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

FERNANDO HENRIQUE DA SILVA COSTA

**Ensemble de agrupamentos para sistemas de recomendação baseados em
conteúdo**

São Paulo

2018

FERNANDO HENRIQUE DA SILVA COSTA

**Ensemble de agrupamentos para sistemas de recomendação baseados em
conteúdo**

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 5 de novembro 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

Costa, Fernando Henrique da Silva

Ensemble de agrupamentos para sistemas de recomendação
baseados em conteúdo / Fernando Henrique da Silva Costa ;
orientadora, Sarajane Marques Peres. – 2018.
150 f. : 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. Mineração de dados. 2. Inteligência artificial. I.
Peres, Sarajane Marques, orient. II. Título.

CDD 22.ed.– 006.312

Dissertação de autoria de Fernando Henrique da Silva Costa, sob o título “**Ensemble de agrupamentos para sistemas de recomendação baseados em conteúdo**”, 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 5 de novembro de 2018 pela comissão julgadora constituída pelos doutores:

Prof. Dr. José Jesús Perez Alcazar

Universidade de São Paulo

Presidente

Prof. Dr. Helton Hideraldo Bísaro

Universidade de São Paulo

Prof. Dr. Marcelo de Souza Lauretto

Universidade de São Paulo

Prof. Dr. Fabrício Olivetti de França

Universidade Federal do ABC

Dedico este trabalho aos meus queridos pais, Cesar Fernando e Mara Regina, pelo apoio incondicional em todas as frentes e pela educação, a qual me possibilitou contribuir para o meio científico.

Agradecimentos

Primeiramente, aos professores da graduação do Centro Universitário de Votuporanga, em especial a dois: Prof. Me. Fernando Kendy Aoki Rizzatto e Profa. Ma. Denise Marin Rodrigues. Estes foram de importância inigualável em relação ao apoio no meu desejo de realizar pós-graduação.

Aos professores do PPgSI por todo o conhecimento que produzi à respeito do mundo científico e acadêmico. Em especial, ao Prof. Dr. Ivandré Paraboni, por ter me auxiliado durante o período de estágio PAE, e à Profa. Dra. Sarajane Marques Peres, não somente por ter sido minha orientadora, mas também por todos os aspectos de sua pessoa, os quais me desenvolveram como pesquisador.

Aos meus amigos que me acompanharam durante a vida, inclusive neste momento. Às amizades realizadas em São Paulo, em especial ao Diego Vieira Neves e Jonas Mendonça Targino, além dos companheiros do grupo de pesquisa: Andrei Martins Silva, Alexandra Katiuska Ramos Diaz, André Paulino de Lima e José Luiz Maturana Pagnossim.

Por fim, à minha companheira, Vanessa Aparecida dos Santos, por estar presente em diversos momentos importantes desde a minha chegada em São Paulo e aos meus familiares em geral, destacando os meus pais Cesar Fernando Soares da Costa e Mara Regina Marcieli da Silva Costa, pelos auxílios a fim de me prover o melhor sempre que possível.

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Código de Financiamento 001.

“Tenho em mim todos os sonhos do mundo.”

(Fernando Pessoa)

Resumo

COSTA, Fernando Henrique da Silva. **Ensemble de agrupamentos para sistemas de recomendação baseados em conteúdo**. 2018. 150 f. Dissertação (Mestrado em Ciências) – Escola de Artes, Ciências e Humanidades, Universidade de São Paulo, São Paulo, 2018.

O crescimento acelerado da internet proporcionou uma quantidade grande de informações acessíveis aos usuários. Ainda que tal quantidade possua algumas vantagens, os usuários que possuem pouca ou nenhuma experiência para escolher uma alternativa dentre as várias apresentadas terão dificuldades em encontrar informações (ou itens, considerando o escopo deste trabalho) úteis e que atendam às suas necessidades. Devido a esse contexto, os sistemas de recomendação foram desenvolvidos para auxiliar os usuários a encontrar itens relevantes e personalizados. Tais sistemas são divididos em diversas arquiteturas. Como exemplo estão as arquiteturas baseadas em: conteúdo, filtro colaborativo e conhecimento. Para este trabalho, a primeira arquitetura foi explorada. A arquitetura baseada em conteúdo recomenda itens ao usuário com base na similaridade desses aos itens que o usuário mostrou interesse no passado. Por consequência, essa arquitetura possui a limitação de, geralmente, realizar recomendações com baixa serendipidade, uma vez que os itens recomendados tendem a ser semelhantes àqueles observados pelo o usuário e, portanto, não apresentam novidade ou surpresa. Diante desta limitação, o aspecto de serendipidade tem destaque nas discussões apresentadas neste trabalho. Assim, o objetivo deste trabalho é minimizar o problema da baixa serendipidade das recomendações por meio da utilização da análise de similaridades parciais implementada usando *ensemble* de agrupamentos. Para alcançar este objetivo, estratégias de recomendação baseadas em conteúdo implementadas usando agrupamento e *ensemble* de agrupamento foram propostas e avaliadas neste trabalho. A avaliação contou com análises qualitativas sobre as recomendações produzidas e com um estudo com usuários. Nesse estudo, quatro estratégias de recomendação de notícias foram avaliadas, incluindo as duas propostas neste trabalho, uma estratégia baseada em recomendação aleatória, e uma estratégia baseada em coagrupamento. As avaliações consideraram aspectos de relevância, surpresa e serendipidade de recomendações. Esse último aspecto é descrito como itens que apresentam tanto surpresa quanto relevância ao usuário. Os resultados de ambas análises mostraram a viabilidade da utilização de agrupamento como base de recomendação, uma vez que o *ensemble* de agrupamentos obteve resultados satisfatórios em todos os aspectos, principalmente em surpresa, enquanto a estratégia baseada em agrupamento simples obteve os melhores resultados em relevância e serendipidade.

Palavras-chaves: Sistemas de recomendação baseados em conteúdo. Serendipidade. Similaridade parcial. *Ensemble* de agrupamento.

Abstract

COSTA, Fernando Henrique da Silva. **Cluster ensemble to content-based recommender systems**. 2018. 150 p. Dissertation (Master of Science) – School of Arts, Sciences and Humanities, University of São Paulo, São Paulo, 2018.

The accelerated growth of the internet has provided a large amount of information accessible to users. Although this amount of information has some advantages, users who have little or no experience in choosing one of several alternatives will find it difficult to find useful information (or items, considering the scope of this work) that meets their needs. Due to this context, recommender systems have been developed to help users find relevant and personalized items. Such systems are divided into several architectures as content-based, collaborative filtering and knowledge-based. The first architecture was explored in this work. The content-based architecture recommends items to the user based on their similarity to items that the user has shown interest in the past. Consequently, this architecture has the limitation of generally making recommendations with low serendipity, since the recommended items tend to be similar to those observed by the user and, therefore, do not present novelty or surprise. Given this limitation, the aspect of serendipity is highlighted in the discussions presented in this work. Thus, the objective of this work is to minimize the problem of the low serendipity of the recommendations through the use of the partial similarity analysis implemented using cluster ensemble. To achieve this goal, content-based recommendation strategies implemented using clustering and cluster ensemble were proposed and evaluated. The evaluation involved qualitative analysis of the recommendations and a study with users. In such a study, four news recommendation strategies were evaluated including the two strategies proposed in this work, a strategy based on random recommendation, and a strategy based on co-clustering. The evaluations considered aspects of relevance, surprise and serendipity of recommendations. This last aspect is described as items that present both surprise and relevance to the user. The results of both analyzes showed the feasibility of using clustering as the basis of recommendation, since cluster ensemble had satisfactory results in all aspects, mainly in surprise, whereas the simple clustering-based strategy obtained the best results in relevance and serendipity.

Keywords: Serendipity. Partial similarity. Cluster ensemble. Content-based recommender systems.

Lista de figuras

Figura 1 – Exemplos de recomendações no portal do G1	21
Figura 2 – Arquitetura de recomendação	24
Figura 3 – Arquitetura dos primeiros sistemas de recomendação	31
Figura 4 – Exemplo de lista de itens ordenados	32
Figura 5 – Cenário de um sistema de recomendação moderno	33
Figura 6 – Construção de um perfil de interesse	35
Figura 7 – Exemplo de agrupamento de dados	41
Figura 8 – Exemplo de um <i>ensemble</i>	42
Figura 9 – Exemplo de um <i>ensemble</i> de agrupamentos	44
Figura 10 – Passos do algoritmo <i>k-means</i>	49
Figura 11 – Esquema gráfico para estratégias para recomendações relevantes	66
Figura 12 – Esquema gráfico para estratégias para recomendações surpreendentes .	68
Figura 13 – Esquema gráfico para estratégias para recomendações serendipitosas .	70
Figura 14 – Exemplo de grupos com termos similares	73
Figura 15 – Binarizando a matriz C	75
Figura 16 – Interseção dos grupos	75
Figura 17 – Plotagem de Silhouette considerando os cadernos das notícias como um resultado de agrupamento	78
Figura 18 – Plotagem do índice Silhouette em execuções X_4 e X_{10} no experimento #1 do algoritmo <i>k-means++</i>	83
Figura 19 – Nuvens de palavras em cada grupo obtida em execução sobre o conjunto X_{10} , no experimento #1 com o algoritmo <i>k-means++</i>	84
Figura 20 – Relação cadernos de notícias nos grupos obtidos em uma execução sobre o conjunto X_{10} no experimento #1 com o algoritmo <i>k-means++</i>	86
Figura 21 – Plotagem do índice Silhouette em execuções X_8 e X_2 no experimento #2 do algoritmo <i>k-means++</i>	87
Figura 22 – Nuvens de palavras em cada grupo obtidas em uma execução sobre o conjunto X_2 no experimento #2 com o algoritmo <i>k-means++</i>	87
Figura 23 – Relação cadernos de notícias nos grupos obtidos em uma execução sobre o conjunto X_2 no experimento #2 com o algoritmo <i>k-means++</i>	88

Figura 24 – Plotagem do índice Silhouette em uma execução X_{30} no experimento #1 do algoritmo <i>ensemble</i> de agrupamentos	90
Figura 25 – Nuvens de palavras em cada grupo obtidas em uma execução sobre o conjunto X_{30} no experimento #1 com o algoritmo <i>ensemble</i> de agrupamentos	91
Figura 26 – Relação cadernos de notícias nos grupos obtidos em uma execução sobre o conjunto X_{30} no experimento #1 com o algoritmo <i>ensemble</i> de agrupamentos	92
Figura 27 – Plotagem do índice Silhouette em uma execução X_{47} no experimento #2 do algoritmo <i>ensemble</i> de agrupamentos	93
Figura 28 – Nuvens de palavras em cada grupo obtidas em uma execução sobre o conjunto X_{47} no experimento #2 com o algoritmo <i>ensemble de agrupamentos</i>	94
Figura 29 – Relação cadernos de notícias nos grupos obtidos em uma execução sobre o conjunto X_{47} no experimento #2 com o algoritmo <i>ensemble</i> de agrupamentos	95
Figura 30 – Plotagem do índice Silhouette em execuções na matrizes X_{42} e X_8 no experimento #3 do algoritmo <i>ensemble</i> de agrupamentos	96
Figura 31 – Nuvens de palavras em cada grupo obtidas em uma execução sobre o conjunto X_8 no experimento #3 com o algoritmo <i>ensemble</i> de agrupamentos	97
Figura 32 – Relação cadernos de notícias nos grupos obtidos em uma execução sobre o conjunto X_8 no experimento #3 com o algoritmo <i>ensemble</i> de agrupamentos	98
Figura 33 – Plotagem do índice Silhouette em execuções nas matrizes X_8 e X_{47} no experimento #4 do algoritmo <i>ensemble</i> de agrupamentos	99
Figura 34 – Nuvens de palavras em cada grupo obtidas em uma execução sobre o conjunto X_{47} no experimento #4 com o algoritmo <i>ensemble</i> de agrupamentos	100
Figura 35 – Relação cadernos de notícias nos grupos obtidos em uma execução sobre o conjunto X_{47} no experimento #4 com o algoritmo <i>ensemble</i> de agrupamentos	101

Figura 36 – Plotagem do índice Silhouette em execuções em X_8 e X_{49} no experimento #5 do algoritmo <i>ensemble</i> de agrupamentos	102
Figura 37 – Nuvens de palavras em cada grupo obtidas em uma execução sobre o conjunto X_{49} no experimento #5 com o algoritmo <i>ensemble</i> de agrupamentos	103
Figura 38 – Relação cadernos de notícias nos grupos obtidos em uma execução sobre o conjunto X_{49} no experimento #5 com o algoritmo <i>ensemble</i> de agrupamentos	104
Figura 39 – Plotagem do índice Silhouette em uma execução X_{47} no experimento #6 do algoritmo <i>ensemble</i> de agrupamentos	105
Figura 40 – Nuvens de palavras em cada grupo obtidas em uma execução sobre o conjunto X_{47} no experimento #6 com o algoritmo <i>ensemble</i> de agrupamentos	106
Figura 41 – Relação cadernos de notícias nos grupos obtidos em uma execução sobre o conjunto X_{47} no experimento #6 com o algoritmo <i>ensemble</i> de agrupamentos	107
Figura 42 – Plotagem do índice Silhouette em uma execução X_{33} no experimento #7 do algoritmo <i>ensemble</i> de agrupamentos	108
Figura 43 – Nuvens de palavras em cada grupo obtidas em uma execução sobre o conjunto X_{33} no experimento #7 com o algoritmo <i>ensemble</i> de agrupamentos	109
Figura 44 – Relação cadernos de notícias nos grupos obtidos em uma execução sobre o conjunto X_{33} no experimento #7 com o algoritmo <i>ensemble</i> de agrupamentos	110
Figura 45 – Plotagem do índice Silhouette em uma execução em X_8 e X_{10} no experimento #8 do algoritmo <i>ensemble</i> de agrupamentos	111
Figura 46 – Nuvens de palavras em cada grupo obtidas em uma execução sobre o conjunto X_{10} no experimento #8 com o algoritmo <i>ensemble</i> de agrupamentos	112
Figura 47 – Relação cadernos de notícias nos grupos obtidos em uma execução sobre o conjunto X_{10} no experimento #8 com o algoritmo <i>ensemble</i> de agrupamentos	113

Figura 48 – Plotagem do índice Silhouette em uma execução X_8 no experimento #9 do algoritmo <i>ensemble</i> de agrupamentos	114
Figura 49 – Nuvens de palavras em cada grupo obtidas em uma execução sobre o conjunto X_{47} no experimento #9 com o algoritmo <i>ensemble</i> de agrupa- mentos	115
Figura 50 – Relação cadernos de notícias nos grupos obtidos em uma execução sobre o conjunto X_8 no experimento #9 com o algoritmo <i>ensemble</i> de agrupamentos	116
Figura 51 – Tela da notícia semente e das instruções	130
Figura 52 – Detalhes sobre instruções e definições usadas no estudo com usuários .	131
Figura 53 – Tela das recomendações	131
Figura 54 – Tela da escolha	132
Figura 55 – Desempenho dos algoritmos em cada aspecto de recomendação	134
Figura 56 – Relação cadernos das notícias sementes com cadernos da notícias esco- lhidas como melhor recomendação no aspecto de relevância	136
Figura 57 – Relação cadernos das notícias sementes com cadernos da notícias esco- lhidas como melhor recomendação no aspecto de surpresa	137
Figura 58 – Relação cadernos da notícias semente com cadernos da notícias escolhi- das como melhor recomendação no aspecto de serendipidade	137

Lista de algoritmos

Algoritmo 1 – Proposta de estratégia de recomendação	25
Algoritmo 2 – Algoritmo <i>k-means</i>	49
Algoritmo 3 – Algoritmo <i>k-means++</i>	50
Algoritmo 4 – Algoritmo hierárquico aglomerativo	51
Algoritmo 5 – Primeira estratégia de recomendação relevante para o algoritmo <i>k-means++</i>	67
Algoritmo 6 – Segunda estratégia de recomendação relevante para o algoritmo <i>k-means++</i>	67
Algoritmo 7 – Primeira estratégia de recomendação surpreendente para o algoritmo <i>k-means++</i>	69
Algoritmo 8 – Segunda estratégia de recomendação surpreendente para o algoritmo <i>k-</i> <i>means++</i>	69
Algoritmo 9 – Primeira estratégia de recomendação serendipitosa para o algoritmo <i>k-means++</i>	71
Algoritmo 10 – Segunda estratégia de recomendação serendipitosa para o algoritmo <i>k-means++</i>	71
Algoritmo 11 – Estratégia de recomendação serendipitosa para o algoritmo <i>ensemble</i> de agrupamentos	76

Lista de quadros

Quadro 1 – Bloco de linhas e colunas	53
--	----

Lista de tabelas

Tabela 1 – Matriz de similaridade	45
Tabela 2 – Matrizes de similaridade de quatro componentes	46
Tabela 3 – Matriz de coassociação	46
Tabela 4 – Lista de trabalhos correlatos e suas respectivas áreas de aplicação . . .	57
Tabela 5 – Lista de trabalhos correlatos e os conjuntos de dados utilizados	57
Tabela 6 – Distribuição dos dados coletados nos conjuntos MovieLeans e Netflix .	58
Tabela 7 – Lista de trabalhos correlatos e os algoritmos de agrupamento utilizados para geração dos componentes de <i>ensemble</i>	62
Tabela 8 – Lista de trabalhos correlatos e as medidas de similaridades aplicadas .	62
Tabela 9 – Lista de trabalhos correlatos, as funções de consenso aplicadas e os algoritmos utilizados para gerar o resultado final	63
Tabela 10 – Conjunto de notícias: <i>Corpus EBC</i>	78
Tabela 11 – Configurações usadas nos experimentos com o ensemble de agrupamentos	82
Tabela 12 – Qualidade de agrupamento para o experimento #1 do algoritmo <i>k-means++</i> , de acordo com índices de validação de agrupamento. Destaque em negrito para o melhor resultado em cada índice	83
Tabela 13 – Qualidade de agrupamento para o experimento #2 do algoritmo <i>k-means++</i> , de acordo com índices de validação de agrupamento. Destaque em negrito para o melhor resultado em cada índice	86
Tabela 14 – Qualidade de agrupamento para o experimento #1 do algoritmo <i>ensemble</i> de agrupamentos, de acordo com índices de validação de agrupamento. Destaque em negrito para o melhor resultado em cada índice	89
Tabela 15 – Qualidade de agrupamento para o experimento #2 do algoritmo <i>ensemble</i> de agrupamentos, de acordo com índices de validação de agrupamento. Destaque em negrito para o melhor resultado em cada índice	92
Tabela 16 – Qualidade de agrupamento para o experimento #3 do algoritmo <i>ensemble</i> de agrupamentos, de acordo com índices de validação de agrupamento. Destaque em negrito para o melhor resultado em cada índice	96

Tabela 17 – Qualidade de agrupamento para o experimento #4 do algoritmo <i>ensemble</i> de agrupamentos, de acordo com índices de validação de agrupamento. Destaque em negrito para o melhor resultado em cada índice	98
Tabela 18 – Qualidade de agrupamento para o experimento #5 do algoritmo <i>ensemble</i> de agrupamentos, de acordo com índices de validação de agrupamento. Destaque em negrito para o melhor resultado em cada índice	102
Tabela 19 – Qualidade de agrupamento para o experimento #6 do algoritmo <i>ensemble</i> de agrupamentos, de acordo com índices de validação de agrupamento. Destaque em negrito para o melhor resultado em cada índice	104
Tabela 20 – Qualidade de agrupamento para o experimento #7 do algoritmo <i>ensemble</i> de agrupamentos, de acordo com índices de validação de agrupamento. Destaque em negrito para o melhor resultado em cada índice	107
Tabela 21 – Qualidade de agrupamento para o experimento #8 do algoritmo <i>ensemble</i> de agrupamentos, de acordo com índices de validação de agrupamento. Destaque em negrito para o melhor resultado em cada índice	110
Tabela 22 – Qualidade de agrupamento para o experimento #9 do algoritmo <i>ensemble</i> de agrupamentos, de acordo com índices de validação de agrupamento. Destaque em negrito para o melhor resultado em cada índice	113

Lista de símbolos

$A, k, l, m, n, \eta, p, u$	número de elementos de um conjunto (conjunto especificado apropriadamente no texto)
a, b, h, i, j, o	contadores
C, λ, M, MC, U	matrizes
E	conjunto de experiências
G	conjunto de grupos
g	grupo
$\vec{g}, \overline{\mathbf{g}}$	coordenada espacial
L	lista de recomendação
$\mathcal{M}$	medida de desempenho
$\aleph, \Re$	espaço
Φ	algoritmo de agrupamento
T	classe de tarefas
θ	limiar
X	conjunto de dados
x	dado
xo	notícia disponível no conjunto
xs	notícia semente
W	vetor de parâmetros

Sumário

1	Introdução	21
1.1	<i>Definição do problema</i>	23
1.2	<i>Hipótese</i>	25
1.3	<i>Objetivos</i>	25
1.4	<i>Método</i>	26
1.5	<i>Organização do documento</i>	27
2	Sistemas de recomendação baseados em conteúdo	30
2.1	<i>Vantagens</i>	35
2.2	<i>Desvantagens</i>	36
2.3	<i>Qualidade de recomendações</i>	37
2.4	<i>Considerações finais</i>	39
3	Ensemble de agrupamento	40
3.1	<i>Geração de membros</i>	43
3.2	<i>Função de consenso</i>	45
3.3	<i>Métrica de similaridade</i>	47
3.4	<i>Métodos de agrupamento</i>	47
3.5	<i>Algoritmos k-means e k-means++</i>	48
3.6	<i>Algoritmo de agrupamento hierárquico aglomerativo</i>	50
3.7	<i>Índices de avaliação de agrupamento</i>	51
3.8	<i>Similaridades parciais</i>	52
3.9	<i>Considerações finais</i>	53
4	Trabalhos correlatos	55
4.1	<i>Serendipidade em sistemas de recomendação</i>	55
4.1.1	<i>Estratégias de recomendação</i>	55
4.1.2	<i>Áreas de aplicação</i>	56
4.1.3	<i>Conjuntos de dados</i>	57
4.1.4	<i>Avaliações das abordagens</i>	58
4.1.5	<i>Resultados</i>	59

4.2	<i>Agrupamentos em sistemas de recomendação</i>	60
4.3	<i>Ensemble de agrupamentos em dados textuais</i>	61
4.3.1	Geração dos componentes de <i>ensemble</i>	61
4.3.2	Medidas de similaridades	62
4.3.3	Funções de consenso	62
4.4	<i>Considerações finais</i>	63
5	Estratégias de recomendação baseadas em agrupamento	64
5.1	<i>Estratégias de recomendação baseadas em agrupamento com k-means++</i> 64	
5.1.1	Estratégias para recomendações relevantes	65
5.1.2	Estratégias para recomendações surpreendentes	67
5.1.3	Estratégia para recomendações serendipitosas	69
5.2	<i>Estratégia de recomendação baseada em ensemble de agrupamentos</i> . .	72
5.2.1	Construindo um <i>ensemble</i> de agrupamentos	73
5.2.2	Construindo a estratégia de recomendação	75
5.3	<i>Considerações finais</i>	76
6	Experimentos com agrupamentos	77
6.1	<i>Conjunto de notícias</i>	77
6.2	<i>Pré-processamento</i>	79
6.3	<i>Configurações de experimento</i>	80
6.3.1	<i>k-means++</i>	80
6.3.2	<i>Ensemble</i> de agrupamentos	81
6.4	<i>Resultados com k-means++</i>	82
6.4.1	Experimento #1	83
6.4.2	Experimento #2	86
6.5	<i>Resultados com ensemble de agrupamentos</i>	88
6.5.1	Experimento #1	89
6.5.2	Experimento #2	91
6.5.3	Experimento #3	95
6.5.4	Experimento #4	98
6.5.5	Experimento #5	101
6.5.6	Experimento #6	104

6.5.7	Experimento #7	107
6.5.8	Experimento #8	109
6.5.9	Experimento #9	112
6.6	<i>Análises</i>	115
6.7	<i>Considerações finais</i>	118
7	Análise das estratégias de recomendação	119
7.1	<i>Recomendações geradas a partir do agrupamento obtido com o k-means++</i>	119
7.1.1	<i>Experimento #1 - análise qualitativa das recomendações</i>	119
7.1.2	<i>Experimento #2 - análise qualitativa das recomendações</i>	122
7.1.3	<i>Discussão geral</i>	125
7.2	<i>Recomendações realizadas a partir do agrupamento obtido com o ensemble</i>	126
7.3	<i>Estudo com usuários</i>	128
7.3.1	<i>Objetivo geral do estudo</i>	128
7.3.2	<i>Sistema para interação com usuário</i>	129
7.3.3	<i>Experimento piloto</i>	131
7.3.4	<i>Dados sobre a condução do estudo</i>	133
7.3.5	<i>Análises</i>	133
7.3.6	<i>Considerações finais</i>	137
8	Conclusão	139
8.1	<i>Contribuições</i>	140
8.2	<i>Limitações</i>	141
8.3	<i>Trabalhos futuros</i>	141
	Referências¹	142

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

1 Introdução

O crescimento acelerado da internet originou uma quantidade grande de dados e informações, além de uma alta disponibilidade para eles. Consequentemente, esse fato gerou dificuldade aos usuários para encontrar os dados e informações apropriados aos seus propósitos (LI et al., 2018). Portanto, devido a essa grande quantidade de opções disponibilizadas aos usuários, surgiu a necessidade de desenvolver formas que ajudassem na descoberta de itens oportunos e personalizados. Um modo de encontrar itens relevantes dentre as várias alternativas apresentadas é por meio de recomendações, as quais podem ser oferecidas por outros usuários ou por meio de sistemas de recomendação (KUMAR et al., 2018).

O objetivo de um sistema de recomendação é providenciar itens, selecionados dentre todas as opções, que sejam os mais atrativos e relevantes para um determinado usuário (WU et al., 2018). Tais sistemas tornaram-se uma ferramenta de negócio importante em empresas como Netflix¹, Amazon², Google³, além de outras empresas (KOUTRIKA, 2018). Alguns exemplos de aplicação nesse contexto são as recomendações de livros na Amazon, de filmes na Netflix e de hotéis no TripAdvisor⁴ (CHE; HÖRNER, 2017).

A figura 1 ilustra um cenário real de recomendação de notícias no portal G1⁵. Nesse cenário, as duas notícias apresentadas na figura foram recomendadas a partir de uma “notícia semente”⁶ intitulada *“Justiça de SP concede liminar que suspende cobrança extra por despacho de bagagem”*⁷.

Figura 1 – Exemplos de recomendações no portal do G1

Fonte: Portal de notícias G1, acessado em 14 de março de 2017

¹ <https://www.netflix.com/br-en/>

² <https://www.amazon.com/>

³ <https://www.google.com.br/>

⁴ <https://www.tripadvisor.com.br/>

⁵ <http://g1.globo.com/>

⁶ O termo “notícia semente” é usado neste texto com o fim de indicar a notícia na qual a estratégia de recomendação se baseia para gerar uma lista de recomendação.

⁷ <http://g1.globo.com/economia/seu-dinheiro/noticia/justica-de-sp-concede-liminar-que-suspende-cobranca-extra-por-despacho-de-bagagem.ghtml>

Existem diversas arquiteturas no contexto de sistemas de recomendação, dentre elas estão as que são baseadas em: conteúdo, filtro colaborativo e conhecimento (BAGHER; HASSANPOUR; MASHAYEKHI, 2017). Este trabalho estuda a primeira arquitetura, a qual sugere itens similares aos itens que foram apreciados pelo usuário alvo (BORATTO et al., 2017). A título de exemplo, considere um usuário que tenha o hábito de ler notícias a respeito do seu clube de futebol favorito. Visto que, muito provavelmente, esse usuário já acessou e apreciou notícias sobre o clube, o sistema de recomendação, muito provavelmente e por consequência desse perfil de leitura, recomendará novas notícias sobre tal clube.

Como mencionado anteriormente, vários cenários de aplicação utilizam um sistema de recomendação. Neste trabalho, o cenário de interesse é o de notícias publicadas em portais na internet, assim, o trabalho trata de recomendação baseada em conteúdo textual. Nesse cenário, o usuário é exposto a um grande volume de informações com o qual ele pode ter dificuldade em lidar e, portanto, ocasionando um certo desconforto (LIU; DOLAN; PEDERSEN, 2010). Tendo isso em vista, um sistema de recomendação pode ajudar os usuários a encontrar notícias que sejam pertinentes ao contexto de suas leituras.

Atualmente, há inúmeros sistemas de recomendação disponíveis aos usuários em diversos contextos. Esses sistemas estão constantemente buscando oferecer boas recomendações. Em relação à avaliação do que pode ser considerado uma boa recomendação, a acurácia foi a primeira medida a ser aplicada pelos desenvolvedores dos sistemas. De acordo com Shani e Gunawardana (2011), as recomendações acuradas, ainda que importantes, são insuficientes para dar suporte às arquiteturas que oferecem boas recomendações. Além disso, há também um consenso entre os pesquisadores atuais sobre o fato que recomendações com alta acurácia nem sempre agradam o usuário (KOTKOV et al., 2018). Com a finalidade de melhorar a qualidade de recomendação, novos aspectos de avaliação em sistemas de recomendação foram elaborados, e exemplos são os aspectos de surpresa e serendipidade. Surpresa pode ser definida como um item recomendado que não está na expectativa do usuário (ADAMOPOULOS; TUZHILIN, 2011). Serendipidade é definida como a experiência de receber recomendações de itens que sejam tanto inesperados quanto relevantes (KAMINSKAS; BRIDGE, 2017). Este último aspecto é o principal interesse de análise neste trabalho.

1.1 Definição do problema

Os sistemas de recomendação baseados em conteúdo possuem algumas vantagens em relação a outras arquiteturas típicas. De acordo com Lops, Gemmis e Semeraro (2011), essas vantagens são: a independência de outros usuários para a geração de recomendações (problema encontrado na arquitetura de filtro colaborativo); a transparência do funcionamento do sistema; e a independência de avaliações sobre novos itens inseridos no sistema (problema de *cold start* de item). Apesar disso, os mesmos autores destacam que há certas deficiências nesses sistemas, as quais são: a limitação da análise de conteúdo; o problema de recomendações com baixa serendipidade; e a dependência de avaliações de itens por um usuário novo no sistema (problema de *cold start* de usuário).

Dentre as deficiências listadas anteriormente, o problema estudado neste trabalho é o problema de recomendações com baixa serendipidade, o qual está relacionado à super especialização (do inglês: *overspecialization*) das recomendações. Essa limitação é alusiva a um sistema de recomendação que recomenda apenas itens similares aos itens que o usuário mostrou interesse (BALABANOVIĆ; SHOHAM, 1997; JAMES; RAJKUMAR, 2017) e, por consequência, a geração de recomendação de itens que não são novos ou surpreendentes ao usuário. Um exemplo de uma recomendação acurada, porém que não apresenta serendipidade, é a recomendação dos filmes atuais da saga *Star Wars* a usuários que já assistiram aos filmes mais antigos.

Dado o contexto de deficiência de sistemas de recomendação baseados em conteúdo em relação às recomendações com baixa serendipidade, este trabalho define como problema de pesquisa a busca por uma estratégia de recomendação que seja capaz de melhorar essa deficiência. A alternativa levantada para resolver esse problema é a adoção da perspectiva de análise de similaridade parcial ou similaridade baseada em partes. Tal perspectiva é comumente difundida em algoritmos de coagrupamento, os quais visam agrupar simultaneamente dados e seus respectivos atributos. Contudo, devido à tendência do espaço de busca de tais algoritmos ser mais complexo quando comparado com a perspectiva de similaridade total (FRANÇA, 2010), este trabalho utiliza *ensemble* de agrupamentos (STREHL; GHOSH, 2002; FERN; BRODLEY, 2003). Dessa maneira, as similaridades parciais são buscadas a partir da estratégia de criação de diversidade de componentes do *ensemble*, usando variação do espaço de atributos ou de dados para a

indução de cada componente. Ainda que o *ensemble* de agrupamentos também seja uma estratégia cara em termos de exploração de todo o espaço de busca, o tamanho desse espaço pode ser tratado como uma decisão de projeto de *ensemble* e a sua arquitetura permite, por exemplo, paralelização.

A figura 2 ilustra a arquitetura de recomendação proposta. Dado um conjunto de dados (neste trabalho, notícias provenientes de um portal de notícias), a primeira tarefa é realizar o pré-processamento dos dados. Após a realização dessa tarefa, uma matriz de n documentos por m termos é gerada. O *ensemble* de agrupamentos recebe a matriz como entrada, e os seus p componentes geram saídas (resultados isolados de agrupamento). Essas saídas são integradas por meio de uma estratégia de coassociação. Por fim, o agrupamento resultante do *ensemble* serve como uma base para implementação de uma estratégia de recomendação.

Figura 2 – Arquitetura de recomendação

Fonte: Fernando Henrique da Silva Costa, 2018

Considerando o resultado de agrupamento obtido com o *ensemble* e a interação do usuários com o sistema de recomendação, uma estratégia de recomendação é proposta. O algoritmo 1 apresenta como é realizado a recomendação proposta neste trabalho.

Em paralelo a esta solução, neste trabalho também é proposta uma estratégia de recomendação mais simples, construída sobre o resultado de agrupamento oferecido por um algoritmo básico, o *k-means++* (ARTHUR; VASSILVITSKII, 2007). Essa estratégia é proposta para fins de comparação com a estratégia baseada em *ensemble* de agrupamentos.

Algoritmo 1 Proposta de estratégia de recomendação

- 1: **procedure** ESTRATEGIARECOMENDACAO
- 2: Obter o resultado de agrupamento por meio do *ensemble* de agrupamentos.
- 3: Encontrar as similaridades parciais entre os grupos com base nos termos (palavras) que esses grupos possuem em comum.
- 4: Identificar o grupo no qual está localizada a notícia semente, ou seja, aquela na qual o usuário mostrou interesse.
- 5: Recomendar notícias aleatórias pertencentes aos grupos que possuem similaridades parciais ao grupo da notícia semente.

Fonte: Fernando Henrique da Silva Costa ,2018

A motivação para uso de *ensemble* de agrupamentos reside nas premissas que: (a) a aplicação do aprendizado não supervisionado direciona o raciocínio de recomendação ao uso exclusivo do conteúdo das notícias, liberando a recomendação de qualquer viés pré existente⁸; b) resultados de *ensemble* de agrupamentos podem ser interpretados em termos de similaridades parciais entre os objetos agrupados. Dadas essas premissa, o interesse deste trabalho é explorar as similaridades totais e similaridades parciais existentes entre as notícias, no que diz respeito ao efeito sobre os agrupamentos que organizarão as notícias e que embasarão a criação das estratégias de recomendação.

1.2 Hipótese

Considerando o escopo de definição do problema e a alternativa do uso de *ensemble* de agrupamentos, a hipótese delineada para este trabalho é: uma estratégia de recomendação construída sobre a estrutura de agrupamento fornecida por *ensemble* de agrupamento gera recomendações mais relevantes, surpreendentes e serendipitosas, do que aquelas construídas por estratégias geradas sobre a estrutura de agrupamento fornecida por um algoritmo que usam similaridade total.

1.3 Objetivos

Com o intuito de produzir resultados que permitam verificar a hipótese delineada para este trabalho, estabelece-se o seu objetivo geral: aplicar *ensembles* de agrupamentos como estratégia de extração de informação de notícias, de forma a prover uma base de

⁸ A informação de organização pré-existente (canais que organizam as notícias no portal de origem) será usada para fins de análise de resultados. No entanto, a abordagem proposta independe da existência de tal informação.

recomendação. A base de recomendação pretendida deve ser capaz de alcançar resultados relevantes, surpreendentes e serendipitosos.

O pleno alcance do objetivo geral proposto para este projeto depende do alcance de dois objetivos específicos que representam problemas particulares no tema aqui tratado:

- propor uma estratégia de recomendação de notícias a partir do resultado apresentado pelo *ensemble* de agrupamentos, que seja capaz de oferecer recomendações de qualidade, considerando os aspectos de relevância, surpresa e serendipidade;
- realizar uma interpretação do resultado apresentado pelo *ensemble* de agrupamentos sob a perspectiva de análise de similaridade baseada em partes.

1.4 Método

Este trabalho é constituído como uma pesquisa experimental, a qual envolve: pesquisa bibliográfica; tratamento e representação de dados textuais (notícias); experimentação computacional e experimentação com usuários; e avaliação dos resultados obtidos em ambas experimentações, sob aspectos quantitativos e qualitativos.

A pesquisa bibliográfica engloba a realização de estudos exploratórios, com o intuito de levantar informações de fundamentação teórica referente aos temas envolvidos na pesquisa, a saber: sistemas de recomendação baseado em conteúdo e *ensemble* de agrupamentos. Além disso, também tem o objetivo de levantar os trabalhos correlatos aos assuntos de interesse abordados nesta pesquisa: serendipidade em sistemas de recomendação, agrupamento como base para desenvolvimento de estratégias de recomendação e *ensemble* de agrupamentos aplicados a dados textuais.

Como o contexto de recomendação estudado neste trabalho é referente a conteúdo textual, as experimentações planejadas necessitam de bases de notícias. Para tal, o conjunto *CorpusEBC*⁹ foi utilizado para servir como conjunto de dados. O pré-processamento nesse conjunto foi realizado por meio de funções disponibilizadas na biblioteca *Quanteda*¹⁰, a qual é intrínseca à linguagem R (R Core Team, 2017). As funções utilizadas foram: remoções de números, símbolos, pontuações e *stopwords*; representações binária, via *tf* e *tf-idf* não normalizadas e normalizadas; seleção de atributos por meio de limiares baseados em mínima frequência nos documentos e na própria medida gerada na representação *tf-idf*. O

⁹ <https://anonimo.anonimo>

¹⁰ <https://www.rdocumentation.org/packages/quanteda/versions/1.0.0>

detalhamento sobre o uso de ferramentas para pré-processamento de textos está detalhado em um relatório técnico associado a este trabalho (DIAZ et al., 2018).

A plataforma Matlab¹¹ foi escolhida para a implementação, testes e análise das estratégias de agrupamento. Nessa mesma plataforma foram implementadas as estratégias de recomendação. Já para a construção do protótipo de sistema de recomendação usado no estudo com usuários, a linguagem R juntamente com a biblioteca Shiny(CHANG et al., 2018) foram usadas.

As avaliações realizadas sobre os artefatos gerados neste trabalho (componentes do *ensemble*, agrupamento de dados e listas de recomendação) precisam ser analisadas de diferentes pontos de vista. Para cada um deles, um conjunto de índices e procedimentos foi adotado. Para avaliação dos agrupamentos gerados nos experimentos, tanto com algoritmo de agrupamento básico quanto com *ensemble*, os índices externos Rand Ajustado (HUBERT; ARABIE, 1985) e Informação Mútua Normalizada (do inglês: *Normalized Mutual Information*) (STREHL; GHOSH, 2002) e o índice interno Silhouette (ROUSSEEUW, 1987) foram aplicados. O intuito de usar tais índices é validar os agrupamentos que podem ser usados dentro da estratégia de recomendação. Para o cálculo da diversidade entre os componentes do *ensemble* foram utilizadas quatro medidas baseadas no índice Rand Ajustado (NALDI; FACELI; CARVALHO, 2009). Para avaliar os aspectos de qualidade de recomendação, as interações dos usuários com um protótipo de recomendação foram avaliadas usando estatística descritiva.

1.5 Organização do documento

O presente trabalho é dividido em oito capítulos, considerando esta introdução. Os demais capítulos estão organizados da seguinte forma:

- Capítulo 2: nesse capítulo são apresentadas informações sobre sistemas de recomendação baseados em conteúdo. No decorrer do capítulo, são discutidas as motivações para criação de um sistema de recomendação, arquiteturas de recomendação com destaque para a arquitetura baseada em conteúdo, as vantagens e desvantagens da arquitetura baseada em conteúdo e, por fim, os aspectos de qualidade de recomendação avaliados neste trabalho.

¹¹ <https://www.mathworks.com/products/matlab.html>

- Capítulo 3: nesse capítulo são apresentadas informações referentes a *ensemble* de agrupamentos. Inicialmente, os paradigmas de *ensemble* e agrupamento em aprendizado de máquina são discutidos. Então, são apresentadas informações específicas a respeito de *ensemble* de agrupamentos. Para isso, conteúdo sobre a geração dos membros do ensemble, cálculo da diversidade entre as partições geradas por cada membro, a função de consenso, métrica de similaridade, métodos de agrupamentos, algoritmos *k-means*, *k-means++* e hierárquicos aglomerativos, índices de avaliação de agrupamento e análise de similaridades parciais são discutidos.
- Capítulo 4: nesse capítulo são apresentados os trabalhos correlatos. Esses trabalhos foram divididos em três tópicos: serendipidade em sistemas de recomendação, agrupamentos em sistemas de recomendação e *ensemble* de agrupamentos em dados textuais. Para discutir os dois primeiros tópicos foram pesquisadas informações sobre as estratégias de recomendação aplicadas nos trabalhos, bem como as áreas de aplicação, os conjuntos de dados utilizados, as medidas de avaliações aplicadas sobre as estratégias e sobre os resultados obtidos. Por fim, para discutir o terceiro tópico, foram abordados conteúdos sobre os algoritmos usados na geração dos componentes de um ensemble, as medidas de similaridades aplicadas e as funções de consenso aplicadas, com vistas à análise de dados textuais.
- Capítulo 5: nesse capítulo são apresentadas as estratégias de recomendação propostas neste trabalho. Seis estratégias são baseadas nos agrupamentos gerados pelo o algoritmo *k-means++*, sendo duas projetadas para tratar relevância em recomendação, outras duas para tratar o aspecto de surpresa e as últimas duas para o aspecto tratar o aspecto de serendipidade. Também é proposta uma estratégia, visando a serendipidade, cuja base para recomendação é o resultado de agrupamento obtido por *ensemble* de agrupamentos.
- Capítulo 6: nesse capítulo são apresentados os resultados obtidos pelos algoritmos *k-means++* e *ensemble* de agrupamentos. Além disso, são descritos o conjunto de dados utilizado (Corpus EBC), as tarefas de pré-processamento aplicada obre o corpus a fim de gerar matrizes documentos por termos, as configurações de dois experimentos para aplicação do *k-means++* e dos nove experimentos para a aplicação do *ensemble* de agrupamentos. No final desse capítulo são apresentadas as análises sobre os resultados obtidos.

- Capítulo 7: o propósito desse capítulo é apresentar análises realizada sobre estratégias de recomendação propostas neste trabalho. A primeira análise diz respeito a uma reflexão sobre a eficiência das estratégias de recomendação propostas em atender os aspectos de qualidade de recomendação para os quais elas foram projetadas. A segunda análise é feita a partir de um estudo com usuários. Sobre esse estudo, são apresentados o objetivo, o sistema de informação que o suporta, o experimento piloto, os dados da condução do estudo, as análises sobre os resultados obtidos e as considerações finais.
- Capítulo 8: nesse capítulo são apresentadas as conclusões do trabalho. Informações a respeito das limitações do trabalho, dos problemas encontrados e das sugestões de trabalhos futuros são discutidas.

2 Sistemas de recomendação baseados em conteúdo

A internet trouxe uma variedade de vantagens, de facilidades e de conteúdo diversificado para os usuários. Exemplos são as notícias em tempo real disponibilizadas nos portais de notícia, inúmeras músicas e filmes disponibilizados via *streaming*, conteúdos para pesquisas em bibliotecas virtuais, dentre outros itens e funcionalidades. Generalizando, o acesso à rede mundial de computadores proporcionou uma quantidade considerável de informações, além de uma alta disponibilidade para elas.

Ainda que os usuários usufruam das inúmeras vantagens propiciadas pela facilidade de acesso à informação, algumas dificuldades surgiram para o usuário devido à abundância de dados disponíveis. Por exemplo, a dificuldade de um usuário com pouca ou nenhuma experiência para escolher itens úteis e que atendam a suas necessidades dentre várias alternativas apresentadas (CAZELLA; NUNES; REATEGUI, 2010). Embora a disponibilidade de opções aparenta ser um benefício, ela também produz um desconforto, uma vez que os usuários podem tomar decisões ruins e equivocadas (RICCI; ROKACH; SHAPIRA, 2011).

Por meio da análise dos problemas surgidos com a quantidade de informação disponível, uma solução inicial elaborada para contorná-los, segundo Shardanand e Maes (1995), foi utilizar o esforço próprio ou de terceiros com a finalidade de encontrar itens relevantes. Pela mesma ótica, os usuários poderiam também confiar em recomendações realizadas por seus amigos ou por outras pessoas de confiança, tais como revisores de filmes e recomendações provenientes de jornais ou outras mídias de informação.

Embora a solução inicial apresentasse alguns méritos, ela não era totalmente eficiente. Em virtude disso, outra solução foi elaborada, na qual era pretendido um sistema capaz de recomendar itens relevantes automaticamente para o usuário, tendo em vista a sua falta de experiência em realizar escolhas dentre as opções existentes (RESNICK; VARIAN, 1997). Essa solução é comumente conhecida como sistemas de recomendação, que são ferramentas de *software* capazes de fornecer sugestões de itens úteis ao usuário, tais como quais músicas ouvir, quais filmes assistir ou quais notícias ler (RICCI; ROKACH; SHAPIRA, 2011).

Historicamente, os primeiros sistemas de recomendação surgiram na década de 1990 suportados por uma abordagem de filtro colaborativo e, desde então, têm sido um campo importante de pesquisa na academia e na indústria (ADOMAVICIUS; TUZHILIN, 2005). Segundo o trabalho de Resnick e Varian (1997), tais sistemas são definidos da

seguinte forma: a entrada está relacionada com as recomendações fornecidas pelos usuários pertencentes a uma comunidade e, a partir delas, procedimentos de integração (construção de listas de recomendação) eram realizados, de forma que, os resultados eram destinados por meio do sistema aos usuários receptores apropriados. A figura 3 ilustra a arquitetura desses sistemas de recomendação. Na primeira parte da figura é encontrada a entrada, ou seja, as recomendações realizadas por u usuários. Essas recomendações são integradas para formar listas de recomendações para os usuários receptores.

Figura 3 – Arquitetura dos primeiros sistemas de recomendação

Fonte: Fernando Henrique da Silva Costa, 2018

As opiniões dos membros de uma comunidade eram utilizadas pelos sistemas de recomendação para auxiliar outros usuários da mesma comunidade na execução da tarefa de localizar informações ou produtos mais relevantes para eles (KONSTAN, 2004). Contudo, tais comunidades possuíam a limitação de conter interesses divergentes entre seus membros, os quais realizavam recomendações de acordo com seus próprios interesses, gerando recomendações que poderiam não ser úteis a outros membros. Em razão dessa deficiência, foi definido que a comunidade ideal seria aquela que possuísse apenas membros com interesses em comum (CAZELLA; NUNES; REATEGUI, 2010).

Nos anos mais recentes, os sistemas de recomendação se tornarão mais abrangentes. O objetivo desses sistemas é produzir recomendações individualizadas ou que tenha o efeito de orientar o usuário de maneira personalizada em relação aos itens relevantes dentro do espaço variado de opções (BURKE, 2002). Diante da personalização das recomendações, surge a possibilidade dos usuários, ou de um grupo de usuários, receberem sugestões aderentes aos seus interesses.

Em concordância com Ricci, Rokach e Shapira (2011), as recomendações individualizadas são fornecidas mediante uma lista de itens ordenados, a qual é construída como uma tentativa de prever os itens mais adequados ao usuário, tendo em vista suas preferências e restrições. Sítios de *e-commerce*, portais de notícias e serviços de filmes e músicas são exemplos de aplicações de sistemas de recomendação encontrados frequentemente. Na figura 4 é apresentado um exemplo de uma lista de filmes e séries ordenados definidos como escolhas prováveis ao usuário no serviço de filmes e séries via *streaming* Netflix.

Figura 4 – Exemplo de lista de itens ordenados

Fonte: Netflix (<https://www.netflix.com/br/>). Acessado em 24/02/2017

Um cenário comumente utilizado para o emprego de um sistema de recomendação moderno é uma aplicação na internet cujos usuários estarão em constante interação com ela (PAZZANI; BILLSUS, 2007). Um exemplo de tal cenário é um portal de notícias que possui títulos sobre diferentes assuntos (política, economia, esportes, entre outros). Nesse ambiente apresentado, o usuário pode escolher um ou mais títulos para dispor de mais detalhes e para acessar o conteúdo completo. Por consequência, será possível recolher informações sobre as preferências do usuário e, dessa forma, construir seu perfil de interesse (SCHAFFER; KONSTAN; RIEDL, 1999). Com o perfil de interesse de um usuário construído, o sistema de recomendação é capaz de recomendar itens relevantes ao usuário em questão.

A figura 5 destaca o cenário atual de um sistema de recomendação. Na primeira parte da figura, o usuário possui seu perfil de interesse e está interagindo com a aplicação. O sistema de recomendação pertencente à aplicação recebe o perfil de interesse por meio da interação realizada. Por conseguinte, o sistema analisa o perfil de interesse, verifica os itens relevantes de acordo com esse perfil (por exemplo, cálculo de similaridade entre os itens que o usuário gostou e itens não vistos por ele) e cria uma lista de recomendação com tais itens (por exemplo, um *ranking* dos l itens mais relevantes). Por fim, a lista de recomendação é apresentada pela aplicação para o usuário.

Figura 5 – Cenário de um sistema de recomendação moderno

Fonte: Fernando Henrique da Silva Costa, 2018

A criação de um perfil de interesse é realizada geralmente de duas maneiras: explícita ou implícita. A criação de forma explícita utiliza as estimativas dos usuários, ou seja, o apontamento deles a respeito dos tópicos sobre os quais ele deseja receber recomendações. A título de exemplo, considere um usuário que explicita o desejo de receber apenas recomendações de filmes de determinados gêneros, como comédia e ação. Já na maneira implícita, há um trabalho do sistema em analisar o comportamento dos usuários sem a informação prévia de tópicos desejados. O comportamento dos usuários pode ser analisado a partir das características das músicas que ele ouve, dos filmes que ele assiste ou das notícias que ele lê.

Segundo Meteren e Someren (2000), os sistemas de recomendação são definidos como um tipo especial de sistema de filtragem de informações, podendo também ser vistos como uma tarefa de classificação. Esses autores ainda afirmam que um perfil de usuário pode ser induzido por meio de dados de treinamento. Tais dados dizem respeito a itens avaliados no passado e permitem a classificação de itens ainda não visualizados pelo usuário em duas classes: a positiva, que representa itens relevantes para o usuário; e a negativa, que representa itens não relevantes para o usuário.

Dentro da literatura de sistemas de recomendação existem três arquiteturas principais. Essas são classificadas em sistemas de recomendação baseados em conteúdo, filtro colaborativo e conhecimento. As arquiteturas e uma breve descrição a respeito delas são apresentadas por Adomavicius e Tuzhilin (2005). Segue um resumo do significado de cada uma:

- **conteúdo:** nessa arquitetura, a recomendação de novos itens é formulada com base na similaridade entre itens que o usuário tenha gostado no passado e os itens

disponíveis para recomendação. O usuário recebe recomendações de novos itens com base na análise de similaridade desses itens com aqueles que ele já tenha gostado ou se interessado no passado;

- **filtros colaborativos:** semelhante aos primeiros sistemas de recomendação, nessa arquitetura os itens recomendados a um usuário são aqueles que outros usuários, com interesses similares ao primeiro, tenham gostado ou se interessado no passado;
- **conhecimento:** essa arquitetura usa um determinado conhecimento de domínio para construir suas recomendações. Devido a isso, há a necessidade de procedimentos de aquisição do conhecimento em questão. Por exemplo, essa abordagem consegue reconhecer que frutos do mar não é uma recomendação adequada a pessoas vegetarianas a partir da análise de informações obtidas no domínio de cozinha, de alimento e do perfil do usuário.

Dentre as arquiteturas de recomendação já discutidas neste texto, a arquitetura utilizada para esta pesquisa é a baseada em conteúdo. O seu objetivo é recomendar itens relevantes com base na similaridade destes com aqueles já visualizados e avaliados positivamente pelo usuário alvo. Segundo Aggarwal (2016), os sistemas de recomendação baseados em conteúdo são concentrados, principalmente, nas avaliações do próprio usuário alvo, aquele que receberá as recomendações, e nas características dos itens apreciados por ele.

O processo de recomendação nessa arquitetura é iniciado a partir de um conjunto de itens avaliados por um usuário e as características desses itens são utilizadas na criação do perfil de interesse desse usuário (LOPS; GEMMIS; SEMERARO, 2011). Após essa etapa, o sistema de recomendação será capaz de explorar o perfil de interesse com a finalidade recomendar itens novos ao usuário correspondente.

A figura 6 ilustra como um perfil de interesse de um usuário é construído. Inicialmente, o usuário interage com n itens e, por meio deles, formam um conjunto. Por fim, as características de cada item desse conjunto são utilizadas para a criação do perfil de interesse do usuário.

Na área de sistemas de recomendação baseados em conteúdo há dois sub problemas: (a) encontrar uma representação significativa dos itens; e (b) criar um perfil que seja capaz de receber recomendações de itens não vistos (PAZZANI, 1999). Por exemplo, no contexto deste trabalho, uma solução para o sub problema (a) é representar as notícias

Figura 6 – Construção de um perfil de interesse

Fonte: Fernando Henrique da Silva Costa, 2018

por meio de uma representação vetorial, em que cada vetor representa uma notícia e cada palavra dessa notícia corresponde a uma coordenada do vetor. Já para o sub problema (b), uma possibilidade de solução é a criação de um perfil correspondente ao armazenamento dos itens que o usuário visualizou e avaliou positivamente. Ademais, esse perfil pode ser enriquecido com informações referentes a interação do usuário com o sistema, como por exemplo, quais foram as notícias acessadas pelo o usuário, ainda que não tenham sido avaliadas.

O aprendizado em um sistema de recomendação baseado em conteúdo ocorre por meio das características explicitadas nos itens em que os usuários tiveram interesse e atribuíram suas avaliações (BURKE, 2002). Como exemplo, considere que um usuário tenha classificado positivamente diversos filmes de comédia. Dessa maneira, o sistema é capaz de aprender que tal gênero de filme é algo relevante a esse usuário e, portanto, recomendará filmes de comédia ainda não vistos por ele.

2.1 Vantagens

Os sistemas de recomendação de arquitetura baseada em conteúdo possuem algumas vantagens quando comparados à arquitetura baseada em filtros colaborativos. Essas vantagens e suas devidas explicações são discutidas no trabalho de Lops, Gemmis e Semeraro (2011), e seguem resumidas aqui:

- **independência de outros usuários:** para construir o perfil de interesses e recomendar itens não vistos e relevantes a um usuário, os sistemas de recomendação baseados em conteúdo necessitam somente das avaliações sobre os itens vistos por esse usuário. Essa maneira de recomendar não é observada na arquitetura de filtros colaborativos, uma vez que esses sistemas necessitam de avaliações de outros usuários para encontrar similaridade entre eles e o usuário alvo e, assim, serem capazes de fazer as recomendações;
- **transparência:** essa vantagem é alusiva a existência de uma explicação bem definida sobre o funcionamento do sistema de recomendação. Tal explicação se manifesta na lista de itens recomendados, a qual também contém as descrições que possibilitaram a sua construção com esses itens;
- **novo item:** há sistemas de recomendação que atualizam os seus conjuntos de itens a partir da inserção de novos e remoção dos mais antigos. Quando os novos itens são introduzidos, eles não possuem avaliação realizada pelos usuários do sistema. Entretanto, a arquitetura baseada em conteúdo não necessita de tais avaliações, uma vez que ela realiza as recomendações com base nas características dos itens. Dessa forma, essa arquitetura não sofre do problema conhecido como “primeira avaliação” ou *cold start* de item.

2.2 Desvantagens

Ainda que existam determinadas vantagens na arquitetura baseada em conteúdo, é importante salientar que ela também possui certas desvantagens. Em conformidade com o que defende Lops, Gemmis e Semeraro (2011), as desvantagens e suas devidas explicações são:

- **análise limitada de características:** há um limite natural na quantidade e no tipo de características associados aos itens de recomendação. O conhecimento do domínio é comumente necessário, em que o domínio de filmes é um exemplo claro. Nesse domínio há a necessidade de conhecer os autores e diretores envolvidos em um filme, o seu gênero, entre outras ontologias. Todavia, algumas representações utilizam somente algumas dessas características para representar o conteúdo como

um todo, o que pode acarretar em falta de informações suficientes para encontrar itens novos, os quais os usuários poderão se interessar ou não;

- **novo usuário:** uma vez que a arquitetura baseada em conteúdo necessita coletar as avaliações dos usuários alvos sobre os itens, a fim de construir seu perfil, ela tende ser inadequada para recomendar itens para os usuários ingressantes, justamente por não possuir informações a respeito de seus interesses. Essa limitação é denominada de *cold start* de usuário;
- **problema da serendipidade:** esse problema está fortemente ligado ao problema da super especialização (do inglês: *over-specialization*) de recomendações. Essa desvantagem é comum na abordagem baseada em conteúdo, a qual não possui recursos para encontrar algo novo e surpreende ao usuário alvo. O sistema é capaz de recomendar apenas itens semelhantes àqueles que o usuário avaliou anteriormente.

2.3 Qualidade de recomendações

De acordo com Herlocker et al. (2004), a relevância em sistemas de recomendação está baseada na probabilidade dos itens recomendados serem de interesse do usuário, ou seja, aqueles que são adequados a sua preferência. Indo em conformidade com essa afirmação, os primeiros sistemas de recomendação buscavam alcançar a relevância por meio da alta acurácia das recomendações.

É importante salientar que a acurácia é um aspecto bastante discutido dentro do campo de sistemas de recomendação e, além disso, a maioria desses sistemas possui um mecanismo de previsão de itens acurados para os usuários (GUNAWARDANA; SHANI, 2015). A fim de realizar esse propósito, o mais comum é a utilização da similaridade entre usuários, itens ou ambos (ZHOU et al., 2010). Por exemplo, dada a similaridade entre os itens visualizados e apreciados pelo usuário e os que ainda não foram vistos por ele, o sistema pode recomendar aqueles que são mais semelhantes com base em seu histórico.

Ainda que as recomendações acuradas sejam desejáveis e importantes, elas não são suficientes (WU; HE; YANG, 2012), uma vez que elas podem ocasionar o problema da baixa serendipidade. A qualidade de uma recomendação pode ser expressa em termos do quão ela supera a expectativa do usuário (GE; DELGADO-BATTENFELD; JANNACH,

2010). Diante desse contexto, novos aspectos de avaliação de sistema de recomendação surgem, como por exemplo, a surpresa e a serendipidade.

Murakami, Mori e Orihara (2007) explicam que a ideia básica de surpresa em sistemas de recomendação pode ser considerada como o desvio de um resultado obtido por um método de predição primitivo, o qual recomenda itens com alta acurácia. Em conformidade com essa ideia, a fim de serem considerados surpreendentes, os itens recomendados devem ser diferentes daqueles que o usuário já demonstrou interesse e devem ser recomendados de forma inesperada. No entanto, as recomendações surpreendentes sofrem o mesmo que as recomendações acuradas, uma vez que somente a surpresa não é o bastante para ser adequada ao usuário (ADAMOPOULOS; TUZHILIN, 2011). Isso ocorre, pois itens bastante diferentes do interesse do usuário podem não ser úteis a ele.

O aspecto de serendipidade surge a fim de suprir a insuficiência de recomendação relacionadas aos dois aspectos anteriores (acurácia e surpresa). O termo serendipidade foi cunhado em 1754 por Horace Walpole com base no conto *“Os Três Príncipes de Serendip”* cuja história relata as viagens de três príncipes que descobriam constantemente novas coisas por acidente e sagacidade (BOSENMAN, 1988). Além disso, esse aspecto pode ser encontrado em diversos casos na história da ciência, como por exemplo, a descoberta da penicilina por Alexander Fleming (ANDRÉ et al., 2009).

No contexto de sistemas de recomendação, a serendipidade está relacionada com a qualidade das recomendações, sendo que em grande parte, ela depende da subjetividade do usuário e, por consequência, um conceito complexo de compreensão (IAQUINTA et al., 2008). No trabalho de Ge, Delgado-Battenfeld e Jannach (2010), os autores destacam três definições de serendipidade em sistemas de recomendação encontradas na literatura da área, sendo elas:

- a serendipidade é definida como uma medida da extensão com que os itens recomendados são atraentes e surpreendentes ao usuário. Isso significa que recomendações com alta serendipidade auxiliam os usuários a encontrar itens surpreendentes e interessantes. Por meio dessa definição, é possível notar dois fatores importantes relacionados com a serendipidade. Primeiro, um item serendipitoso não deve ser de conhecimento do usuário e não ser esperado como recomendação pelo mesmo. Em segundo lugar, o item deve conter relevância ao usuário, ou seja, ser útil para ele;

- semelhante à primeira definição, a serendipidade é definida como uma medida que avalia o quão surpreendentes e bem sucedidas as recomendações são;
- a serendipidade é definida como a experiência de um usuário que recebeu uma recomendação inesperada e fortuita.

2.4 Considerações finais

O propósito deste capítulo foi apresentar conceitos sobre sistemas de recomendação, incluindo as arquiteturas de recomendação. Também foram apresentadas as vantagens e desvantagens da arquitetura baseada em conteúdo, principal interesse deste trabalho, além de noções básicas sobre medidas de qualidade de uma recomendação. Tendo em vista o problema da baixa serendipidade em recomendações e a proposta de tratá-lo usando técnicas de agrupamento, o próximo capítulo apresenta conteúdos a respeito de agrupamento e *ensemble* de agrupamentos.

3 Ensemble de agrupamento

Diversos conjuntos de dados com grande volume de dados e alta dimensionalidade têm sido criados devido aos avanços das tecnologias de produção e armazenamento de dados e, também, ao crescimento do uso de aplicações (JAIN, 2010). Consequentemente, surge a necessidade de melhorar os métodos de análise que permitem processar, compreender e sumarizar automaticamente os dados. Os algoritmos de aprendizado de máquina têm sido utilizados nesse contexto visando melhorar as análises sobre esses dados.

Segundo Michalski, Carbonell e Mitchell (2013), aprendizado é descrito como o processo que inclui a aquisição de novos conhecimentos, o desenvolvimento de habilidades motoras e cognitivas por meio de instruções ou práticas, a organização de conhecimento em geral, as representações efetivas e a descoberta de novos fatos e teorias mediante observações e experimentações. O estudo e a modelagem computacional de processos de aprendizado constitui o contexto de aprendizado de máquinas cuja definição é destacada em Mitchell (1997) como:

“Um programa de computador aprende a partir de um elenco de experiências E , relacionadas a uma classe de tarefas T e dispondo de uma medida de desempenho $\mathcal{M}$, se seu desempenho medido por $\mathcal{M}$ junto à tarefa T melhora com o elenco de experiências E . ”

Existem diversos paradigmas de aprendizado dentro da área de aprendizado de máquina, dentre eles, os paradigmas de aprendizado supervisionado e aprendizado não supervisionado. De forma resumida, a diferença entre tais paradigmas está relacionada a existência ou não de dados rotulados. No aprendizado supervisionado, uma coleção de dados rotulados (pré classificados) é fornecida, enquanto no aprendizado não supervisionado, a coleção de dados fornecida não possuem rótulos (JAIN; MURTY; FLYNN, 1999).

Em conformidade com Nilsson (1996), o objetivo do aprendizado supervisionado é encontrar uma função hipotética que seja capaz de prever, em grande parte, a classe dos dados. Para alcançar tal fim, ele utiliza o resultado obtido pela função e compara esse resultado com os rótulos dos dados. Já o objetivo do aprendizado não supervisionado é atribuir rótulos simbólicos¹ aos dados, todavia, ele apenas utiliza as características dos dados, sem ter conhecimento da classe real dos dados em seu processo.

¹ Nesse contexto, simbólico significa pouca representatividade.

O agrupamento é definido como uma metodologia importante na análise exploratória de dados, sendo possível encontrar diversos métodos de agrupamento reportados na literatura (WU; LEAHY, 1993). Seu objetivo principal é encontrar os padrões gerais de distribuição do conjunto de dados e disponibilizar uma visualização mais eficiente por meio de grupos derivados do conjunto original (ZHANG; RAMAKRISHNAN; LIVNY, 1996).

A figura 7 ilustra o objetivo da tarefa de agrupamento, o qual está representado pela descoberta de grupos divididos em cores. O lado esquerdo representa um conjunto de nove dados X e o lado direito apresenta os três grupos formados: x_1, x_2 e x_3 coloridos em azul; x_4, x_5 e x_6 colorido em verde; e x_7, x_8 e x_9 coloridos em amarelo.

Figura 7 – Exemplo de agrupamento de dados

Fonte: Fernando Henrique da Silva Costa, 2018

A principal função da tarefa de agrupamento é resumir os dados, oferecendo um meio para produzir uma descrição sobre eles. A partir disso, torna-se possível proceder com tarefas de nomeação (todos os objetos de um mesmo grupo recebem o mesmo nome ou rótulo), exibição da organização presente nos dados dos dados, e predição. (HARTIGAN, 1975). Assim, os algoritmos de agrupamento organizam objetos em grupos, de forma que os objetos pertencentes ao mesmo grupo sejam mais semelhantes entre si quando comparados com aqueles que estão organizados em grupos diferentes (IAM-ON et al., 2011).

Segundo Peres et al. (2012), o agrupamento estuda as relações de similaridade entre os dados, de forma que determine quais dados formam quais grupos. Estes são formados visando maximizar a similaridade entre os dados de um grupo (similaridade intra-grupo) e minimizar a similaridade entre dados de grupos diferentes (similaridade inter-grupos). Formalmente, dado o conjunto de dados de entrada X , situado no espaço $\mathbb{R}^m$ ($X \in \mathbb{R}^m$) de m dimensões, é esperado encontrar uma função como apresentada na equação 1, em que W é um vetor de parâmetros ajustáveis por meio de um algoritmo de agrupamento que determina k grupos no conjunto de dados X , $G = \{G_1, \dots, G_k\}$, sendo $k \leq n$ e n o número de dados.

$$F : \mathbb{R}^m \times W \rightarrow G \quad (1)$$

Em Fern e Lin (2008), os autores discutem que algoritmos de agrupamento desenvolvidos com base em diversas medidas de similaridade ou de distância, mesmo quando aplicados aos mesmos dados, produzem resultados que são diferentes entre si. Além disso, esses algoritmos possuem a deficiência em conceder soluções satisfatórias em dados com alta dimensionalidade (JING; TIAN; HUANG, 2015). Diante destes fatos, motiva-se a elaboração de *ensembles* de agrupamento, com o intuito de melhorar os resultados obtidos para tarefas de agrupamento.

Um *ensemble* é definido como um “multi aprendiz” simples e efetivo. A integração de resultados obtidos por componentes, que representam instâncias diferentes de solução para um problema, em apenas um resultado, melhora a acurácia e a robustez da solução global para uma tarefa (STREHL; GHOSH, 2002). Tal integração de resultados é observado em uma pessoa que busca opiniões de terceiros antes de tomar uma decisão importante. As opiniões individuais são ponderadas e depois combinadas chegando a uma decisão final (POLIKAR, 2006).

A figura 8 ilustra um exemplo de um *ensemble*. Inicialmente, um conjunto de entrada é apresentado para os p componentes do *ensemble*, os quais podem ser classificadores, regressores (LIMA, 2004) ou algoritmos de agrupamento (NALDI; FACELI; CARVALHO, 2009), a depender da tarefa sob resolução. Cada um dos componentes gera uma saída diferente. Essas saídas, então, são combinadas por uma estratégia a fim de produzir uma saída final.

Figura 8 – Exemplo de um *ensemble*

Fonte: Fernando Henrique da Silva Costa, 2018

Dois passos distintos, porém correlacionados, são necessários para o desenvolvimento de um *ensemble*, tanto para as tarefas de aprendizado supervisionado quanto para aprendizado não supervisionado. O primeiro passo diz respeito a geração dos componentes, os quais devem ser precisos e diversificados (ZHOU, 2015), e o segundo passo equivale em uma estratégia de combinação das soluções produzidas.

A técnica utilizada nesta pesquisa é o *ensemble* de agrupamento. De acordo com Fern e Brodley (2004), esse paradigma tem como objetivo combinar as soluções de agrupamento de cada componente sobre um determinado conjunto de dados com o propósito de melhorar o desempenho da tarefa proposta e apresentar uma solução geral.

O desenvolvimento de *ensemble* de agrupamentos segue os mesmos estágios que o desenvolvimento de *ensemble* de aprendizado supervisionado, em que há a necessidade de seguir geralmente duas etapas. Os passos a serem seguidos, de acordo com Liu, Liao e Liao (2015), são:

1. a geração de um conjunto de partições de agrupamentos, ou a geração de componentes do *ensemble*, por meio dos algoritmos de agrupamento individuais (diferentes algoritmos ou o mesmo algoritmos com diferentes parâmetros de inicialização);
2. a combinação de tal conjunto por meio de uma função de consenso, a fim de conseguir um resultado final superior.

Com o intuito de definir um *ensemble* de agrupamento, foi utilizada a notação com base no trabalho de Strehl e Ghosh (2002). Tal definição é apresentada da seguinte forma: dado um conjunto de dados $X = \{x_1, x_2, \dots, x_n\}$, um particionamento desses n dados em k grupos é representado por uma matriz de rótulos $\lambda \in \mathbb{N}^n$. Um algoritmo de agrupamento representado por Φ fornece, como saída, uma matriz de rótulos dado as tuplas de dados. A figura 9 ilustra essa definição, considerando λ_{final} a matriz de rótulos gerada pelo *ensemble* de agrupamentos.

3.1 Geração de membros

O primeiro passo de um *ensemble* de agrupamentos visa gerar p partições iniciais, em que cada partição é gerada por um dos p componentes. Posteriormente, essas partições serão integradas por uma estratégia a fim de produzir a partição final. Para isso, é importante ter conhecimento do objetivo da combinação e conhecer bem os algoritmos

Figura 9 – Exemplo de um *ensemble* de agrupamentos

Fonte: Fernando Henrique da Silva Costa, 2018

a serem combinados com a finalidade de garantir que os algoritmos escolhidos possam contribuir para esse objetivo (NALDI; FACELI; CARVALHO, 2009).

Ainda de acordo com Naldi, Faceli e Carvalho (2009), a diversidade entre as partições iniciais é necessária, pois a geração de partições semelhantes por métodos distintos tende a refletir a presença de uma forte estrutura nos dados agrupados. Além disso, os autores citam quatro formas de obter a diversidade, sendo elas:

- utilizar diferentes algoritmos de agrupamentos para gerar as partições iniciais;
- utilizar um mesmo algoritmo de agrupamento, no entanto com diferentes parâmetros de inicialização ou ordem de apresentação dos dados;
- empregar algoritmos de agrupamentos mais simples com o objetivo de diminuir o custo computacional do processo;
- utilizar diferentes conjuntos de dados, os quais podem ser formados por seleção de alguns atributos dos dados do conjunto original, projeção do conjunto original em um espaço de dimensão menor, amostragem e inserção de ruído.

O trabalho de Kuncheva e Hadjitodorov (2004) apresenta formas de calcular a diversidade entre partições. Esse cálculo é baseado em índices de critérios externos, como Índice Rand Ajustado e Informação Mútua Normalizada (apresentados na seção 3.7). Em ambos os índices, quando as partições são idênticas, o resultado será 1, porém quando as partições são totalmente diferentes não há valor constante para o primeiro índice e será igual a 0 para o segundo índice.

Quatro funções para calcular a diversidade propostas e apresentadas no trabalho de Naldi, Faceli e Carvalho (2009) são derivadas a partir do cálculo do Índice Rand Ajustado. Considerado $1 - ARI(\lambda_i, \lambda_j)$ a diversidade entre λ_i e λ_j , as quatro funções são:

- D_1 : cálculo da média entre todas partições iniciais e partição final;
- D_2 : cálculo do desvio padrão entre todas partições iniciais e partição final;
- D_3 : cálculo derivado das duas anteriores, de forma que, $D_3 = \frac{1}{2}(1 - D_1 + D_2)$;
- D_4 : também é derivado das funções D_1 e D_2 , de forma que, $D_4 = \frac{D_2}{D_1}$.

3.2 Função de consenso

O segundo passo de um *ensemble* de agrupamentos visa integrar as soluções geradas pelos componentes por meio de uma estratégia denominada de função de consenso. De acordo com Topchy et al. (2004), tal função é derivada dos princípios da estatística, além de ser fundamentada em grafos e na teoria da informação. Ainda no mesmo trabalho, os autores citam que as funções de consenso mais populares são baseadas em: matriz de coassociação; corte de hipergrafos; informação mútua; mistura de modelos; e votação.

A estratégia de matriz de coassociação, baseada no trabalho de Naldi, Faceli e Carvalho (2009), será empregada neste trabalho. Ela é iniciada pela criação de uma matriz de similaridade para cada um dos p componentes. Essa matriz é binária e simétrica de tamanho $n \times n$, em que cada linha e cada coluna representa um dado do conjunto X . Desta forma, caso x_i e x_j , sendo $i \neq j$, estejam agrupados em um mesmo grupo, então $M_{ij}^a = 1$, caso contrário, $M_{ij}^a = 0$. A tabela 1 apresenta um exemplo de uma matriz de similaridade, em que os dados x_1 e x_2 estão agrupados juntos, ao passo que, x_3 e x_n estão agrupados juntos, porém em um grupo diferente de x_1 e x_2 .

Tabela 1 – Matriz de similaridade

M^a	x_1	x_2	x_3	...	x_n
x_1	1	1	0	...	0
x_2	1	1	0	...	0
x_3	0	0	1	...	1
...	...	...	...	...	...
x_n	0	0	1	...	1

Fonte: Fernando Henrique da Silva Costa, 2018

Dessa forma, cada solução gerada por um componente será denominada de M^a , em que a é o índice do componente. O próximo passo é gerar uma matriz denominada matriz de coassociação. Essa matriz é gerada de acordo com a equação 2 em que MC é a matriz de coassociação e p é a quantidade de componentes.

$$MC = \frac{M^1 + M^2 + \dots + M^p}{p} \quad (2)$$

De acordo com o trabalho de Naldi, Faceli e Carvalho (2009), dado quatro componentes de um *ensemble* de agrupamentos (tabela 2), a matriz de coassociação pode ser gerada conforme a equação 2, resultando na matriz representada pela tabela 3.

Tabela 2 – Matrizes de similaridade de quatro componentes

λ_1	x_1	x_2	x_3	x_4	x_5
x_1	1	1	1	0	0
x_2	1	1	1	0	0
x_3	1	1	1	0	0
x_4	0	0	0	1	1
x_5	0	0	0	1	1

λ_2	x_1	x_2	x_3	x_4	x_5
x_1	1	1	0	0	0
x_2	1	1	0	0	0
x_3	0	0	1	1	1
x_4	0	0	1	1	1
x_5	0	0	1	1	1

λ_3	x_1	x_2	x_3	x_4	x_5
x_1	1	1	0	1	0
x_2	1	1	0	1	0
x_3	0	0	1	0	1
x_4	0	0	0	1	0
x_5	0	0	1	0	1

λ_4	x_1	x_2	x_3	x_4	x_5
x_1	1	1	0	0	1
x_2	1	1	0	0	1
x_3	0	0	1	1	0
x_4	0	0	1	1	0
x_5	1	1	0	0	1

Fonte: Fernando Henrique da Silva Costa, 2018

Tabela 3 – Matriz de coassociação

MC	x_1	x_2	x_3	x_4	x_5
x_1	1	1	0,25	0,25	0,25
x_2	1	1	0,25	0,25	0,25
x_3	0,25	0,25	1	0,5	0,5
x_4	0,25	0,25	0,5	1	0,5
x_5	0,25	0,25	0,5	0,5	1

Fonte: Fernando Henrique da Silva Costa, 2018

Por fim, uma forma de gerar o resultado de agrupamento final é por meio da aplicação de um algoritmo de agrupamento na matriz de coassociação, considerando-a como uma matriz de dados (NALDI; FACELI; CARVALHO, 2009). Neste trabalho, dois algoritmos de agrupamento foram utilizados para essa tarefa, sendo eles: hierárquico aglomerativo (AGNES) e *k-means++*.

3.3 Métrica de similaridade

Um requisito fundamental em algoritmos de agrupamento é a similaridade entre os dados. Nesse contexto, uma medida de similaridade entre dois dados dentro do mesmo espaço é essencial para uma classe dos procedimentos de agrupamento (JAIN; MURTY; FLYNN, 1999). A importância dessa medida está relacionada a construção de uma superfície de decisão que seja capaz de determinar a organização dos dados nos grupos.

De forma geral, as medidas de distância mapeiam a similaridade entre dois dados para um valor numérico, o qual depende de dois fatores: as propriedades dos dados e a métrica propriamente dita (HUANG, 2008). Em relação à esta pesquisa, a medida de distância aplicada é a cosseno. A equação 3 formaliza o cálculo dessa distância, em que $\cos\theta$ é o ângulo formado entre os vetores x_i e x_j (DEZA; DEZA, 2009)². Essa distância é uma métrica de similaridade bastante utilizada para o agrupamento de documentos, ou seja, conteúdo textual (MUFLIKHAH; BAHARUDIN, 2009).

$$dist_{cos}(x_i, x_j) = 1 - \cos\theta = 1 - \frac{\langle x_i, x_j \rangle}{\|x_i\|_2 \cdot \|x_j\|_2} \quad (3)$$

3.4 Métodos de agrupamento

Uma característica importante para a resolução da tarefa de agrupamento é definir qual a classe de algoritmos que deverá ser utilizada. Segundo Han, Pei e Kamber (2011), alguns métodos de algoritmos de agrupamento são: particionamento, hierárquico, baseado em densidade, baseado em grade (do inglês: *grid-based*) e baseado em modelos. Os métodos que estão sendo usados no contexto desta pesquisa são agrupamento por particionamento e hierárquico.

O método de particionamento é definido da seguinte forma: dado n tuplas de dados de entrada não rotulados, serão construídos k grupos formando uma partição (um dado pertence apenas a um grupo e um grupo tenha ao menos um dado) que respeite $k \leq n$ (LIAO, 2005). Além disso, ele apresenta uma solução de agrupamento por meio da maximização da similaridade entre os objetos dentro de um grupo e da minimização das dissimilaridades entre objetos de grupos diferentes (PATERLINI; KRINK, 2006).

² Em Deza e Deza (2009) é apresentada uma discussão aprofundada sobre distâncias e tanto similaridade cosseno como a visão de distância cosseno são explicadas em detalhes.

É de conhecimento que o método de agrupamento por particionamento é adequado para agrupar um conjunto de dados grande, devido à sua baixa necessidade computacional e à complexidade temporal ser quase linear (CUI; POTOK; PALATHINGAL, 2005). O algoritmo *k-means* e suas respectivas variações são exemplos de algoritmos de agrupamento populares que empregam o método de agrupamento por particionamento.

De acordo com Fraley e Raftery (1998), o método hierárquico procede por etapas produzindo uma sequência de partições, cada uma correspondendo a um número diferente de grupos. Esse método pode ser aglomerativo, o qual está no escopo desta pesquisa, ou divisivo. De acordo com Zhao e Karypis (2002), o método hierárquico aglomerativo atribui inicialmente cada dado do conjunto a um grupo, em seguida, seleciona e combina repetidamente pares de grupos até obter um único grupo ou k grupos desejados. Uma estratégia para realizar os últimos dois passos do método é selecionar o par de grupos mais similar e combinar tais grupos em um.

3.5 Algoritmos *k-means* e *k-means++*

O algoritmo *k-means*, além de ser um dos mais famosos em agrupamentos por particionamento, também é simples, computacionalmente eficiente e não exige que o usuário especifique vários parâmetros (FRED; JAIN, 2002). De acordo com Hartigan e Wong (1979), o objetivo do *k-means* é dividir n dados com m dimensões em k grupos, de forma que a soma dos quadrados intra-grupo seja minimizada.

Formalmente, considerando um conjunto de dados $X = \{x_i\}$, com $i = 1, \dots, n$, a ser agrupado em k grupos $G = \{g_j\}$, $j = 1, \dots, k$, o algoritmo busca por uma partição para X que minimize o problema dado pela equação 4. Nela, g_j é um determinado centroide, $U = \{0, 1\}^{k,n}$ é uma matriz indicadora para associação dos n dados aos k grupos e $dist$ é calculado pela distância cosseno apresentada na seção 3.3.

$$P(U, G) = \sum_{j=1}^k \sum_{i=1}^n U_{ji} dist(g_j, x_i), \quad (4)$$

Com a finalidade de realizar o agrupamento por meio do algoritmo *k-means*, há a necessidade de seguir, usualmente, três etapas, as quais são definidas pelo o algoritmo 2.

A figura 10 ilustra os passos do *k-means* intuitivamente. Inicia-se no plano a , que contém os dados de entrada provenientes do conjunto de dados. A seguir, no plano b ,

Algoritmo 2 Algoritmo *k-means*

- 1: **procedure** K-MEANS(X, k)
 - 2: Selecionar uma partição inicial com k centroides aleatórios
 - 3: **while** Repetir até os centroides estabilizarem **do**
 - 4: Atualizar cada centroide calculando a distância média dos dados associados a seu grupo;
 - 5: Gerar uma nova partição associando cada dado ao centroide mais próximo a ele
-

Fonte: Traduzido de Arthur e Vassilvitskii (2007)

três centroides (círculos) são inicializados aleatoriamente e os dados mais próximos a eles formam os três grupos. No plano *c*, o centroide foi atualizado de acordo com o que havia no plano *b* e os dados foram reassociados aos centroides. No plano *d*, uma outra atualização de posição dos centroides ocorre. Por fim, no plano *e*, como não há mais atualização dos centroides, a partição final é obtida e fornecida como resultado para a tarefa de agrupamento.

Figura 10 – Passos do algoritmo *k-means*

Fonte: Jain (2010)

O algoritmo *k-means* possui algumas variações, dentre elas, Jain (2010) cita: ISO-DATA, *fuzzy k-means*, *bisecting k-means*, entre outras. Para esta pesquisa, a variação utilizada foi o algoritmo *k-means++* (ARTHUR; VASSILVITSKII, 2007).

No algoritmo *k-means++*, a inicialização dos centroides é da seguinte forma: inicialmente, escolhe-se um centroide aleatoriamente, na sequência os demais centroides serão escolhidos seguindo uma probabilidade proporcional à sua contribuição ao erro global dado as seleções anteriores (BAHMANI et al., 2012).

Após as escolhas do centroides, o algoritmo *k-means++* segue as mesmas etapas que o *k-means* segue. Em conformidade com Arthur e Vassilvitskii (2007), em que $D(x)$ é a menor distância de um dado x para o centroide já escolhido, as etapas do algoritmo *k-means++* são:

Algoritmo 3 Algoritmo *k-means++*

- 1: Escolher aleatoriamente o centróide g_1
 - 2: $i \leftarrow 2$
 - 3: **while** $i \neq k$ **do**
 - 4: Selecione o próximo centróide g_i , selecionando $x' \in X$ com a probabilidade:

$$\frac{D(x')}{\sum_{x \in X} D(x)^2}$$
 - 5: Proceda como no algoritmo *k-means*
-

Fonte: Traduzido de Arthur e Vassilvitskii (2007)

3.6 Algoritmo de agrupamento hierárquico aglomerativo

Conforme Zhao, Karypis e Fayyad (2005), as soluções de agrupamento hierárquico são representadas na forma de árvores dendrogramas. Essas árvores hierárquicas fornecem uma visão dos dados em diferentes níveis de abstração.

Segundo Linden (2009), os algoritmos de agrupamento hierárquico criam uma hierarquia de relacionamento entre os dados e são dividido em duas versões: aglomerativo e divisivo. Em conformidade com Mingoti e Lima (2006), os algoritmos hierárquicos aglomerativos trabalham da seguinte maneira: no primeiro estágio, cada um dos n dados é considerado como um grupo. Os grupos são comparados entre si por uma medida de distância e o par que possuir a menor distância são combinados. Esse último procedimento é repetido diversas vezes até o número desejável de grupos for alcançado. O algoritmo 4 apresenta o procedimento básico do algoritmo de agrupamento hierárquico aglomerativo, sendo baseado no trabalho de Steinbach et al. (2000).

Um método de *linkage* é usado para comparar os grupos em cada etapa e decidir quais deles devem ser combinados. Para esta pesquisa, os métodos de *linkage* utilizados são: *single-link* e *average-link*. No primeiro método citado, a distância entre dois grupos é calculada de acordo com o par de dados mais próximos entre os dois grupos (SEIFODDINI, 1989). A equação 5 formula o procedimento do método *single-link* em que: g_a e g_b são os grupos; x_i e x_j são os i -ésimo e j -ésimo dados dos grupos g_a e g_b , respectivamente; e n_a e n_b são as quantidades de dados dos grupos g_a e g_b , respectivamente.

Algoritmo 4 Algoritmo hierárquico aglomerativo

- 1: **procedure** AGNES(X, k)
- 2: Computar a similaridade entre todos os pares de grupos, i.e., calcular uma matriz de similaridade em que M_{ij} possui o valor da distância entre o grupo i e j .
- 3: **while** Repetir até restar apenas os k grupos **do**
- 4: Fundir os dois grupos mais similares (mais próximos).
- 5: Atualizar a matriz de similaridade para refletir a similaridade do novo grupo gerado com os outros grupos.

 Fonte: Traduzido de Steinbach et al. (2000)

$$d(g_a, g_b) = \min(dist(x_i, x_j)), i \in (1, \dots, n_a), j \in (1, \dots, n_b) \quad (5)$$

O procedimento realizado no método *average-link* é definido pelo cálculo da similaridade entre dois grupos por meio da média de distância de todos os pares de dados desses grupos (SEIFODDINI, 1988). A equação 6 apresenta como é realizado o cálculo desse procedimento em que as notações são idênticas à equação 5.

$$d(g_a, g_b) = \frac{1}{n_a n_b} \sum_{i=1}^{n_a} \sum_{j=1}^{n_b} dist(x_i, x_j) \quad (6)$$

3.7 Índices de avaliação de agrupamento

A avaliação dos resultados obtidos por um algoritmo de agrupamento é fundamental na análise de agrupamento. Esse procedimento é parte integral do processo de agrupamento, em que é necessário não apenas comparar partições uma com a outra, como também determinar se alguma delas é suficientemente boa (LEWIS; ACKERMAN; SA, 2012).

Geralmente, para realizar a avaliação dos resultados, são utilizados índices externos e internos. Os índices externos só podem ser aplicados quando há o conhecimento *a priori* do rótulo de classe para cada dado (ROSENBERG; HIRSCHBERG, 2007). Já os índices internos são aplicados sem o conhecimento dos rótulos, utilizando informações obtidas estritamente dentro do processo de agrupamento (MILLIGAN, 1981).

Desde que neste trabalho há o conhecimento sobre os rótulos, é possível usar tanto índices externos quanto índices internos. Dentre os disponíveis na literatura (HALKIDI; BATISTAKIS; VAZIRGIANNIS, 2001), foram escolhidos dois índices externos, Índice Rand

Ajustado (HUBERT; ARABIE, 1985) e Informação Mútua Normalizada (STUDHOLME; HILL; HAWKES, 1999) e o índice interno Silhouette (ROUSSEEuw, 1987).

O Índice Rand Ajustado é formulado conforme a equação 7 em que A_1 é o número de pares de dados associados aos mesmos rótulos e também associados ao mesmo grupo no agrupamento; A_2 é o número de pares de dados associados aos mesmos rótulos, porém associados a grupos distintos; A_3 é o número de pares de dados associados a rótulos distintos, porém associados ao mesmo grupo; A_4 é o número de pares de dados associados a rótulos distintos e a grupos distintos; e, por fim, $A_{total} = A_1 + A_2 + A_3 + A_4$.

$$I_{ARI} = \frac{A_1 - \frac{(A_1+A_3)(A_1+A_2)}{A_{total}}}{\frac{(A_1+A_3)+(A_1+A_3)}{2} - \frac{(A_1+A_3)(A_1+A_2)}{A_{total}}} \quad (7)$$

A equação 8 é a formulação da Informação Mútua Normalizada, sendo λ^a e λ^b os rótulos associados aos grupos e os rótulos já conhecidos para dados; respectivamente; η_j e η_h são o número de dados pertencentes ao grupo g_j em λ^a e dados pertencentes à classe g_h em λ^b , respectivamente; e, por fim, $\eta_{j,h}$ é o número de dados que pertencem ao mesmo grupo g_j de acordo com λ^a e à mesma classe g_h de acordo com λ^b .

$$I_{NMI}(\lambda^a, \lambda^b) = \frac{\sum_{j=1}^{k^a} \sum_{h=1}^{k^b} \eta_{j,h} \log\left(\frac{n \cdot \eta_{j,h}}{\eta_j^a \eta_h^b}\right)}{\sqrt{(\sum_{j=1}^{k^a} \eta_j^a \log \frac{\eta_j^a}{n}) (\sum_{h=1}^{k^b} \eta_h^b \log \frac{\eta_h^b}{n})}} \quad (8)$$

A definição do índice Silhouette é apresentada pela equação 9 em que $a(x_i)$ é a distância média do dado x_i para todos os outros dados pertencentes ao mesmo grupo, enquanto $b(x_i)$ é a distância média do dado x_i para todos os outros dados que são pertencentes ao grupo mais próximo de x_i , de acordo com Rousseeuw (1987). Por fim, o resultado final de índice Silhouette para o agrupamento é calculado pela média de todos I_{SIL} de cada dado.

$$I_{SIL}(x_i) = \frac{b(x_i) - a(x_i)}{\max\{a(x_i), b(x_i)\}} \quad (9)$$

3.8 Similaridades parciais

O conjunto de dados é formado por elementos, ou objetos, e cada um deles pode ser representado por um vetor de atributos. Tendo isso em vista, de acordo com França (2010), o procedimento clássico dos algoritmos de agrupamento é definido como similaridade total,

ou global, em que há a utilização integral do vetor de atributos para realizar a tarefa de agrupamento.

Uma outra forma de agrupar os dados é por meio da aplicação de um procedimento baseado apenas em similaridades parciais ou baseado em partes. Em tal procedimento, geralmente aplicado em algoritmos de coagrupamento, é esperado encontrar “blocos” de linhas e colunas que são correlacionados (HUANG et al., 2015), ou seja, os dados são agrupados por subconjuntos de atributos, não sendo necessário o vetor de atributos integral para esse procedimento.

No quadro 1 é apresentado um exemplo de coagrupamento, ou seja, aplicação da perspectiva de similaridade parcial para análise dos dados. Nesse exemplo, o conjunto de dados diz respeito a um conjunto de n documentos. Os documentos estão representados nas linhas e são descritos pelos atributos organizados em colunas. Os blocos de linhas e colunas coloridos (em azul) significam que os documentos x_1 , x_2 e x_3 devem ser agrupados juntos devido às semelhanças encontradas nos atributos 1 e 2. Ainda é possível observar que o documento x_3 também deve ser agrupado junto ao documento x_4 , uma vez que há semelhanças nos atributos 3, 4 e 5.

Quadro 1 – Bloco de linhas e colunas

X	1	2	3	4	5	...	m
x_1	$x_{1,1}$	$x_{2,1}$	$x_{3,1}$	$x_{4,1}$	$x_{5,1}$	...	$x_{m,1}$
x_2	$x_{1,2}$	$x_{2,2}$	$x_{3,2}$	$x_{4,2}$	$x_{5,2}$	...	$x_{m,2}$
x_3	$x_{1,3}$	$x_{2,3}$	$x_{3,3}$	$x_{4,3}$	$x_{5,3}$	...	$x_{m,3}$
x_4	$x_{1,4}$	$x_{2,4}$	$x_{3,4}$	$x_{4,4}$	$x_{5,4}$	...	$x_{m,4}$
x_5	$x_{1,5}$	$x_{2,5}$	$x_{3,5}$	$x_{4,5}$	$x_{5,5}$	...	$x_{m,5}$
...	...	...	...	...	...	...	...
x_n	$x_{1,n}$	$x_{2,n}$	$x_{3,n}$	$x_{4,n}$	$x_{5,n}$	...	$x_{m,n}$

Fonte: Fernando Henrique da Silva Costa, 2017

3.9 Considerações finais

Este capítulo apresentou informações a respeito das técnicas de aprendizado de máquina utilizadas neste trabalho como base para formulação de estratégias para superar o problema de baixa serendipidade nas recomendações: agrupamentos e *ensemble* de agrupamentos. Foram delineadas explicações sobre agrupamento, *ensemble* e *ensemble* de agrupamentos. Para este último algoritmo, foram apresentados os dois passos necessários para seu desenvolvimento: geração de componentes e função de consenso. O próximo

capítulo apresenta os trabalhos correlatos, incluindo a utilização da tarefa de agrupamento como base para formulação de estratégias para recomendação.

4 Trabalhos correlatos

O objetivo deste capítulo é apresentar os trabalhos correlatos a esta pesquisa. Para tal fim, são apresentados artigos que versam sobre a característica de serendipidade no contexto de sistemas de recomendação, a utilização da tarefa de agrupamento como parte intrínseca das estratégias de recomendação e a aplicação de *ensemble* de agrupamento com a finalidade de resolver o agrupamento de dados textuais.

4.1 Serendipidade em sistemas de recomendação

A avaliação do aspecto de serendipidade em sistemas de recomendação tem papel fundamental neste trabalho. Tendo isso em vista, foram buscados na literatura trabalhos que possuem correlação a este, de forma a compreender as estratégias de recomendação aplicadas, as áreas de aplicação do sistema de recomendação, os conjuntos de dados utilizados, as formas de avaliação das abordagens, os resultados e possíveis limitações dos trabalhos.

4.1.1 Estratégias de recomendação

Com a finalidade de gerar listas de recomendação com serendipidade, diversas estratégias de recomendação foram propostas na literatura. Comumente, essas técnicas são baseadas em algoritmos intrínsecos a área de inteligência artificial. Um primeiro exemplo foi a utilização de técnicas de processamento de linguagem natural em Piao e Whittle (2011). O uso dessa técnica está relacionado à análise do conteúdo semântico das mensagens de *tweets* a fim de identificar os interesses dos usuários do Twitter. Os atributos surpreendentes e inesperados que podem apresentar conexões serendipitosas são determinados a partir da identificação dos interesses dos usuários.

Outra técnica que serve de base para estratégias de recomendação são os modelos probabilísticos de tópicos, principalmente o *Latent Dirichlet Allocation* (LDA). Tal técnica é observada nos trabalhos de Taramigkou et al. (2013), Xiao et al. (2014) e Jenders et al. (2015). No primeiro trabalho, os autores utilizam uma modelagem de tópicos, obtidos pelo LDA, em coleções de *tags* e usuários a fim de identificar os tópicos de preferências do usuário, os quais são representados via grafos. A partir do grafo, o objetivo é auxiliar

os usuários a encontrar músicas de sua preferência e, no processo, ajudá-los a descobrir músicas novas e inesperadas. No segundo trabalho, por meio do LDA, é gerado um modelo de ranqueamento de tópicos. Inicialmente, esse modelo reordena a distribuição de tópicos aprendidas com a intenção dos usuários. Por meio desse aprendizado, artigos são classificados, de forma relevante, na biblioteca de acordo com a consulta do usuário, ao passo que, outros tópicos ranqueados buscam fornecer a novidade nos resultados. Por fim, no terceiro trabalho, os autores utilizam um algoritmo baseado em conteúdo que recomenda notícia serendipitosa por meio de um modelo de combinações de tópicos “inesperados”, baseados nas dissimilaridades entre os tópicos latentes, e de um modelo tradicional baseado em similaridade cosseno.

No trabalho visto em Oku e Hattori (2012), os autores utilizam uma estratégia baseada em fusão para recomendar itens. Essa fusão está relacionada a dois dados obtidos pela consulta do usuário. As definições da fusão podem ser do tipo “palavra-palavra”, “palavra-gênero” e “palavra-autor”. A título de exemplo do primeiro tipo fusão, considere que o usuário pesquisou dois livros fictícios, *“A mente do doutor”* e *“O mágico está entre nós”*. Para essa consulta, uma recomendação possível é o livro fictício *“O doutor mágico”*.

Grafos também são utilizados como base para as estratégias de recomendação. Um exemplo é o trabalho de Bhandari et al. (2013). O sistema cria um grafo de similaridades do tipo “*aplicativo-aplicativo*”, incluindo todos os aplicativos do conjunto. Após essa etapa, o sistema analisa a preferência do usuário tendo como base os aplicativos instalados em seu celular. A lista de recomendação é gerada com os aplicativos que possuem pouca similaridade com aqueles que estão instalados no celular do usuário.

Por fim, a utilização de um algoritmo de filtragem colaborativa baseado em um modelo de fator latente PureSVD é utilizado para modelar a utilidade dos itens aos usuários. Tal modelagem serve como base de recomendação. A fim de construir essa base, os autores consideram dois fatores: a “raridade” do item e a distância do item para o perfil do usuário.

4.1.2 Áreas de aplicação

É possível encontrar na internet diversos tipos de itens em grande quantidade de opções. Exemplos desses itens são: músicas, filmes, notícias, livros, jogos e etc. Geralmente, para cada um desses itens, há sistemas de recomendação para auxiliar os usuários a buscar

aqueles que sejam mais apropriados aos seus interesses. A tabela 4 apresenta algumas áreas de aplicação encontradas em trabalhos na literatura.

Tabela 4 – Lista de trabalhos correlatos e suas respectivas áreas de aplicação

Ref.	Aplicação
(PIAO; WHITTLE, 2011)	Conexões de usuários de Twitter
(JENDERS et al., 2015)	Notícias
(XIAO et al., 2014)	Artigos acadêmicos
(TARAMIGKOU et al., 2013)	Músicas
(OKU; HATTORI, 2012)	Livros
(BHANDARI et al., 2013)	Aplicativos de celular
(ZHENG; CHAN; IP, 2015)	Filmes

Fonte: Fernando Henrique da Silva Costa, 2018

4.1.3 Conjuntos de dados

A tabela 5 apresenta os conjuntos de dados utilizados em cada um dos trabalhos correlatos.

Tabela 5 – Lista de trabalhos correlatos e os conjuntos de dados utilizados

Ref.	Conjunto de dados
(PIAO; WHITTLE, 2011)	Tweets de usuários
(JENDERS et al., 2015)	New York Times Corpus
(XIAO et al., 2014)	CiteULike
(TARAMIGKOU et al., 2013)	Last.fm HetRec
(OKU; HATTORI, 2012)	Rakuten Books
(BHANDARI et al., 2013)	Mobilewalla
(ZHENG; CHAN; IP, 2015)	Netflix e MovieLeans

Fonte: Fernando Henrique da Silva Costa, 2018

Em Jenders et al. (2015), foi utilizado um subconjunto do New York Times Corpus¹ publicados entre 2005 e 2007. No trabalho de Xiao et al. (2014), foi utilizado o conjunto CiteULike², o qual contém informações de perfis e leituras de usuários nos anos de 2004 a 2010.

O Last.fm HetRec 2011 é um conjunto de dados obtidos do sistema de músicas *online* Last.fm³ que contém informações de músicas ouvidas e *tagging* de 1892 usuários. Já a base japonesa de livros Rakuten Books⁴, utilizada no trabalho de Oku e Hattori (2012),

¹ <https://catalog.ldc.upenn.edu/LDC2008T19>

² <http://www.citeulike.org/>

³ <https://www.last.fm/pt/>

⁴ <https://books.rakuten.co.jp/>

contém dados de 667.218 livros entre os dias 27 dezembro de 2011 e 10 de fevereiro de 2012.

Os dados coletados para o trabalho de Bhandari et al. (2013) são da Mobilewalla, uma empresa apoiada por capital de risco especializada em coleta, análise e apresentação de dados relacionados a aplicativos móveis em quatro lojas nativas. Dentre os dados coletados, 66.223 são de aplicativos e 22.213 de usuários.

Por fim, os dados coletados do Netflix e MovieLeans⁵ para o trabalho de Zheng, Chan e Ip (2015) são distribuídos conforme apresentado na tabela 6.

Tabela 6 – Distribuição dos dados coletados nos conjuntos MovieLeans e Netflix

Conjunto	# usuários	# itens	# avaliações	Esparsidade
MovieLeans	2.113	10.197	800 mil	3,976%
Netflix	2.000	5.260	635 mil	6,01%

Fonte: Zheng, Chan e Ip (2015)

4.1.4 Avaliações das abordagens

Um passo fundamental nas estratégias de recomendação é a avaliação dos resultados obtidos. Essas avaliações podem ser feitas de diversas formas. A primeira delas é realizada por meio de estudos com usuários. Trabalhos como os de Oku e Hattori (2011), Taramigkou et al. (2013) e Jenders et al. (2015) são exemplos. No primeiro trabalho, para cada recomendação, eram realizadas quatro questões a fim de medir o aspecto de serendipidade. No segundo trabalho, os usuários ouviam as demos das músicas recomendadas e, a partir disso, questões eram realizadas junto com uma escala *likert* variando de 1 a 5 a fim de medir a novidade e serendipidade. Por fim, de forma semelhante aos anteriores, o terceiro trabalho citado mensurava a serendipidade por meio de duas afirmações, as quais os usuários deveriam assinalar se era verdadeira ou falsa: (a) essa recomendação foi relevante; (b) essa recomendação foi surpreendente positivamente.

Em alguns trabalhos, os autores propõem medidas para avaliar suas abordagens. Por exemplo, em Piao e Whittle (2011), duas medidas denominadas “*interessante*” e “*surpreendente*” são propostas a fim, por meio da união de ambas, medir a serendipidade. Semelhantemente, em Xiao et al. (2014), duas medidas também são propostas, denominadas: similaridade entre recomendações e re-recomendar o grau de correspondência. Por fim,

⁵ <https://movielens.org/>

em Bhandari et al. (2013), duas medidas são propostas para avaliar a novidade do item recomendado e a diversidade da lista de recomendações.

O trabalho visto em Zheng, Chan e Ip (2015) utiliza de medidas definidas na literatura para avaliar sua abordagem. Inicialmente, os autores aplicam as medidas de *Precision* e *Recall* para medir a acurácia ou relevância da lista de recomendação. Em termos de serendipidade, os autores aplicam a medida denominada Serendipidade (SRDP) proposta por Murakami, Mori e Orihara (2007). Por fim, aplicam a medida *intra-list diversity* com a finalidade de averiguar a diversidade dos itens da lista de recomendação.

4.1.5 Resultados

Os resultados alcançados em Piao e Whittle (2011) apontam para a possibilidade de que, dado um determinado usuário, as análises de *hashtags* e páginas *web* vinculadas aos *tweets* podem oferecer mais oportunidades para serendipidade do que apenas analisar os próprios *tweets* desse usuário. Em Oku e Hattori (2012), dado os dois *baselines* da Amazon, a abordagem proposta pelos autores conseguiu alcançar o melhor resultado em termos de serendipidade.

O trabalho em Taramigkou et al. (2013) apresentou resultados promissores, em que as médias na escala *likert* de “artistas novos” e serendipidade foram 3.6 e 3.8, respectivamente. No trabalho de Bhandari et al. (2013), em termos de diversidade, o método proposto pelos usuários obteve os melhores resultados quando comparados com o *baseline*. Em termos de novidade, os autores concluírem que quanto maior o número de aplicativos instalados no celular do usuário, maior será a diversidade das recomendações e, portanto, alcançando uma avaliação melhor de novidade.

Em Xiao et al. (2014), os resultados mostraram que o sistema de recomendação com modelo de ranqueamento de tópicos supera aqueles que não aplicam esquemas de ranqueamento. Já em Jenders et al. (2015), os autores concluírem que ao inserir atributos adicionais baseados em conteúdo, por exemplo, autores de conteúdo e data de publicação, esses podem melhorar a qualidade da recomendação e, até mesmo, auxiliar em recomendações serendipitosas para os usuários.

Por fim, em Zheng, Chan e Ip (2015), os resultados alcançados demonstram que o método proposto pelos autores obtiveram os melhores resultados, quando comparados com

outros métodos propostos na literatura, não somente em serendipidade, como em precisão e diversidade.

4.2 Agrupamentos em sistemas de recomendação

A utilização de algoritmos de agrupamento é a estratégia empregada neste trabalho. Dessa maneira, foram buscados na literatura alguns trabalhos que utilizaram algoritmos de agrupamento como parte fundamental na estratégia de recomendação.

Esses trabalhos correlatos têm como objetivo melhorar o aspecto de relevância, uma vez que algoritmos de agrupamentos possuem limitações com dados em alta dimensão. Tendo isso em vista, em Rausch et al. (2014), os autores utilizam técnicas de agrupamento em subespaços, as quais geram grupos de usuário e de dimensões, ou seja, dos itens baseados em quais os usuários são agrupados juntos. Dada a mesma deficiência, em Guo, Zhang e Yorke-Smith (2015), os autores utilizam estratégia de agrupamento em multivisões, a qual agrupa os usuários baseados na similaridade com outros usuário e na confiança em outros usuários. Para lidar com predições presentes em dois grupos, os autores aplicam *support vector regression*. Por fim, aplicam um modelo probabilístico que utiliza as avaliações dos usuários e a confiança para lidar com o problema de *cold start* de usuário. Já em Liao e Lee (2016), os autores propõem uma abordagem que aplica um algoritmo de agrupamento *self-constructing* com a finalidade de reduzir a dimensionalidade relacionada ao número de produtos. Dessa maneira, produtos mais similares serão agrupados juntos enquanto os mais dissimilares serão agrupados separados. A recomendação é realizar por meio desses grupos.

Outra estratégia com a finalidade de melhorar a acurácia das recomendações é apresentando em Zahra et al. (2015). Neste trabalho, os autores realizam um estudo comparando o algoritmo *k-means* com inicialização aleatória dos centroides com outros algoritmos que utilizam maneiras mais otimizadas de inicialização dos centroides. O propósito desse estudo é analisar a implicância da inicialização dos centroides na qualidade da recomendação.

Em relação as avaliações das recomendações, a medida *Mean Absolute Error*(MAE) é destacada (RAUSCH et al., 2014; GUO; ZHANG; YORKE-SMITH, 2015; ZAHRA et al., 2015; LIAO; LEE, 2016). Tal medida é apresentada em todos esses trabalhos a fim de avaliar a acurácia e cobertura da recomendação. A medida *Root Mean Absolute*

Error(RMAE) (RAUSCH et al., 2014; GUO; ZHANG; YORKE-SMITH, 2015; LIAO; LEE, 2016) também é utilizada para medir a acurácia e cobertura. Outras medidas apresentadas são: *Roc-sensitivity* (ZAHRA et al., 2015), F1(ZAHRA et al., 2015) e *degree of agreement* (DOA) (LIAO; LEE, 2016).

Diversos conjuntos de dados são utilizados nos experimentos com sistemas de recomendação. Exemplos são os conjuntos de filmes derivados do MovieLeans (RAUSCH et al., 2014; ZAHRA et al., 2015; LIAO; LEE, 2016), Flixster⁶ (RAUSCH et al., 2014), FilmTrust⁷ (RAUSCH et al., 2014; ZAHRA et al., 2015), Yahoo Movie⁸ (LIAO; LEE, 2016) e Amazon Video⁹ (LIAO; LEE, 2016). Também são encontrados os conjuntos de livros derivados do BookCrossing¹⁰, conjuntos de músicas derivados do Last.fm (ZAHRA et al., 2015) e os conjuntos de produtos derivados do Epinions¹¹ (GUO; ZHANG; YORKE-SMITH, 2015).

4.3 Ensemble de agrupamentos em dados textuais

Os dados textuais, normalmente, são representados em alta dimensão. Os algoritmos de agrupamentos isolados possuem deficiências com dados de alta dimensionalidade. *Ensembles* de agrupamentos são uma alternativa a fim de solucionar essa deficiência. Nesta seção, os trabalhos correlatos apresentam os algoritmos de agrupamento utilizados para gerar os componentes do *ensemble*, as medidas de similaridade ou distância e, por fim, as funções de consenso junto com o algoritmo aplicado para gerar o resultado final.

4.3.1 Geração dos componentes de *ensemble*

A tabela 7 apresenta os trabalhos correlatos junto com os algoritmos de agrupamentos utilizados para gerar os componentes de *ensemble*. É possível verificar que o algoritmo mais aplicado é o *k-means*.

De acordo com TIAN e YANG (2014), o *binary k-means* é uma extensão do algoritmo *k-means* e é descrito da seguinte forma: inicialmente, os dados são particionados

⁶ <https://www.flixster.com/>

⁷ <https://www.librec.net/datasets.html>

⁸ <https://www.yahoo.com/entertainment/movies/>

⁹ <https://www.primevideo.com/>

¹⁰ <https://www.bookcrossing.com/>

¹¹ <http://www.epinions.com/>

Tabela 7 – Lista de trabalhos correlatos e os algoritmos de agrupamento utilizados para geração dos componentes de *ensemble*

Ref.	Algoritmo
(TIAN; YANG, 2014)	<i>Binary k-means</i> e Espectral
(CORRÊA et al., 2015)	<i>k-means</i>
(WAHID; GAO; ANDREAE, 2015)	Hierárquicos e <i>k-means</i>
(SADEGHIAN; NEZAMABADI-POUR, 2015)	<i>k-means</i>
(XU et al., 2016)	<i>k-means</i>

Fonte: Fernando Henrique da Silva Costa, 2018

em dois grupos, e cada um desses grupos são particionados em outros dois, e assim por diante até gerar os k grupos desejados. Em relação ao algoritmo de agrupamento espectral, os autores dizem que esse algoritmo explora as similaridades entre pares de dados e tem sido mais efetivos que os algoritmos de agrupamentos tradicionais em encontrar grupos.

O algoritmo hierárquico aplicado no trabalho de Wahid, Gao e Andreae (2015) é o AGNES com estratégias *single-link*, *complete-link* e *average-link*.

4.3.2 Medidas de similaridades

A tabela 8 apresenta os trabalhos correlatos juntos as medidas de similaridades ou distância aplicadas. É possível observar que a distância cosseno é a mais utilizada.

Tabela 8 – Lista de trabalhos correlatos e as medidas de similaridades aplicadas

Ref.	Medidas de similaridade
(TIAN; YANG, 2014)	Cosseno
(CORRÊA et al., 2015)	Cosseno
(WAHID; GAO; ANDREAE, 2015)	Cosseno
(SADEGHIAN; NEZAMABADI-POUR, 2015)	Euclidiana
(XU et al., 2016)	Cosseno

Fonte: Fernando Henrique da Silva Costa, 2018

4.3.3 Funções de consenso

A tabela 9 apresenta os trabalhos correlatos e as funções de consenso aplicadas, sendo a estratégia de matriz de coassociação a mais aplicada. Além disso, também é apresentado os algoritmos de agrupamentos aplicados para gerar o resultado final do *ensemble*.

Tabela 9 – Lista de trabalhos correlatos, as funções de consenso aplicadas e os algoritmos utilizados para gerar o resultado final

Ref.	Função de consenso	Algoritmo
(TIAN; YANG, 2014)	Matriz de coassociação	<i>k-means</i>
(CORRÊA et al., 2015)	Matriz de coassociação	UPGMA
(WAHID; GAO; ANDREAE, 2015)	Processo evolutivo	MDC
(SADEGHIAN; NEZAMABADI-POUR, 2015)	Matriz de coassociação	CSPA e HGPA
(XU et al., 2016)	Matriz laplaciana	<i>k-means</i>

Fonte: Fernando Henrique da Silva Costa, 2018

De acordo com Corrêa et al. (2015), o UPGMA é um algoritmo hierárquico aglomerativo, o mesmo método de agrupamento que o AGNES com as estratégias de *linkage*.

Em Wahid, Gao e Andreae (2015), o *Multi-objective Document Clustering* (MDC) é iniciado pela geração inicial da população, ou seja, a aplicação dos algoritmos de agrupamento em duas matrizes representadas por duas visões distintas: o conteúdo e o hipertexto das páginas *web*. A representação genética de cada agrupamento é por meio de uma matriz binária igual a matriz de pertinência U . A avaliação do *fitness* é por meio de duas funções, distância intra-grupo e similaridade intra-grupo, as quais devem ser minimizadas. O método de seleção é o torneio binário para determinar os indivíduos. São utilizados dois métodos de *cross-over*, nas linhas e nas colunas, e dois métodos de mutação, a união de dois grupos pequenos em apenas um e o desmembramento de um grupo grande em dois menores. O método de parada é o número de geração.

Em conformidade com Sadeghian e Nezamabadi-pour (2015), *Cluster based Similarity Partitioning* (CSPA) e *Hyper Graph Partitioning Algorithm* (HGPA) são mecanismos eficientes de combinação baseados em grafos e hipergrafos.

4.4 Considerações finais

Este capítulo apresentou os trabalhos correlatos a esta pesquisa. Os trabalhos foram separados em três áreas a fim de otimizar a organização de informações relevantes que auxiliassem o desenvolvimento da pesquisa. A partir das informações apresentadas neste e nos capítulos anteriores, estratégias de recomendação baseadas em agrupamento foram desenvolvidas e seguem apresentadas no próximo capítulo.

5 Estratégias de recomendação baseadas em agrupamento

Neste trabalho são propostas estratégias de recomendação baseada em conteúdo que assumem uma estrutura de agrupamento como base para sua implementação. Sete estratégias de recomendação são propostas. Seis delas são implementadas sobre uma estrutura de agrupamento gerada pelo algoritmo *k-means++*. A sétima é a estratégia de recomendação proposta sobre a estrutura de agrupamento criada pelo ensemble de agrupamentos. As sete estratégias são apresentadas neste capítulo.

5.1 Estratégias de recomendação baseadas em agrupamento com *k-means++*

As seis estratégias propostas sobre o agrupamento gerado pelo *k-means++* foram elaboradas com o intuito de explorar recomendações relevantes (duas estratégias), recomendações surpreendentes (duas estratégias) e recomendações serendipitosas (duas estratégias). Cada uma das estratégias é, presumidamente¹, mais adequada para realizar as recomendações com determinado aspecto. As estratégias foram propostas para um contexto de recomendação baseada em conteúdo de notícias, sendo assim, elas são preparadas para uso em um conjunto de notícias para o qual é pressuposto existir uma partição. A construção da partição deve ser feita a partir da aplicação de um algoritmo de agrupamento (no caso deste trabalho, o algoritmo *k-means++* com o uso da distância cosseno). Nesse esquema, as notícias naturalmente similares estarão associadas a um mesmo grupo na partição, enquanto as notícias dissimilares estarão associadas a grupos diferentes.

De forma geral, a partir dos k grupos gerados pelo algoritmo de agrupamentos, as estratégias de recomendação são aplicadas com a finalidade de criar uma lista L de recomendação de notícias que alcance um dos três aspectos de qualidade de recomendação. A base da recomendação é a existência de uma notícia semente (xs) que é acessada pelo usuário que receberá a recomendação. Na realidade, o acesso à notícia semente é o que caracteriza o perfil do usuário em foco, mitigando o problema de “novo usuário”. Nesse contexto e para análise dos aspectos da recomendação assume-se que:

¹ Diz-se “presumidamente” porque as estratégias são baseadas em hipóteses do que o autor deste trabalho considera ser importante para que uma recomendação seja relevante, surpreendente ou serendipitosa considerando o contexto de análise de dados textuais, ou seja, o contexto do conteúdo das notícias.

- o sistema não conhece outras preferências do usuário, a não ser a preferência pela notícia xs ;
- o usuário não teve acesso prévio a nenhuma das notícias presentes na lista L ;
- a recomendação é estritamente baseada no texto das notícias e na sua organização em grupos.

As próximas subseções apresentam as estratégias de recomendação divididas de acordo com o aspecto de qualidade de recomendação que cada uma é presumidamente preparada para atender.

5.1.1 Estratégias para recomendações relevantes

As duas primeiras estratégias de recomendação são projetadas com o intuito de gerar uma lista de recomendação L de notícias relevantes, ou seja, uma lista com notícias que estejam no hábito de leitura do usuário. Tal hábito é considerado bem representado pela notícia semente e, portanto, quanto mais similar são as notícias (semente e recomendadas), mais relevante é a recomendação. A equação 10 formaliza o objetivo dessas estratégias, assumindo que xo_i é uma notícia no conjunto de notícias disponíveis, l é a quantidade de notícias em L e $a \sim b$ significa que uma notícia a possui conteúdo similar à notícia b .

$$L = \{xo_1, xo_2, \dots, xo_l\} \mid xo_i \sim xs, i = \{1, \dots, l\} \quad (10)$$

Considerando o objetivo geral assumido para a construção de recomendações relevantes, considerando uma partição gerada por um algoritmo de agrupamento, e assumindo a distância cosseno entre duas notícias como sendo inversamente proporcional à similaridade entre elas, as duas estratégias de recomendação propostas para este contexto são definidas da seguinte forma:

- recomendar as notícias mais similares à notícia semente e que sejam pertencentes ao mesmo grupo em que a notícia semente está, i.e.²:

$$\arg \max_i (xo_i \sim xs) \forall xo_i \in L, i = \{1, \dots, l\} \mid xo_i, xs \in g_j \text{ e } xo_i \neq xs;$$

² Assumindo, neste caso e também para as demais estratégias de recomendação, a aplicação iterativa da função de maximização de forma a selecionar i elementos de um conjunto, sem reposição.

- recomendar aleatoriamente notícias pertencentes ao mesmo grupo da notícia semente, i.e:

$$rand(xo_i) \forall xo_i \in L, i = \{1, \dots, l\} \mid xo_i \wedge xs \in g_j \text{ e } xo_i \neq xs.$$

A figura 11 ilustra as duas estratégias de recomendação descritas, em que a parte (a) destaca a primeira estratégia e a parte (b) destaca a segunda estratégia. A notícia semente está destacada na cor azul, enquanto as destacadas na cor amarelo são as notícias que compõem a lista de recomendação. Nesta figura, o aglomerado de documentos representa um único grupo.

Figura 11 – Esquema gráfico para estratégias para recomendações relevantes

Fonte: Fernando Henrique da Silva Costa, 2018

A hipótese que fundamenta a construção destas duas estratégias é que a relevância é expressa pela similaridade de conteúdo. Também é suposto que a similaridade de conteúdo está bem representada na estrutura de grupos descoberta pelo algoritmo de agrupamento. A motivação para a proposição de duas estratégias que usam um único grupo de notícias como base de recomendação é propiciar uma análise sobre a viabilidade, ou não, de usar as notícias de um grupo livremente dentro de uma estratégia de recomendação.

Considere U como uma matriz de pertinências, ou seja, uma matriz binária de dimensões $k \times n$ em que $U_{j,i} = 1$ se o dado i pertence ao grupo j , e $U_{j,i} = 0$ se o dado i não pertence ao grupo j , x como uma notícia no conjunto de dados, l como a quantidade de notícias que serão recomendadas e L como a lista de recomendação de notícias. Os algoritmos 5 e 6 apresentam, respectivamente, os pseudo-códigos das duas estratégias para recomendação com relevância, tomando como base o agrupamento criado pelo algoritmo *k-means++*.

Algoritmo 5 Primeira estratégia de recomendação relevante para o algoritmo *k-means++*

- 1: **procedure** ESTRATEGIARELEVANCIA01(U, x, l)
- 2: Encontrar o grupo ao qual a notícia x pertence, usando a matriz de pertinências U
- 3: Computar a similaridade entre todas notícias pertencentes ao grupo de x com a notícia x
- 4: Selecionar as l notícias mais semelhantes a x
- 5: Formar a lista L de recomendações com tais notícias

Fonte: Fernando Henrique da Silva Costa ,2018

Algoritmo 6 Segunda estratégia de recomendação relevante para o algoritmo *k-means++*

- 1: **procedure** ESTRATEGIARELEVANCIA02(U, x, l)
- 2: Encontrar o grupo ao qual a notícia x pertence usando a matriz de pertinências U .
- 3: Selecionar l notícias aleatórias do grupo da notícia x , exceto x
- 4: Formar a lista L de recomendações com tais notícias

Fonte: Fernando Henrique da Silva Costa ,2018

5.1.2 Estratégias para recomendações surpreendentes

As duas estratégias que visam recomendar notícias surpreendentes são projetadas a partir da diversidade de conteúdo. De acordo com (SHANI; GUNAWARDANA, 2011), a diversidade é definida como o oposto de similaridade. Em alguns casos, recomendar uma lista de itens similares aos que foram apreciados pelo usuário pode não ser interessante, ou mesmo útil, para ele. Essas duas estratégias visam gerar listas de recomendações que tragam diversidade de assuntos para o contexto do usuário. A equação 11 formaliza o objetivo dessas duas estratégias, em que xo_i e l são como definido anteriormente e $a <> b$ significa que uma notícia a possui conteúdo diversificado em relação ao conteúdo da notícia b .

$$L = \{xo_1, xo_2, \dots, xo_l\} \mid xo_i <> xs, i = \{1, \dots, l\} \quad (11)$$

Considerando o objetivo geral assumido para a construção de recomendações surpreendentes, considerando uma partição gerada por um algoritmo de agrupamento, e assumindo o cálculo da distância entre grupos como sendo a distância cosseno entre os seus centroides, as duas estratégias de recomendação propostas para este contexto são definidas da seguinte forma:

- recomendar notícias aleatoriamente escolhidas dentro do grupo mais distante do grupo ao qual a notícia semente pertence, i.e.:

$$rand(xo_i) \forall xo_i \in L, i = \{1, \dots, l\} \mid xo_i \in g_j, xs \in g_o, j, o \in \{1, \dots, k\} \text{ e} \\ \arg \max_j (dist(\vec{g_j}, \vec{g_o}));$$

- recomendar notícias aleatoriamente escolhidas dentro de todos os grupos exceto daquele ao qual a notícia semente pertence, i.e.:

$$rand(xo_i) \forall xo_i \in L, i = \{1, \dots, l\} \mid xo_i \in g_j, xs \in g_o, j = \{1, \dots, k\}, o \in \\ \{1, \dots, k\} \text{ e } j \neq o.$$

A figura 12 ilustra as duas estratégias para recomendações surpreendentes. As notícias estão agrupadas em k grupos. A notícia semente está destacada na cor azul e pertence ao primeiro grupo, enquanto as notícias que formam a lista de recomendação estão destacadas na cor amarelo e: (a) pertencem ao grupo mais distante do grupo da notícia semente (arbitrariamente indexado como grupo k na figura); (b) pertencem a grupos distintos do grupo ao qual a notícia semente pertence.

Figura 12 – Esquema gráfico para estratégias para recomendações surpreendentes

Fonte: Fernando Henrique da Silva Costa, 2018

Para fundamentar essas estratégias, a localização de notícias em grupos diferentes foi assumida como sendo um indicativo de que seus conteúdos são suficientemente diferentes para que a recomendação causasse surpresa. Duas estratégias foram propostas para fins de verificação se a estrutura de grupos é suficiente para embutir surpresa na recomendação ou se é necessário assumir a distância entre os grupos para causar esse efeito.

Considere C como uma matriz de centroides, ou seja, uma matriz $k \times m$, que representa os pontos médios entre os dados de cada grupo do agrupamento, U como uma matriz de pertinências, x como uma notícia no conjunto, l como a quantidade de notícias que serão recomendados e, por fim, L como a lista de recomendações. Tendo isso em vista,

os algoritmos 7 e 8 apresentam, respectivamente, os pseudo-códigos da primeira e segunda estratégias de recomendação surpreendentes do algoritmo *k-means++*.

Algoritmo 7 Primeira estratégia de recomendação surpreendente para o algoritmo *k-means++*

- 1: **procedure** ESTRATEGIASURPRESA01(U, x, l, C)
- 2: Encontrar o grupo ao qual a notícia x pertence usando a matriz de pertinências U
- 3: Calcular a distância do grupo de x para os outros grupos de notícias de acordo com a equação 3 e usando a matriz de centroides C
- 4: Dada as distâncias entre os grupos, selecionar o grupo mais distante do grupo de x
- 5: Selecionar l notícias aleatórias do grupo mais distante
- 6: Formar a lista L de recomendações com tais notícias

Fonte: Fernando Henrique da Silva Costa ,2018

Algoritmo 8 Segunda estratégia de recomendação surpreendente para o algoritmo *k-means++*

- 1: **procedure** ESTRATEGIASURPRESA02(U, x, l)
- 2: Encontrar o grupo ao qual a notícia x pertence usando a matriz de pertinências U
- 3: Selecionar os demais grupos, ou seja, os grupos aos quais x não pertence
- 4: Selecionar l notícias aleatórias pertencentes a estes grupos
- 5: Formar a lista L de recomendações com tais notícias

Fonte: Fernando Henrique da Silva Costa ,2018

5.1.3 Estratégia para recomendações serendipitosas

As duas últimas estratégias de recomendação tem como finalidade gerar listas de recomendação em que as notícias sejam serendipitosas. Para isso, as estratégias são definidas de forma a combinar elementos das estratégias já apresentadas, em conformidade com as definições de serendipidade apresentadas em Shani e Gunawardana (2011) e Ge, Delgado-Battenfeld e Jannach (2010). A equação 12 formaliza o objetivo dessas estratégias, em que xo_i e l são como definido anteriormente e $a \asymp b$ significa que a notícia a possui alguma similaridade com relação à notícia b , porém traz algum conteúdo novo (diferente).

$$L = \{xo_1, xo_2, \dots, xo_l\} \mid xo_i \asymp xs, i = \{1, \dots, l\} \quad (12)$$

Considerando o objetivo geral assumido para a construção de recomendações surpreendentes, considerando uma partição gerada por um algoritmo de agrupamento, assumindo a distância cosseno entre duas notícias como sendo inversamente proporcional à similaridade entre elas, e assumindo o cálculo da distância entre grupos como sendo a

distância cosseno entre os seus centroides, as duas estratégias de recomendação propostas para este contexto são definidas da seguinte forma:

- recomendar as notícias mais similares à notícia semente, considerando notícias pertencentes ao grupo mais próximo do grupo ao qual pertence a notícia semente, i.e.:

$$\arg \max_i (xo_i \sim xs) \forall xo_i \in L, i = \{1, \dots, l\} \mid xo_i \in g_j, xs \in g_o \text{ e} \\ \arg \min_j (dist(\vec{g}_j, \vec{g}_o)), j, o \in \{1, \dots, k\}$$

- recomendar as notícias mais similares à notícia semente, considerando notícias pertencentes ao grupo que representa um balanceamento entre o grupo mais próximo e o grupo mais distante do grupo ao qual pertence a notícia semente, i.e.:

$$\arg \max_i (xo_i \sim xs) \forall xo_i \in L, i = \{1, \dots, l\} \mid xo_i \in g_j, xs \in g_o, \\ \vec{g} \approx \frac{\sum_{j=1 \mid j \neq o}^k \vec{g}_j}{k-1} \text{ e } \arg \min_j (dist(\vec{g}_j, \vec{g})), j \in \{1, \dots, k\}.$$

A figura 13 ilustra as duas estratégias para recomendações serendipitosas. A partição encontrada pelo algoritmo de agrupamento define k grupos. Considere arbitrariamente que o grupo 2 é o mais próximo do grupo 1, o grupo 3 é o mais próximo do grupo 2, o grupo 4 é mais próximo do grupo 3 e o grupo k é o mais próximo do grupo 4. Se a notícia semente pertence ao grupo 1 então, a primeira estratégia para alcance da serendipidade recomenda notícias do grupo 2, enquanto a segunda recomenda notícias do grupo 3. Na figura, a notícia semente está destacada na cor azul e as notícias recomendadas estão destacadas na cor amarela.

Figura 13 – Esquema gráfico para estratégias para recomendações serendipitosas

Fonte: Fernando Henrique da Silva Costa, 2018

A criação dessas duas estratégias fornece condições de analisar duas situações. A primeira, referente à primeira estratégia de recomendação para serendipidade, embute diversidade em um grau menos elevado que a segunda, uma vez que coloca na lista de recomendação notícias de fora do grupo da notícia semente, porém pertencentes ao grupo mais similar ao da notícia semente. A segunda aumenta o grau de diversidade embutida na lista de recomendação em relação à notícia semente. Note que a diversidade em ambas estratégias está sendo considerada como sendo em relação à notícia semente, e não em relação às próprias notícias que compõem a lista. Desta forma, olhando a lista de recomendação internamente, ela não apresentaria diversidade, de acordo com as hipóteses assumidas neste trabalho.

Os algoritmos 7 e 10 apresentam, respectivamente, os pseudo-códigos da primeira e da segunda estratégias de recomendação serendipitosa para o algoritmo *k-means++*. Para tal, considere C como uma matriz de centroides, U como uma matriz de pertinências, x como uma notícia pertencente ao conjunto de dados, l como a quantidade de notícias a ser recomendadas e, por fim, L como a lista de recomendações.

Algoritmo 9 Primeira estratégia de recomendação serendipitosa para o algoritmo *k-means++*

- 1: **procedure** ESTRATEGIASERENDIPIDADE01(U, x, l, C)
- 2: Encontrar o grupo ao qual a notícia x pertence usando a matriz de pertinências U
- 3: Calcular a distância do grupo de x para os outros grupos de acordo com a equação 3 usando a matriz de centroides C
- 4: Dada as distâncias calculadas, selecionar o grupo mais próximo do grupo de x
- 5: Selecionar as l notícias do grupo mais próximo do grupo de x que sejam as mais próximas de x
- 6: Formar a lista L de recomendações com tais notícias

Fonte: Fernando Henrique da Silva Costa ,2018

Algoritmo 10 Segunda estratégia de recomendação serendipitosa para o algoritmo *k-means++*

- 1: **procedure** ESTRATEGIASERENDIPIDADE02(U, x, l, C)
- 2: Encontrar o grupo ao qual a notícia x pertence usando a matriz de pertinências U
- 3: Calcular a distância do grupo de x para os outros grupos de notícias de acordo com a equação 3 usando a matriz de centroides C
- 4: Selecionar o grupo que representa um balanceamento entre o mais distante e mais próximo grupo de x , ou seja, o grupo que tem a distância mais próxima da média entre todas distâncias
- 5: Selecionar as l notícias pertencentes desse grupo que sejam as mais próximas de x
- 6: Formar a lista L de recomendações com tais notícias

Fonte: Fernando Henrique da Silva Costa ,2018

5.2 Estratégia de recomendação baseada em ensemble de agrupamentos

O principal interesse neste trabalho é apresentar alternativas de recomendação que possam alcançar bons níveis de qualidade de recomendação, em termos de relevância, surpresa e principalmente serendipidade. Assim, a estratégia apresentada nesta seção foi construída pensando em atender a este terceiro aspecto.

Semelhantemente às estratégias descritas na seção 5.1, aplicadas a agrupamentos gerados pelo algoritmo *k-means++*, a estratégia a ser aplicada sobre o agrupamento gerado pelo *ensemble* de agrupamentos também pressupõe a existência de uma partição no conjunto de notícias, além de considerar as mesmas diretrizes já apresentadas neste texto para atendimento ao aspecto de serendipidade de recomendação. Desta forma, o propósito é gerar uma lista L de recomendação de notícias serendipitosas para o usuário, tendo como base a notícia semente xs , de acordo com o objetivo já formalizado na equação 12.

Entretanto, o propósito de usar estrutura de grupos formada a partir de um *ensemble* é embutir nessa estrutura a ideia de uso de similaridades parciais entre elementos, o que no presente caso significa considerar similaridades parciais entre notícias. A hipótese que fornece embasamento para o uso de similaridades parciais é que: ainda que duas notícias discutam assuntos diferentes, por exemplo “esporte” e “carnaval”, e estejam alocadas em grupos diferentes dentro de uma partição, elas podem, na realidade, ter algo em comum; o assunto que une as duas notícias pode estar relacionado a um terceiro conceito (“festa” e/ou “local”); desta forma, elas possuem alguma similaridade que é representada por “partes” de seu conteúdo, e não necessariamente pelo seu tema central; e finalmente, recomendar uma notícia de tema central “carnaval”, porém ocorrendo em um determinado local, para um usuário interessado em um evento esportivo também localizado naquele local, pode constituir uma recomendação serendipitosa. O fator surpresa está presente no tema central da notícia recomendada, e o fator relevância está presente na similaridade parcial da notícia recomendada com a notícia de interesse inicial do usuário (a notícia semente).

Considerando uma representação textual baseada em termos, como a representação usada neste trabalho, um subconjunto de termos presente nas duas notícias pode torná-las parcialmente similares. A figura 14 ilustra as similaridades parciais encontradas entre o grupo i e o grupo j . Ainda que o primeiro grupo seja formado por notícias que versam sobre esportes e o segundo seja formado por notícias sobre o carnaval carioca, a similaridade

parcial é observada nos termos “Rio de Janeiro” e “Festa” (destacados em vermelho), os quais estão presentes em ambos grupos.

Figura 14 – Exemplo de grupos com termos similares

Fonte: Fernando Henrique da Silva Costa, 2018

Neste trabalho, o *ensemble* de agrupamento foi usado como uma ferramenta para buscar uma partição que embutisse relevância e diversidade na estrutura dos agrupamentos, fornecendo uma análise sobre similaridades parciais. Então, uma estratégia de recomendação para serendipidade foi implementada sobre a partição gerada.

A fim de explicar toda a abordagem, o restante dessa seção está dividida em duas partes: a primeira apresenta a estratégia de *ensemble* de agrupamento implementada; a segunda apresenta como a estratégia de recomendação utiliza a partição de dados oferecida pelo *ensemble*.

5.2.1 Construindo um *ensemble* de agrupamentos

O passo inicial da construção de um *ensemble* de agrupamentos é a geração de seus p componentes. Neste trabalho, cada um deles é uma solução de agrupamento obtida pelo algoritmo *k-means++* com distância cosseno. No entanto, há uma variação de atributos com reposição do conjunto original para a geração dos componentes. Dessa forma, cada componente é gerado a partir de uma variação do conjunto original com o intuito de buscar similaridades parciais nessa etapa.

A escolha dos atributos é realizada de forma aleatória, em que são selecionados entre 65% a 75% do número total de atributos do conjunto original. Com o intuito de facilitar a aplicação do algoritmo, cada dado tem, obrigatoriamente, um valor de atributo diferente de 0.

O segundo passo é aplicar uma estratégia, denominada função de consenso, que integra todas os agrupamentos gerados pelos componentes em um agrupamento somente. Neste trabalho, essa etapa é realizada pela matriz de coassociação.

O desenvolvimento da matriz de coassociação é iniciada pela representação de cada componente por uma matriz de similaridade. Essa matriz informa se os dados x_i e x_j estão agrupados em um mesmo grupo quando $M_{ij}^a = 1$. Caso $M_{ij}^a = 0$, os dados x_i e x_j estão agrupados em grupos diferentes. Por fim, a matriz de coassociação é gerada pela soma de todas matrizes de similaridades divididas por p (equação 2).

A partir do estabelecimento da matriz de coassociação, essa é representada como uma matriz de dados e, portanto, é aplicado um algoritmo de agrupamento a fim de gerar os grupos finais. A cada um desses grupos há um subconjunto de documentos associados. Tomando os vetores que representam cada documento dos subconjuntos é possível construir o centroide de cada grupo e organizá-los em uma matriz aqui chamada de matriz de centroides C . Nessa matriz, cada linha é um centroide de um grupo, e seu tamanho é $|C|k \times m$. A matriz C é então processada de forma que seja possível encontrar os termos que melhor caracterizam cada grupo de documentos e, então, encontrar os termos que são comuns na caracterização dos grupos.

Para encontrar os termos que melhor caracterizam cada grupo, assume-se um limiar mínimo de representação θ . Usando esse limiar, a matriz C é binarizada, de forma que valores $C_{ij} \geq \theta$ são transformados em 1 ($C_{ij} \leftarrow 1$), pois são considerados suficientemente altos para indicar que o termo j caracteriza o grupo i . Para $C_{ij} \leq \theta$, $C_{ij} \leftarrow 0$, indicando que o termo j não caracteriza o grupo i . O passo de binarização da matriz C é ilustrado na figura 15. Nela, a matriz superior é a matriz C original, $\theta = 0,5$ e a matriz inferior é a matriz C binarizada. Na matriz binarizada é possível observar que o grupo 1 (linha 1) é caracterizado pelos termos 1, 2 e 5 (colunas 1, 2 e 5).

O próximo passo é procurar pelas similaridades parciais, ou seja, realizar uma interseção dos elementos que caracterizam os grupos, destacando se há grupos cujos centroides indicam que há similaridade parcial entre as notícias a eles pertencentes. Na figura 16 são destacadas as similaridades parciais indicadas pelos centroides da matriz C . Segundo o exemplo, o grupo 1 e o grupo 2 possuem similaridade parcial caracterizadas pelos termos 1 e 2. Os grupos 3 e 4 possuem similaridade parcial caracterizada pelos termos 8, 9 e 10. Os grupos 1 e 2 não possuem qualquer similaridade com os grupos 3 e 4.

Figura 15 – Binarizando a matriz C

0,6	0,7	0,2	0,1	0,8	0,3	0,4	0,4	0,1	0,2
0,8	0,9	0,1	0,9	0,4	0,5	0,2	0,1	0,3	0,4
0,1	0,2	0,3	0,4	0,4	0,4	0,9	0,8	0,9	0,7
0,1	0,4	0,4	0,3	0,2	0,1	0,4	0,5	0,8	0,7

↓ $\Theta = 0,5$

1	1	0	0	1	0	0	0	0	0
1	1	0	1	0	1	0	0	0	0
0	0	0	0	0	0	1	1	1	1
0	0	0	0	0	0	0	1	1	1

Fonte: Fernando Henrique da Silva Costa, 2018

Figura 16 – Interseção dos grupos

1	1	0	0	1	0	0	0	0	0
1	1	0	1	0	1	0	0	0	0
0	0	0	0	0	0	1	1	1	1
0	0	0	0	0	0	0	1	1	1

Fonte: Fernando Henrique da Silva Costa, 2018

5.2.2 Construindo a estratégia de recomendação

O resultado da busca pelas similaridades parciais resulta em uma partição diferenciada da partição clássica (usada nas estratégias da seção 5.1), e é usada como base para a estratégia de recomendação baseada em *ensemble* de agrupamento. Assim, o passo final dessa abordagem é implementar a estratégia de recomendação que gera a lista L . Dada a notícia semente e o grupo a qual ela pertence, as notícias que compõem L são selecionadas aleatoriamente dos grupos que possuem similaridade parcial com o grupo da notícia semente, i.e.:

$$rand(xo_i) \forall xo_i \in L, i = \{1, \dots, l\} \mid xo_i \in g_j, xs \in g_o, j, o \in \{1, \dots, k\}, g_j \propto g_o,$$

em que xo_i é uma notícia recomendada e $g_j \propto g_o$ indica que os grupos g_j e g_o apresentam similaridade parcial.

O algoritmo 11 apresenta um pseudo-código de como são realizadas as recomendações serendipitosas com o algoritmo *ensemble* de agrupamentos. Para tal, considere C como uma matriz de centroides, U como uma matriz de pertinências, x como uma notícia pertencente do conjunto, l como a quantidade de notícias que serão recomendadas, L como

a lista de recomendação e, por fim, θ como um limiar de binarização, isto é, se $C_{ij} \geq \theta$, então $C_{ij} = 1$, caso contrário, $C_{ij} = 0$.

Algoritmo 11 Estratégia de recomendação serendipitosa para o algoritmo *ensemble* de agrupamentos

- 1: **procedure** ESTRATEGIAENSEMBLE(U, x, C, θ, l)
- 2: Binarizar a matriz de centroides do *ensemble* de agrupamentos C usando o valor escolhido para θ
- 3: Encontrar os grupos (linhas) que possuem colunas em comum que estejam assinaladas em 1
- 4: Selecionar l notícias aleatórias de qualquer grupo que possua alguma interseção com o grupo de x
- 5: Formar a lista L de recomendações com essas notícias

Fonte: Fernando Henrique da Silva Costa ,2018

5.3 Considerações finais

O objetivo deste capítulo foi apresentar as seis estratégias de recomendação que tomam por base os algoritmos *k-means++* e *ensemble* de agrupamentos. Os dois próximos capítulos apresentam, respectivamente, experimentos para análise quantitativa e experimentos para análise qualitativa. O segundo conjunto de experimentos foi executado com base nos resultados obtidos no primeiro conjunto.

6 Experimentos com agrupamentos

Antes de aplicar as estratégias de recomendação propostas neste trabalho, é necessário analisar a qualidade dos agrupamentos produzidos pelo *k-means++* e pelo *ensemble* de agrupamentos, de forma que seja possível: (a) atestar que as estruturas de agrupamento de notícias tenham qualidade em termos de índices de avaliação de agrupamento; (b) apresentem razoabilidade semântica para serem usadas como base de recomendação.

Neste capítulo são apresentados os resultados alcançados pelos algoritmos de agrupamento, *k-means++* e *ensemble* de agrupamentos. Esses resultados seguem avaliados por meio de índices de avaliação baseados em critérios externos e internos e também por meio de análises de cunho qualitativo sobre os grupos formados. Além disso, informações a respeito do conjunto de notícias usados no trabalho, das tarefas de pré-processamento de textos e das configurações de experimentos são apresentadas.

6.1 Conjunto de notícias

O conjunto de notícias utilizado neste trabalho é o *Corpus EBC*¹, o qual é composto por 1.157 notícias provenientes do portal de notícias Empresa Brasil de Comunicação (EBC)². Cada notícia é pertencente a um dos sete cadernos encontrados no portal, denominados: *Cidadania*, *Cultura*, *Educação*, *Esportes*, *Infantil*, *Notícias* e *Tecnologia*. As notícias são associadas a um caderno de acordo com a interpretação dos produtores de conteúdo do portal.

Anexo ao caderno *Notícias* é possível encontrar alguns “subcadernos”, tais como: *Política*, *Economia*, *Internacional* e *Geral*. É importante salientar que esse último subcaderno corresponde a notícias diversificadas que, sob o ponto de vista dos produtores de conteúdo, não são aderentes aos demais cadernos. Por consequência, o subcaderno é constituinte de notícias que versam sobre assuntos não necessariamente relacionados.

A fim de estabelecer uma verdade fundamental para categorizar as notícias, cada uma delas foi associada a um rótulo correspondente ao caderno ou subcaderno em que ela foi publicada no portal. Tendo isso em vista, dez rótulos foram definidos e as distribuições (absolutas e relativas) das notícias nas categorias são listadas na tabela 10. Como pode ser

¹ <https://anonimo.anonimo>

² <http://www.ebc.com.br/>

observado na tabela, trata-se de um conjunto de notícias com distribuição de categorias desbalanceada.

Tabela 10 – Conjunto de notícias: *Corpus EBC*

Cadernos	Rótulo	# de notícias	%
Educação	3	257	22,2
Esportes	4	200	17,3
Tecnologia	7	125	10,8
Geral	10	125	10,8
Cultura	2	124	10,7
Cidadania	1	96	8,3
Internacional	9	89	7,7
Política	6	86	7,5
Economia	8	41	3,5
Infantil	5	14	1,2
Total		1.157	100

A figura 17 ilustra a plotagem do índice Silhouette considerando os cadernos como grupos. O valor de Silhouette do agrupamento, ou seja, a média entre todos os dados, é 0,050. É possível observar que existe alguma homogeneidade nos grupos. Há grupos bem categorizados, como por exemplo, as notícias que formam o caderno de *Esportes* (grupo 4). Há também grupos com avaliação Silhouette negativa, por exemplo, as notícias que formam o caderno *Geral* (grupo 10).

Figura 17 – Plotagem de Silhouette considerando os cadernos das notícias como um resultado de agrupamento

Fonte: Fernando Henrique da Silva Costa, 2018

6.2 Pré-processamento

De acordo com Hotho, Nürnberger e Paaß (2005) e Sinoara, Antunes e Rezende (2017), para analisar grandes coleções de documentos textuais, é preciso pré-processá-los e armazenar as informações produzidas a partir do pré-processamento em uma estrutura de dados que seja apropriada para a análise pretendida.

A estrutura utilizada neste trabalho é a *bag-of-words* (SALTON; WONG; YANG, 1975). O objetivo dessa estrutura é transformar uma coleção de textos em uma matriz $X^{n \times m}$, de forma que cada linha represente um documento i , com $i = \{1, \dots, n\}$, cada coluna represente um termo j , com $j = \{1, \dots, m\}$ e cada elemento x_{ij} indique a ocorrência de um termo j no documento i . Neste trabalho foram geradas 50 matrizes, $X_1, X_2, \dots, X_{50}$, a partir de 50 configurações de pré-processamento.

Os procedimentos de pré-processamento aplicados ao conjunto de notícias foram:

- **tokenização**: consiste em remover os símbolos, números e pontuações;
- **remoção de stopwords** (opcional): consiste em remover palavras que adicionam pouco valor semântico e, portanto, tem pouco poder de representação do conteúdo de um texto (e.g., “de”, “a”, “com”), a partir de uma lista explícita³;
- **representação vetorial**: consiste em determinar um mapeamento dos textos para uma das representações seguintes representações: binária, *tf* ou *tf-idf*. Na representação binária, caso o termo j esteja presente no documento i , então $x_{ij} = 1$, caso contrário, $x_{ij} = 0$. A segunda representação, denominada como *term-frequency* (*tf*), preenche x_{ij} com o valor da ocorrência absoluta do termo j no documento i . Enquanto essa representação considera apenas a distribuição de frequências em um documento, a representação *tf-idf* (SILVA; PERES; BOSCARIOLI, 2016; BAEZA-YATES; RIBEIRO-NETO, 2011) adiciona a informação da ocorrência de um termo sobre toda coleção. A equação 13 define o cálculo de *idf*, em que j é um termo, n é o número de documentos na coleção e η_j é o número de documentos em que j ocorre.

$$idf(j) = \log \frac{n}{\eta_j} \quad (13)$$

A representação *tf-idf* é a combinação de *tf* com *idf* e, apesar de sua simplicidade, é competitiva a modelos estatísticos mais elaborados (ROBERTSON, 2004). Sendo

³ A lista de *stopwords* utilizada está disponível em: <http://snowball.tartarus.org/algorithms/portuguese/stop.txt>.

$tf(i, j)$ a frequência de um termo j em um documento i , a representação $tf-idf$ é calculada de acordo com a equação 14.

$$tfidf(i, j) = tf(i, j) * idf(j) \quad (14)$$

- **seleção de termos relevantes:** consiste em determinar um procedimento para identificar os termos que mais bem representam os textos da coleção. Neste trabalho, duas técnicas de seleção foram utilizadas: frequência mínima e $tf-idf$. A primeira corresponde à manutenção dos termos que aparecem em ao menos dois documentos. Já a segunda técnica seleciona os q , com $q = \{5, 10, 15, 20\}$, termos com maior escore de $tf-idf$ para representar cada documento.
- **normalização** (opcional): consiste em realizar a normalização euclidiana nas colunas das matrizes com representação vetorial tf e $tf-idf$.

Os parâmetros utilizados na geração das 50 matrizes para este trabalho estão dispostos no arquivo *00 - matriz_geradas.pdf* no Google Drive⁴.

6.3 Configurações de experimento

Para a realização dos experimentos, foram utilizados dois algoritmos de agrupamentos: *k-means++* e *ensemble* de agrupamentos. Ambos algoritmos foram desenvolvidos no ambiente MATLAB, o qual dispõe de diversas funções que auxiliam e facilitam as implementações. Os parâmetros utilizados nos algoritmos são descritos nas subseções seguintes.

6.3.1 *k-means++*

O algoritmo *k-means++* não exige muitos parâmetros em sua execução. Neste caso, são necessários apenas o conjunto de entrada X , o valor de k e a distância para o cálculo da similaridade. Desta forma, foram definidas duas configurações de experimento para esse algoritmo, em que ambas utilizam os 50 conjuntos junto com a distância cosseno, variando apenas o valor de k :

⁴ <https://drive.google.com/drive/folders/1TpyQdcDGMtDjeIGmbnUUn51zEcileCMH?usp=sharing>

- *Experimento #1* ($k = 10$): como o conjunto de notícias está organizado em dez categorias (cf. seção 6.1), a criação de uma partição com dez grupos foi estabelecida para que fosse possível realizar uma avaliação quantitativa, usando dois índices de validação de agrupamento externos, considerando tais categorias como *verdade fundamental*;
- *Experimento #2* ($k = 7$): há uma categorização das notícias que congrega subcategorias de notícias em um único canal (“Notícias” cf. seção 6.1), portanto, como uma segunda análise, um número de grupos reduzido foi aplicado a fim de verificar se os resultados obtidos poderiam trazer resultados aproximados à primeira estratégia de organização de notícias usadas no portal EBC.

6.3.2 *Ensemble* de agrupamentos

Diferentemente do algoritmo *k-means++*, o *ensemble* de agrupamentos exige a determinação de mais parâmetros em sua execução. Os parâmetros são: o conjunto de entrada X , o número de componentes de *ensemble*, o tipo de variação aplicada ao conjunto de entrada para cada componente (amostragem de dados, amostragem de atributos ou uso do conjunto completo), o valor de k para o agrupamento final do *ensemble*, a variação de k em cada componente (caso não aplicável, cada componente utilizará o valor de k definido para o *ensemble*) e o algoritmo utilizado para gerar o agrupamento final por meio da matriz de coassociação (função de consenso aplicada neste trabalho). Os experimentos realizados com esse algoritmo foram divididos em nove combinações de parâmetros, sendo que todos experimentos utilizam a distância cosseno. As configurações dos experimentos executados com o *ensemble* estão organizadas na tabela 11:

Para o *Experimento #6*, uma vez que há variação no valores de k , algumas subconfigurações foram criadas. Outras particularidades encontradas nesse experimento são: a utilização de apenas dois conjuntos de entrada X , sendo um com representação binária e o outro com representação *tf-idf* não normalizado; o uso de 100 componentes de *ensemble*. Subconfigurações foram criadas para o *Experimento #7* devido ao uso de dois procedimentos para geração dos grupos finais. Situações similares geraram as subconfigurações para o *Experimento #8* e para o *Experimento #9*.

Tabela 11 – Configurações usadas nos experimentos com o ensemble de agrupamentos

Exp.	X	# Comp	Variação de X	k	Variação de k	Geração final de grupos
#1	1 à 30	10	Atributos	10	N/A	AGNES- <i>single-link</i>
#2	Todos	10	Atributos	10	Não	AGNES- <i>average-link</i>
#3	Todos	10	Não	10	Não	AGNES- <i>single-link</i>
#4	Todos	10	Não	10	Não	AGNES- <i>average-link</i>
#5	Todos	20	Não	10	Não	AGNES- <i>average-link</i>
#6.1	10 e 47	100	Atributos	7	Não	AGNES- <i>average-link</i>
#6.2	10 e 47	100	Atributos	10	Não	AGNES- <i>average-link</i>
#7.1	Todos	10	Dados	7	Não	AGNES- <i>average-link</i>
#7.2	Todos	10	Dados	7	Não	<i>k-means++</i>
#7.3	Todos	10	Dados	10	Não	AGNES- <i>average-link</i>
#7.4	Todos	10	Dados	10	Não	<i>k-means++</i>
#8.1	Todos	10	Atributos	7	Não	AGNES- <i>average-link</i>
#8.2	Todos	10	Dados	7	Não	AGNES- <i>average-link</i>
#8.3	Todos	10	Atributos	7	Não	<i>k-means++</i>
#8.4	Todos	10	Dados	7	Não	<i>k-means++</i>
#8.5	Todos	10	Atributos	10	Não	AGNES- <i>average-link</i>
#8.6	Todos	10	Dados	10	Não	AGNES- <i>average-link</i>
#8.7	Todos	10	Atributos	10	Não	<i>k-means++</i>
#8.8	Todos	10	Dados	10	Não	<i>k-means++</i>
#9.1	Todos	10	Não	7	Sim	AGNES- <i>average-link</i>
#9.2	Todos	10	Não	7	Sim	<i>k-means++</i>
#9.3	Todos	10	Não	10	Sim	AGNES- <i>average-link</i>
#9.4	Todos	10	Não	10	Sim	<i>k-means++</i>

Fonte: Fernando Henrique da Silva Costa, 2018

6.4 Resultados com *k-means++*

Nesta seção são apresentados os cinco melhores resultados obtidos pelos dois experimentos definidos para o algoritmo de agrupamento *k-means++*. Esses resultados são identificados por identificadores da variação de conjuntos de dados utilizado (X_i) em tabelas nas quais são os maiores valores de média para cada índice de avaliação (I_{ARI} para Índice Rand Ajustado, I_{NMI} para Informação Mútua Normalizada e I_{SIL} para Silhouette). Além da média, sempre calculada sobre dez execuções para cada ID, os resultados também são apresentados em termos de desvio padrão (DP) e coeficiente de variação (CV). Os demais resultados estão apresentados nos arquivos *01 - kmeans_k10.pdf* e *02 - kmeans_k7.pdf* no Google Drive⁵. Para cada caso de análise, nuvens de palavras são apresentadas para ilustrar a categorização semântica dos grupos descobertos.

⁵ <https://drive.google.com/drive/folders/1TpyQdcDGMtDjeIGmbnUUn51zEcileCMH?usp=sharing>

6.4.1 Experimento #1

A tabela 12 apresenta os cinco melhores resultados alcançados no primeiro experimento definido para o algoritmo *k-means++*. Em todas essas execuções, a distância cosseno e $k = 10$ foram utilizados.

Tabela 12 – Qualidade de agrupamento para o experimento #1 do algoritmo *k-means++*, de acordo com índices de validação de agrupamento. Destaque em negrito para o melhor resultado em cada índice

ID	I_{ARI}			I_{NMI}			I_{SIL}		
	Média	DP	CV	Média	DP	CV	Média	DP	CV
2	0,429	0,062	0,144	0,563	0,025	0,045	0,081	0,008	0,015
3	0,400	0,055	0,138	0,520	0,020	0,039	0,085	0,004	0,008
4	0,414	0,047	0,113	0,517	0,015	0,029	0,086	0,005	0,009
8	0,425	0,077	0,182	0,569	0,035	0,062	0,083	0,005	0,008
10	0,428	0,056	0,132	0,578	0,022	0,038	0,082	0,006	0,010

Fonte: Fernando Henrique da Silva Costa, 2018

A figura 18 ilustra a plotagem do índice Silhouette em duas execuções, uma em um experimento com X_4 e outra com X_{10} . Essas matrizes foram escolhidas porque representam os melhores resultados obtidos por I_{SIL} e I_{NMI} conforme visto na tabela 12, respectivamente.

Figura 18 – Plotagem do índice Silhouette em execuções X_4 e X_{10} no experimento #1 do algoritmo *k-means++*

Fonte: Fernando Henrique da Silva Costa, 2018

Na primeira matriz, é possível observar que há grupos bem categorizados, como por exemplo os grupos 5 e 6. No entanto, quando comparado com a segunda matriz, a homogeneidade é menor. Dada as médias de cada matriz, 0,086 e 0,082 respectivamente, é possível observar que na primeira matriz os grupos 5 e 6 possuem altos valores de

Silhouette, enquanto os demais possuem valores menores. Na segunda matriz, os grupos são menores, exceto pelo grupo 7.

A figura 19 mostra nuvens de palavras que descrevem tipos de notícias que são encontradas em cada grupo gerado pelo algoritmo *k-means++*. Para a realização dessas nuvens de palavras, foi utilizada uma execução do resultado obtido no experimento identificado pelo X_{10} , e consideradas as 50 palavras mais frequentes no subconjunto de documentos associado ao grupo.

Figura 19 – Nuvens de palavras em cada grupo obtida em execução sobre o conjunto X_{10} , no experimento #1 com o algoritmo *k-means++*

Fonte: Fernando Henrique da Silva Costa, 2018

Cada nuvem de palavras é criada com respeito a um grupo de notícias. Para sua criação, são consideradas as frequências de cada termo dentro dos documentos que compõem o grupo. Desta forma, quanto maior a frequência do termo em no grupo, maior o seu destaque (tamanho) na nuvem de palavras. Em relação às cores presentes nas nuvens, sua motivação é estritamente estética e determinada pelo *software* utilizado para a geração das nuvens.

Pela caracterização dos grupos via nuvem de palavras, é possível notar que há grupos que possuem termos semelhantes. Por exemplo, os grupos 1, 2 e 6 possuem termos que estão relacionadas à educação. Entretanto, também é observável que cada um desses grupos possui suas particularidades. O primeiro grupo versa, principalmente, sobre educação de nível superior. O segundo grupo disserta sobre educação no Rio de Janeiro. Por fim, o sexto grupo discute sobre educação e política.

Ainda que o terceiro, quarto e quinto grupos versem sobre temas que envolvem política nacional, é possível encontrar algumas particularidades em cada um deles. O terceiro grupo disserta a respeito de questões presidenciais, o quarto grupo discute assuntos sobre segurança e direitos humanos e, por fim, o quinto grupo apresenta conteúdos a respeito de corrupção na política nacional. O oitavo grupo possui termos iguais a esses três grupos, no entanto, as notícias que formam o grupo 8 versam a respeito de política internacional.

Os termos encontrados nos nono e décimo grupos também são semelhantes entre si, e caracterizam notícias sobre eventos culturais. No nono grupo, a localidade do Rio de Janeiro aparece com frequência, podendo indicar que os eventos ocorrem nesta região. No décimo grupo, os eventos são mais abrangentes. O sétimo grupo, embora possua termos em comum com outros grupos, apresenta uma quantidade grande de termos que estão relacionados a esportes.

A figura 20 ilustra a relação entre os cadernos das notícias encontradas em cada grupo de uma execução em X_{10} nesse experimento. Na figura, é possível notar que há grupos que são formados, em sua maioria, por notícias provenientes de apenas um caderno, por exemplo, os grupos 1, 2, 7, 8 e 9. No entanto, a formação dos demais grupos possui um balanceamento entre os cadernos das notícias, com destaque para o grupo 10. Além disso, notícias do caderno *Educação* são encontradas em grande quantidade em diversos grupos. Isso pode indicar que há diversos assuntos descorrelacionados tratados nesse caderno. De forma oposta, notícias do caderno *Esportes* são mais concentradas em um grupo, no caso o grupo 7.

Figura 20 – Relação cadernos de notícias nos grupos obtidos em uma execução sobre o conjunto X_{10} no experimento #1 com o algoritmo $k\text{-means}++$

Fonte: Fernando Henrique da Silva Costa, 2018

6.4.2 Experimento #2

A tabela 13 apresenta os cinco melhores resultados alcançados no segundo experimento definido para o algoritmo $k\text{-means}++$. Em todas essas execuções, a distância cosseno e $k = 7$ foram utilizados.

Tabela 13 – Qualidade de agrupamento para o experimento #2 do algoritmo $k\text{-means}++$, de acordo com índices de validação de agrupamento. Destaque em negrito para o melhor resultado em cada índice

ID	Média	I_{ARI} DP	CV	Média	I_{NMI} DP	CV	Média	I_{SIL} DP	CV
2	0,436	0,062	0,141	0,562	0,028	0,050	0,076	0,006	0,011
8	0,394	0,076	0,193	0,554	0,027	0,048	0,082	0,004	0,008
10	0,369	0,049	0,132	0,544	0,024	0,045	0,082	0,005	0,008
45	0,416	0,065	0,156	0,510	0,033	0,065	0,056	0,003	0,057
49	0,393	0,068	0,173	0,529	0,033	0,063	0,057	0,003	0,054

Fonte: Fernando Henrique da Silva Costa, 2018

A figura 21 ilustra a plotagem do índice Silhouette em duas execuções com as matrizes X_8 e X_2 . A primeira matriz alcançou, como média, o valor 0,082, enquanto a segunda obteve 0,076. Os grupos 4 e 5 da primeira matriz são mais bem categorizados que os demais. No entanto, em termos de homogeneidade, a segunda matriz é melhor que a primeira. Isso ocorre, porque a maior parte dos grupos são formados por valores de Silhouette semelhantes e mais próximos a média, enquanto na segunda matriz há grupos com valores mais altos e outros com valores menores.

Figura 21 – Plotagem do índice Silhouette em execuções X_8 e X_2 no experimento #2 do algoritmo $k\text{-means}++$

Fonte: Fernando Henrique da Silva Costa, 2018

A figura 22 ilustra as nuvens de palavras que descrevem tipos de notícias que são encontradas em cada grupo gerado pelo algoritmo $k\text{-means}++$. Para a realização dessas nuvens de palavras, foi utilizada uma execução obtida no experimento identificado por X_2 .

Figura 22 – Nuvens de palavras em cada grupo obtidas em uma execução sobre o conjunto X_2 no experimento #2 com o algoritmo $k\text{-means}++$

Fonte: Fernando Henrique da Silva Costa, 2018

Devido ao número de grupos encontrados neste experimento ser menor quando comparado ao experimento anterior, os grupos são mais concentrados em assuntos específicos. No entanto, ainda é possível encontrar semelhanças entre grupos. Por exemplo, os grupos 4 e 5 são formados por notícias que dissertam sobre educação, no entanto, o quinto grupo versa sobre educação no Rio de Janeiro, enquanto o quarto é sobre educação em geral.

O primeiro grupo é formado por notícias que versam sobre esportes, enquanto o sétimo é composto por notícias a respeito de eventos culturais. Por fim, os grupos 2, 3 e 6 são compostos por notícias que relatam acontecimentos a respeito de política, porém, é possível encontrar algumas particularidades. O segundo grupo possui notícias que envolvem política internacional, o terceiro é formado por notícias que dissertam sobre investimentos e, por fim, o sexto grupo é composto por notícias mais gerais.

A figura 23 ilustra a relação dos cadernos de notícias encontrados em cada grupo gerado por uma execução em X_2 . Semelhantemente à figura 20, há grupos, que em sua maioria, são formados por notícias provenientes de apenas um caderno. Exemplos são os grupos 1, 2, 4, 5 e 7. Além disso, também é possível verificar que o caderno *Educação* está presente em grande quantidade em diversos grupos, enquanto o caderno *Esportes* concentra-se somente em um grupo. Tais fatores podem indicar que, enquanto o primeiro caderno possui diversos assuntos descorrelacionados, o segundo é composto por notícias que dissertam sobre assuntos semelhantes ou iguais.

Figura 23 – Relação cadernos de notícias nos grupos obtidos em uma execução sobre o conjunto X_2 no experimento #2 com o algoritmo *k-means++*

Fonte: Fernando Henrique da Silva Costa, 2018

6.5 Resultados com ensemble de agrupamentos

Nesta seção são apresentados os cinco melhores resultados obtidos pelos experimento com o *ensemble* de agrupamento. O experimento #6 apresenta todos resultados alcançados, uma vez que foram realizados apenas quatro experimentos.

Esses resultados são identificados por identificadores da variação de conjuntos de dados utilizado (X_i) em tabelas nas quais são os maiores valores de média para cada índice de avaliação (I_{ARI} para Índice Rand Ajustado, I_{NMI} para Informação Mútua Normalizada e I_{SIL} para Silhouette). Além da média, sempre calculada sobre dez execuções para cada ID , os resultados também são apresentados em termos de desvio padrão (DP) e coeficiente de variação (CV). Os demais resultados estão apresentados nos arquivos 03 - *ensemble_exp01.pdf*, 04 - *ensemble_exp02.pdf*, 05 - *ensemble_exp03.pdf*, 06 - *ensemble_exp04.pdf*, 07 - *ensemble_exp05.pdf*, 08 - *ensemble_exp07.pdf*, 09 - *ensemble_exp08.pdf* e 10 - *ensemble_exp09.pdf* no Google Drive⁶.

6.5.1 Experimento #1

A tabela 14 apresenta os cinco melhores resultados alcançados pelo primeiro experimento do algoritmo *ensemble* de agrupamentos. Em todas essas execuções, a distância cosseno, $p = 10$, $k = 10$, o algoritmo *k-means++* para gerar os componentes, o algoritmo AGNES com *single-link* e amostragem de atributos foram utilizados.

Tabela 14 – Qualidade de agrupamento para o experimento #1 do algoritmo *ensemble* de agrupamentos, de acordo com índices de validação de agrupamento. Destaque em negrito para o melhor resultado em cada índice

ID	I_{ARI}			I_{NMI}			I_{SIL}		
	Média	DP	CV	Média	DP	CV	Média	DP	CV
16	0,115	0,002	0,015	0,343	0,003	0,010	0,014	0,002	0,006
18	0,115	0,002	0,015	0,345	0,003	0,009	0,017	0,002	0,007
20	0,114	0,002	0,022	0,351	0,019	0,055	0,018	0,002	0,006
26	0,116	0,002	0,014	0,344	0,003	0,009	0,014	0,002	0,007
30	0,115	0,002	0,016	0,352	0,020	0,055	0,018	0,002	0,006

Fonte: Fernando Henrique da Silva Costa, 2018

A figura 24 ilustra a plotagem dos valores de Silhouette em uma execução na matriz X_{30} . É possível verificar pouca homogeneidade entre os grupos, visto que a média foi 0,018. Isso ocorre, pois há grupos que possuem resultados altos e outros que obtiveram resultados baixos, inclusive totalmente negativos, como por exemplo, o grupo 2.

Em relação a diversidade obtida nas partições geradas pelos componentes e *ensemble* de agrupamentos, conforme equações apresentadas na seção 3.1, os valores são: $D_1 = 0,588$,

⁶ <https://drive.google.com/drive/folders/1TpyQdcDGMtDjeIGmbnUUn51zEcileCMH?usp=sharing>

Figura 24 – Plotagem do índice Silhouette em uma execução X_{30} no experimento #1 do algoritmo *ensemble* de agrupamentos

Fonte: Fernando Henrique da Silva Costa, 2018

$D_2 = 0,130$, $D_3 = 0,270$ e $D_4 = 0,221$. Os valores indicam que os agrupamentos gerados possuem uma quantidade regular de diversidade.

A figura 25 ilustra as nuvens de palavras geradas em cada grupo de uma execução do experimento identificado *ID26*. Os grupos gerados nessa execução são desbalanceados. Por exemplo, os grupos 1, 3, 6, 7 e 9 são formados por apenas uma notícia, enquanto os grupos 2, 5 e 8 são formados por 823, 201 e 122 notícias, respectivamente. Por fim, o grupo 4 é formado por duas notícias e o décimo grupo por quatro notícias. Os valores baixos alcançados pelos índices de avaliação de agrupamento podem ser decorrentes da formação de grupos desbalanceados.

Devido ao desbalanceamento encontrado nesse agrupamento, os grupos com maior número de notícias abrangem diversos conteúdos. No entanto, é possível observar que os termos mais relevantes do segundo grupo são referentes a política nacional. No quinto grupo, os termos mostram que as notícias possuem conteúdo de esporte, como também é observado no quarto grupo que possui poucas notícias. O oitavo grupo é formado por notícias que dissertam a respeito de educação, especificamente, de questões que envolvem o ensino superior.

As notícias que formam os primeiro, terceiro e nono grupos são provenientes do caderno de Educação. A notícia que é encontrada no sexto grupo é classificada como Política, enquanto o sétimo grupo é formado por uma notícia do caderno Infantil. Por fim, o décimo grupo abrange notícias que versam sobre a mega-sena.

Figura 25 – Nuvens de palavras em cada grupo obtidas em uma execução sobre o conjunto X_{30} no experimento #1 com o algoritmo *ensemble* de agrupamentos

Fonte: Fernando Henrique da Silva Costa, 2018

A figura 26 ilustra a relação dos cadernos presentes em cada grupo gerado por uma execução em X_{30} . O gráfico apresenta o desbalanceamento encontrado entre os grupos, uma vez que o grupo 2 possui a maioria das notícias do corpus. Devido a isso, esse grupo reúne notícias em grande quantidade de quase todos os grupos. Além disso, os grupos 5 e 8 são formados, em sua maioria, por notícias provenientes dos cadernos *Esportes* e *Educação*. Em relação ao primeiro caderno, isso pode indicar que as notícias são semelhantes, enquanto o segundo caderno pode indicar que os assuntos tratados são desconcorrelacionados, uma vez que há uma grande quantidade de notícias no grupo 2.

6.5.2 Experimento #2

A tabela 15 apresenta os cinco melhores resultados alcançados pelo segundo experimento do algoritmo *ensemble* de agrupamentos. Em todas essas execuções, a distância cosseno, $p = 10$, $k = 10$, o algoritmo *k-means++* para gerar os componentes, o algoritmo

Figura 26 – Relação cadernos de notícias nos grupos obtidos em uma execução sobre o conjunto X_{30} no experimento #1 com o algoritmo *ensemble* de agrupamentos

Fonte: Fernando Henrique da Silva Costa, 2018

AGNES com *average-link* para gerar o resultado final e a amostragem de atributos foram utilizados.

Tabela 15 – Qualidade de agrupamento para o experimento #2 do algoritmo *ensemble* de agrupamentos, de acordo com índices de validação de agrupamento. Destaque em negrito para o melhor resultado em cada índice

ID	I_{ARI}			I_{NMI}			I_{SIL}		
	Média	DP	CV	Média	DP	CV	Média	DP	CV
2	0,446	0,051	0,114	0,586	0,017	0,029	0,076	0,010	0,125
8	0,467	0,058	0,124	0,586	0,019	0,033	0,079	0,008	0,097
10	0,492	0,030	0,061	0,603	0,009	0,014	0,072	0,006	0,087
47	0,519	0,042	0,081	0,601	0,015	0,025	0,055	0,005	0,099
49	0,510	0,026	0,050	0,585	0,020	0,034	0,054	0,002	0,034

Fonte: Fernando Henrique da Silva Costa, 2018

Em comparação com o experimento #1 do algoritmo *ensemble* de agrupamentos, os resultados nos índices I_{ARI} e I_{NMI} desse experimento são consideravelmente superiores. A diferença desse para o experimento anterior é a estratégia de *linkage* utilizada. Enquanto no primeiro, foi utilizado o *single-link*, esse aplicou o *average-link*.

A figura 27 ilustra a plotagem do índice Silhouette em uma execução nas matrizes X_8 e X_{47} . A primeira matriz representa o melhor resultado alcançado por I_{SIL} , a segunda representa o melhor resultado obtido por I_{NMI} . Os grupos 3 e 4 da primeira matriz são bem categorizadas em comparação aos demais grupos, inclusive da segunda matriz. Tendo em vista os valores médios de Silhouette apresentados na tabela 15, a segunda matriz é

mais homogênea quando comparada com a primeira. Isso ocorre, pois na primeira matriz, dois grupos possuem valores altos de Silhouette, enquanto os demais possuem valores menores.

Figura 27 – Plotagem do índice Silhouette em uma execução X_{47} no experimento #2 do algoritmo *ensemble* de agrupamentos

Fonte: Fernando Henrique da Silva Costa, 2018

Em relação à diversidade obtida em uma execução em X_{47} , os valores obtidos foram $D_1 = 0,547$, $D_2 = 0,0827$, $D_3 = 0,267$ e $D_4 = 0,151$. Tais valores indicam uma regularidade no valor de diversidade nas partições obtidas entre os componentes do *ensemble* e o *ensemble* de agrupamentos.

A figura 28 ilustra as nuvens de palavras geradas para cada grupo de uma execução do ID_{47} . Ainda que em menor escala quando comparada com o experimento anterior, há um desbalanceamento na formação dos grupos, em que os grupos 2 e 3 possuem apenas quatro e três notícias, respectivamente.

As quatro notícias que compõem o segundo grupo são provenientes de dois cadernos, uma de Geral e as demais de Tecnologia. O terceiro grupo reúne três notícias do caderno Geral, ainda que duas relatam sobre acontecimentos no estádio São Januário do Vasco da Gama e uma sobre a violência na cidade do Rio de Janeiro.

Os termos nas nuvens de palavras dos demais grupos apresentam maior consistência. Por exemplo, os termos mais relevantes do primeiro grupo indicam que grande parte das notícias presentes nesse grupo são referentes a esportes, enquanto os termos do quarto grupo indicam que as notícias versam a respeito de política nacional.

Conforme apresentado nas nuvens de palavras, os termos mais relevantes do quinto grupo indicam que a maioria das notícias versam sobre política internacional. Sob a mesma

Figura 28 – Nuvens de palavras em cada grupo obtidas em uma execução sobre o conjunto X_{47} no experimento #2 com o algoritmo *ensemble de agrupamentos*

Fonte: Fernando Henrique da Silva Costa, 2018

ótica, o sexto grupo é formado por notícias que, em grande parte, relatam sobre segurança e violência. Os termos encontrados nos sétimo e décimo grupos são semelhantes, dissertando sobre educação, no entanto o sétimo trata de assuntos mais gerais, enquanto o décimo discute em um nível mais específico, nesse caso, no ensino superior. Por fim, os oitavo e nono grupos são formados por notícias, em grande parte, que relatam sobre cultura e tecnologia, respectivamente.

Este experimento apresentou valores melhores para os índices de avaliação de agrupamento, quando comparado ao anterior. Tais valores são sustentados pela análise da formação dos grupos, ainda que haja um pequeno desbalanceamento na quantidade de notícias por grupo. Dessa maneira, considerando o contexto destes dois experimentos, a utilização do algoritmo AGNES com estratégia *average-link* mostrou ser mais eficiente à utilização de Agnes com *single-link*.

A figura 29 ilustra a relação dos cadernos de notícias presentes nos grupos gerados em uma execução sobre o conjunto X_{47} . Por meio da figura, é possível verificar que

determinados grupos são formados, em sua maioria, por notícias provenientes de apenas um caderno. Exemplos são os grupos 1, 7, 8, 9 e 10. Em relação ao caderno *Educação*, ele aparece em diversos grupos com um certo destaque. Antagonicamente, o caderno *Esportes* aparece em destaque em somente um grupo, no caso, o grupo 1.

Figura 29 – Relação cadernos de notícias nos grupos obtidos em uma execução sobre o conjunto X_{47} no experimento #2 com o algoritmo *ensemble* de agrupamentos

Fonte: Fernando Henrique da Silva Costa, 2018

6.5.3 Experimento #3

Na tabela 16 apresenta os cinco melhores resultados alcançados pelo terceiro experimento do algoritmo *ensemble* de agrupamentos. Em todas essas execuções a distância cosseno, $p = 10$, $k = 10$, o algoritmo *k-means++* para gerar os componentes, o algoritmo AGNES com *single-link* para gerar o resultado final e o conjunto original para cada componente foram utilizados.

Em comparação ao primeiro experimento no algoritmo *ensemble* de agrupamentos, os resultados obtidos em I_{NMI} são superiores. No entanto, são inferiores quando comparados ao segundo experimento.

A figura 30 ilustra a plotagem dos índices Silhouette em uma execução nas matrizes X_{42} e X_8 , respectivamente. Dado os resultados observados na tabela 16, em ambas matrizes houve pouca homogeneidade. No entanto, ainda é possível encontrar grupos com uma boa categorização. Exemplos são os grupos 6, 7 e 8 da primeira matriz.

Tabela 16 – Qualidade de agrupamento para o experimento #3 do algoritmo *ensemble* de agrupamentos, de acordo com índices de validação de agrupamento. Destaque em negrito para o melhor resultado em cada índice

ID	Média	I_{ARI} DP	CV	Média	I_{NMI} DP	CV	Média	I_{SIL} DP	CV
3	0,152	0,065	0,426	0,394	0,065	0,166	0,022	0,014	0,630
8	0,170	0,068	0,404	0,441	0,078	0,176	0,037	0,024	0,639
42	0,124	0,028	0,228	0,407	0,037	0,090	0,056	0,010	0,175
48	0,123	0,022	0,178	0,395	0,041	0,104	0,033	0,027	0,817
50	0,121	0,031	0,253	0,399	0,047	0,118	0,054	0,012	0,223

Fonte: Fernando Henrique da Silva Costa, 2018

Figura 30 – Plotagem do índice Silhouette em execuções na matrizes X_{42} e X_8 no experimento #3 do algoritmo *ensemble* de agrupamentos

Fonte: Fernando Henrique da Silva Costa, 2018

Em relação à diversidade nas partições obtidas entre os componentes do *ensemble* e o *ensemble* de agrupamentos, os valores são: $D_1 = 0,382$, $D_2 = 0,154$, $D_3 = 0,386$ e $D_4 = 0,404$. Esses valores apresentam baixo valor de diversidade entre essas partições. Isso ocorre, pois para esse experimento não foi utilizada técnica de amostragem de dados ou atributos.

A figura 31 apresenta as nuvens de palavras de cada grupo em uma execução no *ID8*. O resultado desse agrupamento trouxe grupos desbalanceados semelhantemente ao *Experimento #1*. Os terceiro, quarto, oitavo e nono grupos possuem apenas uma notícia e o sétimo possui apenas três. As outras notícias são divididas da seguinte forma: 90 para o primeiro grupo, 648 no segundo grupo, 126 no quinto grupo, 82 no sexto grupo e 204 no décimo grupo.

Dado os termos mais relevantes apresentados nas nuvens de palavras, os dois primeiros grupos são compostos por notícias que envolvem política nacional. De forma

Figura 31 – Nuvens de palavras em cada grupo obtidas em uma execução sobre o conjunto X_8 no experimento #3 com o algoritmo *ensemble* de agrupamentos

Fonte: Fernando Henrique da Silva Costa, 2018

similar, os quinto e sexto grupo são formados por notícias que dissertam a respeito de educação, sendo diferenciados apenas pelo contexto. O quinto grupo versa sobre notícias de educação no nível superior e afins, enquanto o sexto grupo trata de assuntos relacionados à educação no Rio de Janeiro. Por fim, o décimo grupo é formado por notícias que argumentam sobre esportes.

Os terceiro e quarto grupos são formados por apenas uma notícia do caderno *Tecnologia*. As três notícias que compõem o sétimo grupo são do caderno *Cidadania*. O oitavo grupo é formado por apenas uma notícia proveniente do caderno *Política*. Por fim, o nono grupo é formado por uma notícia do caderno *Cultura*.

A figura 32 ilustra a relação de cadernos de notícias encontradas em cada grupo de uma execução em X_8 . Devido ao desbalanceamento nesse experimento, o segundo grupo, o qual é composto pela maioria das notícias do corpus, possui notícias de todos os cadernos, exceto por *Esportes*. As notícias desse cadernos são encontradas no último grupo.

Os grupos 5 e 6 são compostos, em sua maioria, por notícias provenientes do caderno *Educação*, o que pode indicar assuntos desconcorrelacionados dentro desse caderno.

Figura 32 – Relação cadernos de notícias nos grupos obtidos em uma execução sobre o conjunto X_8 no experimento #3 com o algoritmo *ensemble* de agrupamentos

Fonte: Fernando Henrique da Silva Costa, 2018

6.5.4 Experimento #4

Na tabela 17 apresenta os cinco melhores resultados alcançados pelo quarto experimento do algoritmo *ensemble* de agrupamentos. Em todas essas execuções a distância cosseno, $p = 10$, $k = 10$, o algoritmo *k-means++* para gerar os componentes, o algoritmo AGNES com *average-link* para gerar o resultado final e o conjunto original para cada componente foram utilizados.

Tabela 17 – Qualidade de agrupamento para o experimento #4 do algoritmo *ensemble* de agrupamentos, de acordo com índices de validação de agrupamento. Destaque em negrito para o melhor resultado em cada índice

ID	I_{ARI}			I_{NMI}			I_{SIL}		
	Média	DP	CV	Média	DP	CV	Média	DP	CV
2	0,476	0,027	0,057	0,588	0,006	0,009	0,068	0,006	0,083
8	0,503	0,031	0,061	0,595	0,008	0,014	0,077	0,005	0,070
10	0,481	0,044	0,091	0,594	0,012	0,020	0,074	0,007	0,092
47	0,530	0,028	0,052	0,600	0,013	0,022	0,058	0,002	0,027
49	0,514	0,035	0,067	0,589	0,025	0,042	0,056	0,004	0,070

Fonte: Fernando Henrique da Silva Costa, 2018

Os resultados alcançados nesse experimento, em termos de I_{ARI} e I_{NMI} , são consideravelmente superiores ao experimento anterior. A diferença entre eles é que o anterior utiliza a estratégia *single-link* no AGNES, enquanto esse aplica o *average-link*.

A figura 33 ilustra as plotagens de Silhouette em duas execuções, uma na matriz X_8 e outra na matriz X_{47} . Os grupos 1 e 2 da primeira matriz são bem categorizados quando comparados aos demais grupos, inclusive da segunda matriz.

Figura 33 – Plotagem do índice Silhouette em execuções nas matrizes X_8 e X_{47} no experimento #4 do algoritmo *ensemble* de agrupamentos

Fonte: Fernando Henrique da Silva Costa, 2018

Em relação à diversidade das partições obtidas pelos componentes do *ensemble* e o *ensemble* de agrupamentos em uma execução em X_{47} , os valores alcançados foram: $D_1 = 0,446$, $D_2 = 0,105$, $D_3 = 0,329$ e $D_4 = 0,236$. Ainda que não foi aplicada amostragem de dados ou atributos, os valores alcançados são regulares e levemente superiores quando comparado ao experimento anterior.

A figura 34 apresenta as nuvens de palavras geradas em cada grupo em uma execução em $ID47$. Diferentemente dos experimentos anteriores, os grupos gerados neste experimento é balanceado. No entanto, o terceiro grupo possui apenas 13 notícias. As outras notícias estão dispostas da seguinte forma: 68 no primeiro grupo, 63 no segundo grupo, 159 no quarto grupo, 141 no quinto grupo, 74 no sexto grupo, 190 no sétimo grupo, 123 no oitavo grupo, 121 no nono grupo e, por fim, 205 no décimo grupo.

Indo em conformidades com os termos mais relevantes apresentados nas nuvens de palavras, os dois primeiros grupos tratam de assuntos políticos nacionais. No entanto, o primeiro disserta particularmente sobre casos de corrupção, enquanto o segundo trata

Figura 34 – Nuvens de palavras em cada grupo obtidas em uma execução sobre o conjunto X_{47} no experimento #4 com o algoritmo *ensemble* de agrupamentos

Fonte: Fernando Henrique da Silva Costa, 2018

de assuntos que envolvem leis e votação. O terceiro grupo é formado por notícias que discutem, em sua maioria, sobre acontecimentos que envolvem o povo indígena no Brasil.

O quarto grupo é formado por notícias que dissertam a respeito de assuntos culturais e o quinto é composto por notícias que versam sobre tecnologia. O sexto grupo contém notícias que apresentam conteúdos a respeito de política internacional. O sétimo grupo apresenta notícias que versam sobre direitos, segurança e violência.

Os oitavo e nono grupos são semelhantes, uma vez que discutem sobre educação. No entanto, há particularidades em cada grupo. Enquanto no oitavo grupo, as notícias apresentam conteúdo sobre acontecimentos envolvendo ensino superior e afins, o nono grupo contém as notícias sobre a educação no Rio de Janeiro. Por fim, o décimo grupo é formado por notícias que versam a respeito de esportes.

A figura 35 ilustra a relação dos cadernos encontrados em cada grupo de uma execução no conjunto X_{47} nesse experimento. É possível observar que os grupos 1, 3, 4, 5, 6, 8, 9 e 10 são formados, em sua maioria, por notícias de apenas um caderno. Destacam-se

os grupos 8 e 9 que apresentam em maior quantidade notícias do caderno *Educação*. Essa situação pode evidenciar assuntos desconcorrelacionados encontrados nesse caderno. Além disso, notícias desse caderno também são encontradas em outros grupos com um certo destaque. O caderno *Esportes* concentra-se no décimo grupo, indicando que as notícias dissertam a respeito de conteúdos semelhantes.

Figura 35 – Relação cadernos de notícias nos grupos obtidos em uma execução sobre o conjunto X_{47} no experimento #4 com o algoritmo *ensemble* de agrupamentos

Fonte: Fernando Henrique da Silva Costa, 2018

6.5.5 Experimento #5

Na tabela 18 apresenta os cinco melhores resultados alcançados pelo quinto experimento do algoritmo *ensemble* de agrupamentos. Em todas essas execuções a distância cosseno, $p = 20$, $k = 10$, o algoritmo *k-means++* para gerar os componentes, o algoritmo AGNES com *average-link* para gerar o resultado final e o conjunto original para cada componente foram utilizados.

Os resultados alcançados nesse experimento são semelhantes aos resultados obtidos no experimento anterior. A diferença entre tais experimentos são a quantidade de componentes de *ensemble*. Enquanto no anterior é $p = 10$, nesse é $p = 20$.

A figura 36 ilustra as plotagens de Silhouette em duas execuções, uma na matriz X_8 e outra em X_{49} . Os dois primeiros grupos obtidos na primeira matriz são bem categorizados, quando comparados com os outros grupos, inclusive os da segunda matriz. No entanto, os

Tabela 18 – Qualidade de agrupamento para o experimento #5 do algoritmo *ensemble* de agrupamentos, de acordo com índices de validação de agrupamento. Destaque em negrito para o melhor resultado em cada índice

ID	I_{ARI}			I_{NMI}			I_{SIL}		
	Média	DP	CV	Média	DP	CV	Média	DP	CV
8	0,496	0,040	0,080	0,597	0,011	0,018	0,078	0,005	0,058
10	0,487	0,035	0,072	0,599	0,009	0,015	0,072	0,009	0,119
41	0,503	0,053	0,105	0,594	0,023	0,038	0,059	0,005	0,092
47	0,523	0,033	0,063	0,603	0,017	0,028	0,059	0,004	0,069
49	0,531	0,051	0,096	0,609	0,027	0,044	0,056	0,004	0,070

Fonte: Fernando Henrique da Silva Costa, 2018

grupos da segunda matriz são mais homogêneos, quando comparados com os grupos da primeira matriz.

Figura 36 – Plotagem do índice Silhouette em execuções em X_8 e X_{49} no experimento #5 do algoritmo *ensemble* de agrupamentos

Fonte: Fernando Henrique da Silva Costa, 2018

Em relação à diversidade entre as partições obtidas nos componentes do *ensemble* e do *ensemble* de agrupamentos em uma execução no conjunto X_{49} , os valores alcançados foram: $D_1 = 0,512$, $D_2 = 0,079$, $D_3 = 0,283$ e $D_4 = 0,155$. Esses valores são levemente superiores aos obtidos no experimento anterior. Ainda que não haja amostragem de dados ou atributos, os valores são regulares.

A figura 37 ilustra as nuvens de palavras obtidas em cada grupo de uma execução em $ID49$. Os segundo e terceiros grupos são desbalanceados em relação aos demais, uma vez que são compostos por 6 e 2 notícias, respectivamente. Os demais grupos são formados da seguinte forma: 170 notícias no primeiro grupo, 130 no quarto grupo, 151 no quinto grupo, 164 no sexto grupo, 84 no sétimo grupo, 127 no oitavo grupo, 118 no nono grupo e 205 no décimo grupo.

Figura 37 – Nuvens de palavras em cada grupo obtidas em uma execução sobre o conjunto X_{49} no experimento #5 com o algoritmo ensemble de agrupamentos

Fonte: Fernando Henrique da Silva Costa, 2018

De acordo com os termos mais relevantes apresentados em cada grupo nas nuvens de palavras, o primeiro grupo é formado por notícias que envolvem assuntos sobre direitos, segurança e violência. O segundo grupo é composto por apenas seis notícias que versam sobre os direitos dos transexuais. O terceiro grupo reúne notícias que versam sobre direitos previdenciários, sendo composto por duas notícias.

O quarto grupo é composto por notícias que dissertam a respeito de política nacional, enquanto o quinto grupo é formado por notícias sobre tecnologia e ciência. O sexto grupo contém notícias sobre cultura, principalmente indígena. O sétimo grupo apresenta notícias sobre política internacional.

Os oitavo e nono grupos são formados por notícias de educação, no entanto o oitavo versa a respeito de acontecimentos no Rio de Janeiro, enquanto o nono versa sobre ensino superior e afins. Por fim, o décimo grupo reúne notícias de cunho esportivo.

A figura 38 ilustra a relação dos cadernos encontrados em cada grupo do agrupamento obtido por uma execução no conjunto X_{49} . Os grupos 5, 6, 7, 8, 9 e 10 são

formados por notícias que, em sua maioria, pertencentes a um caderno. Semelhantemente aos experimentos anteriores, notícias do caderno *Educação* estão dispostos em destaque em diversos grupos, principalmente os grupos 8 e 9, enquanto notícias do caderno *Esportes* estão concentrados em um único grupo, nesse caso o décimo.

Figura 38 – Relação cadernos de notícias nos grupos obtidos em uma execução sobre o conjunto X_{49} no experimento #5 com o algoritmo *ensemble* de agrupamentos

Fonte: Fernando Henrique da Silva Costa, 2018

6.5.6 Experimento #6

A tabela 19 apresenta todos resultados alcançados pelo sexto experimento do algoritmo *ensemble* de agrupamentos. Isso ocorre, pois foram utilizados apenas os conjuntos X_{10} e X_{47} nessa parametrização. Os demais parâmetros são: $p = 100$, $k = 7$ ou $k = 10$, o algoritmo *k-means++* para as gerações dos componentes e do resultado final e amostragem de atributos.

Tabela 19 – Qualidade de agrupamento para o experimento #6 do algoritmo *ensemble* de agrupamentos, de acordo com índices de validação de agrupamento. Destaque em negrito para o melhor resultado em cada índice

ID	Média	I_{ARI} DP	CV	Média	I_{NMI} DP	CV	Média	I_{SIL} DP	CV
1	0,424	0,064	0,151	0,564	0,012	0,021	0,077	0,005	0,060
2	0,432	0,045	0,103	0,587	0,014	0,024	0,081	0,007	0,083
3	0,432	0,070	0,162	0,553	0,025	0,046	0,057	0,004	0,071
4	0,487	0,036	0,074	0,594	0,011	0,018	0,063	0,004	0,070

Fonte: Fernando Henrique da Silva Costa, 2018

Os valores alcançados nesse experimento não apresentam melhores resultados quando comparados com os anteriores. Ainda que o *ensemble* utilizado nesse experimento possuísse um maior número de componentes, os resultados são semelhantes a experimentos anteriores.

A figura 39 ilustra as plotagens de Silhouette de duas execuções, uma no conjunto X_{10} e outra no conjunto X_{47} . Os grupos gerados pelo segundo conjunto são mais homogêneos quando comparados com os grupos do primeiro conjunto. No entanto, os grupos 5 e 6 do primeiro conjunto são bem categorizados, quando comparados com os demais grupos.

Figura 39 – Plotagem do índice Silhouette em uma execução X_{47} no experimento #6 do algoritmo *ensemble* de agrupamentos

Fonte: Fernando Henrique da Silva Costa, 2018

Os valores alcançados no cálculo da diversidade das partições obtidas pelos componentes do *ensemble* e do *ensemble* de agrupamentos em uma execução no conjunto X_{47} foram: $D_1 = 0,568$, $D_2 = 0,076$, $D_3 = 0,253$ e $D_4 = 0,134$. Esses valores indicam uma regularidade no nível de diversidade entre essas partições.

A figura 40 ilustra as nuvens de palavras em cada grupo em uma execução em *ID4*. Os grupos gerados neste experimento são balanceados em comparação aos anteriores. A quantidade de notícias por grupo é: 166 no primeiro grupo, 67 no segundo, 132 no terceiro, 205 no quarto, 162 no quinto, 121 no sexto, 116 no sétimo, 76 no oitavo, 80 no nono e, por fim, 32 no décimo.

Indo em conformidade com os termos mais relevantes ilustrados nas nuvens de palavras, é possível observar que no primeiro grupo as notícias que o compõe versam sobre direitos, segurança e violência. O segundo grupo reúne notícias que dissertam sobre política nacional e o terceiro discute a respeito de assuntos culturais.

Figura 40 – Nuvens de palavras em cada grupo obtidas em uma execução sobre o conjunto X_{47} no experimento #6 com o algoritmo *ensemble* de agrupamentos

Fonte: Fernando Henrique da Silva Costa, 2018

O quarto grupo é formado por notícias que envolvem assuntos esportivos, enquanto o quinto grupo é composto por notícias que tratam de conteúdos sobre ciência e tecnologia. Os sexto e sétimos grupos são formados por notícias de educação, porém divididos por suas particularidades. O sexto trata de assuntos mais especificamente de ensino superior e afins, enquanto o sétimo está mais relacionado ao Rio de Janeiro.

O oitavo grupo é formado majoritariamente por notícias que envolvem assuntos políticos internacionais. Semelhante ao segundo grupo, o nono versa sobre política nacional, no entanto, particularmente em assuntos sobre corrupção e justiça. Por fim, o décimo grupo trata de assuntos culturais.

A figura 41 ilustra a relação de cadernos encontrados em cada grupo em uma execução em X_{47} nesse experimento. Por meio da figura, é possível observar que determinados grupos, como o 3, 4, 6, 7 e 8, são formados, em sua maioria, por notícia de um caderno. Os grupos 6 e 7 são compostos em grande por notícias do caderno de *Educação*, enquanto o quarto grupo é composto por notícias de *Esportes*.

Figura 41 – Relação cadernos de notícias nos grupos obtidos em uma execução sobre o conjunto X_{47} no experimento #6 com o algoritmo *ensemble* de agrupamentos

Fonte: Fernando Henrique da Silva Costa, 2018

6.5.7 Experimento #7

A tabela 20 apresenta os melhores resultados alcançados pelo sétimo experimento do algoritmo *ensemble* de agrupamentos. Em todas essas execuções a distância cosseno, $p = 10$, $k = 10$, o algoritmo *k-means++* para gerar os componentes, o algoritmo AGNES com *average-link* para gerar o resultado final e amostragem de dados foram utilizados.

Tabela 20 – Qualidade de agrupamento para o experimento #7 do algoritmo *ensemble* de agrupamentos, de acordo com índices de validação de agrupamento. Destaque em negrito para o melhor resultado em cada índice

ID	I_{ARI}			I_{NMI}			I_{SIL}		
	Média	DP	CV	Média	DP	CV	Média	DP	CV
36	0,001	0,003	4,339	0,026	0,004	0,136	0,063	0,000	0,000
48	-0,001	0,002	-2,296	0,027	0,005	0,177	0,026	0,000	0,000
84	0,000	0,002	3,862	0,027	0,003	0,102	0,026	0,000	0,000
132	0,000	0,002	-4,321	0,028	0,005	0,179	0,081	0,000	0,000
180	-0,001	0,002	-2,316	0,027	0,004	0,146	0,05	0,000	0,000

Fonte: Fernando Henrique da Silva Costa, 2018

Os resultados alcançados nesse experimento, em termos de I_{ARI} e I_{NMI} , foram significativamente inferiores aos demais, inclusive aos experimentos realizados com o algoritmo *k-means++*.

A figura 42 ilustra a plotagem de Silhouette em uma execução no conjunto X_{33} . A maioria dos dados que formam os dez grupos obtiveram valores negativos nessa execução. Tais valores indicam grupos com desbalanceamentos consideráveis.

Figura 42 – Plotagem do índice Silhouette em uma execução X_{33} no experimento #7 do algoritmo *ensemble* de agrupamentos

Fonte: Fernando Henrique da Silva Costa, 2018

Em relação à diversidade das partições obtidas pelos componentes e *ensemble* de agrupamentos em uma execução de $ID132$, a qual utiliza o conjunto X_{33} , os valores alcançados foram: $D_1 = 1,000$, $D_2 = 0,007$, $D_3 = 0,003$ e $D_4 = 0,007$. Esse experimento alcançou os maiores valores de diversidade entre todos experimentos.

A figura 43 ilustra as nuvens de palavras de cada grupo em uma execução no $ID132$. A formação dos grupos neste experimento é bastante desbalanceada, pois o quarto grupo reúne 1037 notícias. Os outros grupos são formados pelo restante das notícias, divididas em: 19 para o primeiro grupo, 10 para o segundo, 11 para o terceiro, 10 para o quinto, 33 para o sexto, 8 para o sétimo, 14 para o oitavo, 12 para o nono e 3 para o décimo grupo.

Devido ao desequilíbrio considerável encontrado no agrupamento final, a formação dos grupos não contém uma identidade de assuntos tratados. Por exemplo, no primeiro grupo é possível encontrar termos relevantes que tratam de assuntos sobre violência, política e educação. De forma semelhante, o segundo grupo possui termos que indicam notícias sobre esportes, cultura e educação.

O décimo grupo, o qual é formado por três notícias, exemplifica melhor esse desequilíbrio. As notícias que o compõe são: “*PGR marca posse de Raquel Dodge para 18 de setembro*” do caderno *Política*, “*Finep apoiará investimentos de pesquisas de empresas em universidades*” do caderno *Tecnologia* e “*Feira do Livro de Brasília espera receber 300 mil pessoas até 25 de junho*” do caderno *Cultura*.

Figura 43 – Nuvens de palavras em cada grupo obtidas em uma execução sobre o conjunto X_{33} no experimento #7 com o algoritmo *ensemble* de agrupamentos

Fonte: Fernando Henrique da Silva Costa, 2018

A figura 44 ilustra a relação de cadernos encontradas em cada grupo em uma execução no conjunto X_{33} nesse experimento. É possível identificar, por meio da figura, que os grupos são significativamente desbalanceados. O grupo 4 reúne a maioria das notícias do corpus e, devido a isso, notícias de todos os cadernos.

6.5.8 Experimento #8

A tabela 21 apresenta os cinco melhores resultados alcançados pelo oitavo experimento definido para o algoritmo *ensemble* de agrupamento. Em todas essas execuções a distância cosseno, $k = 10$, $p = 10$, o algoritmo *k-means* com inicialização aleatória de centroides para gerar os componentes, os algoritmos AGNES com *average-link* ou *k-means++* para gerar o resultado final e amostragem de atributos foram utilizados.

Figura 44 – Relação cadernos de notícias nos grupos obtidos em uma execução sobre o conjunto X_{33} no experimento #7 com o algoritmo *ensemble* de agrupamentos

Fonte: Fernando Henrique da Silva Costa, 2018

Tabela 21 – Qualidade de agrupamento para o experimento #8 do algoritmo *ensemble* de agrupamentos, de acordo com índices de validação de agrupamento. Destaque em negrito para o melhor resultado em cada índice

ID	I_{ARI}			I_{NMI}			I_{SIL}		
	Média	DP	CV	Média	DP	CV	Média	DP	CV
62	0,450	0,053	0,117	0,579	0,014	0,025	0,079	0,007	0,091
64	0,489	0,023	0,048	0,594	0,012	0,021	0,074	0,005	0,072
80	0,466	0,059	0,127	0,600	0,023	0,038	0,074	0,008	0,109
376	0,505	0,038	0,075	0,581	0,023	0,039	0,054	0,004	0,069
390	0,486	0,055	0,113	0,577	0,016	0,028	0,060	0,004	0,068

Fonte: Fernando Henrique da Silva Costa, 2018

Os resultados alcançados nesse experimento não apresentaram uma melhora quando comparados com experimentos anteriores. Esses resultados são semelhantes aos que obtiveram bons resultados de índices de avaliação.

A figura 45 ilustra a plotagem do índice Silhouette em duas execuções, uma no conjunto X_8 e outra em X_{10} . Há grupos bem categorizados, como por exemplo, os grupos 1 e 7 da primeira matriz. Na segunda matriz, o grupos 4 e 6 apresentaram valores negativos.

Em relação à diversidade entre as partições geradas pelos componentes e pelo *ensemble* de agrupamentos, os valores obtidos foram: $D_1 = 0,467$, $D_2 = 0,106$, $D_3 = 0,319$ e $D_4 = 0,227$. Tais valores indicam uma certa regularidade nos níveis de diversidade.

A figura 46 ilustra as nuvens de palavras de cada grupo em uma execução no $ID80$. Os primeiro e quinto grupos possuem apenas uma notícia, dessa maneira, apresentando um desbalanceamento nesse agrupamento. A formação dos grupos restante é: 91 notícias

Figura 45 – Plotagem do índice Silhouette em uma execução em X_8 e X_{10} no experimento #8 do algoritmo *ensemble* de agrupamentos

Fonte: Fernando Henrique da Silva Costa, 2018

no segundo grupo, 59 no terceiro, 211 no quarto, 138 no sexto, 223 no sétimo, 118 no oitavo, 110 no nono e 205 no décimo.

A notícia que forma o primeiro grupo é intitulada “*Nasa transmite ao vivo o eclipse solar total*” e é classificada como *Tecnologia*. No quinto grupo, a notícia é intitulada “*Quarteto Uirapuru no Blim-blem-blom*” e está no caderno *Infantil*.

Indo em conformidade com os termos apresentados nas nuvens de palavras, o segundo grupo reúne notícias que versam sobre política e afins. O terceiro grupo é composto por notícias que discutem assuntos de violência, direito e segurança. Os termos apresentados no quarto grupo indicam a presença de notícias que dissertam a respeito de ciência e tecnologia.

O sexto grupo é formado por notícias que versam sobre cultura. O sétimo grupo possui termos semelhantes ao segundo grupo, no entanto, é possível destacar que alguns termos indicam notícias a respeito de corrupção e economia. Os oitavo e nono grupos discutem a respeito de educação, de forma que, este é mais intrínseco a ocorridos no Rio de Janeiro, aquele abrange um contexto de educação de nível superior e afins. Por fim, o décimo grupo reúne notícias de cunho esportivo.

A figura 47 ilustra a relação dos cadernos encontrados em cada grupo em uma execução em X_{10} . É possível determinar, por meio da figura, que há grupos que são formados, em maioria, por notícias de um caderno. Exemplos são os grupos 2, 6, 8, 9 e 10. Destacam-se os grupos 8 e 9, uma vez que eles são formados, em sua grande parte,

Figura 46 – Nuvens de palavras em cada grupo obtidas em uma execução sobre o conjunto X_{10} no experimento #8 com o algoritmo *ensemble* de agrupamentos

Fonte: Fernando Henrique da Silva Costa, 2018

por notícias do caderno *Educação*. Além disso, é observável que as notícias do caderno *Esportes* estão concentradas em um único grupo, no caso, o décimo.

6.5.9 Experimento #9

A tabela 22 apresenta os cinco melhores resultados alcançados pelo nono experimento definido para o algoritmo *ensemble* de agrupamento. Em todas essas execuções a distância cosseno, $p = 10$, $k = 10$ o algoritmo *k-means++* para gerar os componentes, o algoritmo AGNES com *average-link*, o conjunto original e a variação do valor de k para cada componente foram aplicados.

Os resultados alcançados por esse experimento não foram superiores aos experimentos anteriores. Esses resultados são semelhantes aos melhores resultados alcançados em experimentos anteriores.

Figura 47 – Relação cadernos de notícias nos grupos obtidos em uma execução sobre o conjunto X_{10} no experimento #8 com o algoritmo *ensemble* de agrupamentos

Fonte: Fernando Henrique da Silva Costa, 2018

Tabela 22 – Qualidade de agrupamento para o experimento #9 do algoritmo *ensemble* de agrupamentos, de acordo com índices de validação de agrupamento. Destaque em negrito para o melhor resultado em cada índice

ID	I_{ARI}			I_{NMI}			I_{SIL}		
	Média	DP	CV	Média	DP	CV	Média	DP	CV
6	0,468	0,036	0,077	0,589	0,009	0,015	0,072	0,011	0,154
30	0,475	0,045	0,096	0,594	0,012	0,021	0,076	0,010	0,135
162	0,504	0,024	0,047	0,579	0,017	0,03	0,056	0,004	0,071
174	0,496	0,040	0,081	0,582	0,019	0,032	0,059	0,005	0,085
186	0,505	0,060	0,119	0,588	0,032	0,055	0,056	0,003	0,057

Fonte: Fernando Henrique da Silva Costa, 2018

A figura 48 ilustra a plotagem do índice Silhouette em uma execução no conjunto X_8 nesse experimento. Os grupos são homogêneos, quando comparado com a média apresentada na tabela 22.

Os valores de diversidade das partições obtidos entre os componentes e o *ensemble* de agrupamentos foram: $D_1 = 0,397$, $D_2 = 0,114$, $D_3 = 0,358$ e $D_4 = 0,288$. Esses valores são ligeiramente inferiores a experimentos anteriores, no entanto, ainda apresenta uma regularidade nos níveis de diversidade.

A figura 49 ilustra as nuvens de palavras geradas em uma execução no $ID30$. O agrupamento gerado dessa execução é balanceado. A formação dos grupos por notícias é: 73 notícias no primeiro grupo, 98 no segundo, 47 no terceiro, 52 no quarto, 209 no quinto, 119 no sexto, 122 no sétimo, 142 no oitavo, 90 no nono e, por fim, 205 no décimo.

Figura 48 – Plotagem do índice Silhouette em uma execução X_8 no experimento #9 do algoritmo *ensemble* de agrupamentos

Fonte: Fernando Henrique da Silva Costa, 2018

Pelos termos apresentados nas nuvens de palavras, o primeiro e o segundo grupos possuem similaridades. No entanto, é possível verificar que o primeiro grupo trata de assuntos relacionados à corrupção, enquanto o segundo grupo reúne notícias generalizadas de política.

Semelhantemente aos dois primeiros grupos, o terceiro e quarto também possuem similaridades, em que são compostos por notícias que versam a respeito de educação. Entretanto, o terceiro grupo reúne notícias que envolvem questões políticas no meio educacional, enquanto o quarto é composto por notícias de educação no Rio de Janeiro. Por fim, o sétimo grupo também é composto por notícias de educação, porém voltadas ao ensino superior e afins.

Os quinto e sexto grupos são compostos por notícia que envolvem direito, segurança e violência, sendo diferenciados pelo contexto. Enquanto no quinto grupo, os termos indicam notícias sobre direitos de um grupo considerado minoria, no sexto, os termos indicam que as notícias relatam acontecimentos no Rio de Janeiro.

Por fim, o oitavo grupo é compostos por notícias que versam a respeito de cultura, o nono disserta sobre política internacional e o décimo é composto por notícias de cunho esportivo.

A figura 50 ilustra a relação de cadernos de notícias encontradas em cada grupo em uma execução no conjunto X_8 . Há grupos que são formados, em grande parte, por notícias de um caderno. Exemplos são os grupos 3, 4, 7, 8, 9 e 10. Nesse experimento, o caderno

Figura 49 – Nuvens de palavras em cada grupo obtidas em uma execução sobre o conjunto X_{47} no experimento #9 com o algoritmo *ensemble* de agrupamentos

Fonte: Fernando Henrique da Silva Costa, 2018

Educação é destacado em três grupos, os quais são formados, em sua maioria, por notícias desse caderno. Semelhante a outros experimentos, o caderno *Esportes* está concentrado em apenas um grupo.

6.6 Análises

Em relação aos resultados alcançados pelo algoritmo *k-means++*, a qualidade dos agrupamentos no *experimento #2* é levemente pior do que a qualidade obtida nos agrupamentos do *experimento #1*, em termos de valor da média para os índices e em termos de sensibilidade às condições iniciais. Porém, o nível de qualidade obtido revela o mesmo comportamento em termos de discordância em relação à verdade fundamental considerada para avaliação baseada em critérios externos e em termos de aderência dos dados aos grupos.

Figura 50 – Relação cadernos de notícias nos grupos obtidos em uma execução sobre o conjunto X_8 no experimento #9 com o algoritmo *ensemble* de agrupamentos

Fonte: Fernando Henrique da Silva Costa, 2018

Os valores obtidos nos índices I_{ARI} e I_{NMI} para ambos experimentos mostram que há discordância em relação à organização das notícias em grupos e a categorização das notícias em cadernos. Essa discordância é um indício que a categorização das notícias pode:

- não seguir um padrão natural, i.e., não representa os reais relacionamentos existente entre as notícias;
- ser inconsistente, no sentido que a interpretação de diferentes produtores de conteúdo sobre os relacionamentos entre notícias e entre notícias e tema dos cadernos pode influenciar, de maneiras diferentes, na decisão sobre a associação das diferentes notícias a cada um dos cadernos.

Todavia, os valores de I_{SIL} mostram que os agrupamentos obtidos não são livres de crítica, uma vez que o valor próximo a 0 mostra que os dados colocados em um grupo são também aderentes ao outro grupo mais próximo de cada dado. Essa última constatação também revela a complexidade de análise oferecida pelo conteúdo do *Corpus EBC*.

Em relação aos resultados de agrupamento alcançados pelo algoritmo *ensemble* de agrupamentos, a qualidade dos agrupamentos obtidos nos experimentos é relativamente melhor quando comparados aos experimentos do *k-means++*, exceto pelos *experimento #1*, *experimento #3* e *experimento #7*. O melhor resultado para o *ensemble* de agrupamentos

(execução no conjunto 49 no experimento #5) é 0,609, o qual é superior ao melhor do algoritmo *k-means++* (execução no conjunto 10 no experimento #1), em termos de I_{NMI} .

Igualmente visto nos resultados alcançados com o *k-means++*, os valores obtidos nos índices I_{ARI} e I_{NMI} em todos os experimentos do *ensemble* de agrupamentos indicam uma discordância da organização das notícias em grupos e a categorização delas em cadernos. Por fim, os valores vistos em I_{SIL} são próximos a 0, indicando que as notícias podem ser aderentes a outro grupo mais próximo dela.

Quanto a diversidade entre os componentes de *ensemble* e o *ensemble* de agrupamentos, altos valores não indicam melhores resultados. Por exemplo, o *experimento #7* apresentou os maiores valores em diversidade, no entanto, teve resultados bastantes inferiores aos demais experimentos, em relação as médias para os índices. De forma similar, baixos valores também trouxeram resultados de agrupamentos inferiores, como por exemplo o *experimento #3*. Porém, é esperado neste trabalho que a diversidade possa contribuir para a qualidade de recomendações, como discutido no capítulo 2.

A aplicação de uma estratégia para diversificar os agrupamentos também não foram o suficiente para alcançar os melhores resultados em termos de média dos índices. Por exemplo, nos *experimento #4* e *experimento #5* não foram utilizadas estratégias para diversificar os agrupamentos e, ainda assim, foram obtidos resultados competitivos em relação aos outros experimentos. Porém, é importante destacar que nesses dois experimentos, houve um desbalanceamento na formação dos grupos.

Dentre as aplicações de estratégias para diversificar componentes, a que varia os dados em cada componente de *ensemble* trouxe resultados bastante inferiores em termos de média dos índices. A utilização do algoritmo *k-means* com inicialização aleatória de centroides, a variação de atributos com reposição nos componentes e variação de k em cada componente foram mais eficientes.

O número de componentes de *ensemble* também não foi determinante para alcançar os melhores resultados. Por exemplo, os resultados obtidos nos *experimento #5* e *experimento #6* são relativamente similares a outros experimentos.

No que diz respeito aos algoritmos aplicados na matriz de coassociação para gerar o agrupamento final, o AGNES com *single-link* resultou em grupos com mais desbalanceamento. O mesmo pode ser encontrado no AGNES com *average-link*, porém em menor escala. Por fim, o algoritmo *k-means++* apresentou resultados com grupos mais

equilibrados, exceto no *experimento #8*, o qual utilizou o *k-means* com iniciação aleatória dos centroides nos componentes.

Por fim, em relação a distribuição dos cadernos de notícias nos grupos gerados nos experimentos, destacam-se os cadernos *Educação* e *Esportes*. O primeiro é relacionado ao fato de estar presente, em destaque, em diversos grupos. Isso pode indicar que os assuntos tratados dentro desse caderno são descorrelacionados. Em relação ao segundo caderno, geralmente, apareceu concentrado em um único grupo.

6.7 Considerações finais

Este capítulo apresentou análises de avaliações quantitativos sobre os resultados dos algoritmos de agrupamento *k-means++* e *ensemble* de agrupamentos. A partir dessas análises, os melhores resultados de cada algoritmo foram selecionados com base na percepção dos autores deste trabalho. O próximo capítulo apresenta as análises das estratégias de recomendação, incluindo um estudo com usuários, que utiliza estes resultados.

7 Análise das estratégias de recomendação

A partir do estabelecimento da estrutura de grupos, proveniente da aplicação de estratégias de agrupamento sobre o Corpus EBC, foi possível aplicar as estratégias de recomendação e analisar a qualidade das recomendações geradas.

Neste capítulo cada uma das estratégias de recomendação apresentada no capítulo 5 são aplicadas sobre os melhores agrupamentos produzidos nos experimentos apresentados no capítulo 6 para os dois experimentos realizados com *k-means++* e para um experimento realizado com o ensemble de agrupamento.

A partir da aplicação das estratégias de recomendação, dois tipos de análises de qualidade são discutidas. A primeira se refere a uma análise qualitativa, de cunho semântico, e realizada sob a interpretação do autor deste trabalho. A segunda análise é proveniente da avaliação de interações de usuários com um sistema no qual os usuários foram levados a atuar como recomendadores de notícias. O intuito do experimento com o usuário foi verificar se as estratégias de recomendação automáticas, incluindo duas que são propostas neste trabalho, seriam capazes de recomendar notícias de forma similar a seres humanos, considerando cada aspecto de qualidade de recomendação (relevância, surpresa e serendipidade).

7.1 Recomendações geradas a partir do agrupamento obtido com o *k-means++*

Nesta seção são apresentadas as análises qualitativas realizadas sobre as estratégias de recomendação visando relevância, surpresa e serendipidade. Tendo isso em vista, foram utilizados dois experimentos com o algoritmo *k-means++* com distância cosseno. O primeiro experimento foi definido 10 como valor de k , enquanto o segundo foi definido $k = 7$.

7.1.1 Experimento #1 - análise qualitativa das recomendações

Para realização das análises qualitativas sobre recomendação no primeiro experimento, as listas de recomendação (compostas por três notícias) foram obtidas sobre o agrupamento realizado sobre a matriz X_{10} . As notícias são apresentadas nesta pesquisa em termos de seus títulos. A notícia semente usada para geração das recomendações foi

escolhida aleatoriamente. Ela é proveniente do caderno *Cidadania* e é intitulada: “*Polícia investiga agressões a LGBTs em show no Rio*”. Seguem as análises qualitativas das listas de recomendação visando: relevância, surpresa e serendipidade.

Recomendação visando relevância

A primeira estratégia de recomendação visando relevância é recomendar notícias mais próximas da notícia semente em que ambas estão no mesmo grupo. A lista de recomendação gerada por ela é composta pelas notícias:

- “*Ministro quer ampliar treinamento de policiais brasileiros com oficiais dos EUA*”.
- “*Idosa é agredida a pedradas no RJ e família denuncia intolerância religiosa*”.
- “*Desligamento de funcionários do projeto Viver preocupa vítimas de abuso sexual*”.

A primeira recomendação pertence ao caderno *Geral*, enquanto as demais pertencem ao mesmo caderno da notícia semente. É possível observar as similaridades entre as notícias recomendadas com a notícia semente, uma vez que todas tratam sobre violência e segurança pública. Especificamente, as duas últimas notícias tratam também de violência e minorias, assim como a notícia semente.

A segunda estratégia de recomendação, visando relevância, toma notícias do mesmo grupo da notícia semente porém as escolhendo de forma aleatória. A lista de recomendação gerada conta com as seguintes notícias:

- “*Auxiliar de creche é baleada dentro de escola no Rio*”.
- “*Estado do Rio ganha força-tarefa social para atuar na redução da violência*”.
- “*Relatores da ONU criticam ”ataques” aos direitos ambientais e indígenas no país*”.

Todas as notícias recomendadas são pertencentes ao caderno *Cidadania*, o mesmo da notícia semente. As duas primeiras recomendações são altamente similares com a notícia semente, pois dissertam sobre a violência e segurança pública no Rio de Janeiro. Entretanto a terceira, ainda que apresente elementos sobre violência, versa sobre um grupo social diferente (indígenas), enquanto mantém alguma característica relacionada a um tipo de violência.

Recomendação visando a surpresa

A primeira estratégia visando o aspecto de surpresa recomenda notícias aleatórias que estejam no grupo mais distante da notícia semente. A lista de recomendação gerada pela primeira estratégia de recomendação visando surpresa é composta pelas notícias:

- “*No Mundo da Bola analisa amistoso entre Brasil e Argentina*”.
- “*Atlético-PR se prepara para o jogo contra o Flamengo pela Libertadores*”.
- “*Chapecoense lidera o Brasileirão depois de quatro rodadas*”.

Todas as notícias que estão na lista pertencem ao caderno *Esportes*. Elas versam sobre um assunto diferente da notícia semente, portanto, apresentam diversidade em relação a ela. No entanto, todas elas versam sobre um mesmo tema, gerando uma lista sem diversidade interna.

A segunda estratégia visando o aspecto de surpresa recomenda notícias aleatórias que não estejam agrupadas junto com a notícia semente. A lista de recomendação gerada por essa estratégia é composta pelas notícias:

- “*Botafogo empata com o Sport e se garante nas quartas da Copa do Brasil*”.
- “*No Mundo da Bola destaca a preparação da Seleção Brasileira em Porto Alegre*”.
- “*Justiça manda soltar militantes do MTST presos durante greve geral*”.

As duas primeiras recomendações são pertencentes ao caderno *Esportes* e a última está presente em *Geral*. Diferentemente da estratégia anterior, essa apresenta uma lista com um pouco mais de diversidade interna, uma vez que há uma notícia que trata sobre um assunto diferente das outras.

Recomendação visando serendipidade

A primeira estratégia visando o aspecto de serendipidade recomenda as notícias mais próximas a notícia semente que estão agrupadas no grupo mais próximo da notícia semente. A lista de recomendação gerada por ela é composta pelas notícias:

- “*Apenas 1% dos brasileiros com deficiência está no mercado de trabalho*”.
- “*Projetos sociais incentivam negócios de impacto no Brasil*”.
- “*Instituto Vital Brazil desenvolve remédio inédito contra veneno de abelha*”.

As notícias da lista de recomendação são pertencentes ao cadernos de *Cidadania*, *Educação* e *Tecnologia*. É possível encontrar diversidade quando comparada com a notícia semente, e ao analisar o conteúdo completo da notícias, tópicos em comum com a notícia semente são encontrados. A primeira recomendação trata de um grupo descrito como minoria (deficientes), igualmente a notícia semente (LGBTs). A segunda disserta sobre projetos sociais que incluem o estado do Rio de Janeiro e alguns grupos de minorias. Por fim, a terceira recomendação tem como similaridade o Instituto Vital Brazil que é ligado à secretaria de saúde do Rio de Janeiro.

A segunda estratégia visando o aspecto de serendipidade recomenda as notícias mais semelhantes a notícia semente que estão agrupadas no grupo que representa um balanceamento entre o grupo mais próximo e mais distante do grupo da notícia semente. A lista de notícias geradas por ela é composta pelas notícias:

- “*Servidores e estudantes da UERJ fazem ato em defesa da instituição*”.
- “*Em crise, UERJ faz vestibular com menos da metade dos candidatos do ano passado*”.
- “*Educadores pedem “blindagem social” em escola onde morreu Maria Eduarda*”.

Todas as notícias recomendadas são pertencentes ao caderno *Educação*. Ainda que apresentem temas diferentes, é possível observar similaridades com a notícia semente. As duas primeiras notícias tratam sobre evento na UERJ, universidade localizada no Rio de Janeiro. A última notícia envolve violência dentro do âmbito escolar e, também, no Rio de Janeiro.

7.1.2 *Experimento #2* - análise qualitativa das recomendações

Para realização das análise qualitativas no segundo experimento, as listas de recomendação obtidas sobre o agrupamento realizado sobre a matriz X_2 foram analisadas. A notícia semente sorteada pertence ao caderno *Economia* e é intitulada: “*Produtor pode solicitar recursos do Plano Safra 2017-2018 a partir de segunda*”. Seguem as análises qualitativas das listas de recomendação visando: relevância, surpresa e serendipidade.

Recomendações visando relevância

A lista de recomendação gerada pela primeira estratégia é formada pelas notícias:

- “*Privatizações devem ajudar no cumprimento da meta fiscal, diz Meirelles*”.
- “*Superintendente do BNDES diz que “pior já passou” para a economia*”.
- “*Pezão diz que Rio de Janeiro deve regularizar salários de servidores em agosto*”.

Todas as notícias que formam a lista de recomendação são pertencentes ao caderno *Economia*. Essas notícias possuem similaridades com a notícia semente, uma vez que elas versam sobre assuntos em comum: o contexto da economia nacional em um determinado período.

A lista gerada pela segunda estratégia de recomendação é formada pelas seguintes notícias:

- “*TRF garante rematrícula de universitária que conseguiu vaga por meio de cota*”.
- “*MPF pede que empresários reparem danos ambientais causados em Angra dos Reis*”.
- “*Prefeitura do Rio inicia nesta terça reordenação de camelôs nas ruas*”.

A primeira notícia recomendada pertence ao caderno *Educação*, enquanto as demais estão presentes em *Geral*. Neste caso, a aleatoriedade na escolha das notícias dentro do grupo inseriu alguma surpresa na lista de recomendação. Porém, percebe-se que a presença de instituições públicas nas notícias é o fator que, provavelmente, as organizou em um mesmo grupo (diferentemente da organização original das notícias, a qual as colocou em cadernos diferentes, incluindo o caderno cuja temática é indefinida - *Geral*).

Recomendações visando surpresa

A lista de recomendação gerada pela primeira estratégia visando surpresa é composta pelas notícias:

- “*Flamengo e Cruzeiro empatam no primeiro jogo da final da Copa do Brasil*”.
- “*Bate Bola Nacional destaca a contratação de Éverton Ribeiro pelo Flamengo*”.
- “*Brasil vence a Austrália de goleada em Melbourne*”.

Todas as notícias que formam a lista de recomendação são pertencentes ao caderno *Esportes*. O comportamento observado é o mesmo obtido para essa estratégia no primeiro experimento, inclusive o grupo com tema “esporte” foi novamente definido como o grupo de maior dissimilaridade ao grupo da notícia semente.

As notícias que compõem a lista de recomendação formada de acordo com segunda estratégia são:

- “*UEA Tabatinga-AM vai realizar terceira Edição dos Jogos Universitários Intercurso*”.
- “*Governo de Cuba critica ”pressão dos EUA” para mudanças de sistema na ilha*”.
- “*Escolas públicas do DF normatizam uso de celular em sala de aula*”.

As notícias recomendadas pertencem aos cadernos *Esportes*, *Internacional* e *Educação*, respectivamente. É possível verificar que os assuntos tratados pelas notícias recomendadas possuem diversidade quando comparados ao assunto sobre o qual a notícia semente disserta.

Recomendações visando serendipidade

A lista de recomendação gerada pela primeira estratégia que visa serendipidade é composta por:

- “*Produção brasileira de cana-de-açúcar pode chegar a 646 milhões de toneladas*”.
- “*Em dez anos, Brasil deve ultrapassar os EUA na produção de soja*”.
- “*Aumento de geração de energia por consumidor pode mudar perfil de distribuidoras*”.

As duas primeiras recomendações são pertencentes ao caderno *Economia*, enquanto a última encontra-se em *Tecnologia*. As duas primeiras notícias apresentam relevância alta já que o conteúdo é relacionado à economia no setor rural, porém estão organizadas em um grupo diferente (vizinho) à da notícia semente. A terceira notícia tem o caráter de surpresa mais acentuado, pois não versa diretamente sobre agricultura, porém está relacionada ao contexto econômico brasileiro.

A lista de recomendação formada pela segunda estratégia é composta pelas notícias:

- “*História Hoje: Há 48 anos, morria o poeta Guilherme de Almeida*”.
- “*Divulgada programação do 50º Festival de Brasília do Cinema Brasileiro, que começa em 15 de setembro*”.
- “*Festival de cinema LGBT exhibe mais de 80 filmes no Rio até dia 16*”.

Todas as recomendações geradas pertencem ao caderno *Cultura*. Essa lista apresenta notícias com alta diversidade comparada a notícia semente. Além disso, os assuntos são concentradas em cultura.

7.1.3 Discussão geral

As estratégias de recomendação baseadas em relevância assumem a premissa de que a similaridade entre os conteúdos das notícias é importante para os usuário. Em ambos os experimentos, as estratégias atenderam à essa premissa. Porém, vale salientar que a segunda estratégia baseada em relevância considera um aspecto de aleatoriedade e, também em ambos os experimentos, essa aleatoriedade foi capaz de inserir um aspecto de surpresa às listas de recomendação. Essa constatação revela que tal estratégia pode ser considerada também quando o aspecto serendipidade está sendo buscado no sistema de recomendação. Para o caso das estratégias de recomendação que visaram surpresa, os resultados não corresponderam às expectativas iniciais. Embora as notícias recomendadas versassem sobre temas diferentes da notícia semente, a primeira estratégia produziu listas sem diversidade interna, i.e., a lista de recomendação continha notícias muito similares entre elas. De fato, tal estratégia, em princípio, tenderia a esse comportamento já que é restrita a recomendar notícias de um mesmo grupo. As estratégias que visam tratar o aspecto de serendipidade produziram resultados qualitativamente difíceis de analisar. No primeiro experimento, ambas estratégias foram capazes de combinar relevância e surpresa, com mais ou menos ênfase a cada um desses aspectos, a depender do ponto de vista em que se analisa as notícias. Já no segundo experimento, a primeira estratégia gerou uma lista com forte aspecto de relevância, e a segunda estratégia gerou uma lista com forte aspecto de surpresa.

Comparando os resultados obtidos nos dois experimentos, percebe-se que no primeiro, os resultados para recomendações relevantes são mais acurados em relação ao perfil do usuário, e que a serendipidade é mais facilmente percebida nas recomendações que usaram estratégias para esse fim. Um fator que pode ter causado essa queda de desempenho no segundo experimento é o uso de um número menor de grupos neste experimento, o que fez com que os grupos fossem formados por notícias menos aderentes ao grupo como um todo.

Por fim, uma observação merece ser realizada quanto à presença de notícias de cadernos diferentes em listas de recomendação geradas por critérios de similaridade (não aleatório), considerando notícias dentro ou fora de um mesmo grupo. No primeiro experimento, uma notícia do caderno *Geral* é recomendada como relevante em relação a

uma notícia do caderno *Cidadania* e, em outro caso, notícias de três cadernos distintos (Cidadania, Educação e Tecnologia) aparecem em uma lista de recomendação gerada sobre notícias de um mesmo grupo. Situações similares ocorrem no segundo experimento para notícias originalmente presentes nos cadernos: (a) *Economia, Educação, Geral*; (b) *Economia e Tecnologia*. Estas constatações ilustram o potencial do uso de estratégias de agrupamento como base para recomendação ou como uma alternativa ao uso da base em cadernos ou como a alternativa quando uma organização prévia de notícias (ou itens) não está disponível.

7.2 Recomendações realizadas a partir do agrupamento obtido com o ensemble

A abordagem de uso de ensemble, neste trabalho, foi proposta com foco no aspecto de qualidade de recomendação referente à serendipidade. Sobre os resultados do ensemble, uma estratégia de recomendação foi aplicada, a qual inicialmente “binariza” os atributos da matriz de centroide a partir de um limiar θ , encontra os atributos assinalados em 1 entre os grupos e, portanto, recomendar notícias de outros grupos que possuem similaridades com o grupo da notícia semente.

Para a realização das análises qualitativas sobre as recomendações geradas para a estratégia definida para o ensemble de *agrupamentos*, foi utilizada uma execução na matriz X_{47} , com $k = 10$, dez membros, variação de atributos na matriz e a utilização do algoritmo Agnes com *average-link* na matriz de coassociação.

Semelhantemente com as estratégias anteriores, as notícias são apresentadas pelos seus títulos. A notícia semente definida para as análises foi escolhida aleatoriamente. Ela é intitulada “*Professores de escolas municipais de Salvador mantêm greve*” e pertence ao caderno *Educação*.

A partir da notícia semente, uma lista de recomendação com cinco notícias foi gerada. Ela contém as seguintes notícias:

- *Inscrições para o Enem começam hoje; veja o que os candidatos devem saber*
- *Contratos do Fies para o segundo semestre já podem ser renovados*
- *Era da Internet das Coisas exige atenção com a segurança, alerta especialista*
- *Enchentes no Amazonas deixam 39 municípios em situação de emergência*
- *SBPC premia pesquisa sobre larvicida contra Aedes aegypti usando óleo de sucupira*

As duas primeiras notícias da lista são pertencentes ao mesmo caderno da notícia semente. A terceira e quinta notícias são pertencentes ao caderno *Tecnologia*. Por fim, a quarta notícia é do caderno *Geral*. É possível observar na lista que o aspecto de surpresa se sobrepõe ao de relevância, pois os conteúdos dissertados nas notícias recomendadas são diversificados quando comparado ao da notícia semente. Outro ponto importante é que a lista possui diversidade interna, apresentando notícias que versam sobre diferentes assuntos.

Em relação as duas primeiras recomendações, elas versam sobre assuntos relacionados sobre educação. No entanto, o contexto que elas estão inseridas são diferentes. A notícia semente disserta a respeito da greve dos professores de escolas municipais em Salvador, enquanto as duas notícias recomendadas discutem a respeito sobre Enem e Fies, assuntos relacionados com educação de nível superior. Isso pode ser um indicativo para o aspecto de serendipidade, pois as notícias estão inseridas em elementos que envolvem educação, configurando a relevância, ao passo que, os assuntos tratados entre a notícia semente e as recomendadas são diferentes, configurando a surpresa.

As últimas três recomendações apresentam mais diversidade quando comparada com a notícia semente. Isso ocorre devido a natureza da estratégia de recomendação aplicada. As notícias recomendadas são escolhidas aleatoriamente de grupos que possuem termos em comum com o grupo da notícia semente. Porém, os termos que são determinados como mais relevantes nos grupos podem não estar presentes em algumas notícias. Dessa maneira, as três últimas notícias recomendadas podem apresentar um grau maior de surpresa ao ser comparada com a relevância.

É importante notar que duas recomendações são pertencentes ao caderno *Educação*, o mesmo da notícia semente, porém foram agrupadas em grupos diferentes. Isso pode significar que, ainda que sejam classificados previamente em um mesmo caderno, o conteúdo dessas notícias não são similares. Tal suposição ilustra a complexidade do *CorpusEBC* e, semelhantemente a análise anterior, mostram o potencial da utilização de estratégias de agrupamento como base de recomendação.

7.3 Estudo com usuários

Esta seção apresenta as informações relacionadas ao estudo com usuários. Ela está organizada da seguinte forma: primeiro, o objetivo do estudo é apresentado; então, são apresentados os detalhes mais importantes do sistema que suportou a interação do usuário com as notícias, em um contexto de recomendação; dados gerais sobre a realização de um experimento piloto são discutidos e finalmente os dados do estudo com os usuários e as análises sobre esses dados são apresentados.

7.3.1 Objetivo geral do estudo

O objetivo deste estudo é avaliar o desempenho de quatro estratégias de recomendação sob os três aspectos relacionados com a qualidade de recomendação (relevância, surpresa e serendipidade) a partir da visão de usuários. Cada estratégia de recomendação sob análise foi implementada a partir de uma base para escolha de notícias a serem recomendadas:

- *aleatoriedade*: neste caso as recomendações são feitas a partir de uma escolha aleatório de notícias, considerando todo o Corpus EBC e considerando uma distribuição uniforme, ou seja, todas as notícias, exceto a notícia semente, tem igual probabilidade de serem escolhidas para compor a recomendação. Usar essa base de recomendação teve sua motivação na intenção de potencializar a geração de recomendações surpreendentes;
- *k-means++*: a base para escolha de notícias é uma partição gerada pelo *k-means++*. Neste experimento, a base é a partição gerada sobre o conjunto de dados X_{10} , discutida na seção 6.4.1. A estratégia de recomendação usada é selecionar uma notícia que esteja no grupo da notícia semente de forma aleatória. Usar tal estratégia teve sua motivação na intenção de potencializar a geração de recomendações relevantes;
- *ensemble* de agrupamentos: a base para escolha das notícias segue a abordagem proposta neste trabalho para pós-processamento da partição gerada pelo *ensemble*. A estratégia de recomendação proposta neste trabalho (seção 5.2) foi aplicada no experimentos. Como a estratégia pressupõe uma lista com cinco notícias recomendadas e no experimento cada estratégia recomenda somente uma, escolheu-se a mais próxima

do grupo da notícia semente pela distância cosseno. Usar tal estratégia teve sua motivação na intenção de potencializar a geração de recomendações serendipitosas;

- *ONMTF*: a base de escolha das notícias é uma copartição gerada pelo algoritmo *Orthogonal Non-Negative Tri-Factorization* (YOO; CHOI, 2010), e a estratégia de recomendação usada foi proposta em um trabalho de dissertação realizado dentro do grupo de pesquisa relacionado ao presente trabalho. Essa estratégia de recomendação segue o mesmo raciocínio atribuído ao algoritmo *k-means++*.

A interação do usuário com o sistema é baseada na realização de três tarefas. Cada uma delas tem como objetivo orientar o usuário para que ele avalie, indiretamente, qual das estratégias de recomendação foi mais eficiente em recomendar uma notícia com um determinado aspecto de qualidade.

Em cada tarefa, o usuário é apresentado a instruções que o guia na realização da tarefa, e a uma notícia escolhida aleatoriamente do Corpus EBC. Nas instruções, o usuário recebe explicações sobre o que define um aspecto de qualidade de recomendação (relevância na primeira tarefa; surpresa na segunda; e serendipidade na terceira). Então, a partir do esclarecimento sobre o aspecto de qualidade em questão, o usuário deve ler o conteúdo da notícia que lhe foi apresentada para seguir com a tarefa. No passo seguinte, quatro notícias são apresentadas a ele. Cada uma delas é uma recomendação de cada estratégia de recomendação sob avaliação. O usuário deve ler as quatro notícias e escolher aquela que ele entende ser a mais relevante, surpreendente ou serendipitosa, a depender da tarefa em questão. Obviamente que o usuário não tem conhecimento de qual estratégia de recomendação apresentou qual notícia.

O sistema gerencia a interação do usuário associando as suas escolhas as estratégias que apresentaram a notícia por ele escolhida em cada tarefa. Posteriormente, essa associação é usada para indicar a estratégia que mais foi capaz de recomendar notícias consideradas relevantes, ou surpreendentes ou serendipitosas, de acordo com a visão dos usuários.

7.3.2 Sistema para interação com usuário

O sistema para interação com o usuário foi desenvolvido em R com o auxílio da biblioteca *Shiny*. Ele foi disponibilizado para acesso via internet e ficou disponível por 40 dias. As interações com os usuários ocorreram: de 10 de junho de 2018 a 16 de junho

de 2018 para o experimento piloto; de 25 de junho de 2018 a 16 de julho de 2018 para o experimento em que as interações foram consideradas para análise das estratégias de recomendação. O projeto para realização do estudo com o usuário foi aprovado pelo Comitê de Ética da EACH sob o número 83105318.4.0000.5390.

Antes iniciar a interação como o sistema, o usuário precisou ler e aceitar o Termo de Consentimento Livre e Esclarecido e informar o nome, e-mail, escolaridade, idade e familiaridades com portais de notícias e sistemas de recomendação. Após essa primeira etapa, foi apresentado ao usuário as instruções iniciais para interação com o sistema.

A figura 51 ilustra a tela usada para apresentar as instruções da tarefa, a definição do aspecto a ser avaliado e a notícia semente. Neste primeiro momento, após a leitura das instruções e, principalmente, da notícia semente, o usuário deveria acionar o botão “*Receber recomendações*” a fim de chamar uma nova tela na qual estavam as as quatro recomendações (quatro notícias recomendadas). Na figura 52, as instruções e definições referentes às tarefas associadas a cada aspecto de qualidade de recomendação sob análise são apresentadas em detalhes.

Figura 51 – Tela da notícia semente e das instruções

Fonte: Fernando Henrique da Silva Costa, 2018

A figura 53 ilustra a tela que apresenta as quatro recomendações. Cada recomendação é acessada por uma *tab*, conforme pode ser observado na figura (Recomendação 3). Além das *tabs* que contêm as notícias recomendadas, a última *tab* é denominada “*Escolha*”, sendo apresentada na figura 54. Após a leitura das quatro notícias recomendadas, o usuário deve acessar a última *tab* com a finalidade de determinar qual foi a recomendação mais

Figura 52 – Detalhes sobre instruções e definições usadas no estudo com usuários

Tarefa 1 - Relevância	Tarefa 2 - Surpresa	Tarefa 3 - Serendipidade
<p>Nesta primeira tarefa, imagine que um usuário virtual tenha lido e gostado da notícia que está apresentada ao lado (esta é a notícia semente). Ele gostaria de continuar lendo outras notícias, então ele espera que o portal de notícias recomende uma outra notícia que seja relevante (útil) para ele.</p> <p>Considerando as instruções e as definições de notícia relevante apresentadas abaixo, sua tarefa é escolher dentre as notícias recomendadas, aquela que tem maior destaque em termos de relevância para o usuário virtual. Lembre-se que o usuário leu e gostou da notícia semente.</p> <p>Instruções:</p> <ul style="list-style-type: none"> A notícia semente está apresentada ao lado. Leia a notícia semente integralmente. Após a leitura, clique no botão Receber recomendações. Você verá uma nova janela com cinco abas. As quatro primeiras abas contêm uma das quatro notícias recomendadas. Leia as quatro notícias integralmente. Escolha a notícia MAIS RELEVANTE para o usuário virtual. Ou seja, de acordo com a notícia semente que ele leu e gostou, escolha a notícia que será mais relevante para ele. <p>Uma notícia relevante é definida como:</p> <ul style="list-style-type: none"> Uma notícia que agrada o usuário, ou seja, uma notícia que é adequada para ele. Uma notícia de fundamental importância para um usuário, de acordo com a motivação que o levou a ler uma notícia (no caso, a notícia semente). 	<p>Na segunda tarefa, imagine que outro usuário virtual tenha lido e se interessado pela notícia apresentada ao lado (agora esta é a notícia semente). Esse usuário, no entanto, gostaria de receber uma recomendação que o surpreendesse quando comparada com a notícia semente que ele acabou de ler.</p> <p>Considerando as instruções e as definições de notícia surpreendente apresentadas abaixo, sua tarefa é escolher dentre as notícias recomendadas, aquela que tem maior destaque em termos de surpresa para o usuário virtual. Lembre-se que o usuário leu e gostou da notícia semente.</p> <p>Instruções:</p> <ul style="list-style-type: none"> A notícia semente está apresentada ao lado. Leia a notícia semente integralmente. Após a leitura, clique no botão Receber recomendações. Você verá uma nova janela com cinco abas. As quatro primeiras abas contêm uma das quatro notícias recomendadas. Leia as quatro notícias integralmente. Escolha a notícia MAIS SURPREENDENTE para o usuário virtual. <p>Uma notícia surpreendente é definida como:</p> <ul style="list-style-type: none"> uma notícia que, possivelmente, o usuário não espera receber como uma recomendação, uma notícia que surpreende, causando espanto, assombro, admiração, estranheza ou maravilhamento uma notícia inesperada. 	<p>Para a última tarefa, imagine que outro usuário virtual tenha lido e se interessado pela notícia apresentada ao lado (esta é a notícia semente para esta tarefa). Esse usuário é um pouco mais exigente quando comparado aos outros dois (das tarefas anteriores). Ele gostaria de receber como recomendação, uma notícia que fosse tanto relevante quanto surpreendente, ou seja, ele gostaria de receber uma notícia serendipitosa.</p> <p>Considerando as instruções e as definições de notícia serendipitosa apresentadas abaixo, sua tarefa é escolher dentre as notícias recomendadas, aquela que tem maior destaque em termos de serendipidade para o usuário virtual. Lembre-se que o usuário leu e se interessou pela notícia semente.</p> <p>Instruções:</p> <ul style="list-style-type: none"> A notícia semente está apresentada ao lado. Leia a notícia semente integralmente. Após a leitura, clique no botão Receber recomendações. Você verá uma nova janela com cinco abas. As quatro primeiras abas contêm uma das quatro notícias recomendadas. Leia as quatro notícias integralmente. Escolha a notícia MAIS SERENDIPITOSA para o usuário virtual. <p>Uma notícia serendipitosa é definida como:</p> <ul style="list-style-type: none"> uma notícia que é, provavelmente, relevante e inesperada; uma notícia que traz descobertas agradáveis e úteis para o usuário; uma notícia que se mostra útil e surpreendente para o usuário.

Fonte: Fernando Henrique da Silva Costa, 2018

adequada para tarefa. Por fim, caso seja do interesse do usuário, ele ainda pode realizar um comentário a respeito da tarefa realizada.

Figura 53 – Tela das recomendações

Fonte: Fernando Henrique da Silva Costa, 2018

7.3.3 Experimento piloto

Com o intuito de melhorar a usabilidade do sistema desenvolvido, e o entendimento dos usuários sobre as tarefas a serem executadas, um experimento piloto foi realizado. Para isso, dez pessoas foram convidadas a proceder com a realização das tarefas no sistema.

Figura 54 – Tela da escolha

Recomendações

Recomendação 1 Recomendado 2 Recomendado 3 Recomendado 4 Escolha

Qual notícia recomendada você escolhe de acordo com o objetivo da tarefa?

☐ Recomendação 1
☐ Recomendação 2
☐ Recomendação 3
☐ Recomendação 4

(opcional) Considerando a tarefa atual, o que você achou sobre as notícias recomendadas?

Avançar

Voltar

Fonte: Fernando Henrique da Silva Costa, 2018

Esse grupo de pessoas foi formado por alunos de graduação e de mestrado em Sistemas de Informação, na Escola de Artes, Ciências e Humanidades da Universidade de São Paulo, além de um professor do mesmo curso. Após a execução do experimento piloto, alguns ajustes foram realizados no protótipo e exemplos são:

- a exclusão de um campo em que o usuário determinava quais os tipos de notícias que estavam no seu hábito de leitura. Isso ocorreu para que não houvesse dúvidas em relação ao objetivo da interação. Desde que o propósito da tarefa é que o usuário pense sobre o aspecto que está sob avaliação e então escolha a notícia que melhor representa uma boa recomendação sob tal aspecto, a informação de preferência do usuário é irrelevante e não deve ser um fator de influência de tomada de decisão. A notícia escolhida como semente é aleatória dentre todas do conjunto, e com a existência da pergunta sobre hábitos de leitura, o usuário criava expectativa de que sua preferência seria considerada na interação com o sistema;
- uma definição mais objetiva dos aspectos de qualidade de recomendação em cada tarefa do experimento. Alguns usuários não se mostraram satisfeitos com uma variabilidade grande de definições, então as definições sobre os aspectos de qualidade de recomendação foram resumidas.
- inserção da definição da figura de um usuário virtual, a ser apresentado pelo usuário que está interagindo com o sistema como sendo a pessoa para a qual a recomendação está sendo feita. O perfil do usuário virtual é definido pela leitura da notícia semente.

O motivo dessa inserção foi melhorar o entendimento que o usuário participante tinha um papel de recomendador e não de usuário leitor, propriamente dito;

- expansão na largura da tela de recomendação a fim acomodar os texto das notícias para que a leitura se tornasse mais confortável.

7.3.4 Dados sobre a condução do estudo

Com a finalização dos *feedbacks* dos usuários do experimento piloto e, por consequência, dos ajustes realizados no sistema, diversas pessoas foram convidadas a participar do estudo oficial.

Os primeiros convites foram enviados via e-mail no dia 25 de junho de 2018 para alunos de graduação e mestrado do curso de Sistemas de Informação da Escola de Artes, Ciências e Humanidades da Universidade de São Paulo e para profissionais da área de tecnologia de informação pertencentes à rede de relacionamento dos membros do grupo de pesquisa ao qual este trabalho está associado. A fim de impulsionar o número de iterações no protótipo, novos convites foram enviado entre os dias 30 de junho de 2018 e 03 de julho de 2018. A solicitação foi enviada para alunos e professores de outras instituições acadêmicas de computação e para novos profissionais da mesma área.

Os dados das iterações dos usuários foram coletados até o dia 16 de julho de 2018. Até esse dia, o número de iterações estava definido em 60. A partir desses dados coletados, foi realizada a análise dos resultados, a qual segue apresentada na próxima subseção.

7.3.5 Análises

A figura 55 mostra o desempenho de cada uma das estratégias de recomendação, considerando os três aspectos de qualidade de recomendação avaliados. O valor representado por cada barra do gráfico é referente a quantidade de vezes que os usuários escolheram como melhor recomendação, aquela sugerida pela estratégia. Por exemplo, a primeira barra de cada trio de barras no gráfico indica quantas vezes a notícia recomendada por cada estratégia foi considerada uma recomendação relevante. A segunda barra diz respeito ao atendimento do aspecto de surpresa, e a terceira diz respeito ao atendimento do aspecto de serendipidade.

Figura 55 – Desempenho dos algoritmos em cada aspecto de recomendação

Fonte: Fernando Henrique da Silva Costa, 2018

Por meio da figura 55, é possível observar que para o aspecto de relevância, a estratégia do algoritmo *k-means++* foi a mais eficiente. Em relação ao aspecto de surpresa, a estratégia do *ensemble* de agrupamentos foi a mais escolhida. Por fim, em serendipidade, novamente a estratégia do *k-means++* teve melhor desempenho.

Esses resultados não confirmam todas as hipóteses iniciais estabelecidas na construção de cada estratégia de recomendação. O desempenho de destaque da estratégia estabelecida sobre o agrupamento *k-means++* no aspecto relevância era esperado, uma vez que a estratégia usada no estudo com o usuário foi baseada na similaridade das notícias presentes no grupo da notícia semente. Assim, a notícia recomendada pela estratégia do *k-means++* tende a ser muito parecida com a notícia semente. Para esse aspecto, a premissa estabelecida pela estratégia foi confirmada. Para o caso da premissa referente à surpresa, o esperado seria que a estratégia aleatória alcançasse o melhor desempenho, entretanto não foi isso que aconteceu. Embora esse aspecto tenha sido o mais proeminente considerando apenas o contexto da estratégia aleatória, considerando o contexto completo do estudo, a surpresa está mais presente nas recomendações baseadas no *ensemble*. A premissa para atendimento da serendipidade estava pautada no fato que *ensemble* de agrupamentos e coagrupamentos seriam os mais eficientes por considerarem similaridades parciais. A estratégia de coagrupamento não alcançou bons resultados e *ensemble* teve um desempenho similar ao *k-means++*. Diante destes resultados, uma hipótese pode ser levantada: a partição criada pelo algoritmo de agrupamento diante do contexto de análise de texto (com representação baseada em *bag-of-words* e uso de distância cosseno) apresenta um balanceamento adequado entre relevância e surpresa, visto que é capaz de alcançar bons índices sob os três aspectos. De forma similar, o *ensemble* de agrupamentos também

alcança esse balanceamento, no entanto, com um viés mais voltado à diversificação nas recomendações, uma vez que é melhor no aspecto de qualidade referente à surpresa. Olhando apenas esses dois casos, é factível avaliar que as premissas estabelecidas para esses dois casos possuem razoabilidade.

Ainda diante desta análise, há um risco neste estudo que pode ser mais bem estudado. Desde que a premissa de atendimento ao aspecto de surpresa estava fortemente pautada na ideia de que a aleatoriedade geraria as melhores recomendações, ou a premissa é realmente falsa, ou os usuários não acataram a ideia de que notícias aleatórias (de conteúdo provavelmente muito diferente do conteúdo da notícia semente) deveriam causar a sensação de surpresa.

As próximas análises relacionam os cadernos da notícia semente e os cadernos das notícias escolhidas como a melhor recomendação pelos usuários, de acordo com a organização usada no portal EBC. Nos eixos dos gráficos apresentados nas figuras 56, 57 e 58 considere: 1 como *Cidadania*, 2 como *Cultura*, 3 como *Economia*, 4 como *Educação*, 5 como *Esportes*, 6 como *Geral*, 7 como *Infantil*, 8 como *Internacional*, 9 como *Política* e 10 como *Tecnologia*.

Na figura 56 é apresentado uma relação entre os cadernos da notícia semente e os cadernos das notícias escolhidas como a melhor recomendação considerando o aspecto relevância. O tamanho das bolhas nos gráficos indica a frequência com que um caderno (o da notícia recomendada - no eixo *y*) oferece uma notícia relevante diante de uma notícia semente - pertencente a um determinado caderno - eixo *x*. Neste gráfico é possível observar, por exemplo, que de 19 vezes em que uma notícia semente do caderno *Educação* foi usada no sistema na tarefa de análise de relevância, 12 vezes a recomendação de notícias também do caderno *Educação* foi considerada a mais adequada pelos usuários. Nas outras sete vezes, notícias dos cadernos *Cidadania*, *Geral*, *Política* e *Tecnologia* foram consideradas relevantes.

Sobre o conteúdo deste gráfico (Figura 56), é importante destacar o comportamento da diagonal principal. Ela mostra que em 33 das 60 interações, os usuários determinaram que o caderno da notícia escolhida como melhor recomendação é o mesmo que da notícia semente. Isso pode indicar que os usuários entenderam que o aspecto de relevância está relacionado a similaridade entre as notícias.

A figura 57 ilustra a relação dos cadernos da notícias sementes com os cadernos das notícias escolhidas como melhor recomendação para o aspecto de surpresa. O destaque

Figura 56 – Relação cadernos das notícias sementes com cadernos da notícias escolhidas como melhor recomendação no aspecto de relevância

Fonte: Fernando Henrique da Silva Costa, 2018

nessa figura vai para o caderno *Educação* (4), uma vez que mais notícias provenientes dele (15 notícias) foram indicadas como surpreendentes. Além disso, esse foi o único caderno envolvido em recomendações surpreendentes provenientes do mesmo caderno da notícia semente. Isso pode indicar que:

- o assunto *Educação* é bem particular, visto que diante de notícias sementes de vários outros cadernos, os usuários indicaram que uma recomendação surpreendente seria aquela que envolve o assunto *Educação*. Ou seja, a partir vários contextos de leituras, receber uma recomendação sobre educação causa surpresa;
- mesmo dentro do contexto de *Educação*, há ainda especificidades, a ponto de duas notícias sobre educação (uma semente e a outra recomendada) serem colocadas em um caso de recomendação surpreendente.

Indicação de recomendações surpreendentes dentro do mesmo caderno da notícia semente, não ocorreu para os demais cadernos, o que pode indicar o entendimento do aspecto de surpresa, sob a visão dos usuários, como uma notícia que apresente alguma diferença em relação a notícia semente.

A figura 58 ilustra uma gráfico de bolhas cuja frequência relaciona aos cadernos da notícias sementes com os cadernos das notícias escolhidas como melhor recomendação. O comportamento visualizado nesse gráfico é mais similar ao gráfico de surpresa quando comparado ao do gráfico de relevância, embora a diagonal principal neste gráfico volte a

Figura 57 – Relação cadernos das notícias sementes com cadernos da notícias escolhidas como melhor recomendação no aspecto de surpresa

Fonte: Fernando Henrique da Silva Costa, 2018

receber mais ocorrências. Isso pode indicar que, sob a visão dos usuários, o aspecto de surpresa tem um maior destaque que a relevância quando se deseja avaliar a serendipidade.

Figura 58 – Relação cadernos da notícias semente com cadernos da notícias escolhidas como melhor recomendação no aspecto de serendipidade

Fonte: Fernando Henrique da Silva Costa, 2018

7.3.6 Considerações finais

Uma limitação do protótipo é que a notícia semente é determinada de forma aleatória entre todas as notícias do conjunto. Isso pode ocasionar desinteresse e tédio por parte do usuário participante. Uma solução para minimizar tal limitação seria questionar

ao usuário quais cadernos estão em seu hábito de leitura e, a partir disso, selecionar notícias aleatórias desses cadernos.

8 Conclusão

Neste trabalho, o uso de *ensemble* de agrupamentos em sistema de recomendação baseado em conteúdo foi discutido na forma da sua utilização como um meio de analisar textos usando a ideia de similaridades parciais, e oferecendo, desta forma, uma base para a apresentação de uma estratégia de recomendação baseada em conteúdo.

De forma mais específica, como descrito na seção 5.2, componentes do *ensemble* foram construídos usando subconjuntos de atributos descritivos dos textos (subconjuntos de palavras), o que representou o uso de similaridades parciais na etapa inicial de análise do conteúdo a ser recomendado. Na sequência, uma matriz de coassociação e um algoritmo de agrupamento hierárquico foram usados na construção da partição final do *ensemble*. Então, novamente similaridades parciais, neste momento em grupos, são encontradas de forma a construir uma nova partição dos dados sobre a qual a estratégia de recomendação foi construída. A estratégia tomou como base grupos similares existentes na partição e escolha aleatória de notícias pertencentes a esses grupos para construir uma estratégia com o intuito de prover recomendações com maior qualidade em termos de serendipidade.

A estratégia de recomendação com base em *ensemble* de agrupamentos foi construída a partir do estudo de outras seis estratégias de recomendação, também propostas neste trabalho, baseadas em partições obtidas pelo algoritmo de agrupamento *k-means++*. Essas estratégias foram projetadas para atender a aspectos específicos de qualidade de recomendação e foram divididas em: duas para o aspecto de relevância, duas para o aspecto de surpresa e, por fim, duas para o aspecto de serendipidade. Elas foram apresentadas na seção 5.1.

A fim escolher partições a serem usadas no teste das estratégias de recomendação, dois experimentos com o algoritmo *k-means++* e nove com o algoritmo *ensemble* de agrupamentos foram realizados (Capítulo 6). Esses experimentos foram avaliados por índices de validação baseados em critérios externos e interno. Grande parte dos experimentos realizados no *ensemble* de agrupamentos alcançaram valores superiores quando comparados com os experimentos do *k-means++*. Em relação ao cálculo da diversidade entre as partições geradas pelos componentes de *ensemble* e o *ensemble* de agrupamentos, valores altos e baixos trouxeram os piores resultados nos índices, enquanto os valores regulares trouxeram os melhores resultados.

A partir das melhores partições obtidas nos experimentos mencionados, as estratégias de recomendação foram aplicadas usando o Corpus EBC e foram avaliadas de duas maneiras: considerando uma interpretação livre do conteúdo das notícias presentes nas listas de recomendação, realizada pelo autor do trabalho (discutida no capítulo 7); e considerando um estudo com usuários (discutido na seção 7.3). Na avaliação livre, foi observado que a relevância de fato é obtida nas duas estratégias, em que a segunda estratégia apresentou diversidade interna, ou seja, notícias recomendadas com assuntos diferentes. A surpresa e a serendipidade também podem ser observadas, em que a segunda estratégia para surpresa apresentou uma maior diversidade interna, quando comparada com as demais.

Os resultados alcançados no estudo com usuários indicam que as estratégias de recomendação geradas sobre os resultados dos algoritmos de agrupamento *k-means++* e *ensemble* de agrupamentos, obtiveram os melhores desempenho no que diz respeito aos aspectos de qualidade: relevância e surpresa. De forma mais específica, a estratégia proposta para o *ensemble* de agrupamento teve o melhor resultado para o aspecto de surpresa, a estratégia baseada no algoritmo *k-means++* obteve os melhores resultados nos aspectos de surpresa e serendipidade. Tendo isso em vista, a hipótese delineada neste trabalho foi parcialmente verificada, uma vez que, em termos gerais, recomendações geradas sobre a partição oferecida pelo algoritmo *k-means++* apresentou resultados mais satisfatórios. Outras duas estratégias foram avaliadas neste estudo: recomendação aleatória e recomendação baseada em partição de coagrupamento. A estratégia baseada na aleatoriedade, como esperado, alcançou resultados importantes quando o aspecto surpresa foi avaliado. A estratégia baseada em coagrupamento não alcançou resultados de destaque.

8.1 Contribuições

Em termos de contribuição científica, este trabalho apresentou a proposição de estratégia de recomendação baseadas em conteúdo, construídas sobre organizações de dados textuais obtidas com algoritmos de agrupamento, usando análise de similaridades tradicional (algoritmo *k-means++*) e análise de similaridades parciais. Essa última análise foi obtida a partir de uma estratégia de interpretação dos resultados do *ensemble* de agrupamentos também proposta neste trabalho.

Este trabalho também apresentou um estudo com usuários que usou uma estratégia inédita para esse tipo de experimento com notícias. Essa estratégia não coloca o participante do experimento na posição de um usuário de um sistema de recomendação, mas o coloca na posição de um recomendador.

A contribuição tecnológica deste trabalho está relacionadas ao desenvolvimento do protótipo de avaliação das recomendações sob a perspectiva do usuário.

8.2 Limitações

Uma limitação deste trabalho está relacionada a não aplicação de uma análise quantitativa sobre os aspectos que avaliam a recomendação. Uma segunda limitação identificada está relacionada com o experimento com usuários, em que eles não determinam qual notícia gostaria de ler. Além disso, as interpretações dos grupos via nuvens de palavras estão sujeitas a imprecisões e dubiedade por meio de interpretações humanas. Por fim, outra limitação presente é que não houve uma estratégia de recomendação baseada em cadernos no estudo com usuários, ou seja, recomendar notícias que são provenientes do mesmo caderno da notícia semente.

8.3 Trabalhos futuros

Os resultados obtidos neste trabalho abrem caminhos para realização de novos projetos de pesquisa. O uso das estratégias de recomendação aqui propostas podem ser adaptadas para aplicação em conjuntos de dados de referência na área de sistemas de recomendação, incorporando, inclusive, informação de histórico sobre os interesses do usuário e avaliação de itens já acessados. Estratégias realizadas sobre esses conjuntos de dados podem simular estudos com usuários, o que pode facilitar a produção de avaliações de estratégias de recomendação. Além disso, pode ser de interesse reexecutar o estudo com o usuário apresentado neste trabalho inserindo uma estratégia de recomendação exclusivamente baseada na divisão de notícias em canais/cadernos já disponível na estrutura do Corpus EBC.

Referências¹

- ADAMOPOULOS, P.; TUZHILIN, A. On unexpectedness in recommender systems: Or how to expect the unexpected. In: *DiveRS@ RecSys*. [S.l.: s.n.], 2011. p. 11–18. Citado 2 vezes nas páginas 22 e 38.
- ADOMAVICIUS, G.; TUZHILIN, A. Toward the next generation of recommender systems: a survey of the state-of-the-art and possible extensions. *Transactions on Knowledge and Data Engineering*, IEEE, v. 17, n. 6, p. 734–749, 2005. Citado 2 vezes nas páginas 30 e 33.
- AGGARWAL, C. C. Content-based recommender systems. In: *Recommender Systems*. [S.l.]: Springer, 2016. p. 139–166. Citado na página 34.
- ANDRÉ, P. et al. Discovery is never by chance: designing for (un) serendipity. In: ACM. *Proceedings of the 7th conference on creativity and cognition*. [S.l.], 2009. p. 305–314. Citado na página 38.
- ARTHUR, D.; VASSILVITSKII, S. k-means++: the advantages of careful seeding. In: SOCIETY FOR INDUSTRIAL AND APPLIED MATHEMATICS. *Proceedings of the 11th annual ACM-SIAM symposium on discrete algorithms*. [S.l.], 2007. p. 1027–1035. Citado 3 vezes nas páginas 24, 49 e 50.
- BAEZA-YATES, R. A.; RIBEIRO-NETO, B. A. *Modern information retrieval - the concepts and technology behind search, 2nd edition*. [S.l.]: Pearson Education Ltd., Harlow, England, 2011. Citado na página 79.
- BAGHER, R. C.; HASSANPOUR, H.; MASHAYEKHI, H. User trends modeling for a content-based recommender system. *Expert Systems With Applications*, Elsevier, v. 87, p. 209–219, 2017. Citado na página 22.
- BAHMANI, B. et al. Scalable k-means++. *Proceedings of the VLDB endowment*, VLDB Endowment, v. 5, n. 7, p. 622–633, 2012. Citado na página 49.
- BALABANOVIĆ, M.; SHOHAM, Y. Fab: content-based, collaborative recommendation. *Communications of the ACM*, ACM, v. 40, n. 3, p. 66–72, 1997. Citado na página 23.
- BHANDARI, U. et al. Serendipitous recommendation for mobile apps using item-item similarity graph. In: SPRINGER. *Asia Information Retrieval Symposium*. [S.l.], 2013. p. 440–451. Citado 4 vezes nas páginas 56, 57, 58 e 59.
- BORATTO, L. et al. Semantics-aware content-based recommender systems: Design and architecture guidelines. *Neurocomputing*, Elsevier, v. 254, p. 79–85, 2017. Citado na página 22.
- BOSENMAN, M. F. Serendipity and scientific discovery. *The Journal of Creative Behavior*, Wiley Online Library, v. 22, n. 2, p. 132–138, 1988. Citado na página 38.
- BURKE, R. Hybrid recommender systems: survey and experiments. *User Modeling and User-Adapted Interaction*, Springer, v. 12, n. 4, p. 331–370, 2002. Citado 2 vezes nas páginas 31 e 35.

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

- CAZELLA, S. C.; NUNES, M.; REATEGUI, E. A ciência da opinião: estado da arte em sistemas de recomendação. *André Ponce de Leon F. de Carvalho; Tomasz Kowaltowski..(Org.). Jornada de Atualização de Informática-JAI*, p. 161–216, 2010. Citado 2 vezes nas páginas 30 e 31.
- CHANG, W. et al. *shiny: Web Application Framework for R*. [S.l.], 2018. R package version 1.2.0. Disponível em: <https://CRAN.R-project.org/package=shiny>. Citado na página 27.
- CHE, Y.-K.; HÖRNER, J. Recommender systems as mechanisms for social learning. *The Quarterly Journal of Economics*, Oxford University Press, v. 133, n. 2, p. 871–925, 2017. Citado na página 21.
- CORRÊA, G. N. et al. Interactive textual feature selection for consensus clustering. *Pattern Recognition Letters*, Elsevier, v. 52, p. 25–31, 2015. Citado 2 vezes nas páginas 62 e 63.
- CUI, X.; POTOK, T. E.; PALATHINGAL, P. Document clustering using particle swarm optimization. In: IEEE. *Proceedings of the Swarm Intelligence Symposium, SIS 2005*. [S.l.], 2005. p. 185–191. Citado na página 48.
- DEZA, M. M.; DEZA, E. Encyclopedia of distances. In: *Encyclopedia of distances*. [S.l.]: Springer, 2009. p. 1–583. Citado na página 47.
- DIAZ, A. K. R. et al. Relatório técnico ppgsi-001/2018 uma análise comparativa das ferramentas de pré-processamento de dados textuais: Nltk, pretext e r. 2018. Citado na página 27.
- FERN, X. Z.; BRODLEY, C. E. Random projection for high dimensional data clustering: A cluster ensemble approach. In: *Proceedings of the 20th international conference on machine learning (ICML-03)*. [S.l.: s.n.], 2003. p. 186–193. Citado na página 23.
- FERN, X. Z.; BRODLEY, C. E. Solving cluster ensemble problems by bipartite graph partitioning. In: ACM. *Proceedings of the 21st International Conference on Machine Learning*. [S.l.], 2004. p. 36. Citado na página 43.
- FERN, X. Z.; LIN, W. Cluster ensemble selection. *Statistical Analysis and Data Mining*, Wiley Online Library, v. 1, n. 3, p. 128–141, 2008. Citado na página 42.
- FRALEY, C.; RAFTERY, A. E. How many clusters? which clustering method? answers via model-based cluster analysis. *The Computer Journal*, Br Computer Soc, v. 41, n. 8, p. 578–588, 1998. Citado na página 48.
- FRANÇA, F. O. de. *Biclusterização na análise de dados incertos*. Tese (Doutorado) — Universidade Estadual de Campinas, 2010. Citado 2 vezes nas páginas 23 e 52.
- FRED, A. L.; JAIN, A. K. Data clustering using evidence accumulation. In: IEEE. *Proceedings of the 16th International Conference on Pattern Recognition*. [S.l.], 2002. v. 4, p. 276–280. Citado na página 48.
- GE, M.; DELGADO-BATTENFELD, C.; JANNACH, D. Beyond accuracy: evaluating recommender systems by coverage and serendipity. In: ACM. *Proceedings of the 4th Conference on Recommender Systems*. [S.l.], 2010. p. 257–260. Citado 2 vezes nas páginas 38 e 69.

- GUNAWARDANA, A.; SHANI, G. Evaluating recommender systems. In: *Recommender Systems Handbook*. [S.l.]: Springer, 2015. p. 265–308. Citado na página 37.
- GUO, G.; ZHANG, J.; YORKE-SMITH, N. Leveraging multiviews of trust and similarity to enhance clustering-based recommender systems. *Knowledge-Based Systems*, Elsevier, v. 74, p. 14–27, 2015. Citado 2 vezes nas páginas 60 e 61.
- HALKIDI, M.; BATISTAKIS, Y.; VAZIRGIANNIS, M. Clustering algorithms and validity measures. In: IEEE. *Proceedings of the 13th International Conference on Scientific and Statistical Database Management, SSDBM 2001*. [S.l.], 2001. p. 3–22. Citado na página 51.
- HAN, J.; PEI, J.; KAMBER, M. *Data mining: concepts and techniques*. [S.l.]: Elsevier, 2011. Citado na página 47.
- HARTIGAN, J. A. *Clustering algorithms*. [S.l.: s.n.], 1975. Citado na página 41.
- HARTIGAN, J. A.; WONG, M. A. Algorithm as 136: a k-means clustering algorithm. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, JSTOR, v. 28, n. 1, p. 100–108, 1979. Citado na página 48.
- HERLOCKER, J. L. et al. Evaluating collaborative filtering recommender systems. *Transactions on Information Systems (TOIS)*, ACM, v. 22, n. 1, p. 5–53, 2004. Citado na página 37.
- HOTH, A.; NÜRNBERGER, A.; PAASS, G. A brief survey of text mining. In: *LDV Forum - GLDV J. for Comp. Linguistics and Lang. Tech.* [S.l.: s.n.], 2005. v. 20, n. 1, p. 19–62. Citado na página 79.
- HUANG, A. Similarity measures for text document clustering. In: *Proceedings of the 6th New Zealand Computer Science Research Student Conference (NZCSRSC2008)*. [S.l.: s.n.], 2008. p. 49–56. Citado na página 47.
- HUANG, S. et al. Spectral co-clustering ensemble. *Knowledge-Based Systems*, Elsevier, v. 84, p. 46–55, 2015. Citado na página 53.
- HUBERT, L.; ARABIE, P. Comparing partitions. *Journal of classification*, Springer, v. 2, n. 1, p. 193–218, 1985. Citado 2 vezes nas páginas 27 e 52.
- IAM-ON, N. et al. A link-based approach to the cluster ensemble problem. *Transactions on Pattern Analysis and Machine Intelligence*, IEEE, v. 33, n. 12, p. 2396–2409, 2011. Citado na página 41.
- IAQUINTA, L. et al. Introducing serendipity in a content-based recommender system. In: IEEE. *8th International Conference on Hybrid Intelligent Systems, HIS'08*. [S.l.], 2008. p. 168–173. Citado na página 38.
- JAIN, A. K. Data clustering: 50 years beyond k-means. *Pattern Recognition Letters*, Elsevier, v. 31, n. 8, p. 651–666, 2010. Citado 2 vezes nas páginas 40 e 49.
- JAIN, A. K.; MURTY, M. N.; FLYNN, P. J. Data clustering: a review. *Computing Surveys (CSUR)*, Acm, v. 31, n. 3, p. 264–323, 1999. Citado 2 vezes nas páginas 40 e 47.

- JAMES, N. T.; RAJKUMAR, K. Product recommendation systems based on hybrid approach technology. 2017. Citado na página 23.
- JENDERS, M. et al. A serendipity model for news recommendation. In: SPRINGER. *Joint German/Austrian Conference on Artificial Intelligence (Künstliche Intelligenz)*. [S.l.], 2015. p. 111–123. Citado 4 vezes nas páginas 55, 57, 58 e 59.
- JING, L.; TIAN, K.; HUANG, J. Z. Stratified feature sampling method for ensemble clustering of high dimensional data. *Pattern Recognition*, Elsevier, v. 48, n. 11, p. 3688–3702, 2015. Citado na página 42.
- KAMINSKAS, M.; BRIDGE, D. Diversity, serendipity, novelty, and coverage: a survey and empirical analysis of beyond-accuracy objectives in recommender systems. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, ACM, v. 7, n. 1, p. 2, 2017. Citado na página 22.
- KONSTAN, J. A. Introduction to recommender systems: algorithms and evaluation. *Transactions on Information Systems (TOIS)*, ACM, v. 22, n. 1, p. 1–4, 2004. Citado na página 31.
- KOTKOV, D. et al. Investigating serendipity in recommender systems based on real user feedback. 2018. Citado na página 22.
- KOUTRIKA, G. Modern recommender systems: from computing matrices to thinking with neurons. In: ACM. *Proceedings of the 2018 International Conference on Management of Data*. [S.l.], 2018. p. 1651–1654. Citado na página 21.
- KUMAR, S. et al. Survey on personalized web recommender system. *International Journal of Information Engineering and Electronic Business*, Modern Education and Computer Science Press, v. 10, n. 4, p. 33, 2018. Citado na página 21.
- KUNCHEVA, L. I.; HADJITODOROV, S. T. Using diversity in cluster ensembles. In: IEEE. *Systems, man and cybernetics, 2004 IEEE international conference on*. [S.l.], 2004. v. 2, p. 1214–1219. Citado na página 44.
- LEWIS, J. M.; ACKERMAN, M.; SA, V. R. de. Human cluster evaluation and formal quality measures: a comparative study. In: *CogSci*. [S.l.: s.n.], 2012. p. 1870–1875. Citado na página 51.
- LI, H. et al. An efficient manifold regularized sparse non-negative matrix factorization model for large-scale recommender systems on gpus. *Information Sciences*, Elsevier, 2018. Citado na página 21.
- LIAO, C.-L.; LEE, S.-J. A clustering based approach to improving the efficiency of collaborative filtering recommendation. *Electronic Commerce Research and Applications*, Elsevier, v. 18, p. 1–9, 2016. Citado 2 vezes nas páginas 60 e 61.
- LIAO, T. W. Clustering of time series data—a survey. *Pattern Recognition*, Elsevier, v. 38, n. 11, p. 1857–1874, 2005. Citado na página 47.
- LIMA, C. A. de M. *Comitê de Máquinas: uma abordagem unificada empregando máquinas de vetores-suporte*. Tese (Doutorado) — Universidade Estadual de Campinas, 2004. Citado na página 42.

- LINDEN, R. Técnicas de agrupamento. *Revista de Sistemas de Informação da FSMA*, v. 1, n. 4, p. 18–36, 2009. Citado na página 50.
- LIU, J.; DOLAN, P.; PEDERSEN, E. R. Personalized news recommendation based on click behavior. In: ACM. *Proceedings of the 15th International Conference on Intelligent User Interfaces*. [S.l.], 2010. p. 31–40. Citado na página 22.
- LIU, L.; LIAO, Z.; LIAO, Z. An efficient clustering ensemble selection algorithm. *International Journal of Autonomous and Adaptive Communications Systems*, Inderscience Publishers (IEL), v. 8, n. 2-3, p. 200–212, 2015. Citado na página 43.
- LOPS, P.; GEMMIS, M. D.; SEMERARO, G. Content-based recommender systems: State of the art and trends. In: *Recommender Systems Handbook*. [S.l.]: Springer, 2011. p. 73–105. Citado 4 vezes nas páginas 23, 34, 35 e 36.
- METEREN, R. V.; SOMEREN, M. V. Using content-based filtering for recommendation. In: *Proceedings of the Machine Learning in the New Information Age: MLnet/ECML2000 Workshop*. [S.l.: s.n.], 2000. p. 47–56. Citado na página 33.
- MICHALSKI, R. S.; CARBONELL, J. G.; MITCHELL, T. M. *Machine learning: an artificial intelligence approach*. [S.l.]: Springer Science & Business Media, 2013. Citado na página 40.
- MILLIGAN, G. W. A monte carlo study of thirty internal criterion measures for cluster analysis. *Psychometrika*, Springer, v. 46, n. 2, p. 187–199, 1981. Citado na página 51.
- MINGOTI, S. A.; LIMA, J. O. Comparing som neural network with fuzzy c-means, k-means and traditional hierarchical clustering algorithms. *European Journal of Operational Research*, Elsevier, v. 174, n. 3, p. 1742–1759, 2006. Citado na página 50.
- MITCHELL, T. M. *Machine learning*. 1. ed. New York, NY, USA: McGraw-Hill, Inc., 1997. ISBN 0070428077, 9780070428072. Citado na página 40.
- MUFLIKHAH, L.; BAHARUDIN, B. Document clustering using concept space and cosine similarity measurement. In: IEEE. *International Conference on Computer Technology and Development, ICCTD'09*. [S.l.], 2009. v. 1, p. 58–62. Citado na página 47.
- MURAKAMI, T.; MORI, K.; ORIHARA, R. Metrics for evaluating the serendipity of recommendation lists. In: SPRINGER. *Annual Conference of the Japanese Society for Artificial Intelligence*. [S.l.], 2007. p. 40–46. Citado 2 vezes nas páginas 38 e 59.
- NALDI, M. C.; FACELI, K.; CARVALHO, A. Uma revisão sobre combinação de agrupamentos. *Revista de Informática Teórica e Aplicada*, v. 16, n. 2, p. 25–52, 2009. Citado 5 vezes nas páginas 27, 42, 44, 45 e 46.
- NILSSON, N. J. Introduction to machine learning. an early draft of a proposed textbook. Citeseer, 1996. Citado na página 40.
- OKU, K.; HATTORI, F. Fusion-based recommender system for improving serendipity. In: ACM. *Proceedings of the Workshop on Novelty and Diversity in Recommender Systems (DiveRS 2011), at the 5th International Conference on Recommender Systems (RecSys 2011)*. [S.l.], 2011. p. 19. Citado na página 58.

OKU, K.; HATTORI, F. User evaluation of fusion-based approach for serendipity-oriented recommender system. In: *Proceedings of the Workshop on Recommendation Utility Evaluation: Beyond RMSE (RUE 2012)*. [S.l.: s.n.], 2012. p. 39–44. Citado 3 vezes nas páginas 56, 57 e 59.

PATERLINI, S.; KRINK, T. Differential evolution and particle swarm optimisation in partitioned clustering. *Computational Statistics & Data Analysis*, Elsevier, v. 50, n. 5, p. 1220–1247, 2006. Citado na página 47.

PAZZANI, M. J. A framework for collaborative, content-based and demographic filtering. *Artificial Intelligence Review*, Springer, v. 13, n. 5-6, p. 393–408, 1999. Citado na página 34.

PAZZANI, M. J.; BILLSUS, D. Content-based recommendation systems. In: *The Adaptive Web*. [S.l.]: Springer, 2007. p. 325–341. Citado na página 32.

PERES, S. M. et al. Tutorial sobre fuzzy-c-means e fuzzy learning vector quantization: abordagens híbridas para tarefas de agrupamento e classificação. *Revista de Informática Teórica e Aplicada*, v. 19, n. 1, p. 120–163, 2012. Citado na página 41.

PIAO, S.; WHITTLE, J. A feasibility study on extracting twitter users' interests using nlp tools for serendipitous connections. In: IEEE. *Privacy, Security, Risk and Trust (PASSAT) and 2011 IEEE Third International Conference on Social Computing (SocialCom), 2011 IEEE Third International Conference on*. [S.l.], 2011. p. 910–915. Citado 4 vezes nas páginas 55, 57, 58 e 59.

POLIKAR, R. Ensemble based systems in decision making. *Circuits and Systems Magazine*, IEEE, v. 6, n. 3, p. 21–45, 2006. Citado na página 42.

R Core Team. *R: A Language and Environment for Statistical Computing*. Vienna, Austria, 2017. Disponível em: <https://www.R-project.org/>. Citado na página 26.

RAUSCH, K. et al. Exploring subspace clustering for recommendations. In: ACM. *Proceedings of the 26th International Conference on Scientific and Statistical Database Management*. [S.l.], 2014. p. 42. Citado 2 vezes nas páginas 60 e 61.

RESNICK, P.; VARIAN, H. R. Recommender systems. *Communications of the ACM*, ACM, v. 40, n. 3, p. 56–58, 1997. Citado na página 30.

RICCI, F.; ROKACH, L.; SHAPIRA, B. *Introduction to recommender systems handbook*. [S.l.]: Springer, 2011. Citado 2 vezes nas páginas 30 e 32.

ROBERTSON, S. Understanding inverse document frequency: on theoretical arguments for IDF. *J. of Doc.*, Emerald Group Publishing Limited, v. 60, n. 5, p. 503–520, 2004. Citado na página 79.

ROSENBERG, A.; HIRSCHBERG, J. V-measure: a conditional entropy-based external cluster evaluation measure. In: *EMNLP-CoNLL*. [S.l.: s.n.], 2007. v. 7, p. 410–420. Citado na página 51.

ROUSSEEUW, P. J. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, Elsevier, v. 20, p. 53–65, 1987. Citado 2 vezes nas páginas 27 e 52.

- SADEGHIAN, A. H.; NEZAMABADI-POUR, H. Document clustering using gravitational ensemble clustering. In: IEEE. *International Symposium on Artificial Intelligence and Signal Processing (AISP)*. [S.l.], 2015. p. 240–245. Citado 2 vezes nas páginas 62 e 63.
- SALTON, G.; WONG, A.; YANG, C.-S. A vector space model for automatic indexing. *Comm. of the ACM*, ACM, v. 18, n. 11, p. 613–620, 1975. Citado na página 79.
- SCHAFER, J. B.; KONSTAN, J.; RIEDL, J. Recommender systems in e-commerce. In: ACM. *Proceedings of the 1st ACM Conference on Electronic Commerce*. [S.l.], 1999. p. 158–166. Citado na página 32.
- SEIFODDINI, H. Machine grouping—expert systems: comparison between single linkage and average linkage clustering techniques in forming machine cells. *Computers & Industrial Engineering*, Elsevier, v. 15, n. 1-4, p. 210–216, 1988. Citado na página 51.
- SEIFODDINI, H. K. Single linkage versus average linkage clustering in machine cells formation applications. *Computers & Industrial Engineering*, Elsevier, v. 16, n. 3, p. 419–426, 1989. Citado na página 50.
- SHANI, G.; GUNAWARDANA, A. Evaluating recommendation systems. In: *Recommender Systems Handbook*. [S.l.]: Springer, 2011. p. 257–297. Citado 3 vezes nas páginas 22, 67 e 69.
- SHARDANAND, U.; MAES, P. Social information filtering: algorithms for automating “word of mouth”. In: ACM PRESS/ADDISON-WESLEY PUBLISHING CO. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. [S.l.], 1995. p. 210–217. Citado na página 30.
- SILVA, L. A.; PERES, S. M.; BOSCARIOLI, C. *Introdução à mineração de dados: com Aplicações em R*. [S.l.]: Elsevier, 2016. 277 p. Citado na página 79.
- SINOARA, R. A.; ANTUNES, J.; REZENDE, S. O. Text mining and semantics: a systematic mapping study. *J. of the Brazilian Computer Society*, Springer, v. 23, n. 1, p. 9, 2017. Citado na página 79.
- STEINBACH, M. et al. A comparison of document clustering techniques. In: BOSTON. *KDD workshop on text mining*. [S.l.], 2000. v. 400, n. 1, p. 525–526. Citado 2 vezes nas páginas 50 e 51.
- STREHL, A.; GHOSH, J. Cluster ensembles: a knowledge reuse framework for combining multiple partitions. *Journal of Machine Learning Research*, v. 3, n. Dec, p. 583–617, 2002. Citado 4 vezes nas páginas 23, 27, 42 e 43.
- STUDHOLME, C.; HILL, D. L.; HAWKES, D. J. An overlap invariant entropy measure of 3d medical image alignment. *Pattern Recognition*, Elsevier, v. 32, n. 1, p. 71–86, 1999. Citado na página 52.
- TARAMIGKOU, M. et al. Escape the bubble: Guided exploration of music preferences for serendipity and novelty. In: ACM. *Proceedings of the 7th ACM conference on Recommender systems*. [S.l.], 2013. p. 335–338. Citado 4 vezes nas páginas 55, 57, 58 e 59.
- TIAN, Y.; YANG, P. Cluster ensemble algorithm using the binary k-means and spectral clustering. *Journal of Computational Information Systems*, v. 10, n. 12, p. 5147–5154, 2014. Citado 3 vezes nas páginas 61, 62 e 63.

TOPCHY, A. P. et al. Analysis of consensus partition in cluster ensemble. In: IEEE. *4th International Conference on Data Mining, ICDM'04*. [S.l.], 2004. p. 225–232. Citado na página 45.

WAHID, A.; GAO, X.; ANDREAE, P. Multi-objective clustering ensemble for high-dimensional data based on strength pareto evolutionary algorithm (spea-ii). In: IEEE. *International Conference on Data Science and Advanced Analytics (DSAA)*. [S.l.], 2015. p. 1–9. Citado 2 vezes nas páginas 62 e 63.

WU, H. et al. Dual-regularized matrix factorization with deep neural networks for recommender systems. *Knowledge-Based Systems*, Elsevier, v. 145, p. 46–58, 2018. Citado na página 21.

WU, W.; HE, L.; YANG, J. Evaluating recommender systems. In: IEEE. *7th International Conference on Digital Information Management (ICDIM)*. [S.l.], 2012. p. 56–61. Citado na página 37.

WU, Z.; LEAHY, R. An optimal graph theoretic approach to data clustering: theory and its application to image segmentation. *Transactions on Pattern Analysis and Machine Intelligence*, IEEE, v. 15, n. 11, p. 1101–1113, 1993. Citado na página 41.

XIAO, Z. et al. Increasing serendipity of recommender system with ranking topic model. *Applied Mathematics & Information Sciences*, Natural Sciences Publishing Corp, v. 8, n. 4, p. 2041, 2014. Citado 4 vezes nas páginas 55, 57, 58 e 59.

XU, S. et al. An integrated k-means-laplacian cluster ensemble approach for document datasets. *Neurocomputing*, Elsevier, v. 214, p. 495–507, 2016. Citado 2 vezes nas páginas 62 e 63.

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 na página 129.

ZAHRA, S. et al. Novel centroid selection approaches for kmeans-clustering based recommender systems. *Information sciences*, Elsevier, v. 320, p. 156–189, 2015. Citado 2 vezes nas páginas 60 e 61.

ZHANG, T.; RAMAKRISHNAN, R.; LIVNY, M. Birch: an efficient data clustering method for very large databases. In: ACM. *Sigmod Record*. [S.l.], 1996. v. 25, n. 2, p. 103–114. Citado na página 41.

ZHAO, Y.; KARYPIS, G. Evaluation of hierarchical clustering algorithms for document datasets. In: ACM. *Proceedings of the 11th International Conference on Information and Knowledge Management*. [S.l.], 2002. p. 515–524. Citado na página 48.

ZHAO, Y.; KARYPIS, G.; FAYYAD, U. Hierarchical clustering algorithms for document datasets. *Data Mining and Knowledge Discovery*, Springer, v. 10, n. 2, p. 141–168, 2005. Citado na página 50.

ZHENG, Q.; CHAN, C.-K.; IP, H. H. An unexpectedness-augmented utility model for making serendipitous recommendation. In: SPRINGER. *Industrial Conference on Data Mining*. [S.l.], 2015. p. 216–230. Citado 3 vezes nas páginas 57, 58 e 59.

ZHOU, T. et al. Solving the apparent diversity-accuracy dilemma of recommender systems. *Proceedings of the National Academy of Sciences*, National Acad Sciences, v. 107, n. 10, p. 4511–4515, 2010. Citado na página 37.

ZHOU, Z.-H. Ensemble learning. *Encyclopedia of Biometrics*, Springer, p. 411–416, 2015. Citado na página 43.