenômeno de interesse. A presente RS segue as diretrizes propostas em (BUDGEN; BRERETON, 2006), que especifica três passos: planejamento da RS, condução da RS e resultados da RS. Nesta subseção serão detalhados os passos de planejamento da RS e a condução da RS, na subseção 2.5 são detalhados os resultados e as análises referentes ao terceiro passo da RS.

A.1 *Justificativa para a RS*

Na literatura existem diferentes estudos teóricos e comparações de heurísticas de biagrupamento aplicados a dados de expressão gênica (MADEIRA; OLIVEIRA, 2004; TANAY; SHARAN; SHAMIR, 2005; BUSYGIN; PROKOPYEV; PARDALOS, 2008; MUKHOPADHYAY; MAULIK; BANDYOPADHYAY, 2010) mas, até onde foi possível identificar, não se encontrou nenhuma RS que aborde os diferentes algoritmos de biagrupamento e coagrupamento aplicados a dados textuais. Esta RS tem como objetivo avaliar e interpretar os estudos primários relevantes das tarefas de biagrupamento e coagrupamento com o propósito de identificar, analisar e avaliar trabalhos correlatos referentes aos algoritmos de biagrupamento e coagrupamento aplicadas aos dados textuais. Também espera-se descobrir os conjuntos de dados utilizados, a modalidade de aprendizado mais utilizada, a natureza dos algoritmos propostos em cada estudo, as medidas de avaliação aplicadas aos resultados das tarefas de biagrupamento e coagrupamento e as aplicações nas quais são utilizadas as tarefas de biagrupamento e coagrupamento.

A.2 *Questões de pesquisa*

Em base às necessidades de uma RS referente ao tema, as questões de pesquisa foram elaboradas:

- **Q1:** Quais são os conjuntos de dados utilizados nestes estudos?.

- **Q2:** Quais são os algoritmos de biagrupamento e coagrupamento utilizados em dados textuais?.
- **Q3:** Como estes algoritmos de biagrupamento e coagrupamento estão caracterizados?.
- **Q4:** Quais são os métodos de avaliação utilizados para estes algoritmos?.
- **Q5:** Quais são as aplicações na área de mineração de texto que podem utilizar o biagrupamento e o coagrupamento?.

O objetivo destas perguntas é determinar se existem conjuntos de dados específicos, úteis para o biagrupamento e coagrupamento (**Q1**); identificar os diferentes algoritmos de biagrupamento e coagrupamento aplicados especificamente em dados textuais (**Q2**); identificar, se estes algoritmos estão baseados em alguns métodos específicos, se existe algumas características entre eles, permitindo ter uma noção geral destes estudos (**Q3**); conhecer se existem na literatura formas de avaliar o biagrupamento e o coagrupamento com a finalidade de determinar sua qualidade (**Q4**); e conhecer as aplicações com respeito à área de mineração de texto em que estão sendo aplicadas as tarefas de biagrupamento e coagrupamento (**Q5**).

A.3 Escolha de fontes de dados e estratégias de busca

Na escolha das bases de dados as fontes de dados utilizadas foram: *IEEE Xplore*, *Science Direct*, *ACM Digital Library*, *Engineering Village* e *Scopus*. Sabe-se que o motor de busca *scopus* indexa todas as outras bases de dados, por tanto o motor de busca usado para todas as bases de dados foi *scopus*. A estratégia de busca foi estabelecida a partir de uma *string* genérica, mostrada na tabela 35 e aplicada no motor de busca. A *string* foi construída usando as palavras-chave derivadas das questões de pesquisa, conectadas através de conectivos lógicos *OR* e *AND*, que tem como finalidade maximizar o escopo da recuperação de informação e trazer estudos primários úteis para a composição das respostas das questões de pesquisa.

A.4 Definição de estratégias para a seleção de estudos primários e avaliação de sua qualidade

As estratégias para a seleção dos estudos primários a serem analisados foram por meio do estabelecimento de um conjunto de critérios de inclusão (CI) e exclusão (CE),

Tabela 35 – *String* de busca

TITLE-ABS-KEY (cocluster* OR "co-cluster*" OR "co cluster*" OR bicluster* OR "bi-cluster*" OR "bi cluster*" OR "bidimensional cluster*" OR "block cluster*" OR "direct cluster*" OR "direct-cluster*" OR "multiway cluster*" OR "multiway-cluster*" OR "multi-way cluster*" OR "simultaneous cluster*" OR "simultaneous-cluster*" OR "two-mode cluster*" OR "two mode cluster*" OR "two-sided cluster*" OR "two sided cluster*" OR "two-way cluster*") OR ("two way cluster*") AND TITLE-ABS-KEY (text* OR word* OR document* OR news OR "information retrieval") AND NOT TITLE-ABS-KEY (texture*) AND (LIMIT-TO (SUBJAREA , "COMP") OR LIMIT-TO (SUBJAREA , "MATH") OR LIMIT-TO (SUBJAREA , "ENGI") OR LIMIT-TO (SUBJAREA , "BUSI") OR LIMIT-TO (SUBJAREA , "MULT")) AND (LIMIT-TO (DOCTYPE , "cp") OR LIMIT-TO (DOCTYPE , "ar") OR LIMIT-TO(DOCTYPE , "ch") OR LIMIT-TO (DOCTYPE , "re") OR LIMIT-TO (DOCTYPE , "ip")) AND (LIMIT-TO (LANGUAGE , "English"))

Fonte: Alexandra Katiuska Ramos Diaz, 2018

tais critérios têm o objetivo de garantir que os estudos incluídos possuem um mínimo de qualidade (CE-4, CE-5), fossem estudos primários (CE-6), acessíveis (CE-2, CE-3) e pertencentes ao escopo da RS (CE-1, CI-1, CI-2, CI-3). Os critérios de inclusão e exclusão aplicados para a RS foram:

- CI-1:** Estudos que apresentem algoritmos de biagrupamento e coagrupamento que estejam relacionados com texto (dados textuais), mostrando explicitamente suas experimentações aplicadas a texto.
- CI-2:** Estudos que apresentem estudos práticos/experimentais que tenha permitido ao autor avaliar as os algoritmos de biagrupamento e o coagrupamento no que diz respeito a texto.
- CI-3:** Estudos pertencentes à área de computação, engenharia ou matemática
- CE-1:** Estudos onde o biagrupamento e o coagrupamento não seja o foco principal do trabalho.
- CE-2:** Estudos cujo texto completo não pode ser acessado via internet e solicitação aos autores.
- CE-3:** Estudos escritos em língua não-inglesa.

CE-4: Estudos incompletos ou mal relatados a ponto de não ser possível identificar como os experimentos foram realizados

CE-5: Estudos não revisados por pares tais como relatórios técnicos, resumos expandidos, capítulos de livros, etc.

CE-6: Estudos secundários ou terciários

Os critérios de inclusão e exclusão foram validados a partir da análise dos metadados, título, resumo e do texto na íntegra dos estudos.

A.5 Método de extração de dados dos estudos selecionados

A extração de dados foi por meio da leitura (mapeamento) dos estudos selecionados depois de aplicar os critérios de inclusão e exclusão. Como guia para a extração de dados foi criado um formulário para as anotações de informações básicas; tais como: título, autores, tipo de documento, ano de publicação; e informações sobre o conteúdo dos estudos (descritos detalhadamente na tabela 36).

Tabela 36 – Formulário de extração da informação dos estudos

Tópico	Descrição
Algoritmo de biagrupamento ou coagrupamento utilizado.	Este atributo contribui a descrever que tipo de algoritmo de biagrupamento ou coagrupamento está sendo utilizado.
Métodos de avaliação usada.	Ajuda a identificar qual ou quais são as medidas de avaliação utilizadas no estudo.
Conjunto de dados.	Este atributo ajuda a identificar como é o conjunto de dados utilizado no estudo, e saber qual é sua natureza.
Natureza do algoritmo proposto	Neste tópico ajuda a identificar se os algoritmos propostos estão baseado em métodos conhecidos na literatura.
Aplicações	Ajuda na identificação de que se os algoritmos propostos estão sendo aplicados à áreas específicas e quais áreas são estas.

Fonte: Alexandra Katiuska Ramos Diaz, 2018

A.6 Condução da RS

Na condução da RS, para a identificação e seleção dos estudos primários, a *string* de busca foi executada no período de julho de 2017 a junho de 2018, os recursos de filtros oferecidos pelo motor de busca foram utilizados para aplicar de forma automática parte dos critérios de exclusão, os filtros considerados foram por tipo de documento, por áreas (computação, matemática, engenharia, negócios e multidisciplinar), retornando 340 estudos. A partir desse ponto os critérios de inclusão e exclusão foram executados

manualmente por meio da análise do título e resumo de cada estudo, sendo selecionados 80 estudos, os quais foram analisados para verificar sua relevância. A extração e síntese seguiram o preenchimento do formulário da tabela 36, e os resultados estão consolidados na seção 2.5.