
Extração de tópicos baseada em
agrupamento de regras de associação

Fabiano Fernandes dos Santos

SERVIÇO DE PÓS-GRADUAÇÃO DO ICMC-USP

Data de Depósito:

Assinatura: _____

Fabiano Fernandes dos Santos

Extração de tópicos baseada em agrupamento de regras de associação

Tese apresentada ao Instituto de Ciências
Matemáticas e de Computação - ICMC-USP, como
parte dos requisitos para obtenção do título de
Doutor em Ciências - Ciências de Computação e
Matemática Computacional. *VERSÃO REVISADA*

Área de Concentração: Ciências de Computação e
Matemática Computacional

Orientadora: Profa. Dra. Solange Oliveira Rezende

USP – São Carlos
Julho de 2015

Ficha catalográfica elaborada pela Biblioteca Prof. Achille Bassi
e Seção Técnica de Informática, ICMC/USP,
com os dados fornecidos pelo(a) autor(a)

d237e dos Santos, Fabiano Fernandes
Extração de tópicos baseado em agrupamento de
regras de associação / Fabiano Fernandes dos Santos;
orientadora Solange Oliveira Rezende. -- São
Carlos, 2015.
129 p.

Tese (Doutorado - Programa de Pós-Graduação em
Ciências de Computação e Matemática Computacional) --
Instituto de Ciências Matemáticas e de Computação,
Universidade de São Paulo, 2015.

1. Extração de tópicos. 2. Agrupamento de Regras
de Associação. 3. Redução de Dimensionalidade. 4.
Mineração de Textos. I. Rezende, Solange Oliveira,
orient. II. Título.

Fabiano Fernandes dos Santos

Topic extraction based on association rule clustering

Doctoral dissertation submitted to the Instituto de Ciências Matemáticas e de Computação - ICMC-USP, in partial fulfillment of the requirements for the degree of the Doctorate Program in Computer Science and Computational Mathematics. *FINAL VERSION*

Concentration Area: Computer Science and Computational Mathematics

Advisor: Profa. Dra. Solange Oliveira Rezende

USP – São Carlos
July 2015

A Vera, Agnaldo, Mônica, Bianca e Rafaela.

– Pois é! – disse Pensador Profundo. – Assim, quando vocês souberem qual é exatamente a pergunta, vocês saberão o que significa a resposta.

Douglas Adams - O guia do mochileiro das galáxias

Agradecimentos

Apesar do resultado final da tese de doutorado ser produto de responsabilidade e esforço de natureza individual, o sentimento que surge neste momento é o de gratidão pelas pessoas que de alguma forma participaram desta jornada, pois sem a ajuda delas nada disso teria sido possível. Ainda que esse espaço seja curto para comportar tantos sentimentos, registro aqui meu agradecimento a essas pessoas.

Primeiramente, agradeço aos meus pais, Vera e Agnaldo, pelo apoio e compreensão nessa jornada tão difícil que é criar um filho para o mundo. E as minhas irmãs Mônica e Bianca, cuja distância se fez necessária nesse momento mas o carinho sempre esteve presente.

A minha noiva, Rafaela, pelo amor, incentivo e compreensão, e por sempre acreditar na minha capacidade quando eu mesmo já pensava em desistir, me estabilizando e me impulsionando continuamente.

A minha orientadora e amiga, Solange Rezende, a quem serei eternamente grato pelas oportunidades oferecidas que me transformaram profissionalmente, e, principalmente, pessoalmente. Seus ensinamentos serão sempre lembrados e aplicados, e certamente serei uma pessoa melhor a cada dia por isso.

À minhas “co-orientadoras” Fernanda e Veronica, cujas contribuições foram fundamentais para os rumos deste trabalho. A amizade recebida de vocês é um presente que levarei comigo e cuidarei para que se torne ainda melhor.

Aos amigos do LABIC, os que já seguiram seu rumo na vida e aos que ainda estão por aqui. Sinto orgulho de fazer parte de um grupo que está sempre disposto a ajudar quem precisa, não importando a dificuldade. Agradeço a vocês, Merley, Igor, RG, Bruno, Rafael, Ricardo, Camila, Víctor, Tatiane, Celso, Newton, Renan, Marcos, Vinícius, Diego, Jorge, Thiago, Roberta, Ivone e tantos outros que fizeram a diferença na minha jornada.

Aos amigos “vocenianos”, que tenho hoje como minha família.

À CAPES pelo auxílio financeiro.

E, por fim, mas não menos importante, à todos os amigos que não tem os seus nomes citados aqui, mas que direta ou indiretamente contribuíram para a minha formação pessoal e profissional que me trouxeram até este momento.

Resumo

Uma representação estruturada dos documentos em um formato apropriado para a obtenção automática de conhecimento, sem que haja perda de informações relevantes em relação ao formato originalmente não-estruturado, é um dos passos mais importantes da mineração de textos, pois a qualidade dos resultados obtidos com as abordagens automáticas para obtenção de conhecimento de textos estão fortemente relacionados à qualidade dos atributos utilizados para representar a coleção de documentos. O Modelo de Espaço de Vetores (MEV) é um modelo tradicional para obter uma representação estruturada dos documentos. Neste modelo, cada documento é representado por um vetor de pesos correspondentes aos atributos do texto. O modelo *bag-of-words* é a abordagem de MEV mais utilizada devido a sua simplicidade e aplicabilidade. Entretanto, o modelo *bag-of-words* não trata a dependência entre termos e possui alta dimensionalidade. Diversos modelos para representação dos documentos foram propostos na literatura visando capturar a informação de relação entre termos, destacando-se os modelos baseados em frases ou termos compostos, o Modelo de Espaço de Vetores Generalizado (MEVG) e suas extensões, modelos de tópicos não-probabilísticos, como o *Latent Semantic Analysis* (LSA) ou o *Non-negative Matrix Factorization* (NMF), e modelos de tópicos probabilísticos, como o *Latent Dirichlet Allocation* (LDA) e suas extensões. A representação baseada em modelos de tópicos é uma das abordagens mais interessantes uma vez que elas fornece uma estrutura que descreve a coleção de documentos em uma forma que revela sua estrutura interna e as suas inter-relações. As abordagens de extração de tópicos também fornecem uma estratégia de redução da dimensionalidade visando a construção de novas dimensões que representam os principais tópicos ou assuntos identificados na coleção de documentos. Entretanto, a extração eficiente de informações sobre as relações entre os termos para construção da representação de documentos ainda é um grande desafio de pesquisa. Os modelos para representação de documentos que exploram a correlação entre termos normalmente enfrentam um grande desafio para manter um bom equilíbrio entre (i) a quantidade de dimensões obtidas, (ii) o esforço computacional e (iii) a interpretabilidade das novas dimensões obtidas. Assim, é proposto neste trabalho o modelo para representação de documentos ***Latent Association Rule Cluster based Model*** (LARCM). Este é um modelo de extração de tópicos não-probabilístico que explora o agrupamento de regras de associação para construir uma representação da coleção de documentos com dimensionalidade reduzida tal que as novas dimensões são extraídas a partir das informações sobre as relações entre os termos. No modelo proposto, as regras de associação são extraídas para cada documento para obter termos correlacionados que formam expressões multi-palavras. Essas relações entre os termos formam o contexto local da relação entre termos. Em seguida, aplica-se um processo de agrupamento em todas as regras de associação para formar o contexto geral das relações entre os termos, e cada grupo de regras de associação obtido formará um tópico, ou seja, uma dimensão da representação. Também é proposto neste trabalho uma metodologia de avaliação que permite selecionar modelos que maximizam tanto os resultados na tarefa de classificação de textos quanto os resul-

tados de interpretabilidade dos tópicos obtidos. O modelo LARCM foi comparado com o modelo LDA tradicional e o modelo LDA utilizando uma representação que inclui termos compostos (*bag-of-related-words*). Os resultados dos experimentos indicam que o modelo LARCM produz uma representação para os documentos que contribui significativamente para a melhora dos resultados na tarefa de classificação de textos, mantendo também uma boa interpretabilidade dos tópicos obtidos. O modelo LARCM também apresentou ótimo desempenho quando utilizado para extração de informação de contexto para aplicação em sistemas de recomendação sensíveis ao contexto.

Palavras-chave: Extração de tópicos, Redução de dimensionalidade, Agrupamento de regras de associação.

Abstract

A structured representation of documents in an appropriate format for the automatic knowledge extraction without loss of relevant information is one of the most important steps of text mining, since the quality of the results obtained with automatic approaches for the text knowledge extraction is strongly related to the quality of the selected attributes to represent the collection of documents. The Vector Space model (VSM) is a traditional structured representation of documents. In this model, each document is represented as a vector of weights that corresponds to the features of the document. The bag-of-words model is the most popular VSM approach because of its simplicity and general applicability. However, the bag-of-words model does not include dependencies of the terms and has a high dimensionality. Several models for document representation have been proposed in the literature in order to capture the dependence among the terms, especially models based on phrases or compound terms, the Generalized Vector Space Model (GVSM) and their extensions, non-probabilistic topic models as Latent Semantic Analysis (LSA) or Non-negative Matrix Factorization (NMF) and still probabilistic topic models as the Latent Dirichlet Allocation (LDA) and their extensions. The topic model representation is one of the most interesting approaches since it provides a structure that describes the collection of documents in a way that reveals their internal structure and their interrelationships. Also, this approach provides a dimensionality reduction strategy aiming to built new dimensions that represent the main topics or ideas of the document collection. However, the efficient extraction of information about the relations of terms for document representation is still a major research challenge nowadays. The document representation models that explore correlated terms usually face a great challenge of keeping a good balance among the (i) number of extracted features, (ii) the computational performance and (iii) the interpretability of new features. In this way, we proposed the *Latent Association Rule Cluster based Model* (LARCM). The LARCM is a non-probabilistic topic model that explores association rule clustering to build a document representation with low dimensionality in a way that each dimension is composed by information about the relations among the terms. In the proposed approach, the association rules are built for each document to extract the correlated terms that will compose the multi-word expressions. These relations among the terms are the local context of relations. Then, a clustering process is applied for all association rules to discover the general context of the relations, and each obtained cluster is an extracted topic or a dimension of the new document representation. This work also proposes in this work an evaluation methodology to select topic models that maximize the results in the text classification task as much as the interpretability of the obtained topics. The LARCM model was compared against both the traditional LDA model and the LDA model using a document representation that includes multi-word expressions (bag-of-related-words). The experimental results indicate that LARCM provides an document representation that improves the results in the text classification task and even retains a good interpretability of the extract topics. The LARCM model also achieved great results as a method to extract contextual information for context-aware recommender systems.

Key-words: Topic extraction, Dimensionality reduction, Association rule clustering.

Sumário

Sumário	xviii
Lista de Figuras	xx
Lista de Tabelas	xxv
Lista de Algoritmos	xxvii
1 Introdução	1
2 Mineração de Dados e Textos	7
2.1 Processo de Mineração de Textos	7
2.2 Agrupamento de Documentos	9
2.2.1 Medidas de Proximidade	9
2.2.2 Métodos de Agrupamento	10
2.3 Regras de Associação	15
2.3.1 Definições e Conceitos	15
2.3.2 Geração de Regras de Associação	17
2.3.3 Medidas de Interesse Objetivas Utilizadas na Avaliação de Regras de Associação	19
2.4 Agrupamento de Regras de Associação	25
2.5 Considerações finais	28
3 Representação de Documentos e Redução de Dimensionalidade	29
3.1 Representação de Documentos	29
3.1.1 Modelo de Espaço de Vetores	30
3.1.2 Construção Automática de Representações Utilizando Mineração de Textos	31
3.1.3 Representação de Documentos Explorando Termos Dependentes	33
3.2 Redução de Dimensionalidade	36
3.2.1 Extração de Atributos	36
3.2.2 Modelos de Extração de Tópicos	37
3.2.3 Modelos de Extração de Tópicos Não-probabilísticos	37
3.2.4 Modelos de Extração de Tópicos Probabilísticos	38
3.3 Considerações Finais	40
4 Extração de Tópicos para Redução de Dimensionalidade	43
4.1 Extração de Tópicos com Termos Dependentes para Representação de Documentos	43
4.2 LARCM: Latent Association Rule Cluster based Model	45
4.2.1 Contexto Local da Relação	47
4.2.2 Contexto Geral da Relação	52
4.2.3 Construção da Representação Documento-Tópico	54

4.3	Metodologia para Avaliação do Modelo Proposto	55
4.3.1	Avaliação na Tarefa de Classificação	55
4.3.2	Avaliação da Interpretabilidade	57
4.4	Considerações Finais	58
5	Avaliação do Modelo Proposto	59
5.1	Preparação para a Avaliação Experimental	60
5.1.1	Seleção de Corpora e Pré-Processamento	60
5.1.2	Configuração dos Modelos para Extração dos Tópicos	61
5.1.3	Configurações para a Avaliação da Tarefa de Classificação	63
5.1.4	Configurações para a Avaliação da Interpretabilidade	63
5.2	Resultados e Discussão da Avaliação do LARCM na Tarefa de Classificação	64
5.3	Resultados e Discussão da Avaliação da Interpretabilidade	66
5.4	Estudo de Caso: Sistemas de Recomendação Sensíveis ao Contexto	70
5.5	Considerações Finais	73
6	Conclusões e Trabalhos Futuros	75
Referências		91
A Coleção de Documentos Usados como Exemplo		93
B Tabelas com os Resultados da Avaliação na Tarefa de Classificação		101
B.1	Comparação dos Modelos de Extração de Tópicos Separados por Coleção de Documentos	101
B.2	Resultados da Acurácia Obtida para Cada Configuração dos Modelos de Extração de Tópicos	107
C Tabelas com os Resultados da Avaliação da Interpretabilidade		117
D Sistemas de Recomendação Sensíveis ao Contexto		127

Lista de Figuras

2.1	Etapas do processo de Mineração de Textos. Fonte: Rezende et al. (2003).	8
2.2	Análise da Estrutura do Dendrograma. Fonte: Metz (2006)	14
2.3	Visão geral do processo de agrupamento de regras de associação no pré-processamento e no pós-processamento. Fonte: Carvalho et al. (2012).	26
3.1	Exemplo do resultado obtido pela rotação dos eixos com a técnica SVD. Adaptado de: Liu (2011)	38
3.2	Modelo gráfico do LDA. Fonte: Blei et al. (2003)	40
4.1	Visão geral do Latent Association Rule Cluster based Model - LARCM. . .	46
4.2	Exemplo do processo de mapeamento dos documentos em transações (Algoritmo 4). (a) A janela é posicionada no início do documento e as palavras são adicionadas ao conjunto de transações até <i>tamanho_janela</i> ser igual a 4. (b) A janela é deslocada para a direita e a sequência encontrada de palavras é adicionada ao conjunto de transações. (c) Esse processo é repetido até o final do documento - <i>tamanho_janela</i> começa a ficar menor. (d) A janela de processamento é reduzida até restar uma palavra.	49
4.3	Exemplo do processo de obtenção das regras de associação para cada documento (Algoritmo 5) aplicado no conjunto de transações obtidos pelo processo apresentado na Figura 4.2. Para o documento em destaque do exemplo, foi calculado o valor de suporte mínimo <i>supmin</i> igual a 4.67%, considerando um valor de α igual a 0,3. No total, foram obtidas 514 regras para o documento <i>aprendizado_de_maquina.txt</i> . As regras do tipo $\emptyset \Rightarrow aprendiz$ são interessantes pois irão representar termos simples ao final do processo do LARCM (Seção 4.2.3).	50
4.4	Exemplo do processo de seleção das regras de associação para cada documento (Algoritmo 6) aplicado no conjunto de regras de associação obtidos pelo processo apresentado na Figura 4.3. A medida objetiva selecionada para o exemplo foi a <i>Odds Ratio</i> . Para o documento <i>aprendizado_de_maquina.txt</i> , o valor corte calculado para a medida foi de 0.07013369, ou seja, as regras de associação com valor maior ou igual a esse foram selecionadas para o conjunto d_S . Para esse documento, foram selecionadas 258 das 514 regras de associação geradas.	52
4.5	Exemplo de construção da matriz regra de associação-termo W_R	53
4.6	Visão geral da metodologia para avaliação do modelo proposto.	55
5.1	Visão geral do processo experimental deste trabalho de doutorado.	60
5.2	Resultados da acurácia média dos classificadores de cada modelo separados por coleção de documentos. O gráfico foi construído a partir dos valores apresentados na Seção B.1.	67

Lista de Tabelas

3.1	Padrão de matriz documento-termo W	30
3.2	Definição das variáveis do modelo LDA utilizadas no modelo gráfico da Figura 3.2.	40
4.1	Visão geral das soluções propostas pelos métodos para construção de representação de documentos que exploram a correlação entre termos relacionadas aos três desafios apresentados.	45
5.1	Descrição das coleções de textos utilizadas nos experimentos.	61
5.2	Valores dos parâmetros utilizados nos modelos avaliados para extração de tópicos.	62
5.3	Comparação múltipla das médias da acurácia dos classificadores para cada modelo de redução de dimensionalidade.	64
5.4	Resultados para as coleções ACM-1, ACM-2, ACM-3 e ACM-4 do valor de Coerência Observada (CO) do tópico com melhor avaliação e do tópico com menor valor de medida entre os 25% melhores tópicos. Os resultados estão ordenados pelo melhor valor de CO obtido. São exibidos os 3 melhores resultados para o LARCM considerando o melhor valor de CO e os 2 modelos obtidos com o LDA.	68
5.5	Resultados para as coleções ACM-5, ACM-6, ACM-7, ACM-8 e Re8 do valor de Coerência Observada (CO) do tópico com melhor avaliação e do tópico com menor valor de medida entre os 25% melhores tópicos. Os resultados estão ordenados pelo melhor valor de CO obtido. São exibidos os 3 melhores resultados para o LARCM considerando o melhor valor de CO e os 2 modelos obtidos com o LDA.	69
5.6	Comparação dos resultados obtidos pelos algoritmos de recomendação sensíveis ao contexto comparados com o algoritmo IBCF. O contexto foi obtido pela extração de tópicos com o modelo LARCM e o modelo LDA, utilizando o valor de K igual a 50.	72
5.7	Comparação dos resultados obtidos pelos algoritmos de recomendação sensíveis ao contexto comparados com o algoritmo IBCF. O contexto foi obtido pela extração de tópicos com o modelo LARCM e o modelo LDA, utilizando o valor de K igual a K=100.	72
5.8	Comparação dos resultados obtidos pelos algoritmos de recomendação sensíveis ao contexto comparados com o algoritmo IBCF. O contexto foi obtido pela extração de tópicos com o modelo LARCM e o modelo LDA, utilizando o valor de K igual a K=150.	73

5.9	Comparação dos melhores resultados obtidos pelos algoritmos de recomendação sensíveis ao contexto considerando todas as variações de parâmetros dos modelos. Foram selecionados os melhores valores obtidos pelo modelo LDA combinado com <i>bag-of-words</i> e <i>bag-of-related-words</i> , pelo modelo LARCM e os obtidos no artigo original de Sundermann et al. (2014).	73
A.1	Textos da coleção de documentos construída para os exemplos apresentados no texto. Cada documento foi construído selecionando o primeiro parágrafo da entrada da Wikipedia, versão em português, referente ao assunto do documento.	94
A.2	Textos pré-processados da coleção de documentos de exemplo apresentada na Tabela A.1.	95
A.3	Transações obtidas pelo mapeamento dos textos pré-processados da coleção de documentos de exemplo apresentada na Tabela A.2.	95
A.4	Transações obtidas pelo mapeamento dos textos pré-processados da coleção de documentos de exemplo apresentada na Tabela A.2.	96
A.5	Transações obtidas pelo mapeamento dos textos pré-processados da coleção de documentos de exemplo apresentada na Tabela A.2.	97
A.6	Transações obtidas pelo mapeamento dos textos pré-processados da coleção de documentos de exemplo apresentada na Tabela A.2.	98
A.7	Valores de quantidade total de regras de associação obtidas e selecionadas para cada documento da coleção de exemplo, e do limiar de corte calculado da medida objetiva <i>Odds Ratio</i> para cada documento da coleção de exemplos.	98
A.8	Tópicos obtidos ao aplicar o modelo LARCM com a medida objetiva <i>Odds Ratio</i> e K igual a 10. Foram selecionadas as 10 regras de associação com maior valor para a medida <i>Odds Ratio</i> como descriptores de cada tópico.	99
B.1	Comparação múltipla das médias da acurácia dos classificadores para cada modelo de redução de dimensionalidade na coleção de documentos ACM-1.	102
B.2	Comparação múltipla das médias da acurácia dos classificadores para cada modelo de redução de dimensionalidade na coleção de documentos ACM-2.	102
B.3	Comparação múltipla das médias da acurácia dos classificadores para cada modelo de redução de dimensionalidade na coleção de documentos ACM-3.	103
B.4	Comparação múltipla das médias da acurácia dos classificadores para cada modelo de redução de dimensionalidade na coleção de documentos ACM-4.	103
B.5	Comparação múltipla das médias da acurácia dos classificadores para cada modelo de redução de dimensionalidade na coleção de documentos ACM-5.	104
B.6	Comparação múltipla das médias da acurácia dos classificadores para cada modelo de redução de dimensionalidade na coleção de documentos ACM-6.	104
B.7	Comparação múltipla das médias da acurácia dos classificadores para cada modelo de redução de dimensionalidade na coleção de documentos ACM-7.	105
B.8	Comparação múltipla das médias da acurácia dos classificadores para cada modelo de redução de dimensionalidade na coleção de documentos ACM-8.	105
B.9	Comparação múltipla das médias da acurácia dos classificadores para cada modelo de redução de dimensionalidade na coleção de documentos Re8.	106
B.10	Acurácia dos classificadores para cada modelo de redução de dimensionalidade, com a quantidade de tópicos obtidos $k = 50$ na coleção de documentos ACM-1.	107
B.11	Acurácia dos classificadores para cada modelo de redução de dimensionalidade, com a quantidade de tópicos obtidos $k = 100$ na coleção de documentos ACM-1.	107

B.12 Acurácia dos classificadores para cada modelo de redução de dimensionalidade, com a quantidade de tópicos obtidos $k = 150$ na coleção de documentos ACM-1.	108
B.13 Acurácia dos classificadores para cada modelo de redução de dimensionalidade, com a quantidade de tópicos obtidos $k = 50$ na coleção de documentos ACM-2.	108
B.14 Acurácia dos classificadores para cada modelo de redução de dimensionalidade, com a quantidade de tópicos obtidos $k = 100$ na coleção de documentos ACM-2.	108
B.15 Acurácia dos classificadores para cada modelo de redução de dimensionalidade, com a quantidade de tópicos obtidos $k = 150$ na coleção de documentos ACM-2.	109
B.16 Acurácia dos classificadores para cada modelo de redução de dimensionalidade, com a quantidade de tópicos obtidos $k = 50$ na coleção de documentos ACM-3.	109
B.17 Acurácia dos classificadores para cada modelo de redução de dimensionalidade, com a quantidade de tópicos obtidos $k = 100$ na coleção de documentos ACM-3.	109
B.18 Acurácia dos classificadores para cada modelo de redução de dimensionalidade, com a quantidade de tópicos obtidos $k = 150$ na coleção de documentos ACM-3.	110
B.19 Acurácia dos classificadores para cada modelo de redução de dimensionalidade, com a quantidade de tópicos obtidos $k = 50$ na coleção de documentos ACM-4.	110
B.20 Acurácia dos classificadores para cada modelo de redução de dimensionalidade, com a quantidade de tópicos obtidos $k = 100$ na coleção de documentos ACM-4.	110
B.21 Acurácia dos classificadores para cada modelo de redução de dimensionalidade, com a quantidade de tópicos obtidos $k = 150$ na coleção de documentos ACM-4.	111
B.22 Acurácia dos classificadores para cada modelo de redução de dimensionalidade, com a quantidade de tópicos obtidos $k = 50$ na coleção de documentos ACM-5.	111
B.23 Acurácia dos classificadores para cada modelo de redução de dimensionalidade, com a quantidade de tópicos obtidos $k = 100$ na coleção de documentos ACM-5.	111
B.24 Acurácia dos classificadores para cada modelo de redução de dimensionalidade, com a quantidade de tópicos obtidos $k = 150$ na coleção de documentos ACM-5.	112
B.25 Acurácia dos classificadores para cada modelo de redução de dimensionalidade, com a quantidade de tópicos obtidos $k = 50$ na coleção de documentos ACM-6.	112
B.26 Acurácia dos classificadores para cada modelo de redução de dimensionalidade, com a quantidade de tópicos obtidos $k = 100$ na coleção de documentos ACM-6.	112
B.27 Acurácia dos classificadores para cada modelo de redução de dimensionalidade, com a quantidade de tópicos obtidos $k = 150$ na coleção de documentos ACM-6.	113
B.28 Acurácia dos classificadores para cada modelo de redução de dimensionalidade, com a quantidade de tópicos obtidos $k = 50$ na coleção de documentos ACM-7.	113

B.29 Acurácia dos classificadores para cada modelo de redução de dimensionalidade, com a quantidade de tópicos obtidos $k = 100$ na coleção de documentos ACM-7.	113
B.30 Acurácia dos classificadores para cada modelo de redução de dimensionalidade, com a quantidade de tópicos obtidos $k = 150$ na coleção de documentos ACM-7.	114
B.31 Acurácia dos classificadores para cada modelo de redução de dimensionalidade, com a quantidade de tópicos obtidos $k = 50$ na coleção de documentos ACM-8.	114
B.32 Acurácia dos classificadores para cada modelo de redução de dimensionalidade, com a quantidade de tópicos obtidos $k = 100$ na coleção de documentos ACM-8.	114
B.33 Acurácia dos classificadores para cada modelo de redução de dimensionalidade, com a quantidade de tópicos obtidos $k = 150$ na coleção de documentos ACM-8.	115
B.34 Acurácia dos classificadores para cada modelo de redução de dimensionalidade, com a quantidade de tópicos obtidos $k = 50$ na coleção de documentos Re8.	115
B.35 Acurácia dos classificadores para cada modelo de redução de dimensionalidade, com a quantidade de tópicos obtidos $k = 100$ na coleção de documentos Re8.	115
B.36 Acurácia dos classificadores para cada modelo de redução de dimensionalidade, com a quantidade de tópicos obtidos $k = 150$ na coleção de documentos Re8.	116
 C.1 Resultados para a coleção ACM-1 do valor de Coerência Observada (CO) do tópico com melhor avaliação e do tópico com menor valor de medida entre os 25% melhores tópicos. Os resultados estão ordenados pelo melhor valor de CO obtido.	118
C.2 Resultados para a coleção ACM-2 do valor de Coerência Observada (CO) do tópico com melhor avaliação e do tópico com menor valor de medida entre os 25% melhores tópicos. Os resultados estão ordenados pelo melhor valor de CO obtido.	119
C.3 Resultados para a coleção ACM-3 do valor de Coerência Observada (CO) do tópico com melhor avaliação e do tópico com menor valor de medida entre os 25% melhores tópicos. Os resultados estão ordenados pelo melhor valor de CO obtido.	120
C.4 Resultados para a coleção ACM-4 do valor de Coerência Observada (CO) do tópico com melhor avaliação e do tópico com menor valor de medida entre os 25% melhores tópicos. Os resultados estão ordenados pelo melhor valor de CO obtido.	121
C.5 Resultados para a coleção ACM-5 do valor de Coerência Observada (CO) do tópico com melhor avaliação e do tópico com menor valor de medida entre os 25% melhores tópicos. Os resultados estão ordenados pelo melhor valor de CO obtido.	122
C.6 Resultados para a coleção ACM-6 do valor de Coerência Observada (CO) do tópico com melhor avaliação e do tópico com menor valor de medida entre os 25% melhores tópicos. Os resultados estão ordenados pelo melhor valor de CO obtido.	123

C.7 Resultados para a coleção ACM-7 do valor de Coerência Observada (CO) do tópico com melhor avaliação e do tópico com menor valor de medida entre os 25% melhores tópicos. Os resultados estão ordenados pelo melhor valor de CO obtido.	124
C.8 Resultados para a coleção ACM-8 do valor de Coerência Observada (CO) do tópico com melhor avaliação e do tópico com menor valor de medida entre os 25% melhores tópicos. Os resultados estão ordenados pelo melhor valor de CO obtido.	125
C.9 Resultados para a coleção Re8 do valor de Coerência Observada (CO) do tópico com melhor avaliação e do tópico com menor valor de medida entre os 25% melhores tópicos. Os resultados estão ordenados pelo melhor valor de CO obtido.	126
D.1 Exemplo de matriz de similaridade de itens para algoritmos de recomendação baseados no IBCF.	128

Lista de Algoritmos

1	Apriori. Fonte: (Agrawal and Srikant, 1994)	18
2	Função <i>apriori-gen</i> . Fonte: (Agrawal and Srikant, 1994)	19
3	Gera Regras de Associação. Fonte: (Agrawal and Srikant, 1994)	20
4	Método Mapeamento-documento-transação	48
5	Método gera-regras	50
6	Método seleciona-regras	51

Introdução

Há aproximadamente 4000 anos, o homem tem organizado a informação para posterior recuperação e uso. O estudo “*The Digital Universe of Opportunities: Rich Data and the Increasing Value of the Internet of Things*” (Turner et al., 2014) publicado pela *International Data Corporation* - IDC¹ - estimou que a humanidade produziu, em 2013, 4.4 Zettabytes de dados (aproximadamente 4 trilhões de gigabytes). Comparativamente, uma estimativa anterior do IDC, realizada em 2008, concluiu que a humanidade produzia 487 exabytes (um Exabyte corresponde a aproximadamente um bilhão de gigabytes) naquele ano (Gantz and Reinsel, 2009). Embora seja lógico pressupor que grande parte desses dados sejam temporários, como *streaming* de vídeos ou comunicação de voz via *VOIP*², o grande aumento na sua geração sugere que as pessoas estejam armazenando mais dados e informações. Esse espantoso volume de dados e informações é tal que extrapola a capacidade humana de, manualmente, analisá-lo e compreendê-lo. Ainda de acordo com o estudo da IDC, apenas 22% dos dados produzidos em 2013 eram candidatos a serem analisados, sendo que apenas 5% foram de fato analisados para extração de algum tipo de conhecimento útil e relevante (Turner et al., 2014). Utilizar esses dados de forma inteligente é de grande interesse para as corporações, pois o conhecimento implícito pode ser valioso para a compreensão de algum processo ou para a obtenção de vantagem competitiva.

A extração automática de conhecimento a partir dessa crescente quantidade de dados armazenada digitalmente tornou-se uma tarefa de grande importância para as corporações. Essa extração pode ser obtida por meio da mineração de dados, o qual possui como principal característica a exploração dos dados, de forma a estruturar e evidenciar padrões, auxiliando na descoberta de conhecimento. Entretanto, nem todos os dados, como os documentos textuais, podem ser estruturados e organizados em um formato bem definido (Rios, 2013). Estudos³ indicam que 80% da informação das corporações no mundo é representada por dados não estruturados, em que uma parte significativa são documentos textuais. Muitos desses textos são armazenados em meio eletrônico e lançados diariamente na *Web*, formando grandes coleções de documentos, como relatórios, especificações de produtos, resumos, notas, correspondência eletrônica e toda variedade de publicações eletrônicas textuais - bibliotecas virtuais e acervos documentais variados (Han and Kamber, 2001; Aggarwal and Zhai, 2012).

¹<http://www.idc.com/>.

²*Voice Over IP* - Tecnologia de comunicação de voz pela rede de computadores que visa substituir uma ligação telefônica tradicional.

³<http://www.delphigroup.com/>.

Quando o conjunto de dados consiste de documentos textuais, utiliza-se uma especialização do processo de mineração de dados, denominada mineração de textos (Aggarwal and Zhai, 2012). A mineração de textos diferencia-se da mineração de dados pela incorporação de atividades que visam a estruturação dos documentos em um formato apropriado para a obtenção automática de conhecimento, sem que haja perda de informações relevantes em relação ao formato originalmente não-estruturado. Uma vez estruturados os documentos, algoritmos de mineração de dados convencionais podem ser aplicados para extrair conhecimento e informação por meio de padrões detectados em toda a coleção de documentos (Aggarwal and Zhai, 2012; Rios, 2013). Nesse contexto, a qualidade dos resultados obtidos com as abordagens automáticas para obtenção de conhecimento de textos estão fortemente relacionados à qualidade dos atributos utilizados para representar a coleção de documentos (Shafiei et al., 2007; Keikha et al., 2009; Aggarwal and Zhai, 2012).

O Modelo de Espaço de Vetores⁴ (MEV) é o modelo mais tradicional para obter uma representação estruturada dos documentos. Neste modelo, cada documento é representado por um vetor de pesos correspondentes aos atributos do texto. O modelo *bag-of-words* é a abordagem de MEV mais utilizada devido a sua simplicidade e aplicabilidade. O modelo *bag-of-words* acabou se mostrando muito eficiente, e muitos outros modelos mais complexos não atingiram uma melhora no desempenho substancial para substituir essa representação (Billhardt et al., 2002; Aggarwal and Zhai, 2012). Entretanto, o modelo *bag-of-words* não trata a dependência entre termos⁵ e possui alta dimensionalidade (Aggarwal and Zhai, 2012). O problema da alta dimensionalidade sempre foi um grande desafio para todos os algoritmos de aprendizagem e a “maldição da dimensionalidade” tem sido estudada há muito tempo (Shafiei et al., 2007; Aggarwal and Zhai, 2012). Além disso, o modelo *bag-of-words* não permite identificar os diferentes sentidos que uma mesma palavra pode ter em diferentes contextos, fenômeno conhecido como polissemia, ou compreender e representar a informação de termos diferentes ocorrendo com mesmo significado, fenômeno conhecido como sinonímia. Ainda, esse modelo não reconhece expressões multi-palavras (por exemplo, “*inteligência artificial*”) (Farahat and Kamel, 2011; Kalogeratos and Likas, 2012; Cheng et al., 2013b).

Existem muitas evidências na literatura de que representações que incorporam alguma informação de relação entre termos melhora os resultados de tarefas como recuperação de informação (Wong et al., 1985; Billhardt et al., 2002; Pôssas et al., 2002), classificação de textos (Lewis, 1992; Keikha et al., 2009; Figueiredo et al., 2011), agrupamento de textos (Beil et al., 2002; Zhang et al., 2010; Rossi and Rezende, 2011; Marcacini et al., 2012a; Kalogeratos and Likas, 2012; Cheng et al., 2013b) ou extração de tópicos (Wallach, 2006; Wu et al., 2010; Zhu et al., 2012; Lau et al., 2013). Diversos modelos para representação dos documentos foram propostos na literatura visando capturar a informação de relação entre termos, destacando-se os modelos baseados em frases ou termos compostos (Pôssas et al., 2002; Figueiredo et al., 2011; Rossi and Rezende, 2011), o Modelo de Espaço de Vetores Generalizado⁶ (MEVG) e suas extensões (Wong et al., 1985; Billhardt et al., 2002; Farahat and Kamel, 2011; Kalogeratos and Likas, 2012; Cheng et al., 2013b), modelos de tópicos não-probabilísticos, como o *Latent Semantic Analysis* (LSA) (Deerwester et al., 1990) ou o *Non-negative Matrix Factorization* (NMF) (Lee and Seung, 1999) e modelos de tópicos probabilísticos e suas extensões (Hofmann, 1999; Blei et al., 2003; Wallach, 2006; Kim et al., 2012; Zhu et al., 2012; Lau et al., 2013; Gao et al., 2013). Os modelos de tópicos são as abordagens mais interessantes uma vez que elas apresentam bons resultados

⁴Do inglês *Vector Space Model* - VSM.

⁵Neste trabalho, **termo**, **atributo** e **palavra** são, por vezes, utilizados com o mesmo significado, podendo se referir a elementos simples ou compostos, como “engenharia”, “engenharia civil” ou “engenharia de software”.

⁶Do inglês *Generalized Vector Space Model* - GVSM.

como modelo de representação de coleções de documentos e fornece uma estrutura que descreve a coleção de documentos em uma forma que revela sua estrutura interna e as suas inter-relações. Essas abordagens de extração de tópicos também fornecem uma estratégia de redução da dimensionalidade visando a construção de novas dimensões que representam os principais tópicos ou assuntos identificados na coleção de documentos. Ainda, uma organização baseada em tópicos agrupa termos com mesmo significado em um mesmo tópico (sinonímia) e permite que um mesmo termo ocorra em mais de um tópico caso ele possa ter significado diferente em diferentes contextos (polissemia). Uma representação construída com os novos atributos extraídos é uma oportunidade de incorporar conhecimento de domínio aos dados (Guyon and Elisseeff, 2003).

Entretanto, a extração eficiente de termos correlacionados para construção da representação de documentos ainda é um grande desafio de pesquisa (Figueiredo et al., 2011; Cheng et al., 2013b; Gao et al., 2013). Os modelos para representação de documentos que exploram a correlação entre termos normalmente enfrentam um grande desafio para manter um bom equilíbrio entre (i) a quantidade de dimensões obtidas, (ii) o esforço computacional e (iii) a interpretabilidade das novas dimensões obtidas. Os desafios (i) e (ii) estão fortemente relacionados, uma vez que a quantidade de possíveis combinações entre os termos de uma coleção contendo T termos é 2^T . Assim, existe um grande esforço computacional para a extração dessas combinações em grandes coleções de documentos e para determinar quais combinações são significativas para a aplicação (Figueiredo et al., 2011). Mesmo em modelos como o *Wallach's Bigram Topic Model* (Wallach, 2006) que possui um tratamento probabilístico teórico bem desenvolvido das correlações de termos, existe uma sobrecarga no esforço computacional uma vez que é preciso estimar T^2K parâmetros para obter K tópicos e T termos.

Por fim, a interpretabilidade das novas dimensões (desafio iii) tem um papel importante nos processos de mineração de textos. Em alguns casos, como nos modelos de extração de tópicos, a obtenção de dimensões facilmente interpretáveis pode ser entendido como o resultado desejado da tarefa. Existe um consenso nos trabalhos de extração de tópicos que a interpretabilidade de um tópico pode ser definida por uma medida conhecida na área como coerência, que normalmente possui resultados mais altos quando o conjunto de termos selecionados como descritores de um tópico se aproxima daquele que o usuário escolheria de forma “natural” para descrever o mesmo tópico. Os autores (Chang et al., 2009) apresentam uma discussão sobre a discordância entre os resultados obtidos por avaliações objetivas quando comparadas aos resultados de avaliações subjetivas realizadas com usuários sobre a qualidade dos tópicos obtidos por diferentes modelos. Os autores apontam que tópicos mais coerentes (com melhores notas nas avaliações subjetivas), são mais interessantes e métodos que melhoram a coerência devem ser usados em detrimento daqueles que otimizam somente os resultados das medidas objetivas.

O agrupamento de regras de associação pode ser utilizado para aproximar ou melhorar os resultados obtidos pelos modelos de extração de tópicos que exploram termos correlacionados considerando os desafios apontados anteriormente (Pôssas et al., 2002; Liu, 2011; Rossi and Rezende, 2011). O agrupamento de regras de associação extraídas dos documentos possibilita o desenvolvimento de uma estratégia para redução de dimensionalidade integrada ao processo de construção da nova representação. Segundo Liu (2011), regras de associação representam correlações ou coocorrências entre itens, e possuem duas vantagens. O algoritmo para geração de regras de associação é eficiente. Como, em geral, são necessárias apenas regras com 2 ou 3 termos, uma vez que termos compostos por mais de 3 termos simples são pouco frequentes, o algoritmo somente percorre a coleção de documentos 2 ou 3 vezes. Outra vantagem é que as regras de associação apresentam boa compreensibilidade. Apesar disso, pouca pesquisa tem sido feita nessa direção.

Sendo assim, a principal hipótese deste trabalho é que a extração de tópicos obtidos com o agrupamento de regras de associação extraídas dos documentos viabiliza a constru-

ção de uma representação dos documentos textuais que possui baixa dimensionalidade, melhora a interpretabilidade das novas dimensões e leva a uma melhor extração de conhecimento nas tarefas de classificação de textos e organização da informação. A seguir, são apresentadas outras hipóteses que dão suporte a hipótese principal:

- Existem termos que ocorrem juntos (um imediatamente após o outro) frequentemente, formando sequências de termos, ou expressões multi-palavras, que expressam um significado ou conceito diferente daquele expresso por cada termo individualmente, sendo essa sequência de termos denominada relação local do termo.
- Documentos de uma coleção que expressam um mesmo tópico levam à escolha de sequências de termos semelhantes para expressá-los, e estes termos influenciam na escolha dos termos da sua vizinhança, sendo essa influência na escolha dos termos denominada relação geral dos termos.
- Essas informações de relação geral e local dos termos podem ser extraídas diretamente dos documentos de uma coleção.
- Os tópicos que deram origem a um documento estão em um espaço latente, que pode ser estimado com a utilização das informações de relação local e geral dos termos da coleção.
- Existe uma representação estruturada dos tópicos encontrados que uma avaliação objetiva e/ou subjetiva possa indicar como melhor quando comparada com a representação obtida pelas técnicas consideradas estado da arte para extração de tópicos de documentos textuais.

O principal objetivo é o desenvolvimento de um modelo de extração de tópicos de documentos textuais que viabilize a representação dos documentos a partir das relações existente entre os termos da coleção, apresentando de forma explícita e estruturada essas relações. Esse modelo deve (i) possuir dimensionalidade reduzida, (ii) reduzir o esforço computacional e (iii) maximizar a interpretabilidade das dimensões obtidas quando comparado com modelos considerados estado da arte na área de extração de tópicos. Para isso, os seguintes objetivos específicos devem ser alcançados:

- Investigar e desenvolver soluções para avaliar, objetivamente e subjetivamente, as dimensões obtidas em relação às técnicas consideradas estado da arte para extração de tópicos de documentos textuais.
- Avaliar os métodos desenvolvidos para identificar o impacto da representação construída no processo de classificação de textos considerando as medidas objetivas da área.
- Avaliar a interpretabilidade das dimensões obtidas para identificar quais modelos apresentam melhores tópicos, considerando as medidas de avaliação da área.

Assim, é proposto neste trabalho o modelo para representação de documentos ***Latent Association Rule Cluster based Model*** (LARCM). Este é um modelo de extração de tópicos não-probabilístico que explora o agrupamento de regras de associação para construir uma representação da coleção de documentos com dimensionalidade reduzida tal que as novas dimensões são extraídas a partir das informações de correlação entre termos. No modelo proposto, as regras de associação são extraídas para cada documento para obter termos correlacionados que formam expressões multi-palavras. As informações obtidas nesse processo formam o contexto local da relação entre termos. Em seguida, aplica-se um processo de agrupamento em todas as regras de associação para formar o contexto

geral das relações entre os termos. Cada grupo de regras de associação obtido formará um tópico, ou seja, uma dimensão da representação. Os documentos são então mapeados nessas novas dimensões, formando uma matriz documento-tópico que pode ser utilizada para diferentes tarefas de mineração de textos e recuperação de informação. Também é proposto neste trabalho uma metodologia de avaliação que permite selecionar modelos que maximizam tanto os resultados na tarefa de classificação de textos quanto os resultados de interpretabilidade dos tópicos obtidos.

Para melhor entendimento da pesquisa descrita nesta tese, no Capítulo 2, são apresentados os principais conceitos da mineração de textos, de agrupamento de dados e textos, e de regras de associação. No Capítulo 3, é realizada uma revisão dos modelos para representação de documentos, bem como abordagens que utilizam a mineração de textos para construir a representação e o uso de termos dependentes para essa tarefa. Adicionalmente, são apresentadas as principais abordagens da literatura para redução de dimensionalidade, com foco na extração de atributos. Também, é apresentada uma discussão do estado da arte dos modelos de extração de tópicos probabilísticos e não-probabilísticos, bem como suas extensões que fazem uso da informação de dependência entre termos. No Capítulo 4, apresentam-se as motivações para a proposição e o uso do LARCM, bem como o próprio modelo e uma metodologia de avaliação dos resultados. Os resultados das avaliações são discutidos no Capítulo 5. Também é apresentado um estudo de caso da aplicação do modelo proposto para sistemas de recomendação sensíveis ao contexto. Por fim as conclusões do trabalho, as limitações dessa pesquisa e os trabalhos futuros são apresentados no Capítulo 6.

Mineração de Dados e Textos

A mineração de textos é uma área de pesquisa cujo objetivo é a busca por padrões, tendências e regularidades em textos escritos em língua natural e pode ser vista como uma especialização do processo de mineração de dados. Enquanto esta última trabalha sobre bases de dados com estrutura pré-definida, a mineração de textos trabalha sobre dados inherentemente não estruturados (Weiss et al., 2005). Este processo pode ser dividido em uma sequência de etapas genéricas, formando um ciclo no qual, ao final, disponibiliza-se o conhecimento acerca dos textos analisados de acordo com o objetivo especificado.

2.1 Processo de Mineração de Textos

A mineração de textos envolve um conjunto de técnicas e processos que busca descobrir conhecimento inovador em textos. Ela se transformou em um importante aliado às mais diversas corporações, agências de informação e entidades que necessitam gerenciar e extrair conhecimento de grandes coleções de documentos textuais que possam ser úteis para algum processo de tomada de decisão (Aggarwal and Zhai, 2012) . É um processo no qual há grande interação por parte do usuário e grande incorporação de conhecimento por parte dos especialistas nas etapas, que visa extrair informação útil por meio da identificação e exploração de padrões interessantes (Feldman and Sanger, 2007).

O processo de mineração de textos pode ser dividido em cinco grandes fases: Identificação do Problema, Pré-processamento, Extração de Padrões, Pós-processamento e Utilização do Conhecimento. O ciclo formado por essas etapas é apresentado na Figura 2.1.

A **Identificação do Problema** é uma etapa muito importante, dado que as decisões tomadas neste ponto guiarão os passos consecutivos e poderão ter reflexo no desempenho da aplicação. Nesta etapa o especialista do domínio identifica e delimita o problema, o subdomínio do problema, a coleção de textos¹ a ser analisada ou sua fonte de busca, se existe algum conhecimento prévio de domínio que possa ser utilizado na análise, o que se espera obter e como os resultados poderão ser utilizados. É uma etapa que demanda muito esforço tanto do especialista do domínio quanto do especialista em Mineração de Textos, pois a mesma fornece subsídios a todo o processo, permitindo identificar requisitos e possíveis técnicas a serem utilizadas em cada passo.

¹Neste trabalho, os termos “coleção de textos” e “base de textos” são considerados sinônimos.



Figura 2.1: Etapas do processo de Mineração de Textos. Fonte: Rezende et al. (2003).

Na etapa de **Pré-Processamento** encontra-se a principal diferença entre o processo de Mineração de Dados e o de Mineração de Textos. O problema resume-se ao fato dos textos não estarem sempre em um formato adequado para a Extração de Padrões; é necessário adequá-los para um formato manipulável por algoritmos utilizados nas etapas seguintes, além de aplicar um processo de tratamento, limpeza e, em geral, redução do volume de textos, sempre preservando as características necessárias para que os objetivos sejam cumpridos.

Na etapa de **Extração de Padrões** são realizadas as tarefas necessárias para o cumprimento dos objetivos delimitados na identificação do problema. Para a realização dessas tarefas pode-se utilizar, por exemplo, modelos de classificação, regressão, regras de associação, agrupamento (*clustering*) e sumarização. Pode-se resumir as principais atividades de extração de padrões em textos em duas grandes tarefas: **preditivas** e **descritivas**.

Para as atividades **preditivas** são utilizados algoritmos de aprendizado de máquina supervisionados. Esses algoritmos, conhecidos como indutores, exigem um conjunto de exemplos de treinamento para os quais o atributo classe tenha valor conhecido (Mitchell, 1997; Monard and Baranauskas, 2003). Esta classe de algoritmos se divide em duas subclasses principais: os algoritmos de **classificação** e os algoritmos de **regressão**. **Classificação** é referente ao processo em que o atributo classe tem valor categórico, enquanto **regressão** remete à predição de variáveis com valores numéricos. A aplicação de tarefas preditivas em mineração de textos tem como principal aplicação a categorização automática de documentos.

As atividades **descritivas**, por sua vez, utilizam algoritmos de aprendizado de máquina não-supervisionados. Esse tipo de atividade consiste na identificação de comportamentos intrínsecos da coleção de dados, sendo que esses dados não possuem rótulos ou são tratados como não rotulados. As principais tarefas deste ramo da mineração de textos são a obtenção de regras de associação, o agrupamento e a sumarização de documentos (Mitchell, 1997; Monard and Baranauskas, 2003). Dessas técnicas, regras de associação e agrupamento serão utilizadas neste trabalho. As **regras de associação** são relações inferidas entre dados correlacionados de uma ou mais bases de dados, analisados conjuntamente (Agrawal and Srikant, 1994). O **agrupamento de dados**, também conhecido como *clustering*, visa agrupar objetos de forma que os objetos de um mesmo grupo sejam similares (ou relacionados) uns ao outros e diferentes (ou não relacionados) aos objetos de outros grupos (Manning et al., 2008).

O **Pós-Processamento** é a etapa de validação das descobertas efetuadas pela etapa

de processamento dos dados e da visualização dos resultados encontrados. Nesta fase, o conhecimento extraído é então utilizado seja por ferramentas de visualização ou simplesmente por tabelas de resultados, não sendo a etapa final do processo de mineração. A análise minuciosa dos resultados obtidos permite que se valide a sua utilidade e até mesmo o próprio processo, determinando a necessidade de retomar passos anteriores, reestruturando-os. Nesta etapa o especialista do domínio e o de Mineração de Textos devem trabalhar juntos, procurando responder a questões como: representatividade do conhecimento obtido; o que há de novo nos resultados encontrados; de que maneira o conhecimento do especialista difere do obtido; validação dos resultados obtidos; identificação da adequação de procedimentos nas etapas anteriores para tentar melhorar os resultados; e de que maneira os resultados obtidos devem ser utilizados.

Na etapa de **Utilização do Conhecimento** os resultados estão validados e aptos a serem utilizados. Dessa forma, o conhecimento extraído pode ser aplicado para apoio à tomada de decisão, conforme objetivo pré-estabelecido na etapa de Identificação do Problema.

Esse processo de mineração de textos será instanciado, nesse trabalho, no processo proposto para construção de uma representação estruturada da coleção de documentos com extração de tópicos. Na Seção 2.2 são apresentados alguns conceitos importantes de técnicas de agrupamento de dados e documentos. Na Seção 2.3 é apresentada a técnica de regras de associação. E na Seção 2.4, são apresentadas as principais estratégias para agrupamento de regras de associação, que serão aplicadas para viabilizar a execução do modelo proposto.

2.2 Agrupamento de Documentos

Agrupamento de documentos é a organização de um conjunto de documentos em grupos, baseado em uma medida de similaridade, no qual documentos de um mesmo grupo são altamente similares entre si mas dissimilares em relação aos documentos de outros grupos (Manning et al., 2008). Em outras palavras, o agrupamento é baseado no princípio de maximizar a similaridade interna dos grupos (intragrupo) e minimizar a similaridade entre os grupos (intergrupos) (Everitt et al., 2001). A análise de agrupamento também é conhecida como aprendizado por observação ou análise exploratória dos dados, pois a organização dos documentos em grupos é realizada apenas pela observação de regularidades nos dados, sem uso de conhecimento externo. Assim, ao contrário de métodos supervisionados, como algoritmos de classificação, em processos de agrupamento não há classes ou rótulos predefinidos para treinamento de um modelo, ou seja, o aprendizado é realizado de forma não-supervisionada.

O processo de agrupamento depende de dois fatores principais: (1) a medida de proximidade e (2) o método de agrupamento escolhido. As **medidas de proximidade** determinam como a similaridade entre dois documentos é calculada. Sua escolha influencia a forma como os grupos são induzidos e depende dos tipos de variáveis ou atributos que representam os documentos. Existe uma variedade de medidas de proximidades e as principais medidas utilizadas em dados textuais são discutidas na Seção 2.2.1. Por fim, na Seção 2.2.2 são discutidos os **métodos de agrupamento**, ou seja, as estratégias adotadas para definição dos grupos.

2.2.1 Medidas de Proximidade

A escolha da medida de proximidade para calcular o quanto similar são dois objetos é fundamental para a análise de agrupamentos. Essa escolha depende das características do conjunto de dados, principalmente dos tipos e escala dos dados. Assim, existem medidas

de proximidade para dados contínuos, discretos e misturas entre dados contínuos e discretos. As medidas de proximidade podem calcular tanto a similaridade quanto distância (ou dissimilaridade) entre objetos. No entanto, as medidas de similaridades podem ser, geralmente, convertidas para medidas de distância, e vice-versa, por exemplo, calculando-se o complemento entre elas: $\text{sim}(x_i, x_j) = 1 - \text{dist}(x_i, x_j)$ (Tan et al., 2005b).

No contexto deste trabalho, pode-se citar duas medidas de proximidades comumente utilizadas em dados textuais: cosseno e jaccard, conforme descrito abaixo.

- **Cosseno:** considera que os documentos estão representados em um espaço m -dimensional, no qual cada atributo (termo ou palavra) representa uma dessas dimensões. Assim, considerando os documentos $x_i = (x_{i1}, x_{i2}, \dots, x_{im})$ e $x_j = (x_{j1}, x_{j2}, \dots, x_{jm})$ e um espaço com m atributos, a similaridade cosseno entre x_i e x_j é calculada pela Equação 2.1.

$$d(x_i, x_j) = \text{cosine}(x_i, x_j) = \frac{x_i \bullet x_j}{|x_i| |x_j|} = \frac{\sum_{l=1}^m x_{il} x_{jl}}{\sqrt{\sum_{l=1}^m x_{il}^2} \sqrt{\sum_{l=1}^m x_{jl}^2}} \quad (2.1)$$

O valor da similaridade representa o cosseno do ângulo formado entre os vetores e é útil quando a magnitude dos dados não é importante. Em documentos textuais, os atributos representam termos (palavras) existentes e os valores podem ser, por exemplo, provenientes da frequência de ocorrência desses termos. Assim, a medida cosseno captura quantos termos são compartilhados por ambos documentos, ponderando-os pelos valores de ocorrência.

- **Jaccard:** em algumas situações os atributos dos documentos são binários, ou seja, indicam a presença ou ausência de alguma característica. Sejam x_i e x_j dois documentos com atributos binários, obtém-se as seguintes contagens:

- f_{00} = número de atributos com valor 0 para ambos objetos;
- f_{11} = número de atributos com valor 1 para ambos objetos;
- f_{01} = número de atributos com valor 0 para x_i e valor 1 para x_j ; e
- f_{10} = número de atributos com valor 1 para x_i e valor 0 para x_j .

A medida Jaccard define a proximidade entre dois objetos com atributos assimétricos, ou seja, a presença de um efeito é mais importante que sua ausência. O coeficiente de Jaccard é definido na Equação 2.2.

$$d(x_i, x_j) = \frac{f_{11}}{f_{01} + f_{10} + f_{11}} \quad (2.2)$$

Uma dúvida pertinente que surge com relação às medidas de proximidade é qual escolher no processo de agrupamento. Não existe uma regra geral para esta escolha. Geralmente, esta decisão é tomada conforme a representação dos dados e é acompanhada de vários testes, seguidos por um processo de validação da qualidade dos grupos obtidos.

2.2.2 Métodos de Agrupamento

Os métodos de agrupamento podem ser classificados considerando diferentes aspectos. Jain et al. (1999) organizam os métodos de agrupamento de acordo com a estratégia adotada para definir os grupos. Uma análise de diferentes métodos de agrupamento considerando o cenário de Mineração de Dados é apresentada em Berkhin (2006). Muitos desses métodos foram avaliados no contexto de mineração de textos em Zhao et al. (2005).

A seguir, são abordados os métodos de agrupamento particional e hierárquico, discutindo os algoritmos comumente citados na literatura.

Métodos Particionais

Os métodos particionais também são conhecidos como métodos de otimização. O objetivo é formar o agrupamento dividindo iterativamente o conjunto de objetos em k grupos, na qual k geralmente é um valor informado previamente pelo usuário. Os grupos são formados otimizando um determinado critério, geralmente uma função baseada na similaridade que busca a compactação e a separação entre os grupos (Everitt et al., 2001).

O algoritmo *k-means* é um dos mais populares dos métodos particionais (Wu et al., 2007). Neste algoritmo, cada grupo possui um representante denominado centroide. O centroide de um grupo é um objeto formado com os valores médios dos atributos dos objetos desse grupo. O *k-means*, em sua proposta original, só é aplicável em dados que possuam atributos numéricos, nos quais a média possa ser calculada. Os passos de execução do algoritmo *k-means* é exibido a seguir.

1. Perguntar ao usuário em quantos grupos (k) o conjunto de documentos será dividido;
2. Selecionar aleatoriamente k documentos como centroides iniciais;
3. Associar cada documento do conjunto ao centroide mais próximo, de acordo com uma medida de proximidade. A partir desta etapa, cada objeto está alocado em um grupo;
4. Atualizar os centroides representantes de cada grupo;
5. Repetir os passos 3 e 4 até um critério de parada, por exemplo, quando a solução convergir ou um determinado número de iterações ocorrer.

O critério de convergência do *k-means* é determinado pela soma dos erros quadráticos, definido como

$$E = \sum_{i=1}^k \sum_{x \in C_i} |x - m_i|^2 \quad (2.3)$$

na qual E é a soma dos erros quadráticos para todos os documentos; x é o vetor de atributos que representa um dado documento; e m_i é o centroide do grupo C_i . Ao minimizar este critério, o *k-means* tenta separar o conjunto de dados em k grupos, otimizando a compactação de cada grupo e a separação entre os grupos. A complexidade do *k-means* é linear em relação ao número de documentos, o que possibilita sua aplicação eficiente em diversos cenários. No entanto, a necessidade de informar com antecedência o número de grupos pode ser visto como uma desvantagem, pois este valor geralmente é desconhecido pelos usuários. Além disso, o método apresenta variabilidade nos resultados, pois a seleção dos centroides iniciais afeta o resultado do agrupamento, inclusive com a possibilidade de que a solução represente um mínimo local. Para minimizar esse efeito, o algoritmo é executado diversas vezes, com várias inicializações diferentes, e o melhor resultado é selecionado.

Métodos Hierárquicos

O diferencial desta abordagem é a obtenção de uma organização hierárquica, com grupos e seus subgrupos, representada por um dendrograma, uma estrutura especial de

árvore. Os algoritmos de agrupamento hierárquico são utilizados em diversos tipos de problemas, uma vez que não exigem informação inicial a respeito do conjunto de dados, por exemplo, o número de grupos (Manning et al., 2008).

Os algoritmos de agrupamento hierárquico são classificados de acordo com sua estratégia de implementação: Aglomerativos (*bottom-up*) e Divisivos (*top-down*). Na primeira, cada objeto é considerado um único grupo. Em seguida, pares de objetos são iterativamente agrupados de acordo com um índice de similaridade, até que todos os objetos pertençam a apenas um grupo. Por outro lado, a estratégia divisiva é iniciada com todos os objetos em um único grupo, que é iterativamente bipartido até que cada objeto seja um único grupo.

A maioria dos trabalhos relacionados com agrupamento hierárquico na literatura referenciam-se às estratégias aglomerativas, mostrando pouco interesse nas estratégias divisivas. Isto pode ser explicado pelo fato de que as estratégias aglomerativas geralmente apresentam complexidade quadrática, ou seja $O(N^2)$, que apesar de relativamente custosa, são muitas vezes aplicáveis. Já a complexidade das estratégias divisivas crescem exponencialmente em relação ao tamanho do conjunto de dados, proibindo sua aplicação em conjuntos de dados grandes (Xu and Wunsch, 2008). No entanto, nos últimos anos foram propostos algoritmos de agrupamento hierárquico divisivos com complexidade semelhante aos aglomerativos, possibilitando sua aplicação em conjuntos de dados maiores, inclusive em coleções textuais (Steinbach et al., 2000; Marcacini et al., 2012b).

Estratégias Aglomerativas

Considerando um conjunto com N documentos, a seguir são descritos os passos executados em um agrupamento hierárquico aglomerativo.

1. Inicialmente, cada documento é considerado um único grupo, ou seja, existem N grupos unitários. Calcule uma matriz de proximidades entre os N grupos;
2. Encontre a distância mínima $d(C_i, C_j) = \min(C_m, C_l)$, com $1 \leq m, l \leq N$, $m \neq l$, em que $d(\bullet, \bullet)$ é um critério de distância entre grupos discutido adiante. Unir os grupos C_i e C_j na formação de um novo grupo C_{ij} ;
3. Atualize a matriz de proximidades calculando as distâncias entre C_{ij} e os grupos restantes;
4. Repetir os passos 2 e 3 até que um único grupo seja formado.

A diferença entre os algoritmos de agrupamento hierárquico aglomerativos está no critério de distância utilizado para unir os grupos C_i e C_j , na formação de um novo grupo C_{ij} (passo 2). Os algoritmos principais são conhecidos como Single Link, Complete Link e Average Link (Manning et al., 2008).

O **Single Link** é um dos algoritmos de agrupamento hierárquico mais simples. Esse método utiliza a técnica do vizinho mais próximo (*Nearest Neighbor*), na qual a distância

entre dois grupos é determinada pela distância do par de objetos mais próximos, sendo cada objeto pertencente a um desses grupos. Esse método de união de grupos apresenta um problema conhecido como “efeito da corrente”, em que ocorre a união indevida de grupos influenciada pela presença de ruídos na base de dados. O critério de distância entre grupos do Single Link é descrita na Equação 2.4.

$$D(C_l, (C_i, C_j)) = \min(D(C_l, C_i), D(C_l, C_j)). \quad (2.4)$$

Já o **Complete Link** utiliza uma técnica conhecida como *Farthest Neighbor*, ou vizinho mais distante. Ao contrário do algoritmo Single Link, esse algoritmo determina a distância entre dois grupos de acordo com a maior distância entre um par de objetos, sendo cada objeto pertencente a um grupo distinto. Esse método dificulta a formação do efeito da corrente, como ocorre no Single Link, e tende a formar grupos mais compactos e em formatos esféricos. O critério de distância entre grupos do Complete Link é definida na Equação 2.5.

$$D(C_l, (C_i, C_j)) = \max(D(C_l, C_i), D(C_l, C_j)). \quad (2.5)$$

O Single Link e Complete Link são opostos em termos do critério utilizado para união dos grupos. Uma abordagem intermediária é o algoritmo Average Link, também conhecido como UPGMA. No algoritmo **Average Link**, a distância entre dois grupos é definida como a média das distâncias entre todos os pares de objetos em cada grupo, cada par composto por um objeto de cada grupo. Esse método elimina muitos problemas relacionados à dependência do tamanho dos grupos, mantendo a próxima variabilidade interna entre eles. O critério de distância entre grupos do Average Link é definida na Equação 2.6.

$$D(C_l, (C_i, C_j)) = \frac{1}{2}(D(C_l, C_i) + D(C_l, C_j)). \quad (2.6)$$

A escolha do critério de união de grupos dos algoritmos aglomerativos depende geralmente do conjunto de dados e dos objetivos da aplicação. Por exemplo, em dados textuais, avaliações experimentais têm mostrado o Average Link como uma boa opção entre os algoritmos que adotam estratégias aglomerativas (Zhao et al., 2005).

Estratégias Divisivas

O principal algoritmo de agrupamento hierárquico divisivo adotado em coleções textuais é conhecido como **Bi-Secting K-Means** (Steinbach et al., 2000) e seu funcionamento será descrito a seguir.

O algoritmo Bi-Secting K-Means é, essencialmente, o uso sucessivo do algoritmo K-Means clássico, de forma a construir um agrupamento hierárquico. No algoritmo K-Means clássico, é necessário informar um número k de grupos desejado e, então, é encontrada a melhor partição (conjunto de grupos não hierárquico) em relação ao k .

Com base no K-Means clássico, o algoritmo Bi-Secting K-Means inicia com um único

grupo de todos os documentos e executa os seguintes passos:

1. Escolha o grupo para dividir;
2. Encontre dois sub-grupos usando o algoritmo K-Means com $k=2$;
3. Repita os passos 1 e 2 até que cada grupo seja um único documento.

As estratégias divisivas baseada no Bi-Secting K-Means têm se mostrado uma alternativa razoável às estratégias aglomerativas Zhao et al. (2005). No entanto, conforme mencionado anteriormente, este método é sensível à escolha inicial dos centroides na etapa do K-Means. Uma forma de contornar este problema é repetir o processo de divisão do grupo em dois sub-grupos, com o K-Means, diversas vezes e escolher a solução mais promissora.

Tanto as estratégias aglomerativas quanto as divisivas organizam os resultados do agrupamento em uma árvore binária conhecida como dendrograma. Esta representação é uma forma intuitiva de visualizar e descrever a sequência do agrupamento e a similaridade com que os grupos foram formados.

Na Figura 2.2 é apresentado um exemplo de dendrograma. Essa estrutura é uma árvore com N folhas e com altura $N - 1$, considerando uma coleção de N documentos. Os documentos são dispostos no eixo horizontal, enquanto que o eixo vertical indica a similaridade entre os grupos. O topo do dendrograma (primeiro nível) corresponde a um grupo contendo todos os documentos da coleção. O último nível contém os grupos unitários (nós folhas), e cada nó² é um documento da coleção. Os nós intermediários e as uniões entre eles representam a sequência do agrupamento (Metz, 2006).

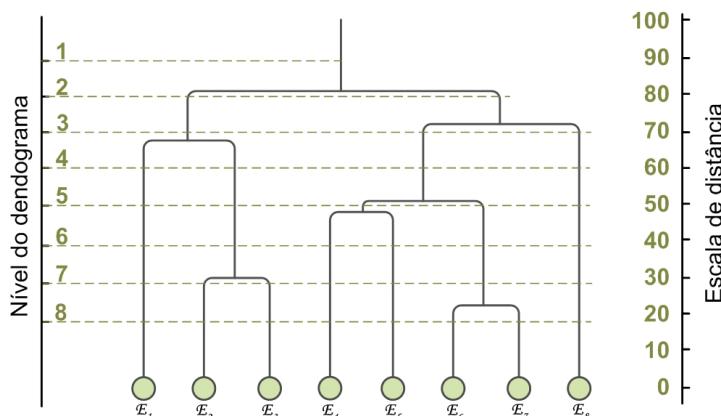


Figura 2.2: Análise da Estrutura do Dendrograma. Fonte: Metz (2006)

O dendrograma também possibilita extrair uma partição, ou seja, um conjunto de grupos não hierárquico. Por exemplo, ainda na Figura 2.2, pode-se extrair uma partição com 3 grupos, com distância em torno de 70, representado pela linha do nível 3 na figura. A extração de uma partição é muito comum quando o interesse é obter uma organização *flat* da coleção.

²Os termos *nó* e *grupo* são utilizados indistintamente neste trabalho.

É possível analisar a qualidade do agrupamento diretamente pelo dendrograma. A altura dos arcos que unem dois subgrupos indicam o grau de compactação do grupo formado por eles. Quanto menor a altura, mais compactos são os grupos. No entanto, também espera-se que os grupos formados sejam distantes entre si, ou seja, que a similaridade de documentos em grupos distintos seja a menor possível. Essa característica é representada quando existe uma grande diferença entre a altura de um arco e os arcos formados abaixo dele. Desta forma, a diferença entre os algoritmos de agrupamento hierárquico são expressados diretamente pelo dendrograma.

2.3 Regras de Associação

Desde que foi proposta por Agrawal et al. (1993) a técnica de regras de associação têm recebido grande atenção. A aplicação direta em problemas de negócios junto com sua comprehensibilidade inerente - não só para especialistas em mineração de dados - tornaram essa técnica muito popular (Hipp et al., 2002).

A mineração de regras de associação surgiu da análise de dados de cestas de compras, e busca construir regras do tipo “clientes que compram os produtos x_1, x_2, \dots, x_n também irão comprar o produto y com probabilidade $c\%$ ”. Apesar de sua ligação com o contexto varejista, as regras geradas apresentam, hoje em dia, uma ampla gama de aplicações em que são bem sucedidas, como na detecção de invasão de redes de computadores (Brauckhoff et al., 2012).

Nesta seção são abordados os conceitos e definições de regras de associação, a forma de geração de regras e medidas de interesse importantes. Também é apresentada a sintaxe para representar regras de associação utilizada neste trabalho.

2.3.1 Definições e Conceitos

Uma regra de associação caracteriza o quanto a presença de um conjunto de itens nos registros de uma Base de Dados implica na presença de algum outro conjunto distinto de itens nos mesmos registros (Agrawal and Srikant, 1994). Desse modo, o objetivo das regras de associação é encontrar tendências que possam ser usadas para entender e explorar padrões de comportamento dos dados. Por exemplo, observando os dados de vendas de um supermercado, sabe-se que 80% dos clientes que compram o produto Q também adquirem, na mesma ocasião, o produto W . Em outras palavras, pode-se dizer que essa regra apresenta confiabilidade de 80%.

O formato de uma regra de associação pode ser representado como uma implicação $LHS \Rightarrow RHS$, em que LHS e RHS são, respectivamente, o lado esquerdo (*Left Hand Side*) e o lado direito (*Right Hand Side*) da regra, definidos por conjuntos disjuntos de itens. As regras de associação podem ser definidas como descrito a seguir (Agrawal and Srikant, 1994):

Seja D uma Base de Dados composta por um conjunto de itens $A = \{a_1, \dots, a_m\}$ ordenados lexicograficamente e por um conjunto de transações $T = \{t_1, \dots, t_n\}$,

na qual cada transação $t_i \in T$ é composta por um conjunto de itens (*itemset*) tal que $t_i \subseteq A$, com $i = 1, \dots, n$.

Em uma Regra de Associação $LHS \Rightarrow RHS$, em que $LHS \subset A$, $RHS \subset A$ e $LHS \cap RHS = \emptyset$. A regra $LHS \Rightarrow RHS$ ocorre no conjunto de transações T com confiança $conf$ se em $conf\%$ das transações de T em que ocorre LHS ocorre também RHS . A regra $LHS \Rightarrow RHS$ tem suporte sup se em $sup\%$ das transações em D ocorre $LHS \cup RHS$.

O valor do suporte mede a força da associação entre LHS e RHS e não relaciona possíveis dependências de RHS com LHS . Por outro lado, a confiança mede a força da implicação descrita pela regra (Zhang and Zhang, 2002). Essas duas medidas são definidas a seguir:

Suporte: quantifica a incidência de um *itemset* X ou de uma regra no conjunto de dados, ou seja, indica a frequência com que X ou com que $LHS \cup RHS$ ocorre no conjunto de dados. Da maneira como foi definido, o suporte para um *itemset* X pode ser representado por:

$$sup(X) = \frac{n(X)}{N}, \quad (2.7)$$

em que $n(X)$ é o número de transações nas quais X ocorre e N é o número total de transações consideradas. Já o suporte de uma regra $LHS \Rightarrow RHS$ pode ser representado por:

$$sup(LHS \Rightarrow RHS) = sup(LHS \cup RHS) = \frac{n(LHS \cup RHS)}{N}, \quad (2.8)$$

em que $n(LHS \cup RHS)$ é o número de transações nas quais LHS e RHS ocorrem juntos e N é o número total de transações consideradas.

Confiança: indica a frequência com que LHS e RHS ocorrem juntos em relação ao número total de transações em que LHS ocorre. Do modo como foi definida, a confiança de uma regra $LHS \Rightarrow RHS$ pode ser representada por:

$$conf(LHS \Rightarrow RHS) = \frac{sup(LHS \cup RHS)}{sup(LHS)} = \frac{n(LHS \cup RHS)}{n(LHS)}, \quad (2.9)$$

em que $n(LHS)$ é o número de transações nas quais LHS ocorre.

Usualmente, valores de suporte e confiança mínimos são definidos pelo usuário antes da mineração das regras de associação. Em geral, a definição de altos valores para esses parâmetros gera apenas regras triviais; já a definição de baixos valores gera, em geral, um grande volume de conhecimento no formato de regras, dificultando a análise do usuário na etapa de Pós-processamento. Uma maneira de superar as dificuldades na análise dessas regras no Pós-processamento é utilizar-se de algoritmos que possibilitam usar taxonomias já durante a etapa de Extração de Padrões.

O problema de obtenção de regras de associação é decomposto em dois sub-problemas (passos) (Agrawal et al., 1993):

1. Encontrar todos os k -*itemsets* (conjunto de k itens) que possuam suporte maior ou igual ao suporte mínimo especificado pelo usuário (**sup-min**). Os *itemsets* com suporte igual ou superior a **sup-min** são definidos como *itemsets* frequentes, os demais conjuntos são denominados de *itemsets* não-frequentes;
2. Utilizar todos os k -*itemsets* frequentes, com $k \geq 2$, para gerar as regras de associação. Para cada *itemset* frequente $l \subseteq A$, encontrar todos os subconjuntos \tilde{a} de itens não vazios de l . Para cada subconjunto $\tilde{a} \subseteq l$, gerar uma regra na forma $\tilde{a} \Rightarrow (l - \tilde{a})$ se a razão de $sup(l)$ por $sup(\tilde{a})$ é maior ou igual a confiança mínima especificada pelo usuário (**conf-min**).

Com um conjunto de *itemsets* frequentes $\{a, b, c, d\}$ e um subconjunto de *itemsets* frequentes $\{a, b\}$, por exemplo, pode-se gerar uma regra do tipo $ab \Rightarrow cd$, desde que $conf(ab \Rightarrow cd) \geq \text{conf-min}$, em que, $conf(ab \Rightarrow cd) = sup(a, b, c, d)/sup(a, b)$.

Para que não haja necessidade de se percorrer todo o espaço de busca à procura dos *itemsets* de fato frequentes, algoritmos modernos para minerar regras de associação utilizam um método que gera e testa *itemsets* candidatos. Esses algoritmos geram conjuntos de *itemsets* potencialmente frequentes chamados conjuntos de *itemsets* candidatos. Utilizam-se da propriedade de linha de fronteira (*downward closure*) do suporte de um *itemset* (Agrawal and Srikant, 1994) que determina que todo subconjunto de um *itemset* frequente deve ser frequente (para remover os *itemsets* com pelo menos um subconjunto de itens não-frequentes). Então, calcula-se o valor de suporte para cada *itemset* candidato (não removido) utilizando a base de dados D, removendo-se, na sequência, os *itemsets* candidatos com suporte inferior ao suporte mínimo (**sup-min**) definido pelo usuário. O algoritmo inicia uma nova iteração utilizando os *itemsets* frequentes gerados na última iteração e tem encerrada sua execução quando não houver nenhum *itemset* potencialmente frequente podendo ser considerado um *itemset* candidato.

Existem diversas formas de implementar o processo de geração de regras de associação, que têm um impacto significativo no desempenho final do processo, e a escolha do algoritmo correto depende do objetivo final da aplicação. Para este trabalho foi escolhido o algoritmo *Apriori* (Agrawal and Srikant, 1994), amplamente utilizado para obtenção de *itemsets* frequentes. Este algoritmo e também um algoritmo simples para obtenção de regras de associação são apresentados na próxima seção.

2.3.2 Geração de Regras de Associação

A obtenção de *itemsets* frequentes para gerar regras de associação pode ser realizada utilizando diversos algoritmos, como: *AIS* (Agrawal et al., 1993), *Apriori* e *AprioriTid* (Agrawal and Srikant, 1994), *SETM* (Houtsma and Swami, 1995), *Opus* (Webb, 1995), *Direct Hashing and Prunning (DHP)* (Park et al., 1997), *Dynamic Set Counting (DIC)* (Brin et al., 1997), *FP-growth* (Han et al., 2000), *Closet* (Pei et al., 2000) e *Charm* (Zaki and Hsiao, 2002). Dentre eles, o *Apriori* (Agrawal and Srikant, 1994), que é amplamente utilizado (Zhang and Zhang, 2002), será apresentado a seguir.

O Algoritmo Apriori

O algoritmo *Apriori* (Algoritmo 1) foi proposto com o objetivo de minerar regras de associação em grandes e complexas bases de dados, e seu grande diferencial está na sua simplicidade original e na versatilidade. Conforme apresentado em Agrawal and Srikant (1994), no primeiro passo o algoritmo simplesmente conta a ocorrência de itens para determinar a largura do conjunto de *1-itemsets* (linha 1 do Algoritmo 1). Em seguida, chamado passo k , temos duas fases:

1. O *itemset* L_{k-1} encontrado no $(k-1)$ -ésimo passo é usado para gerar os *itemsets* candidatos C_k (linha 3 do Algoritmo 1), usando a função *apriori-gen* (Algoritmo 2).
2. A base de dados é percorrida e o valor de suporte dos candidatos em C_k é contado (linhas 4 a 9 do Algoritmo 1).

A solução final é dada pela união de todos os candidatos C_k com um valor de suporte maior que um valor **sup-min** definido pelo usuário (linha 10 do Algoritmo 1).

Algoritmo 1 Apriori. Fonte: (Agrawal and Srikant, 1994)

Entrada: Uma base de dados D composta por um conjunto de itens $A=\{a_1, \dots, a_m\}$ ordenados lexicograficamente e por um conjunto de transações $T=\{t_1, \dots, t_m\}$, na qual cada transação $t_i \in T$ é composta por um conjunto de itens ordenados lexicograficamente tal que $t_i \subseteq A$

```
1:  $L_1 \leftarrow \{\text{1-itemsets frequentes}\}$ 
2: para ( $k = 2; L_{k-1} \neq \emptyset; k = k + 1$ ) faz
3:    $C_k \leftarrow \text{apriori-gen}(L_{k-1})$ 
4:   for all (transações  $t \in T$ ) faz
5:      $C_t \leftarrow \text{subset}(C_k, t)$ 
6:     for all (candidatos  $c \in C_t$ ) faz
7:       c.count++
8:     fim para
9:   fim para
10:   $L_k \leftarrow \{c \in C_k | c.\text{count} \geq \text{sup-min}\}$ 
11: fim para
12: Resposta  $\leftarrow \bigcup_k L_k$ 
```

O conjunto de *itemsets* frequentes encontrado pelo algoritmo 1 é utilizado como entrada para algum algoritmo que gera regras de associação, como por exemplo o Algoritmo 2.

A seguir é apresentado o algoritmo da função ***apriori-gen*** que faz parte do algoritmo ***apriori***. Esta função usa como argumento de entrada o *itemset* L_{k-1} encontrado pelo Algoritmo 1 e retorna um superconjunto do conjunto de todos os $(k-1)$ -*itemsets*. A função executa inicialmente uma operação de união dos elementos dos *itemsets* em L_{k-1} com o último elemento de outros *itemsets*, diferentes do primeiro, em L_{k-1} (linhas 1 a 4 do Algoritmo 2). Em seguida são podados os k -*itemsets* que possuem algum subconjunto de tamanho $(k-1)$ não pertencente a L_{k-1} (linhas 5 a 11 do Algoritmo 2). Essa poda é orientada conforme a propriedade de limite de fronteira de suporte de um *itemset*, descrita na Seção 2.3.1.

Algoritmo 2 Função *apriori-gen*. Fonte: (Agrawal and Srikant, 1994)

Entrada: Um conjunto de *itemsets* frequentes L_{k-1}

Saída: Um conjunto de *itemsets* candidatos C_k

```
1: insert into  $C_k$ 
2: select  $p.item_1, p.item_2, \dots, p.item_{k-1}, q.item_{k-1}$ 
3: from  $L_{k-1}p, L_{k-1}q$ 
4: where  $p.item_1 = q.item_1, \dots, p.item_{k-2} = q.item_{k-2}, p.item_{k-1} < q.item_{k-1}$ 
   //A seguir é realizada a etapa de poda, na qual todos os itemsets  $c \cup C + k$  são
   removidos se algum  $(k - 1)$ -subconjunto de  $c$  não pertencer a  $L_{k-1}$ 
5: for all (itemset  $c \in C_k$ ) faça
6:   for all ( $(k - 1)$ -subconjuntos  $s$  de  $c$ ) faça
7:     se ( $s \notin L_{k-1}$ ) então
8:       delete  $c$  de  $C_k$ 
9:     fim se
10:   fim para
11: fim para
```

Além da função *apriori-gen*, outra importante função apresentada no código do algoritmo *apriori* é a *subset*. Esta função retorna os k -*itemsets* candidatos que estão contidos em uma dada transação t_i . Para isso, os *itemsets* candidatos são armazenados em uma árvore-*hash*. Cada nó da árvore pode conter uma lista de *itemsets* ou uma tabela *hash* (nó folha ou nó intermediário, respectivamente). Partindo do nó raiz, a função encontra todos os *itemsets* candidatos presentes na transação t_i . Se um nó folha é atingido e o *itemset* encontrado está contido na transação t_i , uma referência é adicionada ao conjunto de resposta. Se um nó intermediário é atingido a partir de um item $a_j \in A$, cada item é pesquisado (*hash*) após a_j em t_i . Isso é possível porque os itens estão em ordem lexicográfica. No nó raiz, todos os itens a_j em t_i são pesquisados.

Algoritmo Simples para Gerar Regras de Associação

Existem diversos algoritmos que geram regras de associação a partir dos *itemsets* frequentes obtidos de uma Base de Dados (Domingues, 2004). Um dos algoritmos mais simples foi proposto por Agrawal and Srikant (1994) e é apresentado a seguir.

O algoritmo é executado para os k -*itemsets* frequentes, com $k > 2$. Inicialmente são gerados os subconjuntos não vazios de um *itemset* frequente. Em seguida são geradas regras do tipo $LHS \Rightarrow RHS$ utilizando os subconjuntos definidos, que satisfazem a condição: confiança da regra maior ou igual à confiança mínima especificada pelo usuário (**conf-min**).

2.3.3 Medidas de Interesse Objetivas Utilizadas na Avaliação de Regras de Associação

Além das medidas de interesse mais básicas, Suporte e da Confiança, toda uma série de outras medidas tem sido propostas para avaliar o quanto interessante é uma regra de associação (Tan and Kumar, 2000; Tan et al., 2004; Zhao et al., 2009). Melanda (2004)

Algoritmo 3 Gera Regras de Associação. Fonte: (Agrawal and Srikant, 1994)

Entrada: Um conjunto Resposta contendo todos os k -itemsets frequentes, com $k \geq 2$

```

1: for all ( $k$ -itemset frequente  $l_k, k \geq 2$ ) faça
2:   Call genrules( $l_k, l_k$ )
3: fim para
   //O procedimento genrules gera todas as regras válidas  $\tilde{a} \Rightarrow (l_k - \tilde{a})$ , para todo
    $\tilde{a} \subset a_m$ 
4: procedure GENRULES( $l_k : k - itemset$  frequente,  $a_m : m - itemset$  frequente)
5:    $A \leftarrow \{(m-1)-itemsets a_{m-1} | a_{m-1} \subset a_m\}$ 
6:   for all ( $a_{m-1} \in A$ ) faça
7:      $conf \leftarrow sup(l_k)/sup(a_{m-1})$ 
8:     se ( $conf \geq conf-min$ ) então
9:       Output: regra  $a_{m-1} \Rightarrow (l_k - a_{m-1})$ , com confiança =  $conf$  e suport =
    $sup(l_k)$ 
10:      se ( $m-1 > 1$ ) então
11:        Call genrules( $l_k, a_{m-1}$ )           //Gera regras com subconjuntos de  $a_{m-1}$ 
   como antecedente
12:      fim se
13:    fim se
14:  fim para
15: fim procedure

```

apresenta uma análise crítica sobre o uso exclusivo das medidas de suporte e confiança para avaliação da qualidade das regras, mostrando assim a necessidade de utilização de outras medidas para avaliação das regras de associação geradas. Carvalho (2007) apresenta detalhes das principais medidas objetivas que podem ser utilizadas para este fim. Abaixo, apresenta-se uma definição das medidas que serão utilizadas neste projeto e uma breve descrição da semântica dessas medidas como definidas por Carvalho (2007):

added value: indica o quanto a frequência do consequente aumenta na presença do antecedente, ou seja, mede o ganho de RHS na presença de LHS .

$$added\ value = P(RHS|LHS) - P(RHS) \quad (2.10)$$

certainty factor: mede o aumento da crença em RHS em consequência da observação de LHS , sendo a crença em RHS dada por $P(RHS)$. Em outras palavras, mede a diminuição proporcional na descrença da hipótese RHS como resultado da observação de LHS .

$$certainty\ factor = \frac{P(RHS|LHS) - P(RHS)}{1 - P(RHS)} \quad (2.11)$$

collective strength: mede a correlação entre um conjunto de itens. Para tanto, utiliza os conceitos de taxa de violação e violação esperada. A taxa de violação $v(i)$ é definida pela fração de transações que contém pelo menos um item, mas não todos os itens, de um conjunto de itens I . Em outras palavras, dado um conjunto de itens I , $v(i)$ é igual à fração de transações que contém um subconjunto próprio não vazio

de I . A violação esperada $E[v(i)]$ representa a fração esperada de transações nas quais alguns dos itens, mas não todos, contidos em um conjunto de itens I , ocorrem simultaneamente nas transações ($E[v(i)]$ corresponde a $v(i)$ supondo independência entre LHS e RHS).

$$\begin{aligned} \text{collective strength} &= \frac{1 - v(i)}{1 - E[v(i)]} \times \frac{E[v(i)]}{v(i)} \\ &= \frac{1 - P(LHS)P(RHS) - P(\overline{LHS})P(\overline{RHS})}{1 - P(LHS \cap RHS) - P(\overline{LHS} \cap \overline{RHS})} \times \\ &\quad \frac{P(LHS \cap RHS) + P(\overline{LHS} \cap \overline{RHS})}{P(LHS)P(RHS) + P(\overline{LHS})P(\overline{RHS})} \end{aligned} \quad (2.12)$$

confiança: definida na Seção 2.3.1, Equação 2.9.

$$\text{confiança} = P(RHS|LHS) = \frac{\sup(LHS \cup RHS)}{\sup(LHS)} \quad (2.13)$$

conviction: foi desenvolvida como uma alternativa à *confiança*, uma vez que a *confiança* ignora a probabilidade de $P(RHS)$. A intuição por de trás dessa medida recai sobre a lógica proposicional. Se \Rightarrow for interpretado como uma implicação lógica, tem-se que $LHS \Rightarrow RHS \equiv \neg LHS \vee \neg RHS \equiv \neg(LHS \wedge \neg RHS)$. Portanto, medir a força de $LHS \Rightarrow RHS$ é equivalente a medir a força de $\neg(LHS \wedge \neg RHS)$. A força de $(LHS \wedge \neg RHS)$ pode ser medida em termos de quanto a ocorrência do evento $(LHS \wedge \neg RHS)$ se distancia da ocorrência conjunta dos eventos LHS e $\neg RHS$, considerando que os mesmos são independentes. Essa lógica pode ser modelada pela razão $P(LHS \wedge \neg RHS)/P(LHS)P(\neg RHS)$; quanto maior for a razão, melhor a força da implicação. Entretanto, a razão deve ser invertida para se considerar a negação em $\neg(LHS \wedge \neg RHS)$, o que leva à Equação 2.14.

$$\text{conviction} = \frac{P(LHS)P(\overline{RHS})}{P(LHS \cap \overline{RHS})} \quad (2.14)$$

ϕ -Coefficient: indica o grau de associação (correlação) entre duas variáveis.

$$\phi - \text{Coefficient} = \frac{P(LHS \cap RHS) - P(LHS)P(RHS)}{\sqrt{P(LHS)P(RHS)(1 - P(LHS))(1 - P(RHS))}} \quad (2.15)$$

gini index: frequentemente utilizada como critério de seleção na indução de árvores de decisão, ela é utilizada para medir o decréscimo da impureza ou incerteza de uma determinada classe (variável meta), condicionada ao conhecimento do valor de uma determinada variável (variável preditora). Quanto maior o valor da medida, maior será a associação entre as variáveis. No problema de regras de associação, quanto

maior o valor para esta medida, mais o antecedente e o consequente estão associados. A Equação 2.16 está definida em função das medidas apresentadas anteriormente e $P(LHS|RHS)$ é a probabilidade condicional entre os elementos do antecedente e do consequente. Observa-se que \overline{RHS} e \overline{LHS} são os complementos das medidas RHS e LHS.

$$\text{gini index} = \text{sup}(LHS)[P(RHS|LHS)^2 + P(\overline{RHS}|LHS)^2] + \\ \text{sup}(\overline{LHS})[P(RHS|\overline{LHS})^2 + P(\overline{RHS}|\overline{LHS})^2] - \\ \text{sup}(RHS)^2 - \text{sup}(\overline{RHS})^2 \quad (2.16)$$

IS: representa a média geométrica entre as medidas *lift* e *suporte*. Sendo assim, a medida IS mede tanto o interesse quanto a significância do padrão.

$$IS = \sqrt{\text{lift} \times \text{suporte}} = \sqrt{\frac{P(LHS \text{ RHS})}{P(LHS)P(RHS)} \times P(LHS \text{ RHS})} \\ = \frac{P(LHS \text{ RHS})}{\sqrt{P(LHS)P(RHS)}} \quad (2.17)$$

j-measure: representa o produto de dois termos. O primeiro termo, $P(LHS)$, pode ser visto como uma preferência para a generalidade, ou seja, o antecedente da regra deve ocorrer frequentemente para que uma regra seja considerada útil. O segundo termo, representado entre colchetes na Equação 2.18, mede a diferença entre a probabilidade posterior $P(RHS|LHS)$ e a probabilidade anterior $P(RHS)$, ou seja, o poder discriminativo de *LHS* em *RHS*. Portanto, maximizar o produto entre esses dois termos é equivalente a maximizar simultaneamente tanto a generalidade da hipótese *LHS*, quanto realizar um ajuste entre *LHS* e *RHS*.

$$\text{j-measure} = P(LHS) \times [P(RHS|LHS) \log(\frac{P(RHS|LHS)}{P(RHS)}) + \\ P(\overline{RHS}|LHS)(\frac{P(\overline{RHS}|LHS)}{P(\overline{RHS})})] \quad (2.18)$$

kappa: é um coeficiente de concordância. A medida *kappa* é calculada pela razão $(P(O) - P(E))/(1 - P(E))$, em que $P(O)$ representa a concordância observada entre dois indivíduos e $P(E)$ a concordância esperada, isto é, a proporção de vezes que se espera que os indivíduos concordem por chance.

$$P(O) = P(LHS \text{ RHS}) + P(\overline{LHS} \text{ } \overline{RHS})$$

$$P(E) = P(LHS)P(RHS) + P(\overline{LHS})P(\overline{RHS})$$

$$\kappa = \frac{P(LHS \cap RHS) + P(\overline{LHS} \cap \overline{RHS}) - P(LHS)P(RHS) - P(\overline{LHS})P(\overline{RHS})}{1 - P(LHS)P(RHS) - P(\overline{LHS})P(\overline{RHS})} \quad (2.19)$$

klosgen: representa a combinação entre as medidas *added value* e *suporte*. O *suporte*, neste caso, pondera o quanto *LHS* e *RHS* influenciam o ganho de *RHS* na presença de *LHS*, ou seja, a proporção do ganho encontrado pela medida *added value* quando *LHS* e *RHS* ocorrem com uma determinada probabilidade.

$$klosgen = \sqrt{P(LHS \cap RHS)}(P(RHS|LHS) - P(RHS)) = \sqrt{suporte} \times added\ value \quad (2.20)$$

lambda: mede o decréscimo relativo da probabilidade de um erro em calcular um atributo levando-se em consideração a presença e a ausência de um outro atributo. Em outras palavras, mede a capacidade preditiva de uma variável em relação à outra. Existe o caso assimétrico e o caso simétrico³ dessa medida. Este último será apresentado aqui. No caso simétrico, nenhuma das variáveis é especificamente designada como a variável a ser predita. Pelo contrário, supõe-se que algumas vezes uma variável e outras vezes a outra é fornecida *a priori* e pretende-se predizer a variável não fornecida. Esse coeficiente mostra a redução relativa da probabilidade de um erro em predizer a categoria de uma das variáveis conhecendo e não conhecendo a categoria da outra variável.

$$\lambda = \frac{\max(P(LHS \cap RHS), P(LHS \cap \overline{RHS}))}{2 - \max(P(LHS), P(\overline{LHS})) - \max(P(RHS), P(\overline{RHS}))} + \frac{\max(P(\overline{LHS} \cap RHS), P(\overline{LHS} \cap \overline{RHS}))}{2 - \max(P(LHS), P(\overline{LHS})) - \max(P(RHS), P(\overline{RHS}))} + \frac{\max(P(LHS \cap RHS), P(\overline{LHS} \cap RHS))}{2 - \max(P(LHS), P(\overline{LHS})) - \max(P(RHS), P(\overline{RHS}))} + \frac{\max(P(LHS \cap \overline{RHS}), P(\overline{LHS} \cap \overline{RHS}))}{2 - \max(P(LHS), P(\overline{LHS})) - \max(P(RHS), P(\overline{RHS}))} \quad (2.21)$$

laplace: é uma variação da medida Confiança (Equação 2.9) e penaliza regras muito específicas (regras que cobrem poucas transações). Sua semântica é a mesma daquela apresentada pela Confiança. Na Equação 2.22, *N* é o número total de transações.

³Uma medida *M* é simétrica de $M(LHS \Rightarrow RHS) = M(RHS \Rightarrow LHS)$; caso contrário é assimétrica (Tan et al., 2005a)

$$laplace = \frac{N \times sup(LHS \cup RHS) + 1}{N \times sup(LHS) + 2} \quad (2.22)$$

lift: essa medida indica o quanto a frequência do consequente aumenta na presença do antecedente. Em outras palavras, mede o grau de dependência entre os itens. Regras com $lift = 1$ possuem antecedente e consequente independentes, ou seja, a presença do antecedente não leva ao aumento ou à diminuição da ocorrência do consequente. No caso de regras com $lift > 1$, pode-se dizer que o antecedente influencia positivamente a frequência do consequente (dependência positiva) e, no caso de $lift < 1$, o antecedente e o consequente apresentam dependência negativa.

$$lift = \frac{conf(LHS \Rightarrow RHS)}{sup(RHS)} \quad (2.23)$$

mutual information LHS: é uma medida que indica o grau de associação entre duas variáveis. Essa medida especifica a quantidade de redução na incerteza (entropia) de uma variável RHS quando uma segunda variável LHS é conhecida. Se as duas variáveis forem altamente associadas, então a quantidade de redução na incerteza (entropia) será grande. Para o caso assimétrico apresentado aqui, essa medida é conhecida como *Theil Uncertainty Coefficient* (Blanchard et al., 2005).

$$\text{mutual information LHS} = \frac{P(LHS \cap RHS) \log \frac{P(LHS \cap RHS)}{P(LHS)P(RHS)}}{-P(RHS) \log P(RHS) - P(\overline{RHS}) \log P(\overline{RHS})} + \\ \frac{P(LHS \cap \overline{RHS}) \log \frac{P(LHS \cap \overline{RHS})}{P(LHS)P(\overline{RHS})}}{-P(RHS) \log P(RHS) - P(\overline{RHS}) \log P(\overline{RHS})} + \\ \frac{P(\overline{LHS} \cap RHS) \log \frac{P(\overline{LHS} \cap RHS)}{P(\overline{LHS})P(RHS)}}{-P(RHS) \log P(RHS) - P(\overline{RHS}) \log P(\overline{RHS})} - \\ \frac{P(\overline{LHS} \cap \overline{RHS}) \log \frac{P(\overline{LHS} \cap \overline{RHS})}{P(\overline{LHS})P(\overline{RHS})}}{-P(RHS) \log P(RHS) - P(\overline{RHS}) \log P(\overline{RHS})} \quad (2.24)$$

novelty: calcula a porcentagem de transações adicionais cobertas por uma regra de associação que estão acima do esperado. Em outras palavras, compara o valor observado da ocorrência de LHS e RHS e o valor esperado de ocorrência se LHS e RHS fossem independentes. Também é conhecida como *Rule Interest*, *Piatetsky-Shapiro* e *Leverage*.

$$novelty = P(LHS \cap RHS) - P(LHS)P(RHS) \quad (2.25)$$

odds ratio: indica o grau com que as variáveis LHS e RHS estão associadas uma com a outra. Se RHS ocorre, então a probabilidade de LHS ocorrer em uma mesma transação é $P(LHS \cap RHS)/P(LHS)$. Por outro lado, se RHS não ocorre, então a probabilidade de LHS ocorrer em uma mesma transação é $P(LHS \cap \overline{RHS})/P(\overline{LHS} \cap \overline{RHS})$.

Se não existir nenhuma associação entre LHS e RHS , então a probabilidade de LHS ocorrer em uma determinada transação deve permanecer a mesma, independente de RHS ocorrer ou não na transação.

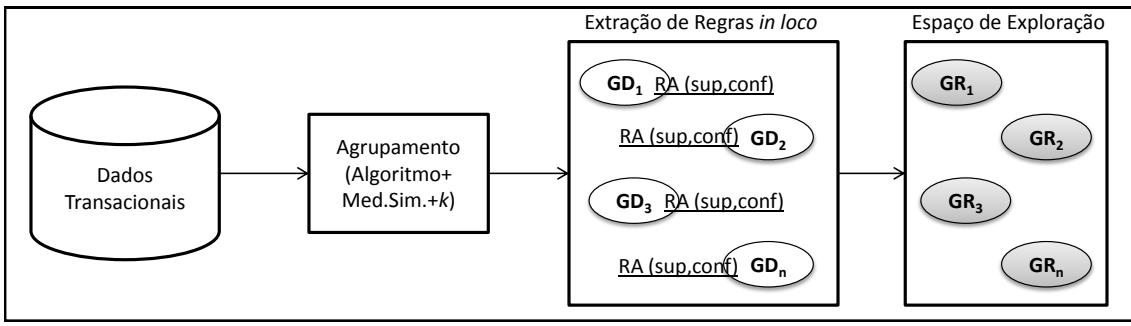
$$\text{odds ratio} = \frac{P(LHS \ RHS)P(\overline{LHS} \ \overline{RHS})}{P(LHS \ \overline{RHS})P(\overline{LHS} \ RHS)} \quad (2.26)$$

2.4 Agrupamento de Regras de Associação

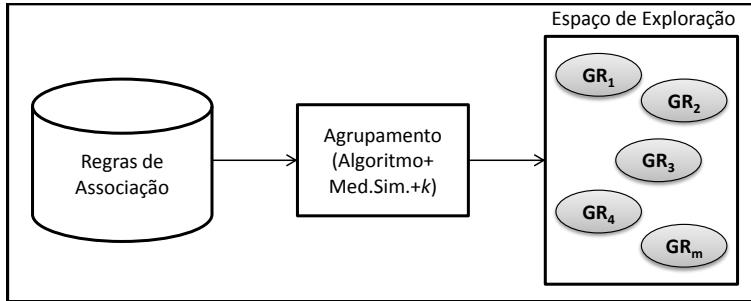
No processo tradicional de geração de regras de associação, independente do domínio, são geradas uma quantidade muito grande de padrões, o que dificulta a exploração por parte do usuário. Três linhas vem sendo adotadas para solucionar esse problema: (i) pré-processar os dados antes de se extraír as regras; (ii) considerar restrições a serem atendidas durante o processo de extração; (iii) explorar as regras de associação na etapa de pós-processamento considerando os interesses do domínio. Entretanto, todas visam o mesmo objetivo: encontrar os padrões associativos relevantes da aplicação. Dentro de cada uma dessas linhas encontram-se várias abordagens.

Uma abordagem bastante utilizada tanto no pré-processamento como no pós-processamento é o agrupamento (*clustering*). O processo de agrupamento nas referidas etapas é apresentado na Figura 2.3. Como observa-se na Figura 2.3(a), no pré-processamento os dados são inicialmente agrupados em n grupos ($GD_1, GD_2, GD_3, \dots, GD_n$). A partir desse agrupamento inicial é que as regras de associação são então extraídas dentro de cada grupo, obtendo-se, assim como no pós-processamento, n grupos de regras ($GR_1, GR_2, GR_3, \dots, GR_n$) (destacados em cinza). No caso da Figura 2.3(b), referente ao pós-processamento, o conjunto de regras de associação, obtido *a priori*, é agrupado e as regras divididas em m grupos ($GR_1, GR_2, GR_3, GR_4, \dots, GR_m$) (destacados em cinza). Contudo, a interpretação do espaço de exploração é diferente em cada uma das etapas. No pré-processamento cada grupo contém um conjunto próprio de regras, ou seja, vários conjuntos de regras existem e, portanto, não há efetivamente um agrupamento. Já no pós-processamento cada grupo contém um subconjunto das regras existentes no conjunto a partir do qual o particionamento foi realizado.

O principal objetivo ao aplicar agrupamento no processo de geração de regras de associação é organizar as regras em grupos que contenham, de algum modo, conhecimento similar. Quando aplicado na etapa de pós-processamento, o objetivo é melhorar a apresentação dos padrões obtidos a fim de fornecer ao usuário uma visão do domínio a ser explorado e, consequentemente, facilitar sua busca por padrões interessantes. Quando aplicado na etapa de pré-processamento, o objetivo é obter regras potencialmente interessantes que não seriam extraídas a partir de conjuntos não particionados por não terem suporte suficiente sem, contudo, sobrecarregar o usuário com uma grande quantidade de padrões. Isso porque, para que esses mesmos padrões sejam descobertos a partir de conjuntos não particionados deve-se configurar o suporte mínimo com um valor muito baixo e abaixo do utilizado nos conjuntos particionados, fazendo com que o número de regras



(a) Pré-processamento



(b) Pós-processamento

Figura 2.3: Visão geral do processo de agrupamento de regras de associação no pré-processamento e no pós-processamento. Fonte: Carvalho et al. (2012).

cresça rapidamente (Carvalho et al., 2012). Neste trabalho de doutorado, será utilizado o agrupamento de regras de associação no pós-processamento, conforme apresentado na metodologia proposta detalhada na Seção 4.2. As regras de associação geradas, neste trabalho, formaram termos compostos que foram agrupados para representar os diferentes tópicos da coleção de documentos. Uma comparação entre as diferenças do uso de agrupamento na etapa de pré-processamento e pós-processamento, bem como suas vantagens e desvantagens, pode ser encontrada em Carvalho et al. (2012).

Agrupamento no Pós-Processamento

Os trabalhos que empregam o agrupamento nessa etapa tem como objetivo melhorar a apresentação dos padrões, obtidos *a priori*, a fim de fornecer ao usuário uma visão do domínio a ser explorado e, consequente, facilitar sua busca por padrões interessantes (Reynolds et al., 2006; Sahar, 2002). Nesse contexto, visando a estruturação do conhecimento, diferentes estratégias de agrupamento vêm sendo utilizadas para pós-processar regras de associação.

Reynolds et al. (2006) propõem agrupar regras de classificação parcial obtidas por dois algoritmos por eles propostos. Nesse caso, todas as regras geradas contêm o mesmo consequente, ou seja, o agrupamento se dá em relação ao antecedente das regras. Embora o tipo de regra por eles considerada não seja de associação, a ideia do trabalho é a mesma dos trabalhos relacionados a associação: a única diferença é que todas as regras contêm o mesmo consequente. Para realizar o agrupamento os autores propõem a utilização de algoritmos particionais (K-means, PAM, CLARANS) e hierárquicos (AGNES) usando a Jaccard como medida de similaridade. Nesse caso, a Jaccard entre duas regras r_1 e r_2 ,

apresentada na Equação 2.27, é calculada considerando as transações t em comum que as regras cobrem. Essa medida de similaridade é referenciada aqui como J-RT (*Jaccard with Rules by Transactions*). Uma regra cobre uma transação t se todos os itens da regra estão contidos em t .

$$J\text{-RT}(r_1, r_2) = \frac{|\{t \text{ cobertas por } r_1\} \cap \{t \text{ cobertas por } r_2\}|}{|\{t \text{ cobertas por } r_1\} \cup \{t \text{ cobertas por } r_2\}|} \quad (2.27)$$

Jorge (2004) propõe agrupar as regras de associação por meio da utilização de algoritmos hierárquicos (Single Linkage, Complete Linkage, Average Linkage) usando a Jaccard como medida de similaridade. Nesse caso, a Jaccard entre duas regras r_1 e r_2 , apresentada na Equação 2.28, é calculada considerando os itens que as regras compartilham. Essa medida de similaridade é referenciada aqui como J-RI (*Jaccard with Rules by Items*).

$$J\text{-RI}(r_1, r_2) = \frac{|\{\text{itens em } r_1\} \cap \{\text{itens em } r_2\}|}{|\{\text{itens em } r_1\} \cup \{\text{itens em } r_2\}|} \quad (2.28)$$

Toivonen et al. (1995) propõem uma medida de similaridade baseada em transação e utiliza um algoritmo baseado em densidade para realizar o agrupamento das regras. Nesse trabalho os autores consideram que todas as regras possuem o mesmo consequente, ou seja, assim como em Reynolds et al. (2006) o agrupamento se dá em relação ao antecedente das regras. Sahar (2002) também propõe um medida de similaridade baseada em transação com base na proposta de Toivonen et al. (1995), embora utilize um algoritmo hierárquico para agrupar as regras. Entretanto, não é mencionado qual o algoritmo utilizado e, diferentemente de Toivonen et al. (1995), estende-se o agrupamento para regras que possuem consequentes distintos.

Uma das dificuldades ao agrupar regras de associação extraídas de textos está no cálculo da similaridade entre duas regras. Em Marcacini et al. (2012a), os autores propuseram uma representação do conjunto de *itemsets* frequentes extraídos no modelo VSM, isto é, cada *itemset* frequente é representado por um vetor em que cada dimensão corresponde aos termos extraídos dos documentos, e o peso de cada termo para cada *itemset* frequente é determinado pelo valor médio do termo nos documentos que contém o *itemset*. Essa representação é interessante por permitir explorar a informação extraída dos textos de forma mais profunda durante o agrupamento das regras de associação extraídas.

O agrupamento se torna útil ao usuário se ele o apoia durante o processo de exploração, apoio esse a ser obtido por meio do direcionamento aos padrões potencialmente interessantes e/ou por meio de bons descritores que forneçam uma visão dos tópicos a serem explorados. Reynolds et al. (2006) e Jorge (2004) selecionam como descritores, em cada grupo, os itens contidos na regra mais similar a todas as outras regras do grupo (o medóide do grupo). Toivonen et al. (1995) não mencionam como os descritores são encontrados, mas fornecem indícios que indicam que os descritores selecionados representam os itens mais frequentes do grupo. Por outro lado, Sahar (2002) propõe uma abordagem para sumarizar cada grupo de regras encontrando, para tanto, padrões $a \Rightarrow c$ que cubram todas as regras contidas no grupo; nesse caso, a e c são itens do domínio e um padrão $a \Rightarrow c$ cobre uma regra $A \Rightarrow C$ se $a \in A$ e $c \in C$. Como observa-se, embora a abordagem

não seja usada para definir descritores, a ideia pode ser utilizada para essa finalidade.

Embora muitos métodos tenham sido propostos para determinar descritores em agrupamentos de documentos nas tarefas de Mineração de Texto (MT) e Recuperação de Informação (RI), como em (Moura and Rezende, 2010; Lopes et al., 2007; Kashyap et al., 2005; Fung et al., 2003; Glover et al., 2002; Popescul and Ungar, 2000; Larsen and Aone, 1999; Cutting et al., 1992), os trabalhos relacionados a agrupamentos de regras de associação não vêm explorando essa questão, nem mesmo a avaliação dos métodos existentes. Entretanto, muitos dos métodos aplicados a documentos são similares aos utilizados em agrupamentos de regras de associação.

2.5 Considerações finais

Neste capítulo foi apresentado o processo de Mineração de Textos, que é fundamental para organizar, gerenciar e extrair conhecimentos de coleções de documentos textuais. Neste capítulo também foram apresentadas as técnicas de agrupamento e de associação, usadas neste trabalho para a obtenção de termos compostos e extração de tópicos para construção de uma representação estruturada da coleção de documentos.

No próximo capítulo, é apresentada uma revisão das principais técnicas de representação de documentos para mineração de textos, e de técnicas de redução de dimensionalidade.

Representação de Documentos e Redução de Dimensionalidade

Uma das etapas mais importantes para as tarefas de mineração textos é a construção da representação dos dados. A quantidade e a qualidade dos atributos para a representação dos documentos são elementos cruciais para a qualidade do conhecimento obtido. Na Seção 3.1 é apresentada uma revisão de alguns modelos para a representação de documentos encontrados na literatura, destacando-se o Modelo de Espaço de Vetores, amplamente utilizado nas tarefas de mineração de textos. Também são apresentados alguns modelos mais sofisticados, que exploram diferentes informações extraídas dos documentos para enriquecer a representação construída, visando melhorar os resultados e ampliar as possibilidades de aplicações para uma coleção de textos. Em grande parte dos modelos, não existe um mecanismo explícito para redução da dimensionalidade da representação. Neste capítulo, na Seção 3.2, também são discutidas técnicas para a redução de dimensionalidade em coleções de documentos encontradas na literatura, bem como serão discutidas algumas técnicas relevantes para extração de atributos.

3.1 Representação de Documentos

Algumas características importantes diferenciam dados textuais de outras formas de dados como os dados relacionais ou dados quantitativos. Como consequência, isto afeta as técnicas de mineração que podem ser aplicadas nesses dados. Dados textuais têm como características serem esparsos e apresentar alta dimensionalidade. Por exemplo, uma coleção de documentos pode conter cerca de cem mil palavras, mas cada um dos documentos deve conter apenas algumas centenas dessas palavras. A representação *bag-of-words* baseada no Modelo de Espaço de Vetores, apresentada na Seção 3.1.1, é uma das mais utilizadas nas tarefas de mineração de textos. Esse modelo de representação

permite a aplicação direta de uma grande quantidade de técnicas de mineração de dados bem estabelecidas na literatura (Aggarwal and Zhai, 2012). Dados textuais podem ser analisados em diferentes níveis de representação. Por exemplo, dados textuais podem facilmente ser representados como uma *bag-of-words*.

Entretanto, em muitas aplicações, é desejável representar as informações textuais utilizando características semanticamente mais ricas pois assim uma análise mais significativa pode ser feita. Infelizmente, as técnicas do estado da arte de processamento de linguagem natural ainda são pouco robustas para serem aplicadas em coleções de documentos sem domínio definido e gerar representações semânticas precisas dos documentos. Assim, a maior parte das técnicas de mineração de textos ainda dependem de representações baseadas em termos, especialmente o modelo *bag-of-words* (Aggarwal and Zhai, 2012).

3.1.1 Modelo de Espaço de Vetores

A representação mais comumente utilizada em Mineração de Textos é a *bag-of-words*, baseada no Modelo de Espaço de Vetores¹ (MEV) (Salton, 1989) dos documentos e termos utilizados. Esse modelo representa cada documento como um vetor de termos distintos que aparecem na coleção. Cada componente do vetor representa o peso de cada termo em cada documento da coleção. Seja $D = \{d_1, d_2, \dots, d_m\}$ uma coleção com m documentos, e $T = \{t_1, t_2, \dots, t_n\}$ o conjunto de n termos distintos da coleção. Cada documento d_j é representado por um vetor de termos $\vec{d}_j = \{w_{1j}, w_{2j}, \dots, w_{nj}\}$, no qual cada peso w_{ij} quantifica a importância do termo $t_i \in T$ para o documento $d_j \in D$. Para os termos da coleção que não estão presentes no documento d_j , $w_{ij} = 0$. Nesse modelo, tradicionalmente o peso w_{ij} representa a medida da frequência do termo no documento ou uma variação desse esquema (Liu, 2011). A coleção de documentos é então representada pela matriz W com dimensões $m \times n$, conhecida na literatura como matriz documento-termo. Cada linha de W corresponde a um documento em D , isto é, o vetor \vec{d}_j , e cada coluna descreve a distribuição de cada termo na coleção de documentos. Caso os documentos possuam classes (rótulos), ainda haverá uma última coluna que corresponde à classe dos documentos (y_i).

Tabela 3.1: Padrão de matriz documento-termo W

	t_1	t_2	\dots	t_n	Y
d_1	w_{11}	w_{12}	\dots	w_{1n}	y_1
d_2	w_{21}	w_{22}	\dots	w_{2n}	y_2
\vdots	\vdots	\vdots	\ddots	\vdots	\vdots
d_m	w_{m1}	w_{m2}	\dots	w_{mn}	y_m

No modelo de espaço de vetores, proposto por Salton et al. (1975), os pesos dos termos extraídos dos documentos para formação dos índices e dos termos da consulta representa um número real positivo, sendo tradicionalmente a medida da frequência do termo no documento ou uma variação desse esquema (Baeza-Yates and Ribeiro-Neto, 1999; Liu,

¹Alguns autores traduzem também como Modelo Espaço Vetorial. Alguns trabalhos também apresentam o termo em inglês *Vector Space Model* (VSM).

2011). Esses pesos podem ser calculados utilizando diversas abordagens, variando conforme necessidades específicas da aplicação ou das tecnologias disponíveis. Entretanto, a abordagem baseada na métrica *tf-idf* e suas extensões (Salton and Buckley, 1987) é a mais utilizada. A *tf-idf* combina os valores das medidas *tf*² (*term frequency*) e *idf*² (*inverse document frequency*). A primeira tem escopo local e indica a importância de um termo para um documento específico. A segunda tem escopo global, e indica a distribuição do termo na coleção. Assim, a *tf-idf* procura balancear os dois efeitos ao calcular o peso de cada termo de indexação (Baeza-Yates and Ribeiro-Neto, 1999; Manning et al., 2008; Liu, 2011; Martins, 2009).

A medida *tf* de um termo representa o número de vezes que o termo t_i ocorre no documento d_j , ou seja a frequência absoluta do termo, representada aqui por $freq_{i,j}$. Defini-se então $tf_{i,j}$ como a frequência do termo t_i no documento d_j como:

$$tf_{i,j} = freq_{i,j} \quad (3.1)$$

Se o termo t_i não aparece no documento d_j , então $tf_{i,j} = 0$;

A medida *idf* de um termo t_i é calculada para toda a coleção, aqui representada por idf_i . Seja N o número total de documentos da coleção D e n_i o número de documentos que contém o termo t_i . A *idf* do termo t_i é dada por:

$$idf_i = \log \frac{N}{n_i} \quad (3.2)$$

Assim, o peso de cada termo t_i para cada documento d_j da coleção é calculado como apresentado na Equação 3.3.

$$w_{ij} = tf_idf_{ij} = tf_{i,j} \times idf_i \quad (3.3)$$

Discussões sobre variações da abordagem *tf-idf* e outras abordagens para obtenção do peso dos termos são apresentadas em Baeza-Yates and Ribeiro-Neto (1999); Manning et al. (2008); Dong et al. (2008) e Turney and Pantel (2010).

3.1.2 Construção Automática de Representações Utilizando Mineração de Textos

A construção de representações de coleções de documentos utilizando mineração de textos é uma área de pesquisa bastante explorada na literatura. A informação extraída para enriquecer a representação permite ampliar as possibilidades de conhecimento extraídos da coleção de documentos. Ainda, essa informação pode fazer parte do conhecimento obtido no processo. Nesta seção, são apresentados diversos trabalhos da literatura para construção (semi-)automática de representações, preferencialmente aqueles relacionados com o modelo de espaço de vetores e suas extensões.

²Como a forma em inglês é muito utilizada na literatura da área, optou-se aqui por não traduzir a expressão para o português.

O modelo de espaço de vetores tem a desvantagem de assumir que os termos de indexação são mutuamente independentes, ou seja, ele não considera qualquer informação relacionada com a ordem em que os termos aparecem no texto e a dependência entre eles (Baeza-Yates and Ribeiro-Neto, 1999). Segundo Rossi and Rezende (2011), diversos conceitos presentes nos documentos são normalmente representados por mais de um termo, como “inteligência artificial” ou “sistema operacional”. Os trabalhos encontrados na literatura que exploram a dependência entre os termos são apresentados e discutidos na Seção 3.1.3.

Kang et al. (2005) propõe o uso de agrupamento de termos para extrair conceitos da coleção, formando a representação denominada *agrupamento conceitual*. Nessa abordagem, cada documento é representado por um vetor formado pelos *grupos conceituais*. Cada *grupo conceitual* é formado por um subconjunto de termos dos documentos ligados por relações lexicais (sinônimos, antônimos, hiperônimos, ...). Cada *grupo conceitual* recebe um peso calculado com base na importância das relações semânticas que compõe o grupo. Com base nesse peso, os grupos mais relevantes para o documento são selecionados formando um conjunto de *grupos conceituais representativos*. Cada termo desses grupos são então utilizados como termo de indexação para o documento. Esse termo é associado a um peso que depende do peso do seu grupo e da sua relação lexical. Kang et al. (2005) aponta que essa construção permite diferenciar cada documento pela sua importância semântica e melhorar o resultado das buscas, com uma grande redução da dimensionalidade da representação. Entretanto, a qualidade dos resultados é limitada pelo recurso linguístico utilizado para encontrar as relações lexicais.

Liu and Croft (2008) avaliam diferentes representações baseadas em agrupamento de documentos para recuperação de informação. Aplica-se um algoritmo de agrupamento na coleção, e cada documento é associado a um grupo. O sistema então trata cada grupo como sendo uma entidade única. Essa entidade é tratada pelo sistema de recuperação de informação utilizando diferentes estratégias encontradas na literatura, como a “concatenação dos documentos”, em que o grupo é representado por todos os termos presentes nos documentos deste grupo. O peso de cada termo é a soma dos pesos do termo em cada documento. O sistema de recuperação então avalia a relevância de cada grupo para a consulta, e apresenta como resposta os documentos do grupo mais relevante.

De maneira geral, as abordagens de mineração de textos e recuperação de informação são processos intrinsecamente subjetivos. No processo de transmissão das necessidades do usuário para o sistema e posterior avaliação da qualidade desses resultados, as abordagens precisam lidar com a incerteza e a imprevisibilidade inerentes às interações (Ingwersen, 1996; Baeza-Yates and Ribeiro-Neto, 1999; Kraft et al., 2007). De acordo com Kraft et al. (2007), a teoria de conjuntos *Fuzzy* tem sido amplamente aplicada em problemas de mineração de textos e recuperação de informação para apoiar o tratamento da imprecisão e da incerteza. Kraft et al. (2007) apresenta algumas estratégias encontradas para utilizar a teoria *Fuzzy* na construção de representações dos documentos. Leite (2009) propõe um modelo de recuperação *Fuzzy* que mapeia os documentos utilizando conceitos de múltiplas ontologias de domínio. Para cada termo de indexação é atribuído um peso que

combina a *tf-idf* do termo ponderado pelo grau de pertinência do termo para a coleção de documentos.

Além disso, o uso de conhecimento semântico “profundo” tem sido explorado com sucesso recentemente. Trabalhos de Fernández et al. (2010); Kara et al. (2012); Dragoni et al. (2012) fazem uso de ontologia para expandir a representação dos documentos utilizando conceito de domínio. O trabalho de Kara et al. (2012) propõe um sistema de recuperação de informação baseado em ontologias de domínio que combina a representação tradicional do modelo de espaço de vetores com representações enriquecidas com uso de ontologias de domínio. Essa proposta tem foco na escalabilidade computacional do processo para viabilizar o uso de ontologias em grandes sistemas de recuperação de informação.

3.1.3 Representação de Documentos Explorando Termos Dependentes

O modelo *bag-of-words* está bem estabelecido na literatura devido aos resultados obtidos com sua aplicação e a simplicidade para a construção da representação. Assim, o desafio de obter um modelo para representação dos documentos é ainda maior, e seus ganhos de qualidade e de conhecimento devem superar a necessidade de esforço extra. O modelo *bag-of-words* possui alta dimensionalidade e não trata a dependência entre termos (Aggarwal and Zhai, 2012). Métodos para redução de dimensionalidade são discutidos na Seção 3.2. A dependência entre termos pode ser determinada por diferentes relações que podem ser identificadas, como sinônima, polissemia, termos compostos, etc. Neste trabalho, o foco será nos modelos para construção de representações com abordagem puramente estatística, isto é, nos quais a aplicação não requer a especificação de qualquer tipo de conhecimento linguístico (por exemplo: morfológico, sintático, etc.) sobre a coleção de documentos. Estes modelos são independentes da língua e das estruturas linguísticas e, normalmente, exigem um menor esforço computacional quando comparados aos modelos que utilizam conhecimento linguístico.

Os modelos apresentados neste trabalho exploram diferentes informações de dependência entre os termos, e essas informações podem ser divididas em dois grandes grupos de relações entre termos. O primeiro grupo se refere às relações entre termos em um **contexto local**, isto é, quando a proximidade dos termos no texto afeta alguma das características desses termos, como seu significado. Uma das principais relações nesse contexto são as formadas pelos termos compostos, cujo aspecto afetado é o significado dos termos, como o termo composto “aprendizado de máquina”. Cada um dos termos tem um significado próprio, porém, quando esses termos ocorrem próximos no texto um novo significado surge relativo ao termo composto que se sobrepõe aos significados de cada termo. O segundo grupo de relações entre termos são as de **contexto geral**. Essas relações existem independente da estrutura ou da posição do termo no documento. São relações normalmente determinadas por estruturas mais gerais da língua, como a sinônima e a polissemia. Também podem ser determinadas por estruturas mais abstratas, como termos específicos de um determinado domínio do conhecimento, como “bisturi” e “cirurgia”, ou por tópicos ou assuntos que reúnem um grande conjunto de conceitos, como a relação dos

termos “futebol” e “atletismo” com o tópico “esporte”. Em geral, os modelos se concentram em um dos grupos de relação entre termos, como no caso dos modelos de extração de termos compostos que se concentra no contexto local ou dos modelos de extração de tópicos que se concentram no contexto geral. A seguir, são apresentados trabalhos relacionados aos modelos para construção da representação de coleções de documentos encontrados na literatura. Eles foram agrupados de acordo com o tipo principal de relação entre termos que eles exploram.

GVSM

A proposta de Wong et al. (1985), Modelo de Espaço de Vetores Generalizado (*Generalized Vector Space Model - GVSM*), é uma das primeiras extensões do modelo VSM que incorpora no processo a correlação dos termos, ainda que não vise à redução da dimensionalidade. Os autores propuseram uma interpretação alternativa ao modelo espaço vetorial na qual os vetores dos termos de indexação são formados por componentes menores, derivados a partir da coleção utilizada, chamados *minterms*. Esse modelo adota como princípio básico a ideia que a coocorrência entre os termos $t_i \in T$ dentro dos documentos induz a dependência entre esses termos (Pôssas et al., 2002; Baeza-Yates and Ribeiro-Neto, 2011). A principal contribuição do modelo é o estabelecimento de um arcabouço formal no qual as dependências entre termos possam ser adequadamente representadas, introduzindo novas ideias que são importantes do ponto de vista teórico (Pôssas et al., 2002; Baeza-Yates and Ribeiro-Neto, 2011). Foram propostas na literatura diversas extensões do modelo que visam suprir algumas das deficiências da proposta original, como em Billhardt et al. (2002); Farahat and Kamel (2011); Kalogeratos and Likas (2012); Cheng et al. (2013b). Segundo os autores Baeza-Yates and Ribeiro-Neto (2011), não fica claro em quais situações o modelo GVSM supera o modelo vetorial clássico. Além disso, ele pode ser bastante complexo e computacionalmente caro para grandes coleções porque, normalmente, o número de *minterms* que precisam ser computados é igual ao número de documentos da coleção.

Frases ou Termos Compostos

Os modelos baseados em frases ou termos compostos, como os apresentados nos trabalhos Pôssas et al. (2002); Figueiredo et al. (2011); Rossi and Rezende (2011), visam construir novos atributos compostos por 2 ou mais termos que, combinados, representam um conceito diferente daquele apresentado pelos termos independentes. Esses atributos podem ser adicionados à representação *bag-of-words* ou podem substituir os atributos originais. De acordo com Rossi and Rezende (2011), os principais modelos propostos na literatura são baseados em n-gramas e em conjunto de palavras. No modelo de n-gramas os atributos são construídos por sequências de n termos adjacentes que aparecem na coleção de textos. No modelo de conjuntos de palavras, os atributos também são compostos por n termos que aparecem na coleção de textos, mas não precisam ser adjacentes. Em geral, os trabalhos da literatura para esse modelo propõe a execução de uma análise de coocorrê-

cia considerando a coleção toda. O modelo proposto em Figueiredo et al. (2011) explora a coocorrência dos termos nos documentos para construir atributos chamados *c-features* (*compound-features* - atributos-compostos). Os atributos são obtidos considerando, na base de treino, a classe de cada documento para obter os atributos mais significativos. Os autores visam obter esses atributos com uma estratégia com baixo custo computacional em relação a outras estratégias, uma vez que o número de combinações possíveis dos atributos originais da coleção com n termos é 2^n . Em seu trabalho, Rossi and Rezende (2011) propõe uma representação denominada *bag-of-related-words*, que extrai dos documentos termos de indexação compostos por palavras simples fortemente correlacionadas e palavras simples que ocorrem frequentemente na coleção, com uso de regras de associação, para formar o vetor que representa o documento. Ainda de acordo com Rossi and Rezende (2011), essa representação apresenta diversas vantagens em relação aos trabalhos da literatura, pois não aumenta significativamente a dimensionalidade da representação como na maioria dos modelos propostos, possui um processo automático para gerar termos compostos significativos, não necessita de coleções rotuladas e não precisa analisar toda a coleção antes de fazer a seleção pois é aplicada em cada documento individualmente. Normalmente, é preciso reduzir a dimensionalidade das representações produzidas por esses modelos aplicando um método de seleção de atributos para tornar o modelo viável para aplicações práticas.

Extração de Tópicos

Dentre as técnicas encontradas na literatura para extração de tópicos de uma coleção de documentos, as propostas baseadas em extração de dimensões latentes, como a *Latent Semantic Analysis* (LSA) (Deerwester et al., 1990) ou a *Non-negative Matrix Factorization* (NMF) (Lee and Seung, 1999) e as de extração de tópicos (Hofmann, 1999; Blei et al., 2003), se destacam pela qualidade dos resultados obtidos e pela boa interpretabilidade das dimensões extraídas. O modelo *Latent Dirichlet Allocation* (LDA) (Blei et al., 2003) é uma das técnicas mais proeminentes para extração de tópicos. No trabalho de Blei et al. (2003), propõe-se um *modelo generativo probabilístico* que descreve uma coleção de documentos a partir de um conjunto reduzido de tópicos. Esse modelo se torna atrativo por descobrir grupos de termos que aparecem frequentemente juntos nos documentos (Zhu et al., 2012; Hu et al., 2014), e é apresentado com mais detalhes na Seção 3.2.4. O trabalho apresentado por Wallach (2006) é um dos primeiros a incorporar termos compostos no processo de extração de tópicos com LDA. Utilizou-se em Wallach (2006) modelos de linguagem hierárquica de Dirichlet para estender o algoritmo original do LDA e considerar, durante a inferência dos tópicos, o peso do termo avaliado condicionalmente ao peso do termo anterior a ele. Os autores de Lau et al. (2013) argumentam que ainda que os resultados dessa e de outras propostas baseadas em modificações do algoritmo de inferência sejam teoricamente interessantes, o algoritmo de inferência fica computacionalmente mais complexo e pode inviabilizar sua aplicação para o usuário final. Os autores de Lau et al. (2013) avaliam a contribuição do uso de termos compostos no modelo de tópicos obtidos aplicando o modelo LDA clássico. Foi proposto por Lau et al. (2013) o uso de um processo

de extração de termos compostos antes da etapa de extração de tópicos. Os termos compostos são incluídos na representação dos documentos e os tópicos são obtidos com o processo tradicional do LDA. Propostas mais recentes integram técnicas de extração de *itemsets* frequentes para explorar a informação da dependência dos termos na extração dos tópicos sem aumentar demais a complexidade do processo, como os trabalhos de Kim et al. (2012); Zhu et al. (2012); Gao et al. (2013), mas pouca avaliação foi feita nestes trabalhos a respeito dos ganhos qualitativos sobre os tópicos obtidos.

3.2 Redução de Dimensionalidade

A redução de dimensionalidade sempre foi um grande desafio para a área de mineração de textos. Construir uma boa representação estruturada dos documentos é crucial para a qualidade do resultado obtido ao final do processo. A redução de dimensionalidade pode ser definida como a tarefa de remover termos (ou dimensões) que podem ser ruídos ou pouco importantes em um documento. O objetivo da redução de dimensionalidade é transformar os documentos representados pelos termos $t_i \in T$ ($n = |T|$) em um novo conjunto de atributos T_R ($p = |T_R|$) tal que $p \ll n$.

Existem duas abordagens principais para redução de dimensionalidade. As técnicas de seleção de atributos selecionam um subconjunto de atributos relevantes do conjunto original, ou seja, $T_R \subset T$. Outra importante família de soluções para redução de dimensionalidade é a das técnicas de extração de atributos, que visa reduzir a dimensionalidade da representação para poucas novas dimensões, as quais são combinações lineares ou não-lineares das dimensões originais. Esta última família de soluções será foco deste trabalho.

3.2.1 Extração de Atributos

O processo de extração de atributos está relacionado à criação de um novo conjunto de atributos, usando para tal uma função de mapeamento entre as representações posicionando os dados em uma dimensão latente de forma a representar informações que não são capturadas com os atributos originais. O novo espaço produzido pode ser obtido pela combinação linear ou não-linear dos atributos originais, por transformações que mapeiam o documento em um espaço de conceitos, ou por técnicas probabilísticas como os modelos gerativos de extração de tópicos. Ainda que seleção de atributos seja computacionalmente mais simples e produza resultados significativos, a construção de novos atributos para coleções de textos se mostra relevante para tarefas em que a interpretabilidade é um fator importante. Além disso, as técnicas de extração de atributos são bem sucedidas em descobrir a estrutura latente da coleção de documentos (Shafiei et al., 2007; Hava et al., 2013). Uma representação construída com novos atributos extraídos é uma oportunidade de incorporar conhecimento de domínio aos dados (Guyon and Elisseeff, 2003; Shafiei et al., 2007; Farahat and Kamel, 2011; Kalogeratos and Likas, 2012).

3.2.2 Modelos de Extração de Tópicos

Os modelos de extração de tópicos referem-se a diferentes abordagens que visam a descoberta da estrutura semântica latente - ou tópicos - de uma coleção de documentos. Essa estrutura latente é desconhecida e só é possível acessar as variáveis observadas geradas a partir dela. No caso de coleções de documentos, as variáveis observadas são os termos em cada documento. Os modelos de extração de tópicos são abordagens não-supervisionadas, isto é, os tópicos “emergem” da análise dos textos originais. Atualmente, os modelos de extração de tópicos probabilísticos como o *Latent Dirichlet Allocation* (LDA) são abordagens amplamente aplicadas (Zhu et al., 2012), sendo referenciadas na grande maioria dos trabalhos da literatura como sinônimos de modelos de extração de tópicos (Steyvers and Griffiths, 2007; O’Callaghan et al., 2015). Os modelos de extração de tópicos não-probabilísticos são, em sua maioria, baseados na decomposição de matrizes, sendo os modelos *Latent Semantic Analysis* (LSA) (Deerwester et al., 1990) e o *Non-negative Matrix Factorization* (NMF) (Lee and Seung, 1999) os principais representantes.

Os modelos de extração de tópicos assumem que um documento é construído após a definição *a priori* de um número pequeno de assuntos ou tópicos a serem abordados, e esses tópicos determinam os termos que serão utilizados neste documento. Considera-se a existência de um processo ou um *modelo generativo* do qual o documento deriva-se, cujos parâmetros utilizados para construção dos documentos são desconhecidos, porém podem ser estimados a partir dos documentos e dos termos (variáveis observadas). O processo de extração de tópicos, então, consiste em encontrar a melhor estimativa possível dos parâmetros que deram origem aos documentos da coleção ao assumir verdadeiro um determinado *modelo generativo*. O resultado do processo de extração de tópico é uma representação documento-tópico que determinam um peso de cada tópico para cada documento e uma representação termo-tópico. Esta última está relacionada com o *modelo generativo* que foi escolhido, e pode representar uma probabilidade da ocorrência do termo quando um tópico ocorre em um documento, a frequência esperada desse termo, ou mesmo um peso estimado matematicamente que não pode ser “traduzido” com algum significado para o contexto linguístico.

3.2.3 Modelos de Extração de Tópicos Não-probabilísticos

Nos modelos não-probabilísticos, um “tópico” pode ser entendido como um grupo de termos com pesos indicando a importância ou significância desses termos para algum assunto. Consequentemente, descobrir tópicos com modelos não-probabilísticos equivale a agrupar termos em conjuntos significativos (Cheng et al., 2013a).

O modelo *Latent Semantic Analysis* (LSA)³ (Deerwester et al., 1990) é considerado um dos primeiros modelos de extração de tópicos propostos. O modelo LSA é um método automático que projeta tanto os documentos quanto os termos em um espaço de baixa dimensionalidade que representa os conceitos semânticos dos documentos (Aggarwal and

³Esse modelo também é conhecido como *Latent Semantic Indexing* (LSI) quando aplicado na área de recuperação de informação.

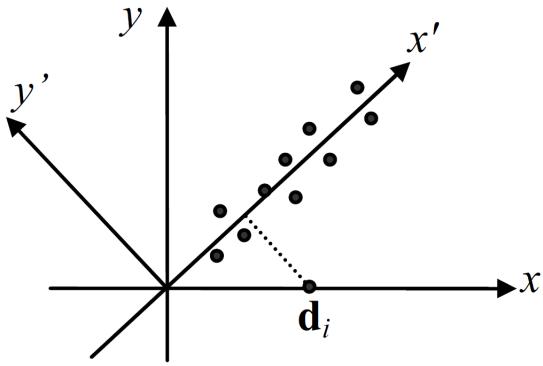


Figura 3.1: Exemplo do resultado obtido pela rotação dos eixos com a técnica SVD. Adaptado de: Liu (2011)

Zhai, 2012). Esse modelo visa reduzir os efeitos adversos causados pela sinonímia e polissemia por meio da identificação de associações estatísticas entre termos. Para isso, aplica-se a técnica *Singular Value Decomposition* (SVD). A ideia dessa técnica é que o espaço original formado pelos termos de W é rotacionado tal que o primeiro eixo aponte para a direção de maior variância dos documentos, o segundo eixo para a direção da segunda maior variância e assim por diante. Na Figura 3.1 é apresentado um exemplo do resultado obtido ao aplicar a rotação dos eixos em um espaço bi-dimensional, tal que o eixo x foi direcionado para apontar na direção de maior variância dos documentos, apontada pelo eixo x' . A redução de dimensionalidade é obtida pela eliminação daqueles eixos do novo espaço obtido que apontam para as direções de menor variância dos documentos.

Um outro modelo importante de extração de tópicos não-probabilísticos é o *Non-negative Matrix Factorization*⁴ (NMF) (Lee and Seung, 1999). O modelo NMF trata o processo de extração de tópicos como o problema de identificar duas matrizes não-negativas Z e A tal que a matriz formada por ZA resulte em uma boa aproximação da matriz documento-termo W . A matriz Z possui dimensão $n \times k$ e corresponde à matriz termo-tópico, e a matriz A corresponde à matriz documento-tópico e possui dimensão $k \times m$, na qual n é a quantidade de termos na coleção de documentos, m é o total de documentos da coleção e k é a quantidade de tópicos desejada. Para gerar as matrizes Z e A , normalmente são consideradas como funções de custo o quadrado da distância Euclidiana e uma generalização da divergência de Kullback-Leibler (KLD) (Arora et al., 2012). Os autores Ding et al. (2008) demonstraram que o modelo NMF aplicado com a função KLD é equivalente à versão probabilística do modelo LSA.

3.2.4 Modelos de Extração de Tópicos Probabilísticos

Modelos de extração de tópicos probabilísticos baseiam-se na premissa de que documentos são misturas de tópicos, e um tópico é uma distribuição probabilística sobre as palavras. Um modelo de tópicos probabilístico propõe um *modelo generativo probabilístico* para os documentos que especifica um procedimento probabilístico pelo qual os documen-

⁴Como a forma em inglês é muito utilizada na literatura da área, optou-se aqui por não traduzir a expressão para o português.

tos podem ser gerados. Para criar um documento, uma distribuição sobre os tópicos é determinada. Em seguida, para cada palavra que será inserida no documento, um tópico é escolhido aleatoriamente com base na sua distribuição definida antes, e uma palavra deste tópico é selecionada. Técnicas estatísticas podem então ser aplicadas para “inverter este processo”, inferindo o conjunto de tópicos que foram responsáveis por gerar a coleção de documentos (Steyvers and Griffiths, 2007). Os modelos probabilísticos se tornaram populares, sendo o modelo *probabilistic Latent Semantic Analysis*⁵ (pLSA) introduzido por (Hofmann, 1999) o primeiro a formalizar a extração de tópicos probabilísticos. Apesar de prover uma boa base para uma análise dos textos, o modelo pLSA apresenta dois problemas. Primeiro, o processo de geração de tópicos para cada documento não é definido, o que exige determinar uma quantidade de parâmetros que cresce linearmente com a quantidade de documentos e pode levar ao “*overfitting*” dos parâmetros estimados. Além disso, o modelo pLSA não determina uma forma natural de calcular as probabilidades relacionadas a um documento que não está no conjunto de treino (Blei et al., 2003; Mauá, 2009; Kim et al., 2012; Aggarwal and Zhai, 2012). Para evitar esses problemas, os autores Blei et al. (2003) propuseram o modelo *Latent Dirichlet Allocation* (LDA).

O modelo LDA é uma extensão do modelo pLSA, que propõe um *modelo gerativo probabilístico* no qual os tópicos são definidos como uma distribuição de probabilidade sobre um vocabulário fixo de termos. Uma característica importante do LDA é que cada documento possui sua própria distribuição de tópicos e, assim, um mesmo documento pode estar relacionado com vários tópicos no qual cada tópico tem sua proporção de relevância. A distribuição dos tópicos em cada documento obedece à distribuição multivariada de *Dirichlet*. O modelo gráfico apresentado na Figura 3.2 é utilizado para representar o modelo LDA. No modelo representado nessa figura, as circunferências indicam as variáveis do modelo e os retângulos indicam a quantidade de vezes que cada variável se repete. Por exemplo, para cada um dos m documentos, a variável w é amostrada N vezes, tal que N é o total de termos do documento d_j tal que $w_{ij} \neq 0$. As variáveis latentes do modelo são de cor branca, enquanto a variável observável, que no caso do LDA são as palavras de cada documento, são destacadas de cor cinza. A notação e o significado correspondente de cada variável do modelo é apresentado na Tabela 3.2. De forma sucinta, pode-se descrever o processo representado no modelo da Figura 3.2 da seguinte forma. Primeiro, amostre k vetores positivos ϕ_k de uma distribuição Dirichlet com parâmetros β . Os vetores ϕ_k , com tamanho n , representam a distribuição dos tópicos para cada um dos n termos $t_i \in T$ da coleção. Então, amostre m vetores positivos θ_j de uma distribuição Dirichlet com parâmetro α . O vetor θ_j que representa a proporção de cada um dos k tópicos para o documento $d_j \in D$ (Blei et al., 2003; Mauá, 2009; Kim et al., 2012; Aggarwal and Zhai, 2012). Na proposta original de Blei et al. (2003) é definido um valor de α para cada um dos k tópicos, porém na literatura utiliza-se uma distribuição simétrica de Dirichlet, tal que $\alpha_1 = \alpha_2 = \dots = \alpha_k = \alpha$ (Kim et al., 2012). Dessa forma, o problema de extração de tópicos para o modelo LDA, dada a coleção de documentos, os termos presentes nessa

⁵Como a forma em inglês é muito utilizada na literatura da área, optou-se aqui por não traduzir a expressão para o português.

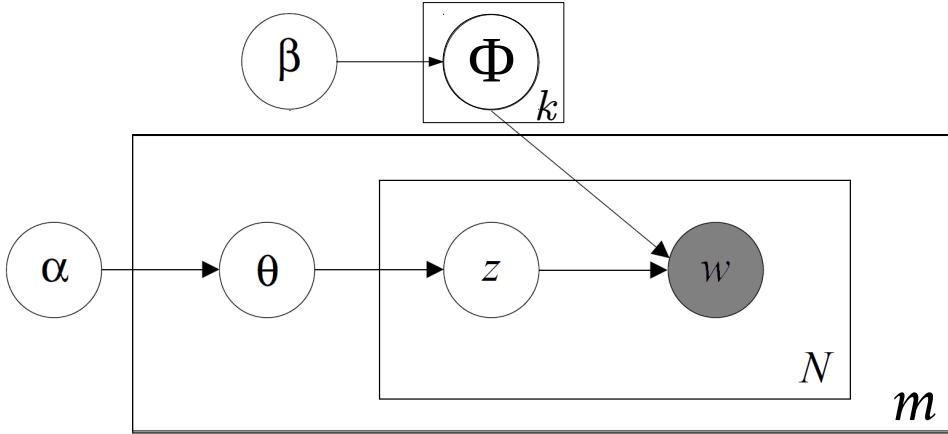


Figura 3.2: Modelo gráfico do LDA. Fonte: Blei et al. (2003)

coleção e o valor k de tópicos da coleção, é inferir as matrizes ϕ e θ . A matriz ϕ é a matriz termo-tópico $n \times k$ tal que cada coluna é formada por um dos vetores $\vec{\phi}_k^T$. A matriz θ corresponde a matriz documento-tópico $k \times m$ tal que cada cada linha é formada por um dos vetores $\vec{\theta}_j$.

Tabela 3.2: Definição das variáveis do modelo LDA utilizadas no modelo gráfico da Figura 3.2.

Variáveis	Significado
k	Número de tópicos
N	Número de termos em um documento d_j
m	Número de documentos
α	Priore da distribuição de Dirichlet em
β	Priore da distribuição de Dirichlet
ϕ	Distribuição dos tópicos sobre os termos de todo o vocabulário T
θ	Distribuição de tópicos por documentos
Z	Tópico associado ao n -ésimo termo gerado para o documento d_j
w	n -ésimo termo gerado para o documento d_j

O processo de inferência necessário para encontrar os parâmetros exatos geradores de uma coleção não pode ser resolvido de forma exata, e alguns métodos de aproximação são aplicados nesse caso, sendo as propostas de uso de algoritmos variacionais e o algoritmo de Gibbs Sampling os mais utilizados (Blei et al., 2003; Mauá, 2009; Kim et al., 2012; Aggarwal and Zhai, 2012). O algoritmo de Gibbs Sampling é o mais amplamente utilizado na literatura, relatado como capaz de obter resultados competitivos aos outros métodos em termos computacionais, com a vantagem de facilidade de entendimento e implementação Mauá (2009). Detalhes sobre métodos para inferência aproximada do modelo LDA podem ser encontrados em Blei et al. (2003); Mauá (2009); Kim et al. (2012); Aggarwal and Zhai (2012).

3.3 Considerações Finais

A obtenção de uma representação estruturada da coleção de documentos é uma etapa importante no processo de mineração de textos. A representação mais utilizadas em mineração de textos é a *bag-of-words*, que representa a coleção por uma matriz documento-

termo em que cada coluna é um termo e, para cada documento, a frequência do termo é contabilizada. Nessa representação, a ordem dos termos e suas relações não são consideradas, e a quantidade de dimensões utilizadas para representar a coleção pode inviabilizar o processo de extração de conhecimento. Nesse sentido, modelos de extração de tópicos fornecem uma estratégia que combina a redução da dimensionalidade e a possibilidade de explorar algumas relações entre termos. Existem muitas evidências na literatura de que representações que incorporam alguma informação de relação entre termos melhoraram os resultados de tarefas como recuperação de informação, classificação de textos, agrupamento de textos e extração de tópicos. Entretanto, a extração eficiente de termos relacionados para construção da representação de documentos ainda é um grande desafio de pesquisa. Os modelos para representação de documentos que explora a correlação entre termos normalmente enfrentam um grande desafio para manter um bom equilíbrio entre a quantidade de dimensões obtidas, o esforço computacional e a interpretabilidade das novas dimensões obtidas.

O uso de regras de associação para identificar dependência entre termos tem se mostrado uma estratégia interessante na literatura, como apresentado nos trabalhos Pôssas et al. (2002); Rossi and Rezende (2011); Kim et al. (2012); Zhu et al. (2012); Gao et al. (2013), levando a bons resultados nas aplicações de mineração de textos e gerando atributos que possuem uma boa interpretabilidade. Entretanto, não existe garantias de que essas soluções resultem em uma representação com menor dimensionalidade quando comparada a representação *bag-of-words*. Os modelos de redução de dimensionalidade que exploram a relação entre termos mais interessantes são os modelos de extração de tópicos. Esses modelos fornecem um mecanismo explícito de redução de dimensionalidade, porém não exploram de forma explícita as possíveis relações entre termos, ou apresentam uma solução que acaba se tornando computacionalmente inviável para grandes coleções.

No próximo capítulo será apresentado o modelo proposto para extração de tópicos baseado em agrupamento de regras de associação para redução de dimensionalidade. São detalhadas todas as etapas do modelo e as formas nas quais cada abordagem apresentada no Capítulo 2 é utilizada para atingir o objetivo do trabalho. Também será apresentada a metodologia proposta para avaliação do modelo.

Extração de Tópicos para Redução de Dimensionalidade

No Capítulo 3, foram discutidos o processo de representação de documentos, e os desafios para a extração de atributos e a importância de utilizar os termos compostos para esse processo. Neste capítulo, são apresentadas as motivações para o modelo proposto LARCM, um modelo de extração de tópicos não-probabilístico que explora o agrupamento de regras de associação para construir uma representação da coleção de documentos com baixa dimensionalidade, na qual cada dimensão é obtida com informações sobre diferentes formas de correlação entre termos.

4.1 Extração de Tópicos com Termos Dependentes para Representação de Documentos

O problema de capturar a dependência entre os termos de forma eficiente para incorporar à representação dos documentos, ainda é um grande desafio de pesquisa, e tem despertado atenção recentemente (Keikha et al., 2009; Figueiredo et al., 2011; Farahat and Kamel, 2011; Kalogeratos and Likas, 2012; Cheng et al., 2013b; Gao et al., 2013). Além disso, diversos modelos para representação de documentos foram propostos na literatura para capturar a dependência entre os termos, como discutido na Seção 3.1.3.

Como apresentado no Capítulo 1, os modelos para representação de documentos que exploram a correlação entre termos normalmente enfrentam um grande desafio para manter um bom equilíbrio entre (i) a quantidade de dimensões obtidas, (ii) o esforço computacional e (iii) a interpretabilidade das novas dimensões obtidas. Existe um grande esforço computacional para a extração das possíveis combinações de termos válidas em grandes coleções de documentos e para determinar quais combinações são significativas para a

aplicação (Figueiredo et al., 2011). Também, os autores Chang et al. (2009) apontam que modelos de extração de tópicos com bons resultados obtidos por avaliações objetivas não levam a melhores resultados em avaliações subjetivas, ou seja, não produzem necessariamente modelos com dimensões com boa interpretabilidade.

Na Tabela 4.1 são apresentados os principais modelos encontrados na literatura para representação de documentos que exploram a correlação entre termos, relacionando quais dos desafios apresentados esses modelos se propõem a resolver. Observa-se que todos os modelos apresentados como alternativa ao modelo *bag-of-words* explorando a correlação entre termos propõem que essa informação de correlação entre termos contribui significativamente para a interpretabilidade das novas dimensões extraídas. Os modelos baseados em *bag-of-words* não apresentam um mecanismo explícito de redução de dimensionalidade e, em suas propostas, precisam construir alguma representação intermediária para processar a correlação entre termos antes de construir a representação final. Os modelos baseados em frases ou termos compostos, por sua vez, exploram o uso de regras de associação como mecanismo para reduzir o esforço computacional da obtenção de termos correlacionados, porém também não apresentam um mecanismo explícito de redução de dimensionalidade. Por exemplo, o modelo *bag-of-related-words*, segundo os autores Rossi and Rezende (2011), apresentam uma quantidade significativamente inferior de dimensões quando comparado ao modelo *bag-of-words*, porém a implementação do método não garante que isso vá ocorrer em todos os cenários, como no mecanismo explícito de definição do número de dimensões existente nos métodos de extração de tópicos. Nos modelos baseados em dimensões latentes, os modelos são orientados a produzir uma quantidade bastante reduzida de dimensões para representar os documentos, e a quantidade de dimensões normalmente é definida *a priori* pelo usuário. Entretanto, cada nova informação a ser considerada pelo modelo durante a extração dos tópicos aumenta substancialmente o esforço computacional do método (Lau et al., 2013).

A dependência entre termos pode ser explorada em diferentes níveis nos documentos. A forma mais comum é a exploração de termos compostos, isto é, quando dois ou mais termos com significados distintos apresentam um novo significado quando ocorrem próximos. Por exemplo, o termo “inteligência artificial” possui um significado enquanto que “inteligência” e “artificial” separadamente podem possuir outros significados. Termos compostos formam uma dependência entre termos em um contexto local, pois o significado dos termos depende apenas de sua proximidade no documento, independente do comportamento do restante da coleção.

Outra forma de dependência encontrada na literatura é a explorada por modelos de extração de tópicos. Nesses modelos, em geral, cada tópico é definido por um conjunto de termos simples utilizados para descrever um determinado assunto. Nesse caso, a dependência entre os termos deve-se ao seu uso em um mesmo contexto, e seu significado é diferente conforme seu uso no contexto, e não depende da proximidade com outros termos. Essa dependência dos termos acontece em um contexto geral, pois o significado dos termos depende dos tópicos existentes na coleção analisada, desconhecidos durante a construção do modelo.

Tabela 4.1: Visão geral das soluções propostas pelos métodos para construção de representação de documentos que exploram a correlação entre termos relacionadas aos três desafios apresentados.

Método	(i) Quantidade de dimensões	(ii) Esforço computacional	(iii) Interpretabilidade das novas dimensões
Baseados no modelo de espaço de vetores			
<i>Context Vector Model</i> - CVM-VSM (Billhardt et al., 2002)			X
Modelo de Espaço de Vetores Generalizado (<i>Generalized Vector Space Model</i> - GVSM) (Wong et al., 1985)			X
<i>Semantic similarity based on term-term correlations</i> (Farahat and Kamel, 2011)			X
<i>Global Term Context Vector-VSM</i> - GTCV-VSM (Kalogeratos and Likas, 2012)			X
<i>Coupled term-term relation model</i> - CRM (Cheng et al., 2013b)			X
Baseados em frases ou termos compostos			
<i>Set-based model</i> (Pôssas et al., 2002)		X	X
<i>c-features</i> (Figueiredo et al., 2011)		X	X
<i>bag-of-related-words</i> (Rossi and Rezende, 2011)		X	X
Baseados em dimensões latentes			
<i>LSI/LSA</i> (Deerwester et al., 1990)	X		X
<i>pLSA</i> (Hofmann, 1999) / <i>LDA</i> (Blei et al., 2003)	X		X
<i>Wallach's Bigram Topic Model</i> (Wallach, 2006)	X		X
<i>frequent pattern-based data enrichment approach</i> (Kim et al., 2012)	X		X
<i>Word-Pair Latent Dirichlet Allocation model</i> - wpLDA (Zhu et al., 2012)	X		X
<i>Two-stage approach topic model</i> (Gao et al., 2013)	X		X

Termos compostos podem aparecer fortemente relacionados em um tópico, entretanto não há garantias de que a ocorrência desses termos em um mesmo tópico esteja relacionada com a sua dependência em um contexto local. Isto é, um tópico descrito pelos termos “inteligência” e “artificial” pode ser encontrado tanto em documentos da área de computação quanto da área de psicologia se referindo a assuntos completamente distintos. Já o termo composto “inteligência artificial” estará mais provavelmente se referindo a técnicas da área de computação.

Considerando os desafios discutidos e os resultados obtidos pelos métodos encontrados na literatura combinando estratégias de extração de tópicos e exploração de termos dependentes, neste trabalho é proposto um novo modelo não-probabilístico para extração de tópicos que explora termos dependentes para representar coleções de documentos e provê uma representação da coleção de documentos com dimensionalidade reduzida.

4.2 LARCM: Latent Association Rule Cluster based Model

O modelo de representação **Latent Association Rule Cluster based Model** (LARCM) proposto explora termos compostos em um contexto local e geral da relação entre termos para identificar tópicos em coleções de documentos. Neste modelo, o **contexto local** da relação entre termos é definido pela identificação de suas co-ocorrências utilizando informações extraídas do documento em que eles ocorrem, ou seja, é independente da

ocorrência dos termos em outros documentos da coleção. Essa informação é explorada neste trabalho para a identificação de termos compostos, como “mineração de dados” ou “inteligência artificial”. Para o **contexto geral** da relação entre termos, identifica-se alguma relação entre os termos extraídos do contexto local. Nos modelos de extração de tópicos, a relação entre os termos normalmente é dada pelo seu agrupamento; em que cada grupo contém termos fortemente relacionados a um assunto ou tópico expresso nos documentos da coleção. Isso permite resolver problemas importantes como a polissemia, o que ainda é um grande desafio na área quando os termos avaliados são compostos. Por exemplo, os termos compostos “mineração de dados” e “extração de conhecimento” podem ser empregados nos textos com significado semelhante, mas são formados por termos distintos.

Na Figura 4.1 é apresentada a visão geral do LARCM. O contexto local captura a correlação entre os termos em cada documento pela coocorrência obtida com a extração de regras de associação. Para capturar o contexto geral, foi proposta uma representação intermediária para as regras de associação que permite representá-las no espaço de vetores original dos documentos. Cada regra de associação extraída é representada por um vetor de termos distintos que aparecem na coleção de documentos, e o peso de cada termo é dado pela sua frequência nas transações cobertas pela regra de associação. Isso permite aplicar medidas de similaridade como a medida de cosseno, para determinar o contexto geral de cada relação obtida considerando todas as outras relações da coleção. A ideia é que a informação da vizinhança dos termos, obtida pela análise do conjunto de transações que cada regra cobre, ajuda a identificar termos diferentes utilizados em um mesmo contexto ou com mesmo sentido e termos idênticos mas que são usados em contextos diferentes ou com significados diferentes. Por exemplo, o termo “cluster” pode estar fazendo referência a uma técnica de mineração de dados quando aparece próximo do termo “algoritmo” ou pode estar fazendo referência a um grupo de computadores quando aparece próximo do termo “rede”.

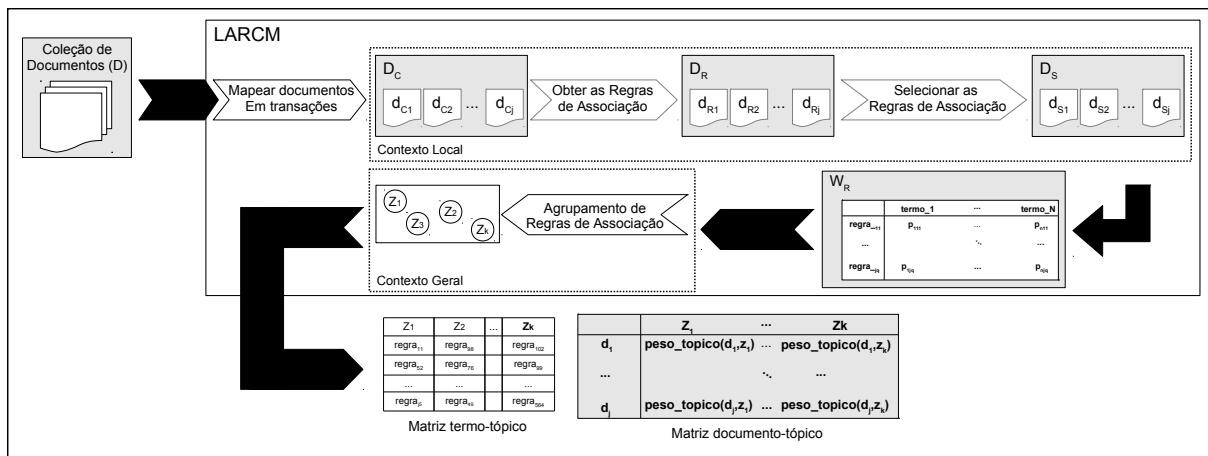


Figura 4.1: Visão geral do Latent Association Rule Cluster based Model - LARCM.

Antes de iniciar o processo do LARCM, recomenda-se o pré-processamento da coleção de documentos como a normalização de palavras, remoção de *stopwords*, e a limpeza e

padronização dos documentos como descritos em Nogueira et al. (2008).

Por fim, para as definições dos passos a seguir, considere $D = \{d_1, d_2, \dots, d_m\}$ uma coleção com m documentos, e $T = \{t_1, t_2, \dots, t_n\}$ o conjunto de n termos distintos da coleção. Cada documento d_j é representado por um vetor de termos $\vec{d}_j = \{w_{1j}, w_{2j}, \dots, w_{nj}\}$, no qual cada peso w_{ij} quantifica a importância do termo $t_i \in T$ para o documento $d_j \in D$. Para os termos da coleção que não estão presentes no documento d_j , $w_{ij} = 0$. A coleção de documentos é então representada pela matriz W com dimensões $m \times n$, conhecida na literatura como matriz documento-termo. Cada linha de W corresponde a um documento em D , isto é, o vetor \vec{d}_j , e cada coluna descreve a distribuição de cada termo na coleção de documentos. Para ilustrar as ideias aqui apresentadas, será utilizada uma coleção de sete textos em português construídos para cobrir alguns tópicos relevantes da área de computação: ciência da computação, engenharia de software, banco de dados, *data warehouse*¹, inteligência artificial, aprendizado de máquina e mineração de dados. Os textos foram criados utilizando o primeiro parágrafo da entrada da Wikipedia referente a cada assunto. Os assuntos foram selecionados por possuírem algum tipo de interseção entre as palavras utilizadas para descrevê-los. Os textos completos e as saídas completas geradas por cada etapa do método para essa coleção de exemplo são apresentados no Apêndice A.

4.2.1 Contexto Local da Relação

Para obter as informações do contexto local da relação entre termos no LARCM, foi proposto um processo adaptado do modelo *bag-of-related-words* (Rossi and Rezende, 2011). O modelo *bag-of-related-words* oferece um processo eficiente para extração de termos compostos utilizando regras de associação. Nesse modelo, as regras de associação são extraídas para cada documento e, após selecionadas, são utilizadas para construir os termos compostos da coleção. Como cada documento é processado de forma independente, os termos compostos são obtidos para o contexto local da relação. Dos quatro passos do modelo *bag-of-related-words*, foram adaptados dois deles para o LARCM: (1) Mapear os documentos em transações; (2) Extrair as regras de associação das transações de cada documento. Esses dois passos foram adaptados para que as informações fornecidas ao final de cada um deles pudessem ser utilizadas para construir a matriz da qual será obtido o contexto geral das relações do LARCM.

O mapeamento dos documentos foi realizado utilizando janelas deslizantes. Nesse mapeamento a primeira transação contém apenas a primeira palavra do documento, a segunda contém as duas primeiras palavras, e assim por diante, até que a janela contenha o número de palavras igual ao tamanho definido “*tamanho_janela*”. Após isso, a janela desliza uma palavra e considera as próximas “*tamanho_janela*” palavras do documento. O resultado desse processo é a coleção $D_C = \{d_{C1}, d_{C2}, \dots, d_{Cm}\}$ em que d_{Cj} corresponde às transações obtidas pelo mapeamento do documento $d_j \in D$. O método apresentado no Algoritmo 4 executa esse mapeamento.

No Algoritmo 4, percorrem-se os documentos pré-processados (linha 2) e, para cada

¹Na Wikipedia em português o artigo referente ao assunto é armazém de dados.

Algoritmo 4 Método Mapeamento-documento-transação

Entrada: Coleção de documentos D ; $tamanho_janela$: tamanho da janela.

Saída: A coleção D_C dos documentos de D mapeados em transações.

```
1:  $D_C \leftarrow \emptyset$ 
2: para cada documento  $d_j \in D$  faça
3:    $d_{Cj} \leftarrow \emptyset$ 
4:   para  $i \leftarrow 1$  to  $|palavras\ em\ d_j|$  faça
5:      $transacao \leftarrow \emptyset$ 
6:      $w \leftarrow i$ 
7:     enquanto  $|transacao| \leq tamanho\_janela$  AND  $i + w \leq |palavras\ em\ d_k|$ 
      faça
8:        $transacao \leftarrow transacao \cup palavra_{i+w}$ 
9:        $w \leftarrow w + 1$ 
10:    fim enquanto
11:     $d_{Cj} \leftarrow d_{Cj} \cup transacao$ 
12:  fim para
13:   $D_C \leftarrow D_C \cup d_{Cj}$ 
14: fim para
15: Retornar  $D_C$ 
```

documento d_j , obtém-se suas palavras na sequência em que elas aparecem no documento (linha 4). Adiciona-se a i -ésima *palavra* encontrada ao conjunto que formará uma nova transação. Em seguida, as palavras consecutivas são adicionadas ao conjunto que formará a transação. Este processo é repetido até que o conjunto tenha “*tamanho_janela*” itens ou até que não seja possível obter palavras do documento processado (linhas 5 a 10). A transação formada é adicionada ao conjunto de transações d_{Cj} relacionadas ao documento d_j (linha 11), e o processo é executado até que todos os documentos sejam mapeados em transações.

Por exemplo, considere o conjunto de documentos do exemplo (Apêndice A) e uma janela de tamanho 4 definida pelo usuário. O processo é executado como mostrado na Figura 4.2. Cada documento é processado individualmente. A janela é posicionada no início do documento (Figura 4.2 - a), e as palavras da janela são adicionadas a lista de transações do documento até formar uma transação de tamanho 4. Em seguida, a janela é deslocada para a esquerda de uma palavra, e a sequência encontrada é adicionada a lista de transações (Figura 4.2 - b). Esse processo é executado até o final do documento (Figura 4.2 - c), e a janela vai reduzindo até obter apenas uma palavra (Figura 4.2 - d). Então, o próximo documento é processado da mesma forma. O processo termina quando todos os documentos forem processados.

Como apresentado no Algoritmo 5, cada conjunto de transações $d_{Cj} \in D_C$ (linha 3) é processado por um algoritmo de geração de regras de associação como o Apriori (Agrawal and Srikant, 1994) (linha 5). Definem-se os valores *supmin* e *confmin* de suporte mínimo e confiança mínima, respectivamente, utilizados pelo algoritmo de geração de regras de associação para cada d_{Cj} . Utilizou-se o valor de $confmin = 0$ (linha 2), permitindo gerar todas as regras de associação possíveis, pois as regras que serão exploradas ao longo do processo são selecionadas utilizando uma medida objetiva na etapa seguinte. Neste

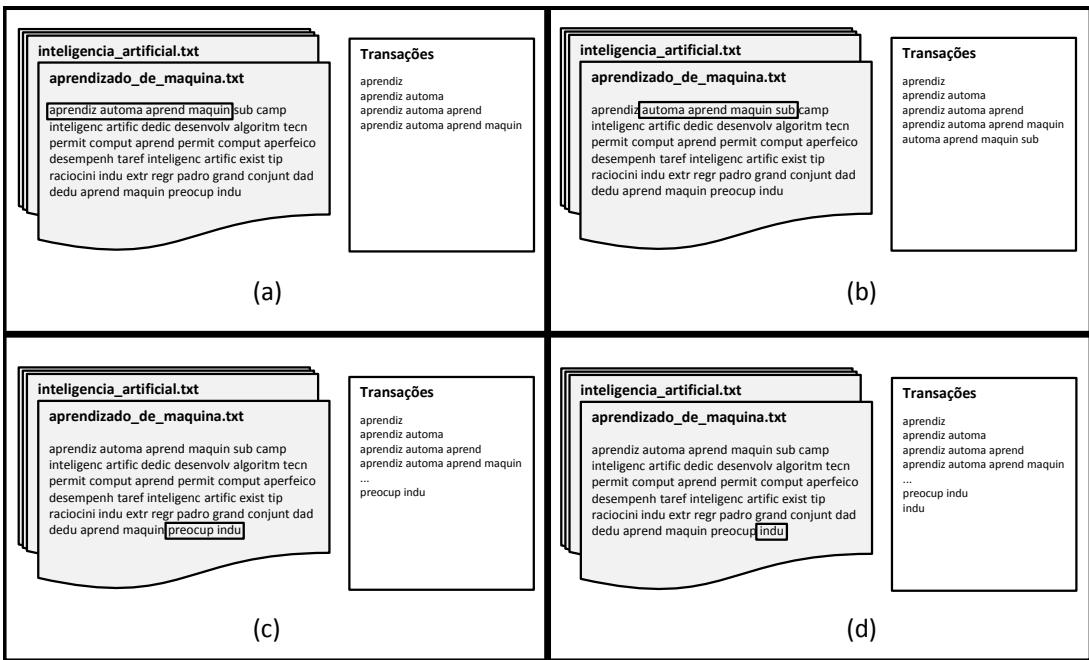


Figura 4.2: Exemplo do processo de mapeamento dos documentos em transações (Algoritmo 4). (a) A janela é posicionada no início do documento e as palavras são adicionadas ao conjunto de transações até $tamanho_janela$ ser igual a 4. (b) A janela é deslocada para a direita e a sequência encontrada de palavras é adicionada ao conjunto de transações. (c) Esse processo é repetido até o final do documento - $tamanho_janela$ começa a ficar menor. (d) A janela de processamento é reduzida até restar uma palavra.

trabalho utilizou-se o cálculo automático do valor $supmin$ para cada documento (linha 4), como proposto por Rossi and Rezende (2011), cuja fórmula é:

$$supmin(d_{Cj}) = \frac{\left(\sum_{\forall t_i \in A} w_{ij} \times tamanho_janela \right) / |A|}{|d_{Cj}|} \times \alpha \quad (4.1)$$

na qual A é o conjunto de termos $t_i \in T$ para o documento d_j tal que $w_{ij} > 0$, $|A|$ é o número total de termos de d_j , $tamanho_janela$ é o valor de tamanho de janela definido para o mapeamento das transações, $|d_{Cj}|$ é o número total de transações do documento d_j e α é um multiplicador que permite suavizar ou intensificar o limiar gerado pelo cálculo e é normalmente definido como 1. O cálculo automático do suporte mínimo possibilita liberar o usuário da responsabilidade de definir esse parâmetro e, segundo os autores Rossi and Rezende (2011), esse cálculo apresenta resultados comparáveis aos obtidos com a definição manual do valor de $supmin$. O resultado desse processo é a coleção $D_R = \{d_{R1}, d_{R2}, \dots, d_{Rm}\}$ em que d_{Rj} corresponde às regras de associação obtidas pelo processamento das transações d_{Cj} do documento d_j . Na Figura 4.3, são apresentadas algumas regras de associação geradas a partir das transações obtidas no exemplo da Figura 4.2.

Considere novamente conjunto de documentos de exemplo (Apêndice A) e o conjunto de transações obtidos como apresentado na Figura 4.2. O Algoritmo 5 recebe o conjunto de documentos mapeados em transações D_C , a coleção de documentos D , a matriz documento-termo W dessa coleção e o conjunto T com todos os termos da coleção. Para

Algoritmo 5 Método gera-regras

Entrada: Coleção de documentos D ; Conjunto de termos distintos da coleção T ; Matriz documento-termo W ; *tamanho_janela*: tamanho da janela definido para o mapeamento das transações; α : multiplicador do suporte automático; Coleção de documentos mapeados em transações D_C .

Saída: A coleção D_R dos documentos de D_C com as regras de associação geradas.

- 1: $D_R \leftarrow \emptyset$
 - 2: $confmin \leftarrow 0$
 - 3: **para** cada documento $d_{Cj} \in D_C$ **faz**
 - 4: $supmin \leftarrow calcula_supmin(d_{Cj}, tamanho_janela, \alpha, D, T, W)$
 - 5: $d_{Rj} \leftarrow apriori(d_{Cj}, supmin, confmin)$
 - 6: $D_R \leftarrow D_R \cup d_{Rj}$
 - 7: **fim para**
 - 8: Retornar D_R
-

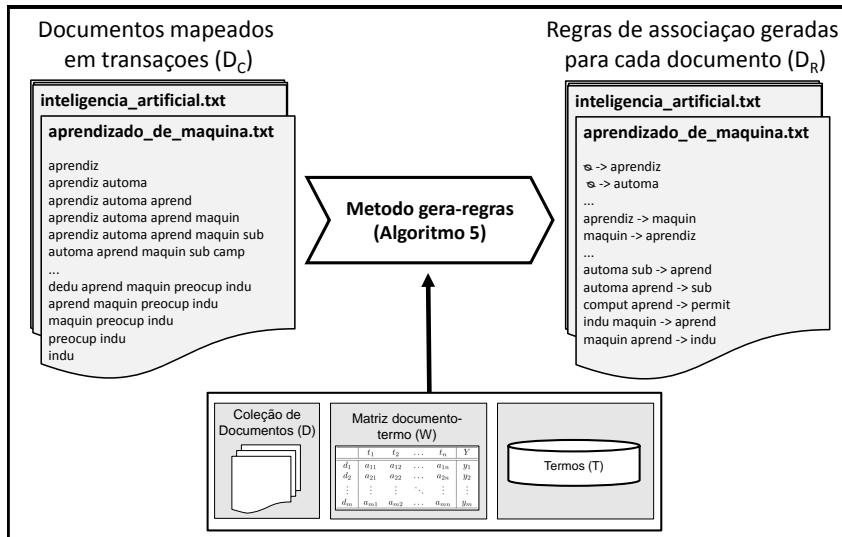


Figura 4.3: Exemplo do processo de obtenção das regras de associação para cada documento (Algoritmo 5) aplicado no conjunto de transações obtidos pelo processo apresentado na Figura 4.2. Para o documento em destaque do exemplo, foi calculado o valor de suporte mínimo $supmin$ igual a 4.67%, considerando um valor de α igual a 0,3. No total, foram obtidas 514 regras para o documento *aprendizado_de_maquina.txt*. As regras do tipo $\emptyset \Rightarrow atributo$ são interessantes pois irão representar termos simples ao final do processo do LARCM (Seção 4.2.3).

cada conjunto de transações d_{Cj} de cada documento d_j , é calculado um valor de suporte mínimo. Aplica-se um algoritmo de geração de regras de associação e o conjunto de regras obtidas é armazenado em d_{Rj} . Além de termos compostos, formados por mais de um termo, é importante que o modelo proposto possa encontrar também termos simples, uma vez que muitos dos termos possuem significado próprio. Para isso, são consideradas regras de associação geradas do tipo $\emptyset \Rightarrow atributo$, que é formada por apenas um item. Essas regras de associação irão representar termos simples importantes para a coleção ao final do processo do LARCM.

Por fim, é feita a seleção das regras de associação de interesse, como apresentado no Algoritmo 6. Para cada conjunto de regras de associação $d_{Rj} \in D_C$ (linha 2) calculase o valor de uma medida objetiva escolhida da literatura para selecionar as regras de

associação (linha 3). A seleção da medida objetiva é um passo importante, dado que cada qual possui uma semântica própria e tem influência no tipo de termo composto que será obtido. Visando reduzir o esforço do usuário na definição dos parâmetros, optou-se aqui por utilizar uma heurística para o cálculo do valor de corte para uma dada medida objetiva em cada documento. Foi utilizada como heurística o valor da média da medida objetiva considerando todas as regras geradas no documento (linha 5), cuja fórmula é:

$$min_medida(d_{Rj}) = \frac{\sum_{r_{jq} \in d_{Rj}} valor_da_medida(r_{jq})}{|d_{Rj}|} \quad (4.2)$$

na qual r_{jq} é a q-ésima regra de associação em d_{Rj} , $valor_da_medida(r_{jq})$ é o valor calculado da medida objetiva para a regra de associação r_{jq} e $|d_{Rj}|$ é o número de regras de associação em d_{Rj} . Essa heurística já foi utilizada com sucesso por Rossi and Rezende (2011). Com o valor de corte definido, todas as regras de associação com o valor de medida objetiva maior ou igual ao valor de corte são selecionadas (linhas 6 a 10). O resultado desse processo é a coleção $D_S = \{d_{S1}, d_{S2}, \dots, d_{Sm}\}$ em que d_{Sj} corresponde às regras de associação selecionadas de d_{Rj} com valor de medida objetiva maior do que min_medida_j .

Algoritmo 6 Método seleciona-regras

Entrada: Coleção de documentos D_R com as regras de associação geradas; $medida_objetiva$: a medida objetiva para seleção das regras de associação.

Saída: A coleção D_S dos documentos de D_R com as regras de associação selecionadas.

```

1:  $D_S \leftarrow \emptyset$ 
2: para cada documento  $d_{Rj} \in D_R$  faz
3:    $d_{Rj} \leftarrow calcula\_valor\_medida\_objetiva(d_{Rj}, medida\_objetiva)$ 
4:    $d_{Sj} \leftarrow \emptyset$ 
5:    $min\_medida \leftarrow calcula\_min\_medida(d_{Rj}, medida\_objetiva)$ 
6:   para cada regra de associação  $r_{jq} \in d_{Rj}$  faz
7:     se valor da medida objetiva de  $r_{jq} > min\_medida$  então
8:        $d_{Sj} \leftarrow d_{Sj} \cup r_{jq}$ 
9:     fim se
10:   fim para
11:    $D_S \leftarrow D_S \cup d_{Sj}$ 
12: fim para
13: Retornar  $D_S$ 
```

Na Figura 4.4 é apresentado um exemplo do processo executado pelo Algoritmo 6, considerando o exemplo da Figura 4.3. Para cada documento, calcula-se o valor da medida objetiva selecionada de cada regra de associação. No nosso exemplo, considerou-se a medida objetiva *Odds Ratio*; detalhes sobre essa medida objetiva e outras medidas objetivas são descritas na Seção 2.3.3. O valor médio para a medida objetiva considerando todas as regras de associação foi calculado como descrito pela Equação 4.2, e o valor calculado foi de 0.07013369. Todas as regras de associação com valor de *Odds Ratio* maior ou igual a esse são selecionadas. O mesmo procedimento é executado para cada um dos documentos, definindo um valor de corte diferente para cada um e selecionando as regras de associação utilizando a mesma medida objetiva selecionada anteriormente.

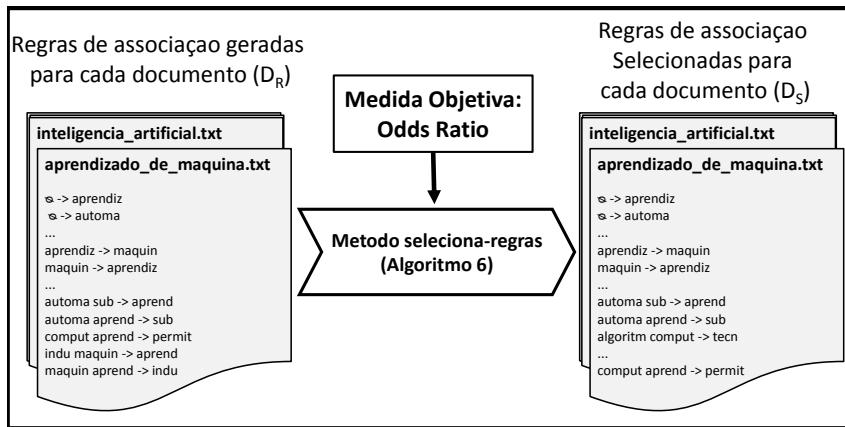


Figura 4.4: Exemplo do processo de seleção das regras de associação para cada documento (Algoritmo 6) aplicado no conjunto de regras de associação obtidos pelo processo apresentado na Figura 4.3. A medida objetiva selecionada para o exemplo foi a *Odds Ratio*. Para o documento *aprendizado_de_maquina.txt*, o valor corte calculado para a medida foi de 0.07013369, ou seja, as regras de associação com valor maior ou igual a esse foram selecionadas para o conjunto d_S . Para esse documento, foram selecionadas 258 das 514 regras de associação geradas.

4.2.2 Contexto Geral da Relação

Uma vez obtidos os termos considerando suas relações no contexto local, esses termos são agrupados com o objetivo de colocar em um mesmo grupo termos que estão relacionados com um mesmo assunto ou tópico. Para formar esses grupos, é preciso identificar uma relação entre os termos no nível da coleção que permita identificar os tópicos da coleção. Entretanto, a informação de quantos e de quais tópicos existem na coleção é desconhecida, bem como a informação de quais termos obtidos estão relacionados com quais tópicos. Considera-se, então, que existe um conjunto de termos mais prováveis de serem utilizados para expressar um determinado tópico ou assunto em cada documento, e que cada termo foi utilizado no documento para expressar um determinado tópico. Também, considera-se aqui que cada documento pode tratar de um ou mais assuntos ou tópicos.

Visando estimar essas dimensões desconhecidas que formam os tópicos, propõe-se o uso de agrupamento das regras de associação obtidas na etapa anterior. O agrupamento de regras de associação tradicional normalmente utiliza como medida de similaridade alguma variação do esquema que conta os itens em comum entre duas regras de associação, seja da própria regra ou das transações que essa regra cobre. Entretanto, no caso das regras de associação obtidas de textos, existe uma informação de contexto que pode ser melhor explorada. A hipótese desse trabalho é que as transações cobertas por cada regra de associação obtidas de documentos de texto corresponde à vizinhança de cada termo e que termos cujas vizinhanças são muito semelhantes possivelmente estão relacionados a um mesmo tópico. De acordo com a hipótese distribucional (Turney and Pantel, 2010), termos que ocorrem em contextos similares tendem a ter significados semelhantes. Assim, seja k a quantidade de tópicos de uma coleção conhecido, as regras de associação são agrupadas em k grupos, de forma que cada grupo obtido corresponde a um tópico e as regras de associação desse grupo formam os termos que são mais prováveis de serem utilizados

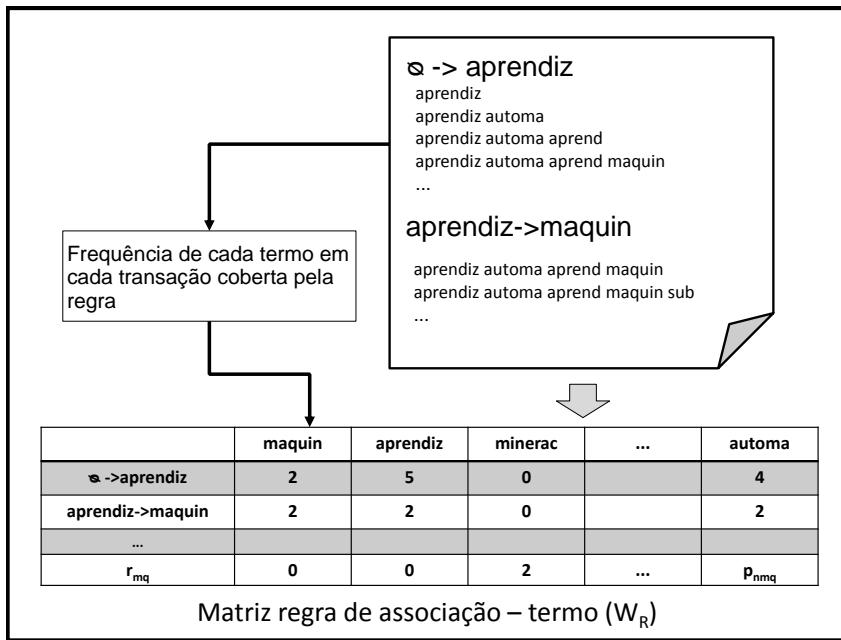


Figura 4.5: Exemplo de construção da matriz regra de associação-termo W_R .

quando um documento trata de um determinado tópico.

Para obter cada um dos grupos de regras de associação, é preciso obter uma representação estruturada que possa capturar as informações relevantes da vizinhança de cada termo. Neste trabalho, propõe-se uma representação intermediária das regras de associação obtidas utilizando os atributos extraídos $T = \{t_1, t_2, \dots, t_n\}$ dos documentos para formar uma matriz regra de associação-termo W_R . Nesta proposta, o peso de cada atributo para cada regra de associação é obtido ao explorar o conjunto de transações cobertas pela regra, ou seja, o conjunto de transações que gerou aquela regra. Assim, considera-se no processo o contexto local da relação e sua vizinhança para encontrar o contexto geral das relações. Cada regra r_{jq} é representada por um vetor de termo $\vec{r}_{jq} = \{p_{1jq}, p_{2jq}, \dots, p_{njq}\}$, no qual cada peso p_{ijq} quantifica a importância do termo $t_i \in T$ para a regra de associação r_{jq} , de forma que p_{ijq} representa a frequência do termo t_i nas transações em $d_{Cj} \in D_C$ cobertas pela regra r_{jq} . Essas regras de associação são então representadas na matriz regra de associação-termo W_R , na qual cada linha corresponde ao vetor da regra de associação \vec{r}_{jq} , e cada coluna descreve a importância p_{ijq} do termo t_i em r_{jq} . Regras de associação idênticas obtidas em documentos diferentes são tratadas como regras diferentes, uma vez que cada uma delas pode estar relacionada a tópicos diferentes. Assim, o modelo proposto apresenta um mecanismo explícito de identificação de polissemias.

Na Figura 4.5, é apresentado um exemplo da construção da matriz regra de associação-termo considerando as regras de associação obtidas ao final do processo apresentado na Figura 4.4. Para cada regra de associação, são encontradas as transações cobertas por essa regra, e cada termo é contabilizado indicando a frequência de cada termo para cada regra de associação. Isso permite não só identificar regras de associação com termos em comum, mas também expressa uma informação de intensidade com que cada termo está presente na vizinhança da regra.

Uma vez obtida a representação W_R , aplica-se um algoritmo de agrupamento tradici-

onal, como apresentado na Figura 4.1. Assim como em outros métodos, o número k de grupos gerados deve ser informado pelo usuário. O resultado desse processo é o conjunto de grupos $Z = \{z_1, z_2, \dots, z_k\}$, em que cada grupo $z_x \in Z$ contém regras de associação obtidas dos diferentes documentos de D_S .

4.2.3 Construção da Representação Documento-Tópico

Por fim, o peso de cada tópico para cada documento é calculado para formar a representação final no formato matriz documento-tópico. Cada grupo de regras obtido forma um tópico. No LARCM, o peso do tópico para cada documento é calculado pela proporção das regras de associação do documento que estão presentes em cada grupo. Ou seja, do total de regras de associação do documento, quantas delas constituem a cada tópico. O uso do valor relativo como peso permite interpretar esse peso como uma ideia da proporção do documento que trata de cada tópico. Assim, o peso de um tópico para cada documento é calculado pela equação a seguir:

$$peso_topico(d_j, z_x) = \frac{|d_{Sj} \cap z_x|}{|d_{Sj}|} \quad (4.3)$$

no qual $|d_{Sj} \cap z_x|$ é a quantidade total de regras de associação de d_{Sj} no tópico z_x e $|d_{Sj}|$ é a quantidade total de regras de associação em d_{Sj} . Este valor pode ser interpretado como a proporção do documento que trata do tópico z_x .

Seleção de Descritores dos Tópicos

Para que o usuário possa analisar e identificar os tópicos obtidos tradicionalmente apresenta-se uma lista dos termos mais relevantes do tópico. Neste trabalho, é proposta a seleção das L melhores regras de associação de cada grupo $z_x \in Z$ de acordo com uma medida objetiva para regras de associação. A medida objetiva é a mesma escolhida para extrair as relações no contexto local (Section 4.2.1). Na avaliação realizada em Santos et al. (2014), concluiu-se que esse processo é capaz de selecionar descritores mais significativos em comparação aos obtidos pelo modelo LDA, considerado o método estado da arte na área. Considere a notação a seguir que define o processo de seleção de descritores:

$$descritores(z_x) = \{melhores_l(z_x), l = 1, \dots, L\},$$

no qual $melhores_l$ é a função que seleciona as L melhores regras de associação de acordo com a medida objetiva escolhida para o k -ésimo tópico. L é um número fixo empiricamente definido como $L = 10$ em muitos trabalhos da literatura. O resultado é representado em uma matriz termo-tópico similar àquelas obtidas em outros modelos de tópicos. Essa matriz representa para cada tópico a lista dos termos relevantes para cada tópico ordenados por um critério de importância definido.

4.3 Metodologia para Avaliação do Modelo Proposto

Uma vez realizada a extração dos tópicos é necessário avaliar o modelo proposto. A avaliação de modelos de extração de tópicos é um grande desafio, uma vez que a definição de quais são os tópicos relevantes de uma coleção é bastante subjetiva e depende da interpretação do usuário. Além disso, os autores Chang et al. (2009) apresentam em seu trabalho a discordância entre os resultados de uma avaliação objetiva e subjetiva de modelos de tópicos, isto é, modelos que produziam tópicos muito bem avaliados com medidas objetivas não tinham bom desempenho na avaliação subjetiva realizada com usuários. Considerando estas questões e os desafios discutidos na Seção 4.1, foi proposto neste trabalho o processo de avaliação apresentado na Figura 4.6. Essa avaliação está dividida em duas partes. A primeira, apresentada na Seção 4.3.1, avalia o desempenho do modelo proposto para representação da coleção de documentos. Para isto, a representação documento-tópico obtida é utilizada na tarefa de classificação de documentos, e os resultados são comparados com os obtidos pelo modelo LDA, considerado o estado da arte na área. A segunda parte da avaliação, apresentada na Seção 4.3.2, se baseia na proposta de Lau et al. (2014) para simular um usuário real avaliando a interpretabilidade dos tópicos obtidos. Os autores de Lau et al. (2014) demonstram em seu trabalho que esse processo de avaliação é capaz de simular de forma muito semelhante a avaliação com usuários reais.

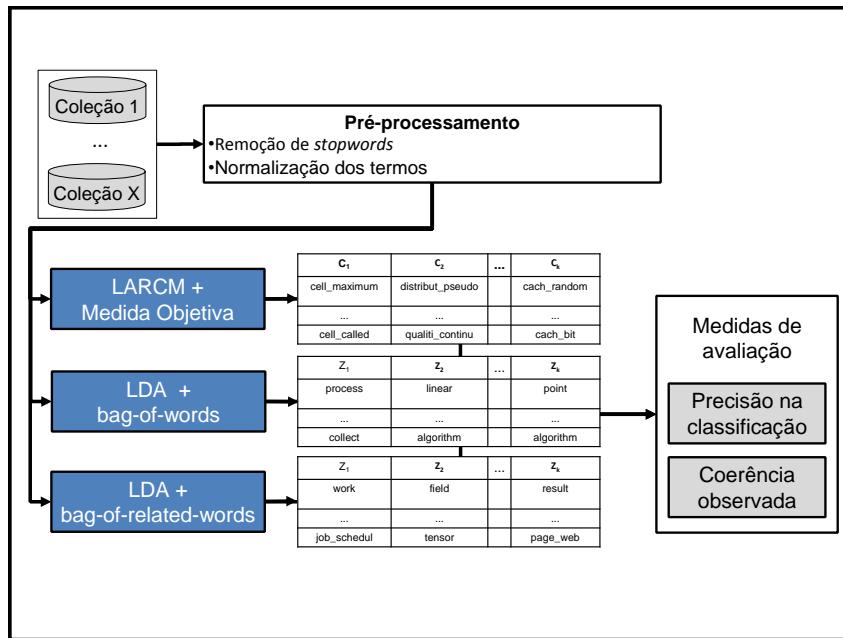


Figura 4.6: Visão geral da metodologia para avaliação do modelo proposto.

4.3.1 Avaliação na Tarefa de Classificação

A representação da coleção de documentos obtida pelo modelo LARCM pode ser utilizada para diversos fins, além do seu uso para exploração da coleção com base nos tópicos extraídos. Uma tarefa importante na mineração de textos é a classificação automática de documentos. A tarefa de classificação de documentos sempre obteve ótimos resulta-

dos em aplicações práticas (Liu, 2011). Devido ao seu uso e sucesso, foi escolhida para avaliação do método proposto. Existe uma grande variedade de coleções de documentos rotuladas disponíveis para avaliação, que permitem uma avaliação objetiva baseada na taxa de acerto e erro dos classificadores induzidos.

O objetivo deste processo de avaliação proposto é identificar qual método pode contribuir para melhorar a acurácia dos modelos obtidos. Após o processo de classificação, são obtidos os valores de documentos corretamente classificados t_p e o total de documentos da coleção $|D|$. Com base nesses valores, a acurácia do modelo construído será dada por:

- Acurácia: $acur = t_p / |D|$

proporção de documentos corretamente classificados em relação ao total de documentos da coleção;

Cada uma das representações obtidas das coleções de documentos para avaliação são apresentadas aos classificadores selecionados. Foi utilizado o protocolo de avaliação cruzada de 10-folds², e cada modelo foi treinado 10 vezes. A média das 100 execuções é obtida e esse valor é utilizado para representar a acurácia do classificador para a representação.

Os resultados experimentais dos valores de acurácia são tabulados para cada classificador avaliado, juntamente com a informação da coleção de documentos avaliada (cd , O = “coleção 1”, “coleção 2”, ...), o valor definido de tópicos para extração (k), o método de redução de dimensionalidade (rm , rm = “método 1”, “método 2”, ...), e o classificador utilizado (cl , cl = “algoritmo 1”, “algoritmo 2”, ...), de maneira que seja possível gerar seu modelo de variância.

Normalmente, se estivesse disponível o resultado de uma medida de avaliação obtida diretamente da saída dos métodos de redução de dimensionalidade, seria possível calcular a média da medida e seu desvio padrão para cada método de redução de dimensionalidade. Em seguida, é possível aplicar o teste *t Student* para avaliar estatisticamente a diferença entre as médias. Entretanto, neste processo de avaliação, estão disponíveis N modelos de redução de dimensionalidade (rm), aplicados a diferentes classificadores (cl), em diferentes coleções de documentos (cd) e com diferentes valores de dimensões extraídas (k). Neste caso, a análise de variância deve ser capaz de avaliar as diferenças entre os 19 modelos de redução de dimensionalidade, considerando os efeitos dos outros componentes. Para que a avaliação seja justa, a variância total deve ser decomposta considerando todos esses fatores (Moura, 2009). Essa decomposição da variância pode ser representada no seguinte modelo linear (para detalhes veja Searle (1971)):

$$m(\widehat{Acur}_{mr}) = \hat{\mu} + \hat{cd} + \hat{k} + \hat{mr} + \hat{cl} + \hat{e} \quad (4.4)$$

Em que:

- $m(\widehat{Acc}_{mr})$: valor estimado da acurácia para o método de redução de dimensionalidade mr ;

²Em inglês *10-fold cross-validation*

- $\hat{\mu}$: valor estimado para a média geral da acurácia;
- \hat{cd} : valor estimado para o efeito da coleção de documentos cd na estimativa do valor da acurácia;
- \hat{k} : valor estimado para o efeito do valor definido de tópicos para extração k na estimativa do valor da acurácia;
- \hat{mr} : valor estimado para o efeito do método de redução de dimensionalidade mr na estimativa do valor da acurácia;
- \hat{cl} : valor estimado para o efeito do classificador cl na estimativa do valor da acurácia;
- \hat{e} : valor estimado para o componente do erro do modelo para o método de redução de dimensionalidade mr , supondo que ele é aleatório.

Para cada representação, uma média geral é obtida considerando todos os resultados em todas as coleções de documentos e para todos os classificadores treinados. Dessa forma, compara-se a influência de cada modelo para representação de documentos e redução de dimensionalidade para a melhora do desempenho de alguma tarefa relevante em relação à média geral. Nesse caso, foi escolhida a tarefa de classificação. Após considerar o efeito de diferentes classificadores em diferentes conjuntos de textos com diferentes quantidades de tópicos compararam-se os ruídos dessas inferências. Dessa forma, é possível apontar para os modelos que estatisticamente vão contribuir para uma melhora na acurácia para a tarefa de classificação.

4.3.2 Avaliação da Interpretabilidade

A avaliação da interpretabilidade dos tópicos é bastante subjetiva, e depende do esforço de muitos avaliadores para se tornar significativa. Com base na proposta de avaliação apresentada por Lau et al. (2014), optou-se por utilizar um processo automático que simula a avaliação de especialistas e fornece uma estimativa bastante confiável da qualidade dos tópicos obtidos.

Os autores Newman et al. (2010) definiram a interpretabilidade de um tópico baseado na avaliação feita por usuários especialistas da coerência observada dos N termos selecionados como descritores do tópico. Na proposta original, os tópicos extraídos são apresentados para um grupo de especialistas que, seguindo um conjunto de instruções padronizadas, devem avaliar a qualidade do tópico quanto à sua utilidade em uma escala de 3 pontos, em que a nota 3 indica um tópico útil (coerente) e a nota 1 indica um tópico pouco útil (menos coerente). Em Lau et al. (2014), os autores estenderam a avaliação para aplicar um processo automático de avaliação da medida de coerência observada. Entre os métodos automáticos avaliados, a medida NPMI (Normalized Pointwise Mutual Information) foi apontada como a que mais se aproxima da avaliação feita pelos especialistas, e pode ser utilizada para automatizar a avaliação da coerência do tópico considerando os termos selecionados como descritores e sua coocorrência em relação a uma coleção de referência. Nessa implementação, a coerência observada (CO) do k -ésimo grupo ou tópico, considerando seu conjunto de descritores, é dada pela soma do valor de NPMI de todas as combinações de pares de termos da lista de descritores. Assim, a medida CO_NPMI é

definida por:

$$CO_NPMI(C_k) = \sum_{j=2}^N \sum_{i=1}^{j-1} \frac{\log(\frac{P(t_j, t_i)}{P(t_i)*p(t_j)})}{-\log(P(t_i, t_j))}$$

A NPMI foi proposta visando reduzir o *bias* da medida PMI para termos pouco frequentes, e ainda os seus valores resultantes estão no intervalo [-1,1]. A medida PMI (Pointwise Mutual Information) havia sido indicada em trabalhos anteriores para avaliação automática de tópicos (Newman et al., 2010). Deste modo, a comparação entre resultados obtidos pelos diferentes métodos de extração de tópicos pode ser mais facilmente obtida. Quanto maior o valor da medida, melhor é a coerência do tópico.

Ainda de acordo com Newman et al. (2010), uma vez que algumas dimensões produzidas pelos métodos podem não ser representativas, já que não foi utilizada nenhuma heurística para determinar um valor ideal de K para os métodos comparados, considerou-se que os tópicos no quartil superior são os mais interessantes. Assim, foram selecionados para comparação um total de 25% dos grupos ou tópicos que foram melhores avaliados pela medida de Coerência Observada.

A avaliação é realizada de forma automática utilizando a metodologia proposta por Lau et al. (2014). A metodologia propõe o uso de uma coleção de referência para calcular a coocorrência entre os termos selecionados como descritores necessária para o cálculo da NPMI. Neste projeto de doutorado, o *corpus* de referência utilizado foi a Wikipédia em inglês. Todos os artigos da coleção de referência devem ser pré-processados seguindo o mesmo procedimento utilizado para os documentos das coleções avaliadas. Para contabilizar as coocorrências necessárias para os cálculos, considera-se os termos que ocorreram em um mesmo artigo, independente da distância entre eles.

4.4 Considerações Finais

Neste capítulo foi apresentado o método proposto LARCM para representação de documentos e redução de dimensionalidade. A ideia geral do método é explorar as relações existentes nos termos dos documentos utilizando regras de associação e, ao agrupar as regras de associação extraídas, obter novas dimensões que representam os tópicos dos documentos. Ainda, foi apresentada a metodologia de avaliação proposta que permite selecionar modelos que maximizam tanto os resultados na tarefa de classificação de textos quanto os resultados de interpretabilidade dos tópicos obtidos. Com essa metodologia de avaliação, é possível comparar diretamente os modelos de extração de tópicos probabilísticos e não-probabilísticos em um mesmo contexto, o que nem sempre é possível com as medidas de avaliação adotada na literatura.

Avaliação do Modelo Proposto

Neste capítulo, apresentam-se e discutem-se os resultados obtidos com a avaliação experimental do modelo LARCM para redução de dimensionalidade em coleções textuais, bem como os recursos e configurações utilizados para essa avaliação. Utilizou-se a metodologia de avaliação proposta neste trabalho, apresentada na Seção 4.3.

Os experimentos realizados visam demonstrar a viabilidade do modelo proposto LARCM para extração de tópicos considerando os objetivos apresentados no Capítulo 1. O objetivo principal da proposta foi desenvolver um modelo que deve (i) possuir dimensionalidade reduzida, (ii) reduzir o esforço computacional e (iii) maximizar a interpretabilidade das dimensões obtidas quando comparado com modelos considerados estado da arte na área de extração de tópicos. O modelo proposto foi comparado com o modelo *Latent Dirichlet Allocation* (LDA), amplamente utilizado para extração de tópicos. Na primeira avaliação (Seção 5.2), os modelos de extração de tópicos são utilizados como estratégias de redução de dimensionalidade, e a nova representação da coleção de documentos obtida é aplicada na tarefa de classificação de documentos. Com isso, avalia-se se as representações obtidas mantêm características importantes para os processos de extração de conhecimento, permitindo alcançar resultados semelhantes ou superiores aos obtidos pelas representações bem estabelecidas na literatura. Uma vez que a qualidade dos descritores de cada dimensão obtida tem impacto no processo, a segunda avaliação (Seção 5.3) visa demonstrar a interpretabilidade geral de cada modelo de tópico. Um estudo de caso na tarefa de sistemas de recomendação sensíveis ao contexto foi realizado (Seção 5.4) para investigar se o modelo LARCM, quando comparado ao modelo LDA, é viável para outros tipos de processos de extração de conhecimento.

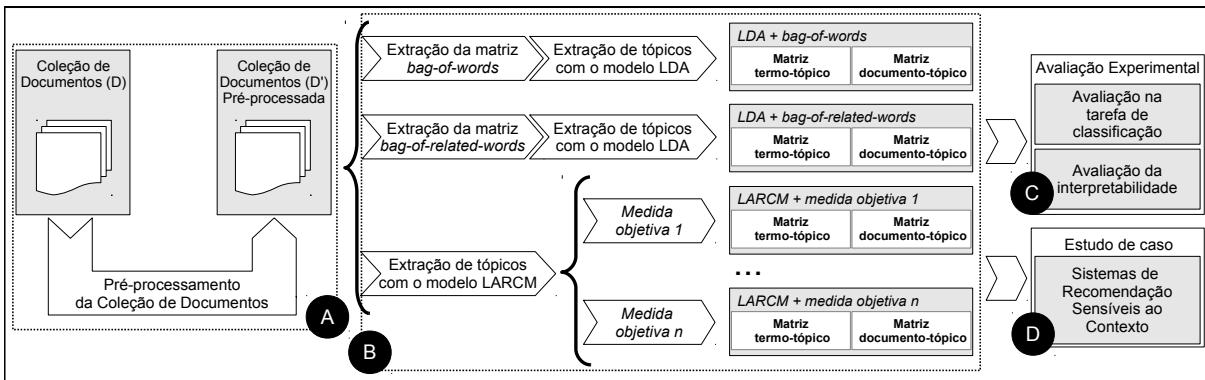


Figura 5.1: Visão geral do processo experimental deste trabalho de doutorado.

5.1 Preparação para a Avaliação Experimental

O desenho do processo experimental para avaliar o modelo LARCM está ilustrado na Figura 5.1. Nas seções seguintes são descritas as etapas do processo experimental, referentes à proposta deste doutorando para extração de tópicos baseado em agrupamento de regras de associação. Na Seção 5.1.1, as coleções de documentos utilizadas na avaliação são detalhadas e é descrito o pré-processamento aplicado nessas coleções - Figura 5.1 - A. Na Seção 5.1.2, são apresentadas as configurações de parâmetros utilizadas neste trabalho para extração dos tópicos utilizando o modelo LDA combinado com as representações *bag-of-words* e *bag-of-related-words*, e para o modelo LARCM - Figura 5.1 - B. Os tópicos obtidos desses modelos são então avaliados. Nas Seções 5.1.3 e 5.1.4 são apresentadas as configurações utilizadas para realizar as avaliações conforme a metodologia de avaliação proposta na Seção 4.3, e os resultados são discutidos nas Seções 5.2 e 5.3 - Figura 5.1 - C. Por fim, na Seção 5.4 é apresentado o estudo de caso da aplicação em sistemas de recomendação sensíveis ao contexto - Figura 5.1 - D.

No Apêndice A é apresentada uma coleção de documentos de exemplo que será utilizada para ilustrar as saídas esperadas dos principais processos executados nas próximas seções. A coleção de documentos inicial é apresentada na Tabela A.1. As saídas esperadas de cada etapa são apresentadas ao longo do Apêndice A.

5.1.1 Seleção de Corpora e Pré-Processamento

Para os experimentos, utilizaram-se três coleções de documentos em inglês¹. Um sumário das principais características das coleções de documentos utilizadas nas avaliações é fornecido na Tabela 5.1. A coleção Re8 é composta por textos jornalísticos, os quais representam uma porção significativa dos tipos de documentos digitais disponíveis. A coleção ACM² é composta por textos de artigos científicos completos, que também representam documentos digitais de grande interesse.

Todas as bases de textos foram pré-processadas utilizando o mesmo método. Foi

¹Estas coleções podem ser acessadas em http://sites.labic.icmc.usp.br/text_collections/.

²Association for Computing Machinery - <http://dl.acm.org/>.

Tabela 5.1: Descrição das coleções de textos utilizadas nos experimentos.

Coleção	# classes	# docs	# termos	Descrição
ACM-1	5	399	40918	Coleção de artigos científicos de conferências de diferentes áreas da computação extraídos do repositório digital da ACM
ACM-2	5	410	47907	
ACM-3	5	416	40181	
ACM-4	5	394	53474	
ACM-5	5	471	40732	
ACM-6	5	437	54088	
ACM-7	5	469	55015	
ACM-8	5	495	50486	
Re8	8	7674	17335	Coleção de notícias das 8 classes mais frequentes da coleção Reuters-21578

realizada uma padronização dos textos, removendo-se números, símbolos e também as *stopwords*. Os termos foram obtidos pela redução das palavras ao seu *stem* aplicando-se o algoritmo de Porter (Porter, 1997; Nogueira et al., 2008). Uma vez que os métodos de extração de tópicos visam a redução da dimensionalidade, não foi aplicado nenhum processo de seleção de atributos, como cortes de termos mais frequentes. Na coleção de documentos de exemplo do Apêndice A, os documentos originais, apresentados na Tabela A.1 foram pré-processados conforme o procedimento descrito aqui e o resultado final é apresentado na Tabela A.2.

5.1.2 Configuração dos Modelos para Extração dos Tópicos

Para comparar a abordagem LARCM com o modelo produzido pelo LDA considerando a *bag-of-words* tradicional como entrada e com o modelo produzido pelo LDA considerando a *bag-of-related-words* como entrada, foram selecionadas todas as medidas objetivas utilizadas na avaliação de regras de associação apresentadas na Seção 2.3.3. O uso da *bag-of-related-words* como entrada para o modelo LDA permite comparar o impacto do uso de termos compostos em relação ao modelo tradicional do LDA. A extração dos termos compostos antes da execução do LDA é apontado por Lau et al. (2013) como uma alternativa viável para inclusão de dependência de termos no processo e que obtém bons resultados.

O modelo *bag-of-related-words* foi obtido utilizando a ferramenta FEATuRE³. Com base nos resultados apresentados por Rossi and Rezende (2011), optou-se pelo uso do suporte automático e pelo valor médio da medida objetiva para selecionar os atributos. De acordo com os resultados por Rossi and Rezende (2011), foi utilizada a medida objetiva *Kappa* para construir a representação *bag-of-related-words* para essa avaliação. Para a construção das transações, adotou-se a janela de tamanho 5, que apresentou os melhores resultados segundo Rossi and Rezende (2011).

Como é sugerido por Liu (2011), em geral, são necessárias apenas regras com 2 ou 3 termos, uma vez que termos compostos por mais de 3 termos simples são pouco frequentes. Assim, optou-se pela geração de regras de associação com, no máximo, três itens. Por fim,

³Disponível em <http://sites.labic.icmc.usp.br/feature/>.

Tabela 5.2: Valores dos parâmetros utilizados nos modelos avaliados para extração de tópicos.

LDA	
Representação de entrada	<i>bag-of-words</i>
	<i>bag-of-related-words</i> (medida objetiva: <i>kappa</i> / tamanho da janela: 5)
Hiperparâmetros α e β	estimados automaticamente pela ferramenta
Iterações para a amostragem de Gibbs	1000
Número de tópicos (k)	50, 100, 150
LARCM	
Tamanho da janela	5
Suporte/confiança mínimos por documento	supmin Automático (Equação 4.1) / confmin = 0
Medidas objetivas	Added Value, Certainty Factor, Collective Strength, Confiança, Conviction, ϕ -Coefficient, Gini Index, IS, J-Measure, Kappa, Klosken, Lambda, Laplace, Lift, Mutual Information LHS, Novelty, Odds Ratio
Valor de corte para a medida objetiva por documento	Automático (Equação 4.2)
Algoritmo de agrupamento	Bi-Secting K-Means com medida de similaridade de cosseno
Número de grupos (k)	50, 100, 150

as regras de associação avaliadas aqui podem apresentar alguns itens no antecedente mas apenas um item no consequente devido às otimizações de implementação do algoritmo para geração de regras de associação⁴.

Conforme apresentado na Tabela 5.2, para avaliar as descrições obtidas, foram utilizados os valores de $k = 50$, $k = 100$ e $k = 150$ tanto para o LDA quanto para o LARCM, como utilizado em Chang et al. (2009). Ainda, os modelos LDA foram obtidos utilizando a ferramenta MALLET⁵. Para cada tópico, foram selecionados os dez termos com maior probabilidade para formar o conjunto de descritores. Para a abordagem proposta, as transações foram obtidas utilizando uma janela de tamanho 5, mesmo valor utilizado para obter a representação *bag-of-related-words*. Um exemplo da saída esperada para a obtenção das transações é apresentado no Apêndice A. Para cada documento pré-processado apresentado na Tabela A.2 é gerado seu conjunto de transações correspondentes, apresentados nas Tabelas A.3, A.4 e A.5.

As regras de associação foram obtidas para cada documento da coleção, e um subconjunto dessas regras de associação foram selecionadas utilizando uma das medidas objetivas apresentadas na Tabela 5.2, conforme o processo apresentado na Seção 4.2.1. Detalhes sobre as características das medidas objetivas utilizadas e suas propriedades podem ser encontrados na Seção 2.3.3. Para o agrupamento das regras, foi aplicado o algoritmo *Bi-Secting K-Means*⁶, e a medida de similaridade utilizada foi a cosseno. Para cada grupo, foram selecionadas as dez melhores regras, de acordo com a medida objetiva utilizada no passo anterior, para formar o conjunto de descritores. Para efeito de comparação, os itens da regra foram unidos de forma a apresentar um formato semelhante aos de outros modelos. Por exemplo, a regra *inteligencia* \Rightarrow *artificial* forma o termo “*inteligencia_artificial*”. Ainda, é possível obter a regra *artificial* \Rightarrow *inteligencia* que formará

⁴Disponível em <http://www.borgelt.net/apriori.html>.

⁵Disponível em <http://mallet.cs.umass.edu/>.

⁶Implementado na ferramenta Cluto. Disponível em <http://glaros.dtc.umn.edu/gkhome/views/cluto>.

o termo “artificial_inteligencia” e será tratado de forma independente do termo anterior.

5.1.3 Configurações para a Avaliação da Tarefa de Classificação

Para que a avaliação dos modelos de extração de tópicos para a tarefa de classificação não seja influenciada pelo desempenho dos classificadores, nesta pesquisa, foram utilizados algoritmos tradicionais de classificação que se baseiam em diferentes paradigmas (Sebastiani, 2002; Aggarwal and Zhai, 2012). Os algoritmos são: i) *Naive Bayes* e *Multinomial Naive Bayes* (paradigma probabilístico); ii) J48⁷ (paradigma simbólico); iii) SMO⁸ (paradigma estatístico); e iv) IBk⁹ (paradigma baseado em instâncias).

A implementação desses métodos está disponível na ferramenta Weka¹⁰ (Hall et al., 2009), cujo módulo *Experimenter* da versão 3.7.10 foi utilizada para realização dos experimentos de avaliação dos modelos de extração de tópicos. Para todos os algoritmos, foram adotados os parâmetros padrão que são sugeridos pelo Weka. Os algoritmos SMO e IBk são bastante sensíveis à inicialização de seus parâmetros (Batista and Silva, 2009; Braga, 2014), e por isso optou-se por executar esses algoritmos com algumas variações de parâmetros. Para o algoritmo SMO, foram executados experimentos utilizando o parâmetro de generalização “c” com os valores 1 (padrão da ferramenta Weka) e 10, escolhidos empiricamente com base na análise dos resultados de classificação obtidos por Rossi et al. (2013) em diversas coleções de documentos. Para o algoritmo IBk, seguindo as indicações de Batista and Silva (2009), os valores de k foram definidos em 3, 5 e 7. Os algoritmos foram executados e avaliados conforme o processo descrito na Seção 4.3.1, e a acurácia média de cada modelo de extração de tópicos foi obtida considerando os possíveis efeitos de cada algoritmo de classificação e das coleções no processo.

5.1.4 Configurações para a Avaliação da Interpretabilidade

Como foi apresentado na Seção 4.3.2, a avaliação é realizada de forma automática utilizando a metodologia proposta por Lau et al. (2014), disponível na ferramenta *topic-interpretability*¹¹. O *corpus* de referência utilizado foi a Wikipédia em inglês, versão extraída em 15 de janeiro de 2015. Todos os artigos do *corpus* de referência foram pré-processados utilizando o mesmo processo apresentado na Seção 5.1.1. A ferramenta foi configurada para considerar termos que ocorreram em um mesmo artigo, independente da distância entre eles.

Após o cálculo da medida Coerência Observada de cada tópico obtido, para cada modelo, seleciona-se o conjunto dos tópicos melhor avaliados tal que esse conjunto corresponda a 25% do total de tópicos extraídos para o modelo. Por exemplo, para um modelo obtido com valor de k igual a 100 tópicos, o conjunto será formado pelos 25 tópicos com

⁷A implementação do algoritmo de árvore de decisão C4.5 no Weka.

⁸A implementação do algoritmo SVM do Weka.

⁹A implementação do KNN do Weka.

¹⁰Disponível em <http://www.cs.waikato.ac.nz/ml/weka/>

¹¹Disponível em https://github.com/jhlau/topic_interpretability/

melhor valor de Coerência Observada. Desse conjunto, os tópicos com maior valor e o com menor valor da medida são considerados para avaliação.

5.2 Resultados e Discussão da Avaliação do LARCM na Tarefa de Classificação

O principal objetivo dessa avaliação é verificar o desempenho do LARCM quando comparado ao LDA quando aplicados como métodos de redução de dimensionalidade para construção de uma nova representação da coleção de documentos. Os tópicos obtidos pelos modelos devem ter boa interpretabilidade sem sacrificar a capacidade de extração de conhecimento. Como a tarefa de classificação é um dos problemas mais fundamentais na literatura de mineração de dados e textos, é importante que qualquer representação da coleção de documentos proposta seja capaz de melhorar os resultados obtidos por outras representações bem estabelecidas, ou pelo menos se aproximar desses resultados, oferecendo outras vantagens indiretas.

Tabela 5.3: Comparação múltipla das médias da acurácia dos classificadores para cada modelo de redução de dimensionalidade.

(gl=4068, $\widehat{V}(E)=12,5626$, $\alpha=0,05$)		
Método	acurácia	grupo
LARCM + Odds Ratio	84.900	<i>a</i>
LDA + <i>bag-of-words</i>	83.680	<i>b</i>
LARCM + ϕ -Coefficient	81.133	<i>c</i>
LARCM + Novelty	80.554	<i>c d</i>
LARCM + Gini Index	80.302	<i>c d e</i>
LARCM + Certainty Factor	80.263	<i>c d e</i>
LARCM + Confiança	80.231	<i>c d e</i>
LARCM + Laplace	80.055	<i>d e</i>
LARCM + Kappa	79.917	<i>d e</i>
LARCM + Added Value	79.840	<i>d e</i>
LARCM + Klosgen	79.811	<i>d e</i>
LARCM + J-Measure	79.623	<i>d e</i>
LARCM + Collective Strength	79.533	<i>d e</i>
LARCM + Mutual Information LHS	79.469	<i>d e</i>
LARCM + IS	79.457	<i>d e</i>
LARCM + Lift	79.209	<i>e</i>
LDA + <i>bag-of-related-words</i>	78.378	<i>f</i>
LARCM + Lambda	77.894	<i>f</i>
LARCM + Conviction	68.830	<i>g</i>

Na Tabela 5.3 são apresentados os resultados da acurácia média de cada modelo de extração de tópicos para a tarefa de classificação, como descrito na Seção 4.3.1. O valor de acurácia apresentado na tabela corresponde à média obtida considerando todos os algoritmos de classificação, o valor referente à quantidade de tópicos obtidos k , e coleções de documentos avaliadas para cada modelo de extração de tópicos. Essa média é calculada pela decomposição de todos os fatores envolvidos no processo utilizando o modelo linear apresentado também na Seção 4.3.1. Por ser calculada com esse processo, a média obtida já contabiliza o possível efeito de viés do algoritmo de classificação, do impacto das características das coleções nas representações e dos possíveis efeitos da escolha da quantidade de tópicos. Na tabela, a primeira coluna identifica o modelo avaliado e a segunda refere-se à média como descrita anteriormente. A terceira coluna é obtida pelo

resultado do teste estatístico aplicado. Modelos que estão em um mesmo grupo indicado por cada uma das letras (a , b , c , d , ...) não apresentam diferença estatística significativa, e portanto, considera-se que apresentam desempenho semelhante. Quando os modelos estão relacionados com letras diferentes nessa última coluna, indica que houve diferença estatística significativa entre eles e, nesse caso, há indícios para identificar qual modelo é melhor para a tarefa.

De acordo com a avaliação realizada, o modelo proposto LARCM combinado com a medida objetiva *Odds Ratio* obteve um valor médio de acurácia, considerando todos os efeitos, superior a todos os outros modelos com diferença estatística significativa. A medida *Odds Ratio* avalia o grau de associação entre o antecedente e o consequente da regra de associação. Dessa forma, favorece a extração de termos que são fortemente dependentes. As medidas objetivas com melhores resultados na sequência, grupo “c”, apresentam uma semântica semelhante a da medida objetiva *Odds Ratio*, com exceção da medida objetiva Confiança. Ou seja, as medidas objetivas com desempenho semelhantes, do grupo “c”, dão preferência por selecionar regras de associação nas quais o antecedente e o consequente possuem alta dependência ou correlação. A maioria das medidas objetivas avaliadas apresentam comportamento semelhante no impacto do resultado de classificação, considerando que a maioria das medidas estão no grupo grupo “d”, isto é, não existe diferença estatística significativa entre elas.

Para as representações obtidas com o modelo LDA combinado com a *bag-of-words* e *bag-of-related-words*, a representação mais tradicional que utiliza como entrada a *bag-of-words* obteve os melhores resultados de acurácia, sendo melhor com diferença estatística significativa em relação a quase todas as outras combinações, isto é, é o único modelo no grupo “b” na Tabela 5.3. Ainda, o modelo cuja entrada foi a *bag-of-related-words* não teve um bom desempenho para a tarefa de classificação quando comparado os outros modelos. Nesses experimentos, a adição de termos relacionados no modelo LDA parece não trazer tantos benefícios, isto é, a adição da informação de relação local entre os termos não trouxe um impacto positivo no processo. Esse resultado fornece alguns indícios de que, para que as técnicas tenham ganhos significativos com o uso de termos compostos, elas devem utilizar de uma forma explícita a informação de que existe dependência entre os termos durante a identificação dos tópicos da coleção.

Na Figura 5.2 são mostrados os valores obtidos de acurácia média de para cada modelo de extração de tópicos avaliados separados por coleção de documentos. Os valores de acurácia média e os resultados da análise de variância são apresentados na Seção B.1. Pode-se observar que o modelo LARCM combinado com a maioria das medidas objetivas parece apresentar um comportamento semelhante, ou seja, não apresentam grande diferença no valor médio de acurácia. O uso da medida *Conviction* com o modelo LARCM apresentou resultados significativamente inferiores aos obtidos pelos outros modelos. Destaca-se aqui os resultados obtidos pelo modelo LARCM combinado com a medida objetiva *Odds Ratio*, que foi superior aos outros modelos em 5 das 8 coleções avaliadas. Na coleção de documentos ACM-5, não houve diferença estatística significativa entre o modelo LARCM combinado com a medida objetiva *Odds Ratio* e o modelo LDA combinado com a *bag-of-*

words. Também observa-se que o modelo LDA combinado com a *bag-of-related-words* não apresentou desempenho expressivo na avaliação, sendo em alguns casos superior apenas ao resultado obtido pelo modelo LARCM combinado com a medida objetiva *Conviction*. Por fim, destaca-se também o desempenho do modelo LARCM combinado com a medida objetiva ϕ -*Coefficient*, que obteve acurácia média superior ao modelo LDA combinado com a *bag-of-words* em 2 coleções com diferença estatística significativa (Tabelas B.2 e B.7), e desempenho semelhante em outras 2 coleções (Tabelas B.3 e B.4).

5.3 Resultados e Discussão da Avaliação da Interpretabilidade

Nessa avaliação, o objetivo é verificar se o modelo LARCM produz tópicos com melhor interpretabilidade quando comparados aos obtidos com o LDA. Tópicos com melhor interpretabilidade são aqueles que facilitam a interação do usuário com a coleção de documentos de maneira exploratória. O processo de avaliação foi aplicado como descrito na Seção 4.3.2. A avaliação da interpretabilidade de um tópico é um processo bastante subjetivo, que está relacionado com o grupo de usuários que estiver avaliando o processo. Na metodologia adotada neste trabalho, a avaliação automática aplicada visa simular o usuário e seu processo subjetivo, pois a quantidade de tópicos obtidos é tal que inviabiliza o trabalho manual mesmo quando um grupo de voluntário está disposto a participar.

Nas Tabelas 5.4 e 5.5 são apresentados os valores de Coerência Observada para o melhor tópico e o menor valor da medida para o tópico do quartil superior. Nessa avaliação não foi possível apontar uma medida como sendo a mais interessante para todos os casos. De maneira geral, o modelo LARCM obteve tópicos com melhor valor de CO comparado com os modelos LDA em 17 dos 27 casos avaliados (3 valores de k e 9 conjuntos de textos). Nos 10 casos em que o LDA tem resultados de melhor CO superiores, o modelo proposto tende a ficar com valores de melhor CO muito próximos. Entretanto, quando o modelo proposto é superior ao LDA, tende a obter valores de melhor CO significativamente superiores. O método LARCM combinado com a medida objetiva Lift foi o modelo que obteve a maior diferença para o melhor valor de CO comparado com o melhor tópico obtido pelo LDA, na coleção ACM-2 com $k=150$. Ainda, para o valor de Coerência Observada em 25%, o modelo LDA foi ligeiramente superior aos resultados obtidos pelas melhores medidas objetivas com a abordagem proposta, mas não são suficientes para compensar a diferença de resultados para o melhor tópico avaliado. Em geral, as medidas objetivas que visam regras de associação com maior cobertura das transações parecem ser mais apropriadas para a seleção dos descritores.

Também na avaliação de Coerência Observada, considerando apenas os resultados obtidos pelo modelo LDA, a representação que utiliza como entrada a *bag-of-words* foi superior ao modelo que utiliza a *bag-of-related-words* comparando o melhor valor de CO. Entretanto, a diferença aqui não é tão significativa quanto na avaliação da tarefa de classificação. Ainda, considerando o valor de Coerência Observada em 25%, o modelo que utiliza a *bag-of-related-words* teve resultados semelhantes ao LDA tradicional. Entretanto, esse resultado também aponta para poucos ganhos com o uso de termos compostos como

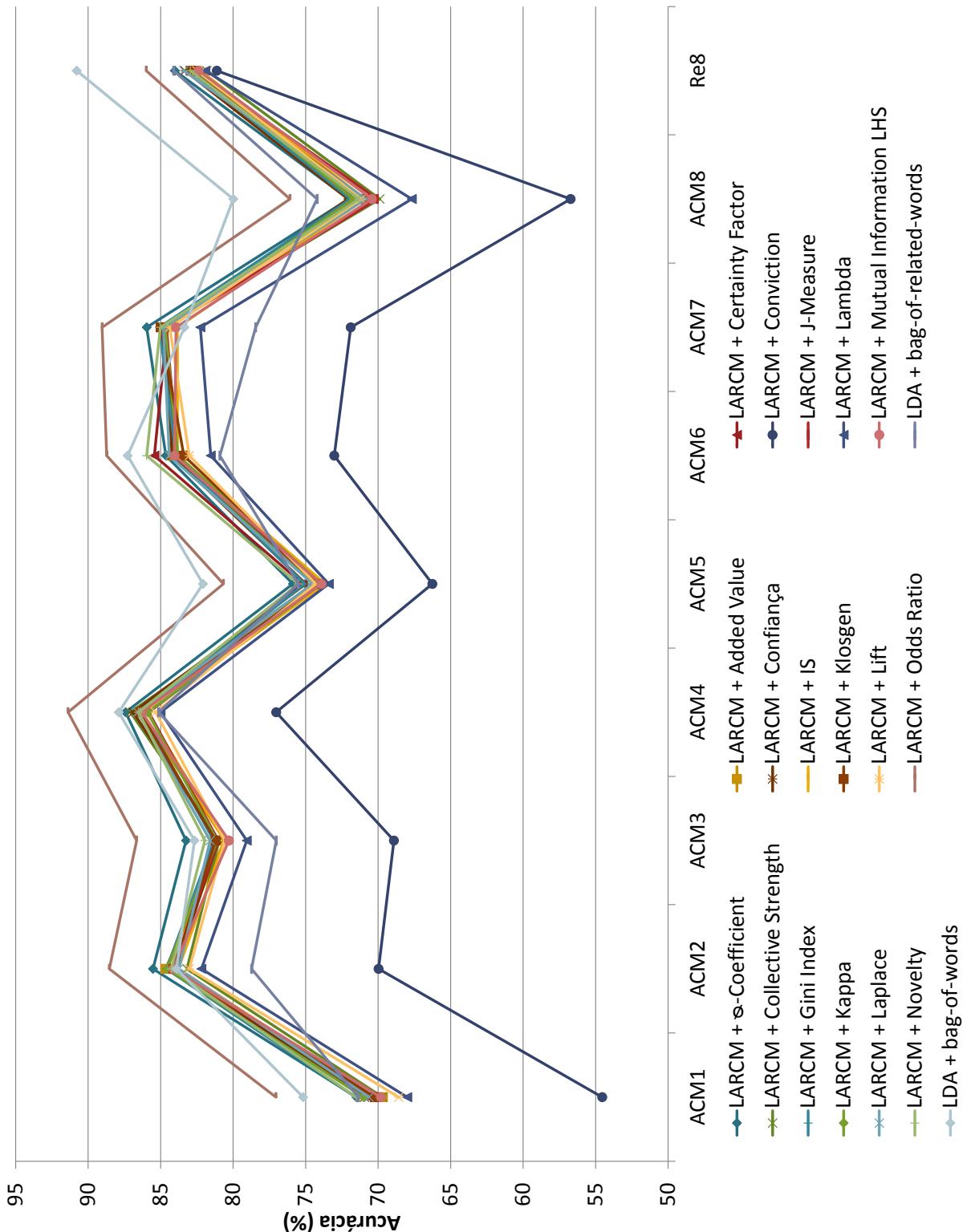


Figura 5.2: Resultados da acurácia média dos classificadores de cada modelo separados por coleção de documentos. O gráfico foi construído a partir dos valores apresentados na Seção B.1.

entrada para o processo do LDA, uma vez que o custo para obter essas representações pode ser muito superior ao da obtenção da representação *bag-of-words*.

Tabela 5.4: Resultados para as coleções ACM-1, ACM-2, ACM-3 e ACM-4 do valor de Coerência Observada (CO) do tópico com melhor avaliação e do tópico com menor valor de medida entre os 25% melhores tópicos. Os resultados estão ordenados pelo melhor valor de CO obtido. São exibidos os 3 melhores resultados para o LARCM considerando o melhor valor de CO e os 2 modelos obtidos com o LDA.

ACM-1					
K=50			K=100		
Configuração	Melhor CO	CO em 25%	Configuração	Melhor CO	CO em 25%
LDA + <i>bag-of-words</i>	0.35	0.27	LDA + <i>bag-of-words</i>	0.36	0.26
LDA + <i>bag-of-related-words</i>	0.32	0.26	LDA + <i>bag-of-related-words</i>	0.35	0.26
LARCM + IS	0.27	0.19	LARCM + Lift	0.33	0.12
LARCM + Mutual Information LHS	0.26	0.21	LARCM + Kappa	0.31	0.09
LARCM + J-Measure	0.25	0.20	LARCM + Collective Strength	0.30	0.11
ACM-2					
K=50			K=100		
Configuração	Melhor CO	CO em 25%	Configuração	Melhor CO	CO em 25%
LARCM + Kappa	0.38	0.10	LARCM + Added Value	0.48	0.11
LDA + <i>bag-of-words</i>	0.34	0.28	LARCM + Lift	0.44	0.15
LDA + <i>bag-of-related-words</i>	0.34	0.28	LARCM + Certainty Factor	0.43	0.12
LARCM + Klosgen	0.29	0.09	LDA + <i>bag-of-words</i>	0.36	0.27
LARCM + J-Measure	0.29	0.22	LDA + <i>bag-of-related-words</i>	0.34	0.26
ACM-3					
K=50			K=100		
Configuração	Melhor CO	CO em 25%	Configuração	Melhor CO	CO em 25%
LDA + <i>bag-of-words</i>	0.34	0.27	LARCM + IS	0.37	0.17
LARCM + Mutual Information LHS	0.34	0.20	LARCM + Kappa	0.36	0.11
LARCM + Lift	0.31	0.08	LARCM + Klosgen	0.36	0.10
LDA + <i>bag-of-related-words</i>	0.30	0.27	LDA + <i>bag-of-words</i>	0.34	0.26
LARCM + J-Measure	0.29	0.19	LDA + <i>bag-of-related-words</i>	0.33	0.26
ACM-4					
K=50			K=100		
Configuração	Melhor CO	CO em 25%	Configuração	Melhor CO	CO em 25%
LARCM + Conviction	0.38	0.12	LARCM + Conviction	0.37	0.13
LARCM + Kappa	0.36	0.08	LARCM + Kappa	0.36	0.10
LDA + <i>bag-of-words</i>	0.33	0.26	LDA + <i>bag-of-words</i>	0.33	0.25
LARCM + Collective Strength	0.33	0.10	LARCM + Collective Strength	0.33	0.11
LDA + <i>bag-of-related-words</i>	0.31	0.26	LDA + <i>bag-of-related-words</i>	0.31	0.25
K=150					
Configuração	Melhor CO	CO em 25%	Configuração	Melhor CO	CO em 25%
LARCM + Mutual Information LHS	0.40	0.17	LARCM + Lambda	0.46	0.07
LDA + <i>bag-of-related-words</i>	0.35	0.26	LARCM + Kappa	0.36	0.11
LDA + <i>bag-of-words</i>	0.35	0.25	LDA + <i>bag-of-words</i>	0.34	0.25
LARCM + Certainty Factor	0.35	0.10	LARCM + Collective Strength	0.33	0.11
LARCM + Novelty	0.33	0.11	LDA + <i>bag-of-related-words</i>	0.32	0.25

Tabela 5.5: Resultados para as coleções ACM-5, ACM-6, ACM-7, ACM-8 e Re8 do valor de Coerência Observada (CO) do tópico com melhor avaliação e do tópico com menor valor de medida entre os 25% melhores tópicos. Os resultados estão ordenados pelo melhor valor de CO obtido. São exibidos os 3 melhores resultados para o LARCM considerando o melhor valor de CO e os 2 modelos obtidos com o LDA.

ACM-5				
K=50			K=100	
Configuração	Melhor CO	CO em 25%	Configuração	Melhor CO
LARCM + Gini Index	0.39	0.09	LARCM + Kappa	0.43
LDA + <i>bag-of-words</i>	0.34	0.27	LARCM + Added Value	0.41
LARCM + J-Measure	0.34	0.20	LARCM + Collective Strength	0.39
LDA + <i>bag-of-related-words</i>	0.33	0.28	LDA + <i>bag-of-related-words</i>	0.35
LARCM + Novelty	0.33	0.08	LDA + <i>bag-of-words</i>	0.34
K=150				
K=150				
Configuração	Melhor CO	CO em 25%		
LARCM + Added Value	0.41	0.10		
LARCM + ϕ -Coefficient	0.40	0.08		
LARCM + Klosgen	0.39	0.11		
LDA + <i>bag-of-related-words</i>	0.36	0.25		
LDA + <i>bag-of-words</i>	0.35	0.26		
ACM-6				
K=50			K=100	
Configuração	Melhor CO	CO em 25%	Configuração	Melhor CO
LARCM + Confiança	0.38	0.05	LARCM + Certainty Factor	0.40
LDA + <i>bag-of-words</i>	0.36	0.28	LARCM + Laplace	0.38
LDA + <i>bag-of-related-words</i>	0.34	0.27	LARCM + Lift	0.36
LARCM + IS	0.33	0.21	LDA + <i>bag-of-words</i>	0.33
LARCM + Gini Index	0.33	0.08	LDA + <i>bag-of-related-words</i>	0.32
K=150				
K=150				
Configuração	Melhor CO	CO em 25%		
LARCM + IS	0.42	0.15		
LARCM + Collective Strength	0.42	0.11		
LARCM + Gini Index	0.42	0.11		
LDA + <i>bag-of-words</i>	0.35	0.26		
LDA + <i>bag-of-related-words</i>	0.35	0.25		
ACM-7				
K=50			K=100	
Configuração	Melhor CO	CO em 25%	Configuração	Melhor CO
LDA + <i>bag-of-related-words</i>	0.34	0.27	LARCM + Odds Ratio	0.41
LDA + <i>bag-of-words</i>	0.33	0.27	LARCM + Conviction	0.36
LARCM + IS	0.29	0.19	LDA + <i>bag-of-words</i>	0.33
LARCM + Added Value	0.28	0.05	LDA + <i>bag-of-related-words</i>	0.31
LARCM + Gini Index	0.27	0.08	LARCM + Novelty	0.30
K=150				
K=150				
Configuração	Melhor CO	CO em 25%		
LARCM + Conviction	0.36	0.10		
LARCM + Novelty	0.36	0.10		
LARCM + Klosgen	0.35	0.09		
LDA + <i>bag-of-words</i>	0.33	0.25		
LDA + <i>bag-of-related-words</i>	0.31	0.25		
ACM-8				
K=50			K=100	
Configuração	Melhor CO	CO em 25%	Configuração	Melhor CO
LDA + <i>bag-of-words</i>	0.33	0.27	LDA + <i>bag-of-words</i>	0.34
LDA + <i>bag-of-related-words</i>	0.31	0.26	LDA + <i>bag-of-related-words</i>	0.33
LARCM + J-Measure	0.29	0.19	LARCM + Mutual Information LHS	0.31
LARCM + Collective Strength	0.28	0.07	LARCM + Collective Strength	0.28
LARCM + IS	0.27	0.18	LARCM + ϕ -Coefficient	0.28
K=150				
K=150				
Configuração	Melhor CO	CO em 25%		
LDA + <i>bag-of-words</i>	0.32	0.25		
LDA + <i>bag-of-related-words</i>	0.32	0.25		
LARCM + ϕ -Coefficient	0.31	0.06		
LARCM + J-Measure	0.28	0.16		
LARCM + Lift	0.28	0.11		
Re8				
K=50			K=100	
Configuração	Melhor CO	CO em 25%	Configuração	Melhor CO
LDA + <i>bag-of-related-words</i>	0.30	0.24	LDA + <i>bag-of-related-words</i>	0.31
LARCM + Collective Strength	0.29	0.15	LDA + <i>bag-of-words</i>	0.30
LARCM + Novelty	0.29	0.15	LARCM + Added Value	0.28
LARCM + Odds Ratio	0.29	0.09	LARCM + Collective Strength	0.28
LDA + <i>bag-of-words</i>	0.28	0.22	LARCM + Certainty Factor	0.26
K=150				
K=150				
Configuração	Melhor CO	CO em 25%		
LARCM + Collective Strength	0.44	0.16		
LARCM + Added Value	0.35	0.16		
LARCM + Novelty	0.34	0.16		
LDA + <i>bag-of-words</i>	0.32	0.24		
LDA + <i>bag-of-related-words</i>	0.31	0.22		

5.4 Estudo de Caso: Sistemas de Recomendação Sensíveis ao Contexto

Uma das grandes vantagens do modelo LDA é seu sucesso em diferentes tipos de aplicações e também em diferentes tipos de dados. Trabalhos nas áreas de processamento de imagens e som, por exemplo, são beneficiados com a redução de dimensionalidade obtida por esse modelo. Assim, é interessante que o modelo proposto LARCM também apresente bons resultados em diferentes cenários de aplicação. Para entender o comportamento do modelo proposto LARCM em outras aplicações, foi realizado o estudo de caso no qual os tópicos obtidos pelo modelo LARCM foram utilizados como informação de contexto para sistemas de recomendação sensíveis ao contexto. Uma breve revisão sobre sistemas de recomendação sensíveis ao contexto e os algoritmos utilizados nesta avaliação é apresentada no Apêndice D.

Em seu trabalho, os autores de Sundermann et al. (2014) propõe o uso de hierarquias de tópicos para extração de contexto de sistemas de recomendação sensíveis ao contexto visando melhorar a recomendação de páginas web baseada tanto na informação de *logs* de acessos dos usuários quanto no conteúdo textual das páginas acessadas. Uma visão geral do processo de extração de contexto para obter as recomendações é apresentada na Figura 5.3. Após o tratamento adequado da base de *logs* de acesso e aquisição do conteúdo das páginas web, aplica-se o processo de extração de contexto. Neste processo, consideram-se as informações contextuais de uma página web os tópicos associados a esse documento. Documentos que estão associados a um mesmo tópico estão relacionados a um mesmo contexto. Essas informações contextuais são processadas por um sistema de recomendação sensível ao contexto, que gera uma lista das páginas web que podem ser de maior interesse do usuário.

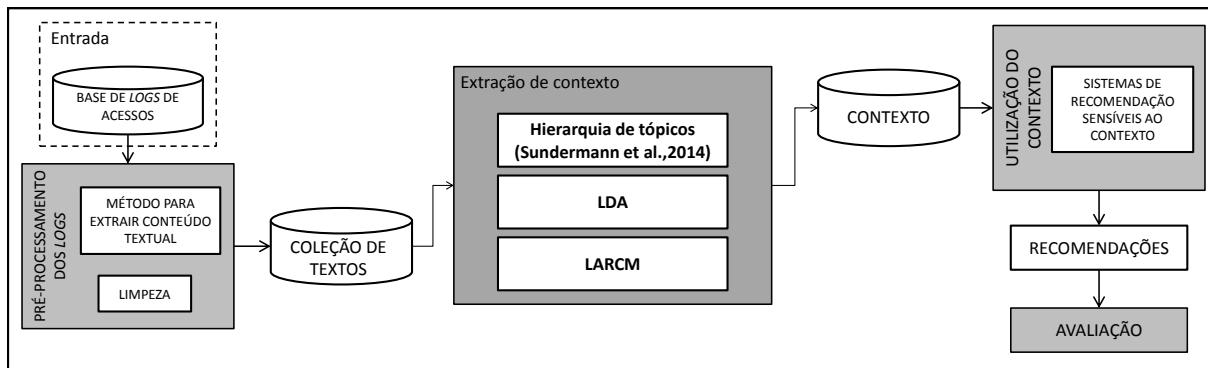


Figura 5.3: Etapas do processo de extração de contexto. Adaptado de: Sundermann (2015)

Foram utilizados neste trabalho os modelos de extração de tópicos LDA e LARCM. Os tópicos foram obtidos pelo modelo LARCM combinados com as medidas objetivas, e pelo modelo LDA cujas entradas são as representações *bag-of-words* e a *bag-of-related-words*. As configurações dos parâmetros utilizados para extração destes tópicos são as mesmas

apresentadas na Tabela 5.2, utilizadas nos experimentos anteriores. Os modelos foram obtidos da coleção de páginas Web do site da Agência Embrapa de Informação Tecnológica¹² que foram acessadas no período de novembro de 2012, utilizada nas avaliações em Sundermann et al. (2014). A coleção contém 1633 páginas da Web, que foram acessadas por 4659 usuários que totalizaram 83729 acessos.

Segundo o processo de avaliação proposto por Sundermann et al. (2014), foram avaliados 3 algoritmos de recomendação sensíveis ao contexto, *cReduction*, *weightPoF* e *filterPoF*, que foram comparados com o algoritmo de recomendação *Item-Based Collaborative Filtering* (IBCF). Para avaliar os sistemas de recomendação foi utilizado o protocolo *All But One* com avaliação cruzada de *10-folds* e a métrica *Mean Average Precision* (MAP). Para comparação, utiliza-se a variação da métrica MAP@N, em que N é a quantidade de recomendações utilizadas para calcular a média. Foram utilizados os valores de N 5 e 10. Seleciona-se na partição de teste uma sequência de acessos de um visitante, isto é, as páginas que um visitante acessou em um determinado período de acesso. O sistema de recomendação recebe a informação que relaciona um usuário e uma das páginas web que ele acessou, e retorna uma lista com as N páginas web recomendadas. Uma recomendação é considerada positiva se, nessa lista de páginas web recomendadas, encontra-se uma das outras páginas web acessadas pelo usuário na mesma seção.

Resultados e discussão

Os resultados da avaliação para a MAP@5 e MAP@10 para cada um dos modelos obtidos são apresentados nas Tabelas 5.6, 5.7 e 5.8. Os melhores resultados para cada algoritmo estão com asterisco (“*”). Nota-se, analisando todos os resultados, que a extração de tópicos apresentam resultados estatisticamente melhores que o baseline IBCF para quase todos os casos dos algoritmos *cReduction* e *weightPoF*. Já para o algoritmo *filterPoF*, os resultados não são tão estáveis com a variação dos parâmetros dos modelos de extração de tópicos. Nota-se, também, que todos os resultados obtidos pelo modelo LARCM com todas as medidas objetivas avaliadas foram superiores aos obtidos pelo modelo LDA.

Com relação aos resultados obtidos pelo modelo proposto LARCM, os melhores resultados para todos os algoritmos de recomendação foram obtidos com a medida objetiva ϕ -Coefficient e com o valor de K igual a 150. A medida ϕ -Coefficient também foi a melhor avaliada na comparação de Coerência Observada apresentada anteriormente, e se mostrou competitiva na avaliação de classificação de documentos. A medida Odds Ratio, que foi a melhor avaliada na tarefa de classificação, não apresentou resultados expressivos na avaliação de sistemas de recomendação sensíveis ao contexto. Observando os resultados para cada valor de K independentemente, para K igual a 50 destaca-se a medida Odds Ratio para o algoritmo *filterPoF*, e a medida Kappa para os outros algoritmos. Para K igual a 100, a medida Mutual Information LHS obteve o melhor valor para a medida objetiva em todos os algoritmos. Nos dois casos, para K igual a 50 e 100, os

¹²<http://www.agencia.cnptia.embrapa.br/>

Tabela 5.6: Comparação dos resultados obtidos pelos algoritmos de recomendação sensíveis ao contexto comparados com o algoritmo IBCF. O contexto foi obtido pela extração de tópicos com o modelo LARCM e o modelo LDA, utilizando o valor de K igual a 50.

	MAP @ 5				MAP @ 10			
	IBCF	cReduction	weightPoF	filterPoF	IBCF	cReduction	weightPoF	filterPoF
LARCM + ϕ -Coefficient	0,2991	0,3727	0,3933	0,1095	0,3089	0,3875	0,4061	0,1142
LARCM + Added Value	0,2991	0,3826	0,3916	0,1405	0,3089	0,3944	0,4033	0,1437
LARCM + Certainty Factor	0,2991	0,3670	0,3937	0,1415	0,3089	0,3819	0,4049	0,1445
LARCM + Collective Strength	0,2991	0,3845	0,3947	0,1401	0,3089	0,3967	0,4059	0,1448
LARCM + Confiança	0,2991	0,3745	0,3880	0,1382	0,3089	0,3884	0,3992	0,1417
LARCM + Conviction	0,2991	0,3723	0,3789	0,1035	0,3089	0,3876	0,3932	0,1093
LARCM + Gini Index	0,2991	0,3714	0,3837	0,1053	0,3089	0,3852	0,3965	0,1104
LARCM + IS	0,2991	0,3821	0,3881	0,1380	0,3089	0,3958	0,4013	0,1414
LARCM + J-Measure	0,2991	0,3658	0,3762	0,1206	0,3089	0,3797	0,3898	0,1242
LARCM + Kappa	0,2991	0,3976*	0,4149*	0,1332	0,3089	0,4069*	0,4249*	0,1382
LARCM + Klogen	0,2991	0,3700	0,3890	0,1249	0,3089	0,3799	0,3995	0,1284
LARCM + Lambda	0,2991	0,3932	0,4019	0,1589	0,3089	0,4059	0,4143	0,1617
LARCM + Laplace	0,2991	0,3827	0,3855	0,1109	0,3089	0,3958	0,3954	0,1138
LARCM + Lift	0,2991	0,3841	0,3941	0,1478	0,3089	0,3967	0,4058	0,1529
LARCM + Mutual Information	0,2991	0,3893	0,3966	0,1335	0,3089	0,3989	0,4064	0,1374
LARCM + Novelty	0,2991	0,3826	0,3894	0,1143	0,3089	0,3904	0,4009	0,1188
LARCM + Odds Ratio	0,2991	0,3702	0,4046	0,1641*	0,3089	0,3933	0,4150	0,1660*
LDA + bag-of-words	0,2991	0,2979	0,3053	0,0378	0,3089	0,3076	0,3150	0,0394
LDA + bag-of-related-words	0,2991	0,3780	0,3903	0,1138	0,3089	0,3933	0,4017	0,1178

Tabela 5.7: Comparação dos resultados obtidos pelos algoritmos de recomendação sensíveis ao contexto comparados com o algoritmo IBCF. O contexto foi obtido pela extração de tópicos com o modelo LARCM e o modelo LDA, utilizando o valor de $K=100$.

	MAP @ 5				MAP @ 10			
	IBCF	cReduction	weightPoF	filterPoF	IBCF	cReduction	weightPoF	filterPoF
LARCM + ϕ -Coefficient	0,2991	0,4752	0,4899	0,1699	0,3089	0,4821	0,4962	0,1742
LARCM + Added Value	0,2991	0,4302	0,4749	0,3892	0,3089	0,4379	0,4789	0,3892
LARCM + Certainty Factor	0,2991	0,3724	0,4312	0,4066	0,3089	0,3912	0,4433	0,4069
LARCM + Collective Strength	0,2991	0,4908	0,5044	0,1889	0,3089	0,4939	0,5067	0,1950
LARCM + Confiança	0,2991	0,4231	0,4580	0,4013	0,3089	0,4328	0,4612	0,4033
LARCM + Conviction	0,2991	0,3740	0,3812	0,1199	0,3089	0,3890	0,3965	0,1247
LARCM + Gini Index	0,2991	0,5045	0,5199	0,3881	0,3089	0,5093	0,5297	0,3911
LARCM + IS	0,2991	0,3530	0,4163	0,2606	0,3089	0,3776	0,4288	0,2633
LARCM + J-Measure	0,2991	0,3940	0,4348	0,2195	0,3089	0,4173	0,4482	0,2201
LARCM + Kappa	0,2991	0,4845	0,5038	0,2721	0,3089	0,4916	0,5088	0,2762
LARCM + Klogen	0,2991	0,3851	0,4132	0,1792	0,3089	0,4033	0,4250	0,1822
LARCM + Lambda	0,2991	0,4909	0,5270	0,3671	0,3089	0,4923	0,5291	0,3704
LARCM + Laplace	0,2991	0,4121	0,4277	0,1459	0,3089	0,4219	0,4372	0,1493
LARCM + Lift	0,2991	0,4269	0,4721	0,3595	0,3089	0,4351	0,4798	0,3619
LARCM + Mutual Information	0,2991	0,5085*	0,5301*	0,4497*	0,3089	0,5112*	0,5325*	0,4522*
LARCM + Novelty	0,2991	0,3934	0,4119	0,2315	0,3089	0,4015	0,4238	0,2342
LARCM + Odds Ratio	0,2991	-	0,4429	0,2236	0,3089	-	0,4493	0,2257
LDA + bag-of-words	0,2991	0,2987	0,3076	0,0477	0,3089	0,3014	0,3176	0,0492
LDA + bag-of-related-words	0,2991	0,3783	0,3874	0,1193	0,3089	0,3889	0,3987	0,1234

valores obtidos para a medida objetiva ϕ -Coefficient estão próximos dos melhores valores obtidos para os algoritmos cReduction e weightPoF. Assim como nos resultados obtidos pelos autores Sundermann et al. (2014), o algoritmo filterPoF parece ser o mais sensível a variações de parâmetros quando comparado aos outros algoritmos que apresentam uma menor diferença entre o menor valor e o maior valor obtido de MAP para cada valor de K .

Observando todos os resultados apresentados, a informação contextual obtida pela aplicação dos modelos de extração de tópicos contribuem para um melhor resultado de recomendação. Quando comparados os melhores resultados obtidos pelos modelos LDA e LARCM com os melhores valores obtidos por Sundermann et al. (2014), apresentados na Tabela 5.9, observa-se também uma melhora significativa nos resultados obtidos pelo LARCM, dando indícios de que o modelo proposto pode ser aplicados em outras tarefas

Tabela 5.8: Comparação dos resultados obtidos pelos algoritmos de recomendação sensíveis ao contexto comparados com o algoritmo IBCF. O contexto foi obtido pela extração de tópicos com o modelo LARCM e o modelo LDA, utilizando o valor de K igual a K=150.

	MAP @ 5				MAP @ 10			
	IBCF	cReduction	weightPoF	filterPoF	IBCF	cReduction	weightPoF	filterPoF
LARCM + ϕ -Coefficient	0,2991	0,6309*	0,6506*	0,6115*	0,3089	0,6324*	0,6522*	0,6120*
LARCM + Added Value	0,2991	0,4890	0,5147	0,3991	0,3089	0,4777	0,5184	0,4003
LARCM + Certainty Factor	0,2991	-	0,4703	0,4170	0,3089	-	0,4744	0,4206
LARCM + Collective Strength	0,2991	0,4897	0,5324	0,4088	0,3089	0,4927	0,5360	0,4097
LARCM + Confiança	0,2991	-	0,5584	0,4420	0,3089	-	0,5619	0,4427
LARCM + Conviction	0,2991	0,3562	0,3626	0,1060	0,3089	0,3713	0,3779	0,1106
LARCM + Gini Index	0,2991	0,5269	0,5450	0,5081	0,3089	0,5332	0,5469	0,5091
LARCM + IS	0,2991	0,4846	0,5446	0,2854	0,3089	0,4902	0,5487	0,2892
LARCM + J-Measure	0,2991	-	0,6486	0,5274	0,3089	-	0,6498	0,5282
LARCM + Kappa	0,2991	0,5903	0,6037	0,5753	0,3089	0,5928	0,6062	0,5758
LARCM + Klosgen	0,2991	-	0,5722	0,4209	0,3089	-	0,5748	0,4209
LARCM + Lambda	0,2991	-	0,5024	0,4045	0,3089	-	0,5078	0,4081
LARCM + Laplace	0,2991	-	0,6405	0,3752	0,3089	-	0,6432	0,3768
LARCM + Lift	0,2991	0,3746	0,4504	0,2565	0,3089	0,3729	0,4519	0,2605
LARCM + Mutual Information	0,2991	0,5286	0,5697	0,4992	0,3089	0,5336	0,5725	0,5015
LARCM + Novelty	0,2991	0,5208	0,5544	0,2522	0,3089	0,5313	0,5580	0,2535
LARCM + Odds Ratio	0,2991	0,3846	0,4461	0,3109	0,3089	0,3973	0,4517	0,3122
LDA + <i>bag-of-words</i>	0,2991	0,3046	0,3117	0,0563	0,3089	0,3151	0,3218	0,0577
LDA + <i>bag-of-related-words</i>	0,2991	0,3687	0,3860	0,1181	0,3089	0,3868	0,3965	0,1222

Tabela 5.9: Comparação dos melhores resultados obtidos pelos algoritmos de recomendação sensíveis ao contexto considerando todas as variações de parâmetros dos modelos. Foram selecionados os melhores valores obtidos pelo modelo LDA combinado com *bag-of-words* e *bag-of-related-words*, pelo modelo LARCM e os obtidos no artigo original de Sundermann et al. (2014).

	MAP @ 5			MAP @ 10		
	cReduction	weightPoF	filterPoF	cReduction	weightPoF	filterPoF
LARCM + ϕ -Coefficient	0,6308	0,6506	0,6115	0,6324	0,6522	0,6120
LDA + <i>bag-of-words</i>	0,3046	0,3117	0,0563	0,3151	0,3218	0,0577
LDA + <i>bag-of-related-words</i>	0,3783	0,3903	0,1193	0,3933	0,4017	0,1234
Melhor avaliação em Sundermann et al. (2014)	0,4216	0,4456	0,4165	0,4274	0,4508	0,4180

com sucesso. Para o modelo LARCM, O algoritmo *filterPoF* se mostrou mais sensível a mudança da medida objetiva, porém, existem algumas configurações em que os resultados obtidos com uso de informação de contexto foram melhores que os obtidos pelo algoritmo IBCF, que não utiliza informação de contexto. Os resultados obtidos com os algoritmos *cReduction* e *weightPoF* com a informação de contexto fornecida pelo LARCM foram sempre melhores que os do algoritmo IBCF.

5.5 Considerações Finais

Neste capítulo foi apresentada a avaliação experimental do modelo proposto LARCM, comparando seu desempenho com o modelo de extração de tópicos LDA, utilizando a metodologia de avaliação apresentada na Seção 4.3. Os resultados experimentais evidenciam que o modelo LARCM é competitivo quando comparado com os resultados obtidos pelo LDA, sendo viável sua aplicação como modelo de extração de tópicos e para redução de dimensionalidade. Na tarefa de classificação de textos, o modelo LARCM utilizando a medida objetiva *Odds Ratio* superou com diferença estatística significativa os resultados obtidos pelo modelo LDA utilizando a entrada *bag-of-words*, considerando o valor médio de acurácia obtido em diferentes algoritmos de classificação e coleções de documentos. Os

resultados obtidos pelo modelo LARCM utilizando as medidas objetivas avaliadas, com exceção da medida objetiva *Odds Ratio*, foram próximos aos obtidos pelo LDA utilizando a entrada *bag-of-words*, destacando-se as medidas objetivas ϕ -*Coefficient*, *Novelty*, *Gini Index*, *Certainty Factor* e *Confiança*. Na avaliação da interpretabilidade dos tópicos obtidos, o modelo LARCM superou o modelo LDA em 17 dos 27 casos avaliados, e apresentou resultados próximos ao do LDA nos outros 10 casos. Nesta avaliação, observou-se que as medidas objetivas para o modelo LARCM melhor avaliadas variam de acordo com os parâmetros como o número de tópicos extraídos k e da coleção. No estudo de caso, o modelo LARCM apresentou resultados significativamente melhores que os obtidos pelo LDA, destacando-se os resultados obtidos pelo LARCM utilizando a medida objetiva ϕ -*Coefficient*. De maneira geral, as medidas objetivas *Odds Ratio* e ϕ -*Coefficient* parecem ser as configurações mais indicadas para aplicação com o modelo proposto LARCM.

No capítulo a seguir, são apresentadas as conclusões e as principais contribuições alcançadas com o desenvolvimento deste trabalho, bem como sugestões de trabalhos futuros.

Conclusões e Trabalhos Futuros

Neste trabalho foi proposto o modelo não-supervisionado para extração de tópicos LARCM baseado no agrupamento de regras de associação, explorando explicitamente a dependência entre os termos para obter uma representação dos documentos com boa interpretabilidade dos tópicos extraídos e com baixa dimensionalidade. A proposta do modelo LARCM foi motivada pela hipótese principal desta tese de doutorado de que *a extração de tópicos obtidos com o agrupamento de regras de associação extraídas dos documentos viabiliza a construção de uma representação dos documentos textuais que possui baixa dimensionalidade, melhora a interpretabilidade das novas dimensões e leva a uma melhor extração de conhecimento nas tarefas de classificação de textos e organização da informação.* Em linhas específicas, buscou-se com o modelo proposto confirmar as hipóteses levantadas no Capítulo 1. O modelo LARCM se relaciona com as hipóteses levantadas da seguinte maneira:

Existem termos que ocorrem juntos (um imediatamente após o outro) frequentemente, formando sequências de termos, ou expressões multi-palavras, que expressam um significado ou conceito diferente daquele expresso por cada termo individualmente, sendo essa sequência de termos denominada relação local do termo.

Essa informação de relação local do termo é obtida pelo modelo pela extração de regras de associação de cada documento, conforme o processo descrito na Seção 4.2.1.

Documentos de uma coleção que expressam um mesmo tópico levam à escolha de sequências de termos semelhantes para expressá-los, e estes termos influenciam na escolha dos termos da sua vizinhança, sendo essa influência na escolha dos termos denominada relação geral dos termos.

Essa informação de relação geral do termo é obtida pelo agrupamento das regras de associação extraídas no contexto local da relação dos termos. Para esse agrupamento,

foi proposto neste trabalho uma representação intermediária das regras de associação extraídas em uma **matriz regra de associação-termos**, que permite identificar de forma eficiente a vizinhança de cada termo e utilizar essa informação para obter os grupos, conforme o processo apresentado na Seção 4.2.2.

Essas informações de relação geral e local dos termos pode ser extraída diretamente dos documentos de uma coleção.

Como foi apresentado anteriormente, é possível obter as informações da relação geral e local dos termos diretamente dos documentos de uma coleção. Neste trabalho, a informação local do termo corresponde às regras de associação obtidas, e a relação geral dos termos corresponde aos grupos formados por essas regras de associação.

Os tópicos que deram origem a um documento estão em um espaço latente, que pode ser estimado com a utilização das informações de relação local e geral dos termos da coleção.

No modelo LARCM, os grupos obtidos da relação geral dos termos correspondem aos tópicos que deram origem aos documentos da coleção. Os documentos são associados aos tópicos pela quantidade de regras de associação deste documento presente em cada grupo, como apresentado na Seção 4.2.3.

Existe uma representação estruturada dos tópicos encontrados que uma avaliação objetiva e/ou subjetiva possa indicar como melhor quando comparada com a representação obtida pelas técnicas consideradas estado da arte para extração de tópicos de documentos textuais.

Para avaliação do modelo LARCM foi proposto a metodologia de avaliação apresentada na Seção 4.3, que permite comparar diferentes modelos de extração de tópicos, probabilísticos ou não-probabilísticos, em um mesmo cenário. Essa metodologia considera a contribuição dos tópicos obtidos para a tarefa de classificação de textos e quanto a interpretabilidade desses tópicos. Essas avaliações dão suporte a escolha do modelo mais adequado conforme a tarefa de interesse para aplicação dos tópicos obtidos.

O LARCM foi comparado com o modelo LDA tradicional e o modelo LDA utilizando uma representação que inclui termos compostos (*bag-of-related-words*), ambas estado da arte na área. Os experimentos, apresentados no Capítulo 5, indicam que o modelo proposto produz uma representação para os documentos que contribui significativamente para a melhora dos resultados na tarefa de classificação de textos, mantendo também uma boa interpretabilidade dos tópicos obtidos. O modelo LARCM também apresentou ótimo desempenho quando utilizado para extração de informação de contexto para aplicação em sistemas de recomendação sensíveis ao contexto.

Como contribuições deste trabalho, a primeira consiste na proposta e desenvolvimento do modelo de extração de tópicos não-probabilístico LARCM. O modelo se mostrou viável

tanto na redução de dimensionalidade quanto na extração de tópicos com boa interpretabilidade.

A segunda contribuição é a proposta da representação das regras de associação extraídas dos documentos em uma matriz regra de associação-termo. Essa representação armazena informações sobre a vizinhança dos termos e, indiretamente, mantém parte da informação estrutural do documento.

A terceira contribuição é a metodologia de avaliação proposta. Com essa metodologia, é possível comparar modelos de extração de tópicos probabilísticos e não-probabilísticos em um mesmo contexto, avaliando a qualidade dos tópicos obtidos de forma objetiva com a tarefa de classificação e de forma subjetiva com a avaliação da Coerência Observada. Este último processo é realizado de forma automática, o que possibilita a aplicação em cenários em que as coleções de textos disponíveis sejam enormes, e consequentemente gerando uma grande quantidade de tópicos extraídos, ou que essas coleções sejam dinâmicas, sofrendo alterações significativas em um curto espaço de tempo.

Durante a realização deste trabalho foram publicados os seguintes artigos:

- Santos, F. F., Rezende, S.O, Carvalho, V. O. Identificando o Assunto dos Documentos em Coleções Textuais Utilizando Termos Compostos In: XI Encontro Nacional de Inteligência Artificial e Computacional (ENIAC 2014), 2014, São Carlos. Proceedings of XI Encontro Nacional de Inteligência Artificial e Computacional (ENIAC 2014). 2014. p.550 - 557
- Carvalho, V. O., dos Santos, F. F., Rezende, S.O. Metrics to Support the Evaluation of Association Rule Clustering In: 15th International Conference - DaWaK, 2013, Praga. Proceedings of Data Warehousing and Knowledge Discovery - 15th International Conference, DaWaK 2013. Berlin: Springer-Verlag Berlin Heidelberg, 2013. v.8057. p.248 - 259
- Padua, R., dos Santos, F. F., da Silva Conrado, M., Carvalho, V. O., Rezende, S.O. Subjective Evaluation of Labeling Methods for Association Rule Clustering In: 12th Mexican International Conference on Artificial Intelligence - MICAI, 2013, Cidade do México. Proceedings of Advances in Soft Computing and Its Applications - 12th Mexican International Conference on Artificial Intelligence, MICAI Part II. Berlin: Springer-Verlag Berlin Heidelberg, 2013. v.8266. p.289 - 300
- Carvalho, V. O., Santos, F. F., Rezende, S.O, Padua, R. PAR-COM: A New Methodology for Post-processing Association Rules. Lecture Notes in Business Information Processing. v.102, p.66 - 80, 2012.
- Carvalho, V. O., Biondi, D.S., Santos, F. F., Rezende, S.O. Labeling methods for Association Rule Clustering In: 14th International Conference on Enterprise Information Systems (ICEIS), 2012, Wroclaw. Proceedings of the 14th International Conference on Enterprise Information Systems. SciTePress, 2012. v.1. p.105 - 109

- Carvalho, V. O., Santos, F. F., Rezende, S.O. Post-processing association rules with clustering and objective measures In: 13th International Conference on Enterprise Information Systems (ICEIS), 2011, Pequim. Proceedings of the 13th International Conference on Enterprise Information Systems. SciTePress, 2011. v.1. p.54 - 63

Um trabalho relacionado ao processo de avaliação de tópicos proposto neste trabalho, aplicado na área de hierarquias de tópicos, está em processo de revisão:

Moura, M.F., Santos, F. F., Rezende, S.O An experimental comparison of label selection methods for hierarchical document clusters. p. 1 - 49. Submetido ao: ACM Transactions on Information Systems

Além disso, dois trabalhos estão em fase de preparação para serem submetidos a revistas da área. O primeiro trabalho abrange as principais contribuições obtidas com a tese, incluindo o modelo de extração de tópicos, a metodologia de avaliação e os resultados obtidos no contexto de sistemas de recomendação. O segundo trabalho contempla a investigação da viabilidade de aplicação do modelo LARCM em dados não-estruturados no domínio de séries temporais de arquivos de música para classificação de gênero musical.

Santos, F. F., Carvalho, V. O., Moura, M.F., Rezende, S.O. Finding Topics with Latent Association Rule Cluster based Model. Em preparação.

Santos, F. F., Silva, D. F., Rezende, S.O. Using Latent Association Rule Cluster based Model to Extract Topics for Music Classification. Em preparação.

Uma limitação deste trabalho está relacionada a análises comparativas em cenários distintos. Os modelos de extração de tópicos são também aplicados em áreas como processamento de imagem e de séries temporais com bastante sucesso. Neste trabalho de doutorado, o foco esteve no uso dos tópicos para processos de mineração de textos. Assim, é necessário investir esforços em realizar experimentos que permitam uma avaliação do modelo LARCM com relação a sua aplicação em dados não-textuais. Embora oneroso, estes experimentos podem reforçar ainda mais a utilizada do modelo proposto.

Na proposta do LARCM apresentada na Seção 4.2, a medida objetiva escolhida é utilizada em dois momentos distintos do processo. No primeiro, discutido na Seção 4.2.1, a medida objetiva é utilizada para selecionar as regras de associação que são potencialmente mais interessantes. No segundo momento, apresentado na Seção 4.2.3, a medida objetiva é utilizada para ordenar o conjunto de candidatos a descritores de cada um dos tópicos. Foi proposto neste trabalho o uso da mesma medida objetiva para ambos os processos. Entretanto, como esses processos são independentes, é interessante avaliar a possibilidade de escolhas de diferentes medidas objetivas e seu impacto na qualidade final dos tópicos extraídos. Por exemplo, a medida objetiva *Odds Ratio* teve um desempenho significativamente superior na tarefa de classificação de textos, porém não teve tanto destaque na avaliação da interpretabilidade dos tópicos.

Devido ao processo de extração de regras de associação utilizado pelo LARCM, o processamento de cada documento pode ser feito individualmente. Esta propriedade permite

que o modelo possa ser implementado de forma paralela e distribuída e sem necessidade de modificações significativas. Por exemplo, considerando o paradigma *MapReduce* (Dean and Ghemawat, 2008), seria possível realizar a distribuição tanto da extração das regras de associação quanto o agrupamento dessas regras entre os nós de um *cluster* de computadores na fase *map*, enquanto que a fase *reduce* realizaria a associação de cada documento com os seus possíveis tópicos.

Referências Bibliográficas

- Adomavicius, G., Sankaranarayanan, R., Sen, S., and Tuzhilin, A. (2005). Incorporating contextual information in recommender systems using a multidimensional approach. *ACM Trans. Inf. Syst.*, 23(1):103–145. Citado na página 128.
- Aggarwal, C. C. and Zhai, C., editors (2012). *Mining Text Data*. Springer. Citado nas páginas 1, 2, 7, 30, 33, 37, 39, 40, e 63.
- Agrawal, R., Imielinski, T., and Swami, A. N. (1993). Mining association rules between sets of items in large databases. In Buneman, P. and Jajodia, S., editors, *Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data*, pages 207–216, Washington, D.C. Citado nas páginas 15, 16, e 17.
- Agrawal, R. and Srikant, R. (1994). Fast algorithms for mining association rules in large databases. In *VLDB*, pages 487–499, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc. Citado nas páginas xxvii, 8, 15, 17, 18, 19, 20, e 48.
- Arora, S., Ge, R., and Moitra, A. (2012). Learning topic models - going beyond SVD. In *53rd Annual IEEE Symposium on Foundations of Computer Science, FOCS 2012, New Brunswick, NJ, USA, October 20-23, 2012*, pages 1–10. IEEE Computer Society. Citado na página 38.
- Baeza-Yates, R. and Ribeiro-Neto, B. (2011). *Modern Information Retrieval: The Concepts and Technology behind Search (2nd Edition) (ACM Press Books)*. Addison-Wesley Professional, 2 edition. Citado na página 34.
- Baeza-Yates, R. A. and Ribeiro-Neto, B. (1999). *Modern Information Retrieval*. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA. Citado nas páginas 30, 31, e 32.
- Batista, G. E. A. P. A. and Silva, D. F. (2009). How k-nearest neighbor parameters affect its performance. In *X Argentine Symposium on Artificial Intelligence (ASAII)*, pages 95–106, Mar del Plata, Argentina. Publicado em CD-ROM. Citado na página 63.
- Beil, F., Ester, M., and Xu, X. (2002). Frequent term-based text clustering. In *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and*

Data Mining, July 23-26, 2002, Edmonton, Alberta, Canada, pages 436–442. ACM.
Citado na página 2.

Berkhin, P. (2006). A survey of clustering data mining techniques. In Kogan, J., Nicholas, C., and Teboulle, M., editors, *Grouping Multidimensional Data*, chapter 2, pages 25–71. Springer-Verlag, Berlin, Heidelberg. Citado na página 10.

Billhardt, H., Borrajo, D., and Maojo, V. (2002). A context vector model for information retrieval. *J. Am. Soc. Inf. Sci. Technol.*, 53(3):236–249. Citado nas páginas 2, 34, e 45.

Blanchard, J., Guillet, F., Gras, R., and Briand, H. (2005). Using information-theoretic measures to assess association rule interestingness. In *Proceedings of the fifth IEEE International Conference on Data Mining, ICDM 2005*, pages 66–73. IEEE Computer Society. Citado na página 24.

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022. Citado nas páginas xix, 2, 35, 39, 40, e 45.

Braga, I. A. (2014). *Stochastic density ratio estimation and its application to feature selection*. PhD thesis, Instituto de Ciências Matemáticas e de Computação (ICMC) - USP - São Carlos, São Carlos, SP, Brasil. Citado na página 63.

Brauckhoff, D., Dimitropoulos, X., Wagner, A., and Salamatian, K. (2012). Anomaly extraction in backbone networks using association rules. *IEEE/ACM Trans. Netw.*, 20(6):1788–1799. Citado na página 15.

Brin, S., Motwani, R., Ullman, J. D., and Tsur, S. (1997). Dynamic itemset counting and implication rules for market basket data. In *SIGMOD '97: Proceedings of the 1997 ACM SIGMOD international conference on Management of data*, pages 255–264, New York, NY, USA. ACM. Citado na página 17.

Carvalho, V. O. (2007). *Generalização de regras de associação utilizando conhecimento de domínio e avaliação do conhecimento generalizado*. Doutorado em ciências da computação e matemática computacional, USP - São Carlos, São Carlos, SP, Brasil. Citado na página 20.

Carvalho, V. O., Santos, F. F., and Rezende, S. O. (2012). Agrupamento de regras de associação no pré-processamento e no pós-processamento: o que vale mais a pena? Technical Report 381, ICMC - USP - São Carlos. Disponível em: http://www.icmc.usp.br/CMS/Arquivos/arquivos_enviados/BIBLIOTECA_113_RT_381.pdf. [04/03/2015]. Citado nas páginas xix e 26.

Chang, J., Boyd-Graber, J. L., Gerrish, S., Wang, C., and Blei, D. M. (2009). Reading tea leaves: How humans interpret topic models. In *NIPS*, pages 288–296. Curran Associates, Inc. Citado nas páginas 3, 44, 55, e 62.

Cheng, X., Guo, J., Liu, S., Wang, Y., and Yan, X. (2013a). Learning topics in short texts by non-negative matrix factorization on term correlation matrix. In *Proceedings of the 13th SIAM International Conference on Data Mining, May 2-4, 2013. Austin, Texas, USA.*, pages 749–757. SIAM. Citado na página 37.

Cheng, X., Miao, D., Wang, C., and Cao, L. (2013b). Coupled term-term relation analysis for document clustering. In *IJCNN*, pages 1–8. IEEE. Citado nas páginas 2, 3, 34, 43, e 45.

Cutting, D. R., Karger, D. R., Pedersen, J. O., and Tukey, J. W. (1992). Scatter/gather: a cluster-based approach to browsing large document collections. In *SIGIR '92: Proceedings of the 15th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 318–329, New York, NY, USA. ACM. Citado na página 28.

Dean, J. and Ghemawat, S. (2008). Mapreduce: Simplified data processing on large clusters. *Commun. ACM*, 51(1):107–113. Citado na página 79.

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., and Harshman, R. (1990). Indexing by latent semantic analysis. *J. Am. Soc. Info. Sci.*, 41(6):391–407. Citado nas páginas 2, 35, 37, e 45.

Ding, C. H. Q., Li, T., and Peng, W. (2008). On the equivalence between non-negative matrix factorization and probabilistic latent semantic indexing. *Computational Statistics & Data Analysis*, 52(8):3913–3927. Citado na página 38.

Domingues, M. A. (2004). Generalização de regras de associação. Mestrado em ciências da computação e matemática computacional, USP - São Carlos, São Carlos, SP, Brasil. Citado na página 19.

Dong, H., Hussain, F., and Chang, E. (2008). A survey in traditional information retrieval models. *Digital Ecosystems and Technologies, 2008. DEST 2008. 2nd IEEE International Conference on*, pages 397–402. Citado na página 31.

Dragoni, M., da Costa Pereira, C., and Tettamanzi, A. G. (2012). A conceptual representation of documents and queries for information retrieval systems by using light ontologies. *Expert Systems with Applications*, 39(12):10376 – 10388. Citado na página 33.

Everitt, B. S., Landau, S., and Leese, M. (2001). *Cluster Analysis*. Arnold Publishers. Citado nas páginas 9 e 11.

Farahat, A. K. and Kamel, M. S. (2011). Statistical semantics for enhancing document clustering. *Knowl. Inf. Syst.*, 28(2):365–393. Citado nas páginas 2, 34, 36, 43, e 45.

Feldman, R. and Sanger, J. (2007). *The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data*. Cambridge University Press. Citado na página 7.

Fernández, M., Cantador, I., López, V., Vallet, D., Castells, P., and Motta, E. (2010). Semantically enhanced information retrieval: An ontology-based approach. *Journal of Web Semantics: Science, Services and Agents on the World Wide Web*, In Press, Corrected Proof:1 – 19. Citado na página 33.

Figueiredo, F., Rocha, L., Couto, T., Salles, T., Gonçalves, M. A., and Meira Jr., W. (2011). Word co-occurrence features for text classification. *Inf. Syst.*, 36(5):843–858. Citado nas páginas 2, 3, 34, 35, 43, 44, e 45.

Fung, B. C., Wang, K., and Ester, M. (2003). Hierarchical document clustering using frequent itemsets. In *Proceedings of the 3rd SIAM International Conference on Data Mining*, pages 59–70. Citado na página 28.

Gantz, J. F. and Reinsel, D. (2009). As the economy contracts, the digital universe expands. *External Publication of IDC (Analyse the Future) Information and Data*, pages 1–10. Citado na página 1.

Gao, Y., Xu, Y., Li, Y., and Liu, B. (2013). A two-stage approach for generating topic models. In *PAKDD*, volume 7819, pages 221–232. Springer. Citado nas páginas 2, 3, 36, 41, 43, e 45.

Glover, E. J., Pennock, D. M., Lawrence, S., and Krovetz, R. (2002). Inferring hierarchical descriptions. In *CIKM*, pages 507–514. ACM. Citado na página 28.

Guyon, I. and Elisseeff, A. (2003). An introduction to variable and feature selection. *Journal of Machine Learning Research*, 3:1157–1182. Citado nas páginas 3 e 36.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., and Witten, I. H. (2009). The weka data mining software: An update. *SIGKDD Explor. Newsl.*, 11(1):10–18. Citado na página 63.

Han, J. and Kamber, M. (2001). *Data mining concepts and techniques*. San Diego, CA: Academic. Citado na página 1.

Han, J., Pei, J., and Yin, Y. (2000). Mining frequent patterns without candidate generation. In *Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data*, SIGMOD '00, pages 1–12, New York, NY, USA. ACM. Citado na página 17.

Hava, O., Skrbek, M., and Kordík, P. (2013). Supervised two-step feature extraction for structured representation of text data. *Simulation Modelling Practice and Theory*, 33:132–143. Citado na página 36.

Hipp, J., Güntzer, U., and Nakhaeizadeh, G. (2002). data mining of association rules and the process of knowledge discovery in databases. In *Advances in Data Mining*, pages 15–36. Citado na página 15.

- Hofmann, T. (1999). Probabilistic latent semantic indexing. In *SIGIR*, pages 50–57, New York, NY, USA. ACM. Citado nas páginas 2, 35, 39, e 45.
- Houtsma, M. A. W. and Swami, A. N. (1995). Set-oriented mining for association rules in relational databases. In *ICDE '95: Proceedings of the Eleventh International Conference on Data Engineering*, pages 25–33, Washington, DC, USA. IEEE Computer Society. Citado na página 17.
- Hu, Y., Boyd-Graber, J., Satinoff, B., and Smith, A. (2014). Interactive topic modeling. *Machine Learning*, 95(3):423–469. Citado na página 35.
- Ingwersen, P. (1996). Cognitive perspectives of information retrieval interaction: elements of a cognitive ir theory. *Journal of Documentation*, 52:3–50. Citado na página 32.
- Jain, A. K., Murty, M. N., and Flynn, P. J. (1999). Data clustering: a review. *ACM Computing Surveys*, 31(3):264–323. Citado na página 10.
- Jorge, A. (2004). Hierarchical clustering for thematic browsing and summarization of large sets of association rules. In Berry, M. W., Dayal, U., Kamath, C., and Skillicorn, D., editors, *SIAM'04: Proceedings of the 4th SIAM International Conference on Data Mining*. 10p. Citado na página 27.
- Kalogeratos, A. and Likas, A. (2012). Text document clustering using global term context vectors. *Knowl. Inf. Syst.*, 31(3):455–474. Citado nas páginas 2, 34, 36, 43, e 45.
- Kang, B.-Y., Kim, D.-W., and Lee, S.-J. (2005). Exploiting concept clusters for content-based information retrieval. *Information Sciences*, 170(2-4):443–462. Citado na página 32.
- Kara, S., Özgür Alan, Sabuncu, O., Akpinar, S., Cicekli, N. K., and Alpaslan, F. N. (2012). An ontology-based retrieval system using semantic indexing. *Information Systems*, 37(4):294 – 305. <ce:title>Semantic Web Data Management</ce:title>. Citado na página 33.
- Kashyap, V., Ramakrishnan, C., Thomas, C., and Sheth, A. (2005). Taxaminer: an experimentation framework for automated taxonomy bootstrapping. *International Journal of Web and Grid Services*, 1(2):240–266. Citado na página 28.
- Keikha, M., Khonsari, A., and Orumchian, F. (2009). Rich document representation and classification: An analysis. *Knowledge-Based Systems*, 22(1):67–71. Citado nas páginas 2 e 43.
- Kim, H. D., Park, D. H., Lu, Y., and Zhai, C. (2012). Enriching text representation with frequent pattern mining for probabilistic topic modeling. *Proc. Am. Soc. Info. Sci. Tech.*, 49(1):1–10. Citado nas páginas 2, 36, 39, 40, 41, e 45.

Kraft, D. H., Pasi, G., and Bordogna, G. (2007). Vagueness and uncertainty in information retrieval: how can fuzzy sets help? In *Proceedings of the 2006 international workshop on Research issues in digital libraries*, IWRIDL '06, pages 3:1–3:10, New York, NY, USA. ACM. Citado na página 32.

Larsen, B. and Aone, C. (1999). Fast and effective text mining using linear-time document clustering. In *KDD '99: Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 16–22, New York, NY, USA. ACM. Citado na página 28.

Lau, J. H., Baldwin, T., and Newman, D. (2013). On collocations and topic models. *ACM Trans. Speech Lang. Process.*, 10(3):10:1–10:14. Citado nas páginas 2, 35, 44, e 61.

Lau, J. H., Newman, D., and Baldwin, T. (2014). Machine reading tea leaves: Automatically evaluating topic coherence and topic model quality. In *Proceedings of the European Chapter of the Association for Computational Linguistics*. Citado nas páginas 55, 57, 58, e 63.

Lee, D. D. and Seung, H. S. (1999). Learning the parts of objects by non-negative matrix factorization. *Nature*, 401(6755):788–791. Citado nas páginas 2, 35, 37, e 38.

Leite, M. A. A. (2009). *Modelo Fuzzy para Recuperação de Informação Utilizando Múltiplas Ontologias Relacionadas*. Tese de doutorado, Faculdade de Engenharia Elétrica e de Computação - Universidade Estadual de Campinas, Campinas, SP, Brasil. Citado na página 32.

Lewis, D. D. (1992). An evaluation of phrasal and clustered representations on a text categorization task. In Belkin, N. J., Ingwersen, P., and Pejtersen, A. M., editors, *Proceedings of the 15th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. Copenhagen, Denmark, June 21-24, 1992*, pages 37–50. ACM. Citado na página 2.

Liu, B. (2011). *Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data (Data-Centric Systems and Applications)*. Springer, 2nd edition. Citado nas páginas xix, 3, 30, 31, 38, 56, e 61.

Liu, X. and Croft, W. B. (2008). Evaluating text representations for retrieval of the best group of documents. In *Proceedings of the IR research, 30th European conference on Advances in information retrieval*, ECIR'08, pages 454–462, Berlin, Heidelberg. Springer-Verlag. Citado na página 32.

Lopes, A. A., Pinho, R., Paulovich, F. V., and Minghim, R. (2007). Visual text mining using association rules. *Computers & Graphics*, 31(3):316–326. Citado na página 28.

Manning, C. D., Raghavan, P., and Schütze, H. (2008). *Introduction to Information Retrieval*. Cambridge University Press, 1 edition. Citado nas páginas 8, 9, 12, e 31.

Marcacini, R. M., Correa, G. N., and Rezende, S. O. (2012a). An active learning approach to frequent itemset-based text clustering. In *ICPR*, pages 3529–3532. IEEE. Citado nas páginas 2 e 27.

Marcacini, R. M., Hruschka, E. R., and Rezende, S. O. (2012b). On the use of consensus clustering for incremental learning of topic hierarchies. In de Barros, L. N., Finger, M., Pozo, A. T. R., Lugo, G. A. G., and Castilho, M. A., editors, *Advances in Artificial Intelligence - SBIA 2012 - 21th Brazilian Symposium on Artificial Intelligence, Curitiba, Brazil, October 20-25, 2012. Proceedings*, volume 7589 of *Lecture Notes in Computer Science*, pages 112–121. Springer. Citado na página 12.

Martins, D. S. (2009). Uma abordagem para recuperação de informações sensível ao contexto usando retroalimentação implícita de relevância. Mestrado em ciências da computação, UFSCAR - Universidade Federal de São Carlos, São Carlos, SP, Brasil. Citado na página 31.

Mauá, D. D. (2009). Modelos de tópicos na classificação automática de resenhas de usuário. Mestrado em engenharia. Área de concentração: Engenharia de controle e automação mecânica, Escola Politécnica/USP - São Paulo, São Paulo, SP, Brasil. Citado nas páginas 39 e 40.

Melandi, E. A. (2004). *Pós-processamento de Regras de Associação*. Doutorado em ciências da computação e matemática computacional, USP - São Carlos, São Carlos, SP, Brasil. Citado na página 19.

Metz, J. (2006). Interpretação de clusters gerados por algoritmos de clustering hierárquico. Mestrado em ciências da computação e matemática computacional, USP - São Carlos, São Carlos, SP, Brasil. Citado nas páginas xix e 14.

Mitchell, T. M. (1997). *Machine Learning*. McGraw-Hill Science/Engineering/Math. Citado na página 8.

Monard, M. C. and Baranauskas, J. A. (2003). Conceitos sobre aprendizado de máquina. In Rezende, S. O., editor, *Sistemas Inteligentes: Fundamentos e Aplicações*, chapter 4, pages 89–114. Manole, 1 edition. Citado na página 8.

Moura, M. F. (2009). *Contribuições para a construção de taxonomias de tópicos em domínios restritos utilizando aprendizado estatístico*. Doutorado em ciências da computação e matemática computacional, USP - São Carlos, São Carlos, SP, Brasil. Citado na página 56.

Moura, M. F. and Rezende, S. O. (2010). A simple method for labeling hierarchical document clusters. In *Proceedings of AIA 2010 - Artificial Intelligence and Applications*, Innsbruck, Austria. Citado na página 28.

Newman, D., Lau, J. H., Grieser, K., and Baldwin, T. (2010). Automatic evaluation of topic coherence. In *HLT*, pages 100–108, Stroudsburg, PA, USA. ACL. Citado nas páginas 57 e 58.

Nogueira, B. M., Moura, M. F., Conrado, M. S., Rossi, R. G., Marcacini, R. M., and Rezende, S. O. (2008). Winning some of the document preprocessing challenges in a text mining process. In *SBBD*, page 10–18. Porto Alegre : SBC, Porto Alegre : SBC. Citado nas páginas 47, 61, e 93.

O’Callaghan, D., Greene, D., Carthy, J., and Cunningham, P. (2015). An analysis of the coherence of descriptors in topic modeling. *Expert Systems with Applications*, 42(13):5645 – 5657. Citado na página 37.

Panniello, U. and Gorgoglione, M. (2012). Incorporating context into recommender systems: An empirical comparison of context-based approaches. *Electronic Commerce Research*, 12(1):1–30. Citado na página 128.

Park, J. S., syan Chen, M., and Yu, P. S. (1997). Using a hash-based method with transaction trimming for mining association rules. *IEEE Transactions on Knowledge and Data Engineering*, 9:813–825. Citado na página 17.

Pei, J., Han, J., and Mao, R. (2000). Closet: An efficient algorithm for mining frequent closed itemsets. In *ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery*, pages 21–30. Citado na página 17.

Popescul, A. and Ungar, L. (2000). Automatic labeling of document clusters, unpublished manuscript. <http://www.cis.upenn.edu/~popescul/Publications/popescul00labeling.pdf>. Citado na página 28.

Porter, M. F. (1997). An algorithm for suffix stripping. *Readings in Information Retrieval*, pages 313–316. Citado nas páginas 61 e 93.

Pôssas, B., Ziviani, N., Meira, Jr., W., and Ribeiro-Neto, B. (2002). Set-based model: A new approach for information retrieval. In *SIGIR*, pages 230–237, New York, NY, USA. ACM. Citado nas páginas 2, 3, 34, 41, e 45.

Reynolds, A. P., Richards, G., de la Iglesia, B., and Rayward-Smith, V. J. (2006). Clustering rules: a comparison of partitioning and hierarchical clustering algorithms. *Journal of Mathematical Modelling and Algorithms*, 5(4):475–504. Citado nas páginas 26 e 27.

Rezende, S. O., Pugliesi, J. B., Melanda, E. A., and Paula, M. F. (2003). Mineração de dados. In Rezende, S. O., editor, *Sistemas Inteligentes: Fundamentos e Aplicações*, chapter 12, pages 307–335. Manole, 1 edition. Citado nas páginas xix e 8.

Rios, T. N. (2013). *Organização flexível de documentos*. Doutorado em ciências da computação e matemática computacional, USP - São Carlos, São Carlos, SP, Brasil. Citado nas páginas 1 e 2.

Rossi, R. G., Marcacini, R. M., and Rezende, S. O. (2013). Benchmarking text collections for classification and clustering tasks. Technical Report 395, ICMC - USP - São Carlos. Disponível em: http://www.icmc.usp.br/CMS/Arquivos/arquivos_enviados/BIBLIOTECA_113_RT_395.pdf. [04/03/2015]. Citado na página 63.

- Rossi, R. G. and Rezende, S. O. (2011). Building a topic hierarchy using the bag-of-related-words representation. In *DocEng*, pages 195–204, New York, NY, USA. ACM. Citado nas páginas 2, 3, 32, 34, 35, 41, 44, 45, 47, 49, 51, e 61.
- Sahar, S. (2002). Exploring interestingness through clustering: A framework. In *Proceedings of the IEEE International Conference on Data Mining*, pages 677–680. Citado nas páginas 26 e 27.
- Salton, G. (1989). *Automatic text processing: the transformation, analysis, and retrieval of information by computer*. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA. Citado na página 30.
- Salton, G. and Buckley, C. (1987). Term weighting approaches in automatic text retrieval. Technical report, Cornell University, Ithaca, NY, USA. Citado na página 31.
- Salton, G., Wong, A., and Yang, C. S. (1975). A vector space model for automatic indexing. *Commun. ACM*, 18(11):613–620. Citado na página 30.
- Santos, F. F., Rezende, S. O., and de Carvalho, V. O. (2014). Identificando o assunto dos documentos em coleções textuais utilizando termos compostos. In *XI Encontro Nacional de Inteligência Artificial e Computacional (ENIAC 2014)*, pages 550–557. Citado na página 54.
- Searle, S. R. (1971). *Linear models*. J. Wiley, New York, NY. Citado na página 56.
- Sebastiani, F. (2002). Machine learning in automated text categorization. *ACM Comput. Surv.*, 34(1):1–47. Citado na página 63.
- Shafiei, M., Wang, S., Zhang, R., Milius, E., Tang, B., Tougas, J., and Spiteri, R. (2007). Document representation and dimension reduction for text clustering. In *Proceedings of the 2007 IEEE 23rd International Conference on Data Engineering Workshop, ICDEW '07*, pages 770–779, Washington, DC, USA. IEEE Computer Society. Citado nas páginas 2 e 36.
- Steinbach, M., Karypis, G., and Kumar, V. (2000). A comparison of document clustering techniques. In *KDD workshop on text mining*, volume 400, pages 525–526. Citeseer. Citado nas páginas 12 e 13.
- Steyvers, M. and Griffiths, T. (2007). *Probabilistic Topic Models*. Lawrence Erlbaum Associates. Citado nas páginas 37 e 39.
- Sundermann, C. V. (2015). *Extração de informação contextual utilizando mineração de textos para sistemas de recomendação sensíveis ao contexto*. Mestrado em ciências da computação e matemática computacional, USP - São Carlos, São Carlos, SP, Brasil. Citado nas páginas xx, 70, e 127.

Sundermann, C. V., Domingues, M. A., Marcacini, R. M., and Rezende, S. O. (2014). Using topic hierarchies with privileged information to improve context-aware recommender systems. In *2014 Brazilian Conference on Intelligent Systems, BRACIS 2014, São Paulo, Brazil, October 18-22, 2014*, pages 61–66. IEEE. Citado nas páginas xxii, 70, 71, 72, 73, 127, e 128.

Tan, P.-N. and Kumar, V. (2000). Interestingness measures for association patterns: A perspective. In *6th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD 2000*, page 9p. Disponível em: <http://www.cas.mcmaster.ca/~bruha/kdd2000/kddrep.html> [22/07/2004]. Citado na página 19.

Tan, P.-N., Kumar, V., and Srivastava, J. (2004). Selecting the right objective measure for association analysis. *Information Systems*, 29(4):293–313. Citado na página 19.

Tan, P.-N., Steinbach, M., and Kumar, V. (2005a). Association analysis: Basic concepts and algorithms. In *Introduction to Data Mining*, chapter 6, pages 327–414. Addison-Wesley. Citado na página 23.

Tan, P.-N., Steinbach, M., and Kumar, V. (2005b). *Introduction to Data Mining*. Addison-Wesley. Citado na página 10.

Toivonen, H., Klemettinen, M., Ronkainen, P., Hätkönen, K., and Mannila, H. (1995). Pruning and grouping discovered association rules. Workshop Notes of the ECML'95 Workshop on Statistics, Machine Learning, and Knowledge Discovery in Databases, 47–52, MLnet. Citado na página 27.

Turner, V., Gantz, J. F., Reinsel, D., and Minton, S. (2014). The digital universe of opportunities: Rich data and the increasing value of the internet of things. *External Publication of IDC (Analyse the Future) Information and Data Sponsored by EMC Corporation*, pages 1–10. Citado na página 1.

Turney, P. D. and Pantel, P. (2010). From frequency to meaning: vector space models of semantics. *J. Artif. Int. Res.*, 37(1):141–188. Citado nas páginas 31 e 52.

Wallach, H. M. (2006). Topic modeling: Beyond bag-of-words. In *ICML*, pages 977–984, New York, NY, USA. ACM. Citado nas páginas 2, 3, 35, e 45.

Webb, G. I. (1995). Opus: An efficient admissible algorithm for unordered search. *Journal of Artificial Intelligence Research*, pages 431–465. Citado na página 17.

Weiss, S. M., Indurkhya, N., Zhang, T., and Damerau, F. J. (2005). *Text Mining - Predictive Methods for Analyzing Unstructured Information*. Springer Science+Business Media, Inc. Citado na página 7.

Wong, S. K. M., Ziarko, W., and Wong, P. C. N. (1985). Generalized vector spaces model in information retrieval. In *Proceedings of the 8th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '85*, pages 18–25, New York, NY, USA. ACM. Citado nas páginas 2, 34, e 45.

- Wu, M.-S., Lee, H.-S., and Wang, H.-M. (2010). Exploiting semantic associative information in topic modeling. In *SLT*, pages 384–388. IEEE. Citado na página 2.
- Wu, X., Kumar, V., Ross Quinlan, J., Ghosh, J., Yang, Q., Motoda, H., McLachlan, G. J., Ng, A., Liu, B., Yu, P. S., Zhou, Z.-H., Steinbach, M., Hand, D. J., and Steinberg, D. (2007). Top 10 algorithms in data mining. *Knowl. Inf. Syst.*, 14(1):1–37. Citado na página 11.
- Xu, R. and Wunsch, D. (2008). *Clustering (IEEE Press Series on Computational Intelligence)*. Wiley-IEEE Press, illustrated edition edition. Citado na página 12.
- Zaki, m. and Hsio, c. (2002). CHARM: An Efficient Algorithm for Closed Itemset Mining. In *Proceedings of the 2nd SIAM International Conference on Data mining*. Citado na página 17.
- Zhang, C. and Zhang, S. (2002). *Association Rule Mining: Models and Algorithms*, volume 2307 of *Lecture Notes in Artificial Intelligence*. Springer-Verlag New York, Inc. Citado nas páginas 16 e 17.
- Zhang, W., Yoshida, T., Tang, X., and Wang, Q. (2010). Text clustering using frequent itemsets. *Knowl.-Based Syst.*, 23(5):379–388. Citado na página 2.
- Zhao, Y., Karypis, G., and Fayyad, U. (2005). Hierarchical clustering algorithms for document datasets. *Data Mining and Knowledge Discovery*, 10(2):141–168. Citado nas páginas 10, 13, e 14.
- Zhao, Y., Zhao, Y., Zhang, C., and Cao, L. (2009). *Post-mining of Association Rules: Techniques for Effective Knowledge Extraction*. Information Science Reference - Imprint of: IGI Publishing, Hershey, PA. Citado na página 19.
- Zhu, D., Fukazawa, Y., Karapetsas, E., and Ota, J. (2012). Intuitive topic discovery by incorporating word-pair's connection into lda. In *Web Intelligence*, pages 303–310. IEEE. Citado nas páginas 2, 35, 36, 37, 41, e 45.

Coleção de Documentos Usados como Exemplo

Para demonstrar a execução e os resultados de cada etapa do modelo proposto LARCM, foi construída uma coleção de documentos de exemplo. Foram selecionados sete assuntos importantes da área de computação, procurando por assuntos que apresentam algum tipo de sobreposição quanto ao seu conceito. Por exemplo, os temas “mineração de dados” e “aprendizado de máquina” possuem grande sobre posição quanto a técnicas aplicadas e conceitos básicos. Os documentos, apresentados na Tabela A.1, foram construídos selecionando o primeiro parágrafo da entrada da Wikipedia¹, versão em português, correspondente ao assunto que dá nome ao arquivo.

Na Tabela A.2 são apresentados os documentos após o processamento realizado na coleção. Foram removidas as *stopwords* e os números. Os caracteres que representam letras acentuadas e o cedilha foram substituídos pelos caracteres correspondentes sem o símbolo. Ainda, todos os caracteres foram substituídos pela sua versão minúscula. Os termos foram substituídos pelo seu radical obtido com a aplicação do algoritmo de Porter (Porter, 1997; Nogueira et al., 2008). Os símbolos que não são alfanuméricos foram removidos, com exceção do símbolo de ponto final (“.”). Este último foi mantido para compatibilidade com as ferramentas que utilizam este símbolo para identificar frases nos documentos.

Os documentos pré-processados são então mapeados em transações. Para o exemplo, foi utilizada a janela de tamanho cinco. As transações obtidas por esse processo são apresentadas nas Tabelas A.3, A.4, A.5 e A.6.

Aplicando-se o algoritmo de extração de regras de associação como apresentado na Seção 4.2.1, foram obtidas 3659 regras de associação. Desse total, 1935 regras de associação foram selecionadas utilizando a medida objetiva *Odds Ratio*. O resultado final da

¹<http://pt.wikipedia.org/>

aplicação do modelo LARCM combinado com a medida objetiva *Odds Ratio* na coleção de exemplos é apresentado na Tabela A.8.

Tabela A.1: Textos da coleção de documentos construída para os exemplos apresentados no texto. Cada documento foi construído selecionando o primeiro parágrafo da entrada da Wikipedia, versão em português, referente ao assunto do documento.

aprendizado_de_maquina.txt
A aprendizagem automática ou aprendizado de máquina é um sub-campo da inteligência artificial dedicado ao desenvolvimento de algoritmos e técnicas que permitam ao computador aprender, isto é, que permitam ao computador aperfeiçoar seu desempenho em alguma tarefa. Enquanto que na inteligência artificial existem dois tipos de raciocínio - o indutivo, que extrai regras e padrões de grandes conjuntos de dados, e o dedutivo - o aprendizado de máquina só se preocupa com o indutivo.
banco_de_dados.txt
Bancos de dados (português brasileiro) ou bases de dados (português europeu) são coleções organizadas de dados que se relacionam de forma a criar algum sentido (Informação) e dar mais eficiência durante uma pesquisa ou estudo. ¹ ² ³ São de vital importância para empresas e há duas décadas se tornaram a principal peça dos sistemas de informação. ⁴ ² ⁵ Normalmente existem por vários anos sem alterações em sua estrutura. ⁶ ⁷
ciencia_da_computacao.txt
Ciência da Computação é a ciência que estuda as técnicas, metodologias e instrumentos computacionais, que automatiza processos e desenvolve soluções baseadas no uso do processamento digital. Não se restringe apenas ao estudo dos algoritmos, suas aplicações e implementação na forma de software, extrapolando para todo e qualquer conhecimento pautado no computador, que envolve também a telecomunicação, o banco de dados e as aplicações tecnológicas que possibilitam atingir o tratamento de dados de entrada e saída, de forma que se transforme em informação. Assim, a Ciência da Computação também abrange as técnicas de modelagem de dados e os protocolos de comunicação, além de princípios que abrangem outras especializações da área.
data_warehouse.txt (Observação: A entrada para este assunto da Wikipedia em português é “Armazém de Dados”)
Um armazém de dados, ou ainda depósito de dados, é utilizado para armazenar informações relativas às atividades de uma organização em bancos de dados, de forma consolidada. O desenho da base de dados favorece os relatórios, a análise de grandes volumes de dados e a obtenção de informações estratégicas que podem facilitar a tomada de decisão.
engenharia_de_software.txt
Engenharia de Software é uma área da computação voltada à especificação, desenvolvimento e manutenção de sistemas de software, com aplicação de tecnologias e práticas de gerência de projetos e outras disciplinas, visando organização, produtividade e qualidade. ²
inteligencia_artificial.txt
Inteligência artificial (IA) é a inteligência similar a humana exibida por mecanismos ou software. Também é um campo de estudo acadêmico. Os principais pesquisadores e livros didáticos definem o campo como "o estudo e projeto de agentes inteligentes", onde um agente inteligente é um sistema que percebe seu ambiente e toma atitudes que maximizam suas chances de sucesso. John McCarthy, quem cunhou o termo em 1956 ("numa conferência de especialistas celebrada em Darmouth College" Gubern, Román: O Eros Eletrônico), a define como "a ciência e engenharia de produzir máquinas inteligentes".
mineracao_de_dados.txt
Prospecção de dados (português europeu) ou mineração de dados (português brasileiro) (também conhecida pelo termo inglês data mining) é o processo de explorar grandes quantidades de dados à procura de padrões consistentes, como regras de associação ou sequências temporais, para detectar relacionamentos sistemáticos entre variáveis, detectando assim novos subconjuntos de dados.

Tabela A.2: Textos pré-processados da coleção de documentos de exemplo apresentada na Tabela A.1.

aprendizado_de_maquina.txt
aprendiz automa apred maquin sub camp intelligenc artific dedic desenvolv algoritm tecn permit comput apred permit comput aperfeico desempenh taref . intelligenc artific exist tip raciocini indu extr regr padro grand conjunt dad dedu apred maquin preocup indu .
banco_de_dados.txt
banc dad portugu brasil bas dad portugu europ sao coleco organizad dad relacion form cri sent informaca dar eficienc pesquis estud . sao vital importanc empr dec torn princip pec sistem informaca . norm exist ano alteraco estrut .
ciencia_da_computacao.txt
cienc computaca cienc estud tecn metodolog instrument computac automatiz process desenvolv soluco base uso process digit . restring estud algoritm aplicaco implementaca form softw extrapol conhec paut comput envolv telecomunicaca banc dad aplicaco tecnolog possibilid ating trat dad entr said form transform informaca . cienc computaca abrang tecn model dad protocol comunicaca principi abrang especializaco are .
data_warehouse.txt
armaz dad deposit dad util armazen informaco relativ atividad organizaca banc dad form consolid . desenh bas dad favorec relato analis grand volum dad obtenca informaco estrateg pod facil tom decis .
engenharia_de_software.txt
engenh softw are computaca volt especificaca desenvolv manutenca sistem softw aplicaca tecnolog pra gerenc projet disciplin vis organizaca produt qualidad .
inteligencia_artificial.txt
intelligenc artific intelligenc simil human exib mecan softw . camp estud academ . princip pesquis livr dida defin camp estud projet agent intelig agent intelig sistem percep ambi tom atitud maximiz chanc sucess . john mccarthy cunh term conferenc especi celebr darmouth coleb gubern roman ero eletron defin cienc engenh produz maquin intelig .
mineracao_de_dados.txt
prospecca dad portugu europ mineraca dad portugu brasil conhec term ingl dat mining process explor grand quant dad proc padro consist regr associaca sequenc tempor detect relacion sistema varia detect nov subconjunt dad .

Tabela A.3: Transações obtidas pelo mapeamento dos textos pré-processados da coleção de documentos de exemplo apresentada na Tabela A.2.

aprendizado_de_maquina.txt	banco_de_dados.txt
aprendiz aprendiz automa aprendiz automa apred aprendiz automa apred maquin aprendiz automa apred maquin sub automa apred maquin sub camp apred maquin sub camp intelligenc maquin sub camp intelligenc artific sub camp intelligenc artific dedic camp intelligenc artific dedic desenvolv intelligenc artific dedic desenvolv algoritm artific dedic desenvolv algoritm tecn dedic desenvolv algoritm tecn permit desenvolv algoritm tecn permit comput algoritm tecn permit comput apred tecn permit comput apred permit permit comput apred permit comput comput apred permit comput aperfeico apred permit comput aperfeico desempenh permit comput aperfeico desempenh taref comput aperfeico desempenh taref intelligenc aperfeico desempenh taref intelligenc artific desempenh taref intelligenc artific exist taref intelligenc artific exist tip intelligenc artific exist tip raciocini artific exist tip raciocini indu exist tip raciocini indu extr tip raciocini indu extr regr raciocini indu extr regr padro indu extr regr padro grand extr regr padro grand conjunt regr padro grand conjunt dad padro grand conjunt dad dedu grand conjunt dad dedu apred conjunt dad dedu apred maquin dad dedu apred maquin preocup dedu apred maquin preocup indu apred maquin preocup indu maquin preocup indu preocup indu indu	banc banc banc dad portugu banc dad portugu brasil banc dad portugu brasil bas dad portugu brasil bas dad portugu brasil bas dad portugu brasil bas dad portugu europ bas dad portugu europ sao dad portugu europ sao coleco portugu europ sao coleco organizad europ sao coleco organizad dad sao coleco organizad dad relacion coleco organizad dad relacion form organizad dad relacion form cri dad relacion form cri sent relacion form cri sent informaca form cri sent informaca dar cri sent informaca dar eficienc sent informaca dar eficienc pesquis informaca dar eficienc pesquis estud dar eficienc pesquis estud sao eficienc pesquis estud sao vital pesquis estud sao vital importanc estud sao vital importanc empr sao vital importanc empr dec vital importanc empr dec torn importanc empr dec torn princip empr dec torn princip pec dec torn princip pec sistem torn princip pec sistem informaca princip pec sistem informaca norm pec sistem informaca norm exist sistem informaca norm exist ano informaca norm exist ano alteraco norm exist ano alteraco estrut exist ano alteraco estrut ano alteraco estrut alteraco estrut estrut

Tabela A.4: Transações obtidas pelo mapeamento dos textos pré-processados da coleção de documentos de exemplo apresentada na Tabela A.2.

ciencia_da_computacao.txt	data_warehouse.txt
cienc	armaz
cienc computaca	armaz dad
cienc computaca cienc	armaz dad deposit
cienc computaca cienc estud	armaz dad deposit dad
cienc computaca cienc estud tecn	armaz dad deposit dad util
computaca cienc estud tecn metodolog	dad deposit dad util armazen
cienc estud tecn metodolog instrument	deposit dad util armazen informaco
estud tecn metodolog instrument computac	dad util armazen informaco relativ
tecn metodolog instrument computac automatiz	util armazen informaco relativ atividad
metodolog instrument computac automatiz process	armazen informaco relativ atividad organizaca
instrument computac automatiz process desenvolv	informaco relativ atividad organizaca banc
computac automatiz process desenvolv soluco	relativ atividad organizaca banc dad
automatiz process desenvolv soluco base	atividad organizaca banc dad form
process desenvolv soluco base uso	organizaca banc dad form consolid
desenvolv soluco base uso process	banc dad form consolid desenh
soluco base uso process digit	dad form consolid desenh bas
base uso process digit restrin	form consolid desenh bas dad
uso process digit restrin estud	consolid desenh bas dad favorec
process digit restrin estud algoritm	desenh bas dad favorec relato
digit restrin estud algoritm aplicaco	bas dad favorec relato analis
restrin estud algoritm aplicaco implementaca	dad favorec relato analis grand
estud algoritm aplicaco implementaca form	favorec relato analis grand volum
algoritm aplicaco implementaca form softw	relato analis grand volum dad
aplicaco implementaca form softw extrapol	analis grand volum dad obtenca
implementaca form softw extrapol conh	grand volum dad obtenca informaco
form softw extrapol conh paut	volum dad obtenca informaco estrateg
softw extrapol conh paut comput	dad obtenca informaco estrateg pod
extrapol conh paut comput envolv	obtenca informaco estrateg pod facil
conh paut comput envolv telecomunicaca	informaco estrateg pod facil tom
paut comput envolv telecomunicaca banc	estrateg pod facil tom decisa
comput envolv telecomunicaca banc dad	facil tom decisa
envolv telecomunicaca banc dad aplicaco	tom decisa
telecomunicaca banc dad aplicaco tecnolog	decisa
banc dad aplicaco tecnolog possibil	
dad aplicaco tecnolog possibil ating	
aplicaco tecnolog possibil ating trat	
tecnolog possibil ating trat dad	
possibil ating trat dad entr	
ating trat dad entr said	
trat dad entr said form	
dad entr said form transform	
entr said form transform informaca	
said form transform informaca cienc	
form transform informaca cienc computaca	
transform informaca cienc computaca abrang	
informaca cienc computaca abrang tecn	
cienc computaca abrang tecn model	
computaca abrang tecn model dad	
abrang tecn model dad protocol	
tecn model dad protocol comunicaca	
model dad protocol comunicaca principi	
dad protocol comunicaca principi abrang	
protocol comunicaca principi abrang especializaco	
comunicaca principi abrang especializaco are	
principi abrang especializaco are	
abrang especializaco are	
especializaco are	
are	

Tabela A.5: Transações obtidas pelo mapeamento dos textos pré-processados da coleção de documentos de exemplo apresentada na Tabela A.2.

engenharia_de_software.txt	inteligencia_artificial.txt
engenh engenh softw engenh softw are engenh softw are computaca engenh softw are computaca volt softw are computaca volt especificaca are computaca volt especificaca desenvolv computaca volt especificaca desenvolv manutenca volt especificaca desenvolv manutenca sistem especificaca desenvolv manutenca sistem softw desenvolv manutenca sistem softw aplicaca manutenca sistem softw aplicaca tecnolog sistem softw aplicaca tecnolog pra softw aplicaca tecnolog pra gerenc aplicaca tecnolog pra gerenc projet tecnolog pra gerenc projet disciplin pra gerenc projet disciplin vis gerenc projet disciplin vis organizaca projet disciplin vis organizaca produt disciplin vis organizaca produt qualidad vis organizaca produt qualidad organizaca produt qualidad produt qualidad qualidad	inteligenc inteligenc artific inteligenc artific intelligenc inteligenc artific intelligenc simil inteligenc artific intelligenc simil human artific intelligenc simil human exib inteligenc simil human exib mecan simil human exib mecan softw human exib mecan softw camp exib mecan softw camp estud mecan softw camp estud academ softw camp estud academ princip camp estud academ princip pesquis estud academ princip pesquis livr academ princip pesquis livr dida princip pesquis livr dida defin pesquis livr dida defin camp livr dida defin camp estud dida defin camp estud projet defin camp estud projet agent camp estud projet agent intelig estud projet agent intelig agent projet agent intelig agent intellig agent intellig agent intellig sistem intelig agent intellig sistem perceb agent intellig sistem perceb ambi intellig sistem perceb ambi tom sistem perceb ambi tom atitud perceb ambi tom atitud maximiz ambi tom atitud maximiz chanc tom atitud maximiz chanc sucess atitud maximiz chanc sucess john maximiz chanc sucess john mccarthy chanc sucess john mccarthy cunh sucess john mccarthy cunh term john mccarthy cunh term conferenc mccarthy cunh term conferenc especi cunh term conferenc especi celebr term conferenc especi celebr darmouth conferenc especi celebr darmouth cole especi celebr darmouth cole gubern celebr darmouth cole gubern roman darmouth cole gubern roman ero cole gubern roman ero eletron gubern roman ero eletron defin roman ero eletron defin cienc ero eletron defin cienc engenh eletron defin cienc engenh produz defin cienc engenh produz maquin cienc engenh produz maquin intellig engenh produz maquin intellig produz maquin intellig maquin intellig intelig

Tabela A.6: Transações obtidas pelo mapeamento dos textos pré-processados da coleção de documentos de exemplo apresentada na Tabela A.2.

mineracao_de_dados.txt	
prospecca	
prospecca dad	
prospecca dad portugu	
prospecca dad portugu europ	
prospecca dad portugu europ mineraca	
dad portugu europ mineraca dad	
portugu europ mineraca dad portugu	
europ mineraca dad portugu brasil	
mineraca dad portugu brasil conhec	
dad portugu brasil conhec term	
portugu brasil conhec term ingl	
brasil conhec term ingl dat	
conhec term ingl dat mining	
term ingl dat mining process	
ingl dat mining process explor	
dat mining process explor grand	
mining process explor grand quant	
process explor grand quant dad	
explor grand quant dad proc	
grand quant dad proc padro	
quant dad proc padro consist	
dad proc padro consist regr	
proc padro consist regr associaca	
padro consist regr associaca sequenc	
consist regr associaca sequenc tempor	
regr associaca sequenc tempor detect	
associaca sequenc tempor detect relacion	
sequenc tempor detect relacion sistema	
tempor detect relacion sistema varia	
detect relacion sistema varia detect	
relacion sistema varia detect nov	
sistema varia detect nov subconjunt	
varia detect nov subconjunt dad	
detect nov subconjunt dad	
nov subconjunt dad	
subconjunt dad	
dad	

Tabela A.7: Valores de quantidade total de regras de associação obtidas e selecionadas para cada documento da coleção de exemplo, e do limiar de corte calculado da medida objetiva *Odds Ratio* para cada documento da coleção de exemplos.

Documento	Regras de associação geradas	Regras de associação selecionadas	limiar de corte
aprendizado_de_maquina.txt	514	258	0.07013369
banco_de_dados.txt	506	273	0.073221266
ciencia_da_computacao.txt	793	424	0.049963307
data_warehouse.txt	413	233	0.08688247
engenharia_de_software.txt	285	147	0.11666617
inteligencia_artificial.txt	706	358	0.052093197
mineracao_de_dados.txt	442	242	0.080592036

Tabela A.8: Tópicos obtidos ao aplicar o modelo LARCM com a medida objetiva *Odds Ratio* e *K* igual a 10. Foram selecionadas as 10 regras de associação com maior valor para a medida *Odds Ratio* como descriptores de cada tópico.

protocol_tecn_model especializaco_comunicaca_principi computaca_tecn_model abrang_dad_tecn model_tecn_protocol comunicaca_especializaco comunicaca_model comunicaca_dad_principi model_tecn_computaca tecn_protocol
brasil_banc coleco_sao_portugu bas_dad_europ bas_europ_dad banc_brasil coleco_dad_europ banc_brasil_dad portugu_coleco organizad_dad_sao relacion_coleco_organizad
conferenc especi_darmouth gubern_celebr eletron_gubern produz_intelig_engenh coleg_especi especi_darmouth_coleg cole_gubern_ero coleg_ero_gubern especi_celebr_coleg engenh_eletron_cienc
sucess_atitud_maximiz chanc_tom_maximiz perceb_tom_sistem cunh_term_john perceb_agent_sistem john_chanc_maximiz tom_atitud_chanc cunh_sucess_mccarthy sistem_perceb_tom sistem_tom
metodolog_estud_instrument form_cienc cienc_tecn_abrang transform_informaca_computaca computac_tecn estud_tecn_computaca transform_form_cienc informaca_abrang_cienc abrang_cienc_tecn metodolog_estud_cienc
estud_aplicaco_implementaca uso_restrin digit_estud_algoritm estud_implementaca algoritm_estud_implementaca algoritm_aplicaco_restring estud_process_digit algoritm_digit implementaca_aplicaco_estud algoritm_estud_digit
form_dad_entr banc_aplicaco_telecomunicaca banc_tecnolog_dad aplicaco_telecomunicaca tecnolog_dad_banc form_dad_said aplicaco_dad_telecomunicaca dad_form envolv_telecomunicaca_dad ating_entr
regr_indu_padro conjunt_grand_regr indu_padro conjunt_regr_grand extr_grand grand_regr_conjunt padro_regr_indu padro_indu_regr conjunt_regr extr_raciocini_regr
artific_inteligenc_human human_artific artific_human_inteligenc artific_human_exib_mecan_simil simil_human_mecan inteligenç_human_artific inteligenc_exib inteligenc_exib_simil intelligenc_simil_exib
instrument_computac_process conhec_form desenvolv_uso_process form_algoritm conhec_form_extrapol extrapol_form_form_conhec instrument_automatiz_metodolog automatiz_metodolog extrapol_implementaca envolv_conhec_comput

Tabelas com os Resultados da Avaliação na Tarefa de Classificação

Neste trabalho de doutorado, foi proposta uma metodologia de avaliação (Seção 4.3) que está dividida em duas partes. A primeira parte avalia o desempenho do modelo proposto na tarefa de classificação de textos. Neste apêndice, são apresentados os resultados da comparação da acurácia média dos modelos para cada coleção de documentos na Seção B.1 e os resultados detalhados da acurácia obtida para cada configuração de parâmetros escolhida na Seção B.2.

B.1 Comparação dos Modelos de Extração de Tópicos Separados por Coleção de Documentos

Os resultados apresentados nas Tabelas B.1, B.2, B.3, B.4, B.5, B.6, B.7, B.8 e B.9, correspondem a análise de variância da acurácia média obtida pelos modelos de extração de tópicos considerando o efeito da quantidade de tópicos extraídos k e do classificador escolhido, similar ao processo descrito na Seção 4.3.1. Para essa avaliação, a decomposição da variância pode ser representada no seguinte modelo linear:

$$\widehat{m(Acur_{mr})} = \hat{\mu} + \hat{k} + \hat{mr} + \hat{cl} + \hat{e} \quad (\text{B.1})$$

Em que:

- $\widehat{m(Acur_{mr})}$: valor estimado da acurácia para o método de redução de dimensionalidade mr ;
- $\hat{\mu}$: valor estimado para a média geral da acurácia;
- \hat{k} : valor estimado para o efeito do valor definido de tópicos para extração k na estimativa do valor da acurácia;

- \hat{m}_r : valor estimado para o efeito do método de redução de dimensionalidade m_r na estimativa do valor da acurácia;
- \hat{c}_l : valor estimado para o efeito do classificador c_l na estimativa do valor da acurácia;
- \hat{e} : valor estimado para o componente do erro do modelo para o método de redução de dimensionalidade m_r , supondo que ele é aleatório.

Tabela B.1: Comparação múltipla das médias da acurácia dos classificadores para cada modelo de redução de dimensionalidade na coleção de documentos ACM-1.

(gl=428, $\widehat{V}(E)=8,0847$, $\alpha=0.05$)		
Método	acurácia	grupo
LARCM + Odds Ratio	77,041	<i>a</i>
LDA + <i>bag-of-words</i>	75,171	<i>b</i>
LARCM + Gini Index	71,522	<i>c</i>
LARCM + Novelty	71,510	<i>c</i>
LARCM + ϕ -Coefficient	71,492	<i>c</i>
LDA + <i>bag-of-related-words</i>	71,357	<i>c</i>
LARCM + Kappa	70,995	<i>c d</i>
LARCM + Confiança	70,936	<i>c d</i>
LARCM + Certainty Factor	70,631	<i>c d</i>
LARCM + Laplace	70,624	<i>c d</i>
LARCM + IS	70,322	<i>c d e</i>
LARCM + Klosgen	70,230	<i>c d e</i>
LARCM + J-Measure	70,230	<i>c d e</i>
LARCM + Mutual Information LHS	69,812	<i>c d e</i>
LARCM + Collective Strength	69,692	<i>c d e</i>
LARCM + Added Value	69,673	<i>c d e</i>
LARCM + Lift	68,576	<i>d e</i>
LARCM + Lambda	67,991	<i>e</i>
LARCM + Conviction	54,542	<i>f</i>

Tabela B.2: Comparação múltipla das médias da acurácia dos classificadores para cada modelo de redução de dimensionalidade na coleção de documentos ACM-2.

(gl=428, $\widehat{V}(E)=6,7492$, $\alpha=0.05$)		
Método	acurácia	grupo
LARCM + Odds Ratio	88,539	<i>a</i>
LARCM + ϕ -Coefficient	85,525	<i>b</i>
LARCM + Added Value	84,666	<i>b c</i>
LARCM + Gini Index	84,597	<i>b c</i>
LARCM + Kappa	84,567	<i>b c</i>
LARCM + Certainty Factor	84,558	<i>b c</i>
LARCM + Novelty	84,267	<i>b c</i>
LARCM + J-Measure	84,165	<i>b c</i>
LARCM + Mutual Information LHS	84,139	<i>b c</i>
LARCM + Klosgen	84,110	<i>b c</i>
LARCM + Confiança	83,930	<i>b c</i>
LDA + <i>bag-of-words</i>	83,909	<i>b c</i>
LARCM + IS	83,775	<i>b c</i>
LARCM + Laplace	83,750	<i>b c</i>
LARCM + Collective Strength	83,185	<i>b c</i>
LARCM + Lift	83,029	<i>b c</i>
LARCM + Lambda	82,190	<i>c</i>
LDA + <i>bag-of-related-words</i>	78,727	<i>d</i>
LARCM + Conviction	69,968	<i>e</i>

Tabela B.3: Comparação múltipla das médias da acurácia dos classificadores para cada modelo de redução de dimensionalidade na coleção de documentos ACM-3.

(gl=428, $\widehat{V}(E)=7,1358$, $\alpha=0.05$)		
Método	acurácia	grupo
LARCM + Odds Ratio	86,654	<i>a</i>
LARCM + ϕ -Coefficient	83,245	<i>b</i>
LDA + <i>bag-of-words</i>	82,669	<i>b c</i>
LARCM + Novelty	81,940	<i>b c</i>
LARCM + Laplace	81,582	<i>b c d</i>
LARCM + Gini Index	81,327	<i>b c d</i>
LARCM + Klosgen	81,215	<i>b c d</i>
LARCM + Certainty Factor	81,177	<i>b c d</i>
LARCM + Kappa	81,102	<i>b c d</i>
LARCM + Confiança	81,022	<i>b c d</i>
LARCM + J-Measure	81,017	<i>b c d</i>
LARCM + Collective Strength	80,971	<i>b c d</i>
LARCM + IS	80,590	<i>c d</i>
LARCM + Added Value	80,570	<i>c d</i>
LARCM + Lift	80,568	<i>c d</i>
LARCM + Mutual Information LHS	80,290	<i>c d</i>
LARCM + Lambda	79,060	<i>d</i>
LDA + <i>bag-of-related-words</i>	77,021	<i>e</i>
LARCM + Conviction	68,910	<i>f</i>

Tabela B.4: Comparação múltipla das médias da acurácia dos classificadores para cada modelo de redução de dimensionalidade na coleção de documentos ACM-4.

(gl=428, $\widehat{V}(E)=5,9388$, $\alpha=0.05$)		
Método	acurácia	grupo
LARCM + Odds Ratio	91,386	<i>a</i>
LDA + <i>bag-of-words</i>	87,888	<i>b</i>
LARCM + ϕ -Coefficient	87,416	<i>b c</i>
LARCM + Collective Strength	87,086	<i>b c</i>
LARCM + Confiança	86,897	<i>b c</i>
LARCM + Certainty Factor	86,580	<i>b c</i>
LARCM + Laplace	86,572	<i>b c</i>
LARCM + Novelty	86,522	<i>b c</i>
LARCM + Klosgen	86,422	<i>b c</i>
LARCM + Added Value	86,421	<i>b c</i>
LARCM + Mutual Information LHS	86,299	<i>b c</i>
LARCM + Gini Index	86,299	<i>b c</i>
LARCM + IS	85,949	<i>b c</i>
LARCM + Kappa	85,886	<i>b c</i>
LARCM + J-Measure	85,809	<i>b c</i>
LARCM + Lift	85,348	<i>c</i>
LDA + <i>bag-of-related-words</i>	85,132	<i>c</i>
LARCM + Lambda	85,008	<i>c</i>
LARCM + Conviction	77,029	<i>d</i>

Tabela B.5: Comparação múltipla das médias da acurácia dos classificadores para cada modelo de redução de dimensionalidade na coleção de documentos ACM-5.

(gl=428, $\widehat{V(E)}=8,2924$, $\alpha=0.05$)		
Método	acurácia	grupo
LDA + <i>bag-of-words</i>	82,0802	<i>a</i>
LARCM + Odds Ratio	80,6677	<i>a</i>
LARCM + ϕ -Coefficient	75,8750	<i>b</i>
LDA + <i>bag-of-related-words</i>	75,5028	<i>b</i>
LARCM + Novelty	75,5009	<i>b</i>
LARCM + Certainty Factor	75,1953	<i>b</i>
LARCM + Gini Index	75,0750	<i>b</i>
LARCM + Confiança	74,9900	<i>b</i>
LARCM + Laplace	74,5149	<i>b</i>
LARCM + Collective Strength	74,3976	<i>b</i>
LARCM + Added Value	74,3531	<i>b</i>
LARCM + Lift	74,2821	<i>b</i>
LARCM + J-Measure	74,1900	<i>b</i>
LARCM + Klosgen	74,0943	<i>b</i>
LARCM + Kappa	73,9191	<i>b</i>
LARCM + Mutual Information LHS	73,8995	<i>b</i>
LARCM + IS	73,6459	<i>b</i>
LARCM + Lambda	73,4017	<i>b</i>
LARCM + Conviction	66,2597	<i>c</i>

Tabela B.6: Comparação múltipla das médias da acurácia dos classificadores para cada modelo de redução de dimensionalidade na coleção de documentos ACM-6.

(gl=428, $\widehat{V(E)}=4,8945$, $\alpha=0.05$)		
Método	acurácia	grupo
LARCM + Odds Ratio	88,716	<i>a</i>
LDA + <i>bag-of-words</i>	87,274	<i>b</i>
LARCM + Novelty	85,963	<i>c</i>
LARCM + Certainty Factor	85,428	<i>c d</i>
LARCM + ϕ -Coefficient	84,636	<i>c d e</i>
LARCM + Laplace	84,400	<i>c d e</i>
LARCM + Confiança	84,216	<i>c d e</i>
LARCM + Added Value	84,191	<i>c d e</i>
LARCM + Mutual Information LHS	84,037	<i>d e</i>
LARCM + Collective Strength	84,026	<i>d e</i>
LARCM + Gini Index	83,941	<i>d e</i>
LARCM + IS	83,840	<i>d e</i>
LARCM + J-Measure	83,836	<i>d e</i>
LARCM + Kappa	83,813	<i>d e</i>
LARCM + Klosgen	83,393	<i>d e</i>
LARCM + Lift	83,023	<i>e</i>
LARCM + Lambda	81,545	<i>f</i>
LDA + <i>bag-of-related-words</i>	80,913	<i>f</i>
LARCM + Conviction	73,025	<i>g</i>

Tabela B.7: Comparação múltipla das médias da acurácia dos classificadores para cada modelo de redução de dimensionalidade na coleção de documentos ACM-7.

(gl=428, $\bar{V}(E)=3,4782$, $\alpha=0.05$)		
Método	acurácia	grupo
LARCM + Odds Ratio	89,033	<i>a</i>
LARCM + ϕ -Coefficient	85,958	<i>b</i>
LARCM + Gini Index	85,086	<i>b c</i>
LARCM + Novelty	85,016	<i>b c</i>
LARCM + Klosgen	84,993	<i>b c</i>
LARCM + Laplace	84,806	<i>b c</i>
LARCM + Collective Strength	84,755	<i>b c</i>
LARCM + Confiança	84,691	<i>b c</i>
LARCM + Kappa	84,682	<i>b c</i>
LARCM + J-Measure	84,607	<i>b c</i>
LARCM + Added Value	84,569	<i>b c</i>
LARCM + Certainty Factor	84,556	<i>b c</i>
LARCM + Lift	84,373	<i>b c</i>
LARCM + Mutual Information LHS	83,940	<i>c</i>
LARCM + IS	83,837	<i>c</i>
LDA + <i>bag-of-words</i>	83,356	<i>c</i>
LARCM + Lambda	82,260	<i>d</i>
LDA + <i>bag-of-related-words</i>	78,446	<i>e</i>
LARCM + Conviction	71,887	<i>f</i>

Tabela B.8: Comparação múltipla das médias da acurácia dos classificadores para cada modelo de redução de dimensionalidade na coleção de documentos ACM-8.

(gl=428, $\bar{V}(E)=5,4408$, $\alpha=0.05$)		
Método	acurácia	grupo
LDA + <i>bag-of-words</i>	79,988	<i>a</i>
LARCM + Odds Ratio	76,074	<i>b</i>
LDA + <i>bag-of-related-words</i>	74,209	<i>c</i>
LARCM + Confiança	72,070	<i>d</i>
LARCM + ϕ -Coefficient	72,032	<i>d</i>
LARCM + Gini Index	71,913	<i>d</i>
LARCM + Kappa	71,686	<i>d</i>
LARCM + Certainty Factor	71,437	<i>d</i>
LARCM + Added Value	71,350	<i>d</i>
LARCM + Novelty	71,318	<i>d</i>
LARCM + Lift	71,182	<i>d</i>
LARCM + Klosgen	70,909	<i>d</i>
LARCM + IS	70,896	<i>d</i>
LARCM + Laplace	70,842	<i>d</i>
LARCM + Mutual Information LHS	70,429	<i>d</i>
LARCM + J-Measure	70,059	<i>d</i>
LARCM + Collective Strength	69,883	<i>d</i>
LARCM + Lambda	67,675	<i>e</i>
LARCM + Conviction	56,724	<i>f</i>

Tabela B.9: Comparação múltipla das médias da acurácia dos classificadores para cada modelo de redução de dimensionalidade na coleção de documentos Re8.

(gl=428, $\widehat{V}(E)=0,9622$, $\alpha=0.05$)		
Método	acurácia	grupo
LDA + <i>bag-of-words</i>	90,788	<i>a</i>
LARCM + Odds Ratio	85,992	<i>b</i>
LDA + <i>bag-of-related-words</i>	84,101	<i>c</i>
LARCM + ϕ -Coefficient	84,015	<i>c</i>
LARCM + Laplace	83,404	<i>d</i>
LARCM + Confiança	83,324	<i>d e</i>
LARCM + Gini Index	82,957	<i>d e f</i>
LARCM + Novelty	82,947	<i>d e f</i>
LARCM + Klosgen	82,931	<i>d e f</i>
LARCM + Certainty Factor	82,807	<i>d e f</i>
LARCM + Added Value	82,773	<i>d e f</i>
LARCM + J-Measure	82,695	<i>d e f g</i>
LARCM + Kappa	82,603	<i>d e f g h</i>
LARCM + Lift	82,502	<i>e f g h</i>
LARCM + Mutual Information LHS	82,374	<i>f g h</i>
LARCM + IS	82,260	<i>f g h</i>
LARCM + Lambda	81,915	<i>g h</i>
LARCM + Collective Strength	81,807	<i>h</i>
LARCM + Conviction	81,124	<i>i</i>

B.2 Resultados da Acurácia Obtida para Cada Configuração dos Modelos de Extração de Tópicos

Nesta seção são apresentados os resultados obtidos para cada configuração utilizada nos experimentos, isto é, para cada coleção de documentos e quantidade de tópicos extraídos. Nas tabelas, são apresentados os valores de acurácia e o desvio padrão obtidos após a aplicação do processo de avaliação apresentado na Seção 4.3.1. Os algoritmos de classificação avaliados, descritos na Seção 5.1.3, são: (NB) *Naive Bayes*; (MNB) *Multinomial Naive Bayes*; (J48) J48; (SMO) SMO utilizando o parâmetro de generalização “c” com valor 1 ($c=1$) e 10 ($c=10$); e (IBk) IBk com os valores de k definidos em 3, 5 e 7;

Tabela B.10: Acurácia dos classificadores para cada modelo de redução de dimensionalidade, com a quantidade de tópicos obtidos $k = 50$ na coleção de documentos ACM-1.

Método	NB	MNB	J48	SMO c=1	SMO c=10	IBk 3	IBk 5	IBk 7
LDA + bag-of-related-words	78.42±4.93	79.48±5.79	77.48±6.63	76.97±5.62	80.33±5.69	68.35±6.21	69.50±6.80	68.25±6.88
LDA + bag-of-words	86.04±4.98	84.77±5.44	82.03±5.06	85.57±5.02	84.29±5.01	80.05±6.39	81.25±6.18	81.10±6.50
LARCM + Added Value	73.18±6.51	74.39±5.73	66.99±6.98	74.26±6.08	77.09±5.72	66.98±6.74	68.71±7.08	69.37±6.71
LARCM + Certainty Factor	71.03±6.66	74.56±6.54	71.33±6.58	72.50±6.93	73.33±6.36	62.98±7.19	64.98±6.80	63.93±7.24
LARCM + Collective Strength	75.08±7.23	75.72±7.10	69.69±6.56	72.79±7.07	75.69±6.75	65.89±7.11	64.74±6.96	67.77±6.58
LARCM + Confiança	75.29±6.37	78.02±6.14	66.11±7.56	75.64±6.12	78.47±5.84	65.24±6.78	66.59±7.41	66.87±7.50
LARCM + Conviction	55.95±8.19	60.70±7.99	55.90±8.11	60.03±7.45	60.08±7.57	50.83±7.11	53.69±7.50	53.82±7.58
LARCM + ϕ -Coefficient	71.71±6.81	75.34±6.60	67.70±7.05	72.61±6.73	77.27±6.28	65.40±5.97	64.99±7.37	65.59±7.09
LARCM + Gini Index	73.24±6.50	74.60±7.17	66.45±7.29	72.40±7.29	75.35±6.57	66.12±6.39	66.05±6.31	67.51±6.49
LARCM + IS	75.44±6.02	75.14±5.36	70.43±6.04	74.53±6.76	76.24±6.28	64.03±7.34	64.74±6.96	65.41±6.68
LARCM + J-Measure	72.86±6.43	72.96±6.41	64.31±7.18	71.20±6.70	74.94±6.28	64.51±7.09	65.14±6.35	65.31±6.78
LARCM + Kappa	74.41±7.56	75.79±6.14	69.38±6.02	74.37±6.21	79.27±6.31	64.79±6.70	67.33±6.59	65.90±6.98
LARCM + Klosgen	71.65±6.86	75.34±6.10	64.35±7.54	73.07±7.06	75.74±7.08	63.20±6.95	64.53±7.19	66.53±7.49
LARCM + Lambda	67.11±6.45	74.60±5.84	61.43±7.10	74.28±5.96	75.73±5.93	64.00±7.57	64.88±7.37	66.21±7.43
LARCM + Laplace	74.27±6.29	76.62±5.26	69.35±6.94	75.39±5.05	75.54±5.90	64.85±5.87	66.02±6.22	66.77±6.16
LARCM + Lift	72.08±5.25	73.48±5.97	65.44±7.27	72.66±5.83	76.24±6.21	61.53±6.54	64.33±6.08	65.36±6.42
LARCM + Mutual Information LHS	69.82±6.94	72.92±6.51	66.81±7.17	71.12±6.52	74.53±6.92	62.65±6.49	64.75±6.98	64.93±6.35
LARCM + Novelty	71.94±6.93	77.12±6.65	69.15±7.08	75.46±6.60	76.29±6.52	67.12±6.92	68.87±6.53	68.87±6.39
LARCM + Odds Ratio	79.40±6.57	79.97±6.22	75.92±6.95	79.26±6.80	82.27±6.38	72.37±6.93	73.95±7.51	73.65±7.78

Tabela B.11: Acurácia dos classificadores para cada modelo de redução de dimensionalidade, com a quantidade de tópicos obtidos $k = 100$ na coleção de documentos ACM-1.

Método	NB	MNB	J48	SMO c=1	SMO c=10	IBk 3	IBk 5	IBk 7
LDA + bag-of-related-words	77.07±6.88	73.64±7.52	69.56±6.54	75.16±6.98	78.25±6.54	64.59±7.06	67.55±6.83	71.74±7.26
LDA + bag-of-words	78.55±6.46	80.03±5.76	72.71±7.04	77.47±6.27	79.80±6.28	72.09±7.16	72.46±6.71	71.88±6.64
LARCM + Added Value	76.09±6.10	73.53±6.13	63.84±6.11	73.55±5.93	75.71±6.65	61.60±6.77	62.91±7.55	63.48±7.59
LARCM + Certainty Factor	78.90±6.30	77.29±6.52	69.94±6.20	74.84±6.54	76.57±6.98	65.42±6.53	68.35±6.57	70.06±6.27
LARCM + Collective Strength	77.07±6.41	74.11±6.87	67.89±7.14	72.56±7.57	75.49±6.18	61.15±6.82	60.60±7.53	62.58±7.57
LARCM + Confiança	80.33±5.79	75.19±6.30	73.43±7.00	74.09±5.97	75.21±6.24	59.67±7.46	62.66±7.02	63.47±6.05
LARCM + Conviction	54.61±7.15	59.25±7.88	55.64±7.80	57.17±7.39	59.17±7.44	44.76±7.52	46.87±7.56	48.21±8.12
LARCM + ϕ -Coefficient	79.07±6.57	77.27±6.76	72.28±7.33	75.14±6.93	78.27±7.03	67.10±7.38	66.60±6.83	68.12±6.64
LARCM + Gini Index	77.94±6.61	75.63±6.44	71.27±7.44	74.71±6.83	77.64±7.06	67.09±6.45	68.42±6.97	69.60±7.39
LARCM + IS	76.06±7.26	74.64±5.82	71.25±6.83	73.98±6.27	76.21±6.42	60.81±7.03	61.95±7.25	63.33±6.62
LARCM + J-Measure	76.85±6.34	76.13±6.41	69.22±6.82	73.82±5.94	76.98±5.51	63.83±7.46	64.84±6.81	66.42±7.07
LARCM + Kappa	75.92±6.87	74.37±6.25	65.63±6.55	72.55±6.92	74.08±5.79	64.36±6.13	65.31±6.78	64.94±6.78
LARCM + Klosgen	78.37±6.07	73.33±6.62	72.46±6.93	71.43±6.74	75.74±6.66	63.25±7.07	64.10±7.19	63.88±6.48
LARCM + Lambda	72.25±7.71	71.62±6.61	66.97±7.60	71.12±7.57	73.83±6.08	59.63±7.75	61.55±7.88	62.60±7.61
LARCM + Laplace	78.32±6.76	76.04±7.00	71.75±7.00	74.14±7.28	77.40±7.20	64.51±7.54	66.59±6.38	67.74±7.83
LARCM + Lift	74.81±5.73	74.29±6.19	68.04±6.89	74.14±6.28	76.54±6.45	59.44±6.94	60.50±6.71	62.28±6.04
LARCM + Mutual Information LHS	78.25±6.40	74.48±6.60	72.63±7.43	72.86±6.91	75.57±6.43	62.90±7.19	67.19±7.61	68.12±7.83
LARCM + Novelty	78.97±5.77	74.91±5.66	74.36±6.74	74.31±5.85	75.39±6.15	64.01±6.49	66.84±5.78	68.07±6.15
LARCM + Odds Ratio	80.85±4.91	79.76±6.02	74.81±6.31	78.31±6.17	84.03±5.01	74.28±6.41	73.25±6.59	75.20±5.77

Tabela B.12: Acurácia dos classificadores para cada modelo de redução de dimensionalidade, com a quantidade de tópicos obtidos $k = 150$ na coleção de documentos ACM-1.

Método	NB	MNB	J48	SMO c=1	SMO c=10	IBk 3	IBk 5	IBk 7
LDA + bag-of-related-words	73.81±6.34	66.54±7.38	62.59±6.81	69.40±6.82	70.63±6.72	61.41±7.32	64.82±6.78	67.05±7.07
LDA + bag-of-words	71.54±6.86	71.36±6.35	62.34±8.02	53.09±6.36	79.58±6.23	68.90±7.32	64.74±7.55	62.43±9.50
LARCM + Added Value	77.61±6.23	74.36±6.60	66.59±7.34	70.58±5.90	75.36±6.86	59.16±7.54	63.10±7.48	63.70±6.80
LARCM + Certainty Factor	78.64±5.99	73.93±6.85	67.18±6.67	71.06±7.31	76.13±6.29	63.06±6.66	64.55±6.83	64.57±7.57
LARCM + Collective Strength	80.88±6.34	72.69±7.07	66.35±7.47	70.21±7.38	77.95±7.29	61.07±7.05	61.96±7.95	62.68±8.02
LARCM + Confiança	77.36±6.11	72.93±5.92	72.83±7.45	70.23±7.72	75.06±7.20	64.63±7.50	65.47±8.09	67.68±7.20
LARCM + Conviction	57.25±7.65	60.40±6.62	53.41±7.63	58.18±7.30	60.55±7.16	45.52±7.93	47.24±8.20	49.78±7.83
LARCM + ϕ -Coefficient	79.50±6.95	75.51±6.37	70.38±6.37	72.25±6.81	78.51±5.24	60.84±6.51	66.58±6.23	67.77±6.37
LARCM + Gini Index	77.74±6.25	75.84±5.96	70.31±7.05	72.80±6.44	78.27±6.00	62.46±7.33	66.89±6.18	68.19±7.19
LARCM + IS	78.75±7.06	74.09±5.59	67.52±6.78	72.11±5.98	76.40±6.14	62.14±6.82	64.09±6.89	68.45±6.50
LARCM + J-Measure	80.10±6.07	74.49±6.78	70.49±7.08	72.51±6.72	78.20±5.64	61.67±7.19	63.41±7.68	65.34±6.77
LARCM + Kappa	78.89±5.90	75.33±6.59	71.48±6.63	72.57±6.65	76.67±6.92	65.03±6.99	66.83±6.57	68.66±6.54
LARCM + Klosgen	78.34±6.42	76.00±6.20	66.89±6.64	73.38±6.25	76.41±6.14	63.90±6.42	64.99±6.86	68.62±6.46
LARCM + Lambda	74.19±6.96	72.88±5.24	63.16±7.77	69.35±5.65	75.81±6.33	61.26±7.01	62.99±6.75	64.34±7.05
LARCM + Laplace	79.72±6.05	71.40±6.83	71.05±7.19	70.90±6.67	75.14±6.50	57.97±6.98	61.17±6.93	62.35±6.88
LARCM + Lift	76.98±6.27	69.34±6.69	66.01±7.20	67.86±7.13	75.93±6.79	61.84±8.62	62.53±8.01	64.17±8.40
LARCM + Mutual Information LHS	78.09±6.72	73.24±7.22	64.97±8.00	72.76±7.00	76.27±6.21	60.57±7.45	63.89±6.74	66.16±6.49
LARCM + Novelty	76.28±6.98	74.21±6.55	70.32±6.54	71.65±6.44	76.66±6.34	63.91±6.90	64.94±7.62	66.59±7.52
LARCM + Odds Ratio	82.73±5.92	78.62±6.26	79.23±6.59	75.08±7.11	82.73±6.04	67.92±5.89	71.40±6.76	73.98±6.72

Tabela B.13: Acurácia dos classificadores para cada modelo de redução de dimensionalidade, com a quantidade de tópicos obtidos $k = 50$ na coleção de documentos ACM-2.

Método	NB	MNB	J48	SMO c=1	SMO c=10	IBk 3	IBk 5	IBk 7
LDA + bag-of-related-words	87.78±4.85	88.61±4.79	85.39±5.43	84.41±5.36	87.17±5.20	77.83±6.36	77.22±6.47	79.68±6.02
LDA + bag-of-words	94.93±3.09	95.29±3.01	85.83±5.17	94.00±3.33	94.37±3.23	93.49±3.36	92.98±3.77	93.02±3.54
LARCM + Added Value	83.15±5.33	89.00±4.81	82.88±5.63	88.29±4.48	89.73±4.21	79.80±5.57	81.61±5.57	82.85±5.49
LARCM + Certainty Factor	82.73±6.05	89.85±4.26	82.93±5.83	88.61±4.72	88.90±4.59	77.15±6.09	80.66±6.10	81.76±5.40
LARCM + Collective Strength	82.15±5.01	89.44±4.44	79.51±5.44	89.10±4.57	90.05±4.42	78.73±5.84	79.20±5.93	79.90±5.64
LARCM + Confiança	81.61±5.54	87.27±4.58	82.17±5.48	87.90±4.41	88.00±4.70	78.71±5.18	80.34±5.60	80.44±5.00
LARCM + Conviction	70.98±7.07	74.39±6.74	72.41±6.66	73.49±7.08	72.61±6.70	67.12±6.81	68.61±7.42	70.07±7.21
LARCM + ϕ -Coefficient	81.66±5.18	90.44±4.21	81.88±5.43	90.56±4.02	91.07±3.56	83.37±5.67	83.88±5.11	84.98±5.04
LARCM + Gini Index	84.39±5.52	88.98±4.90	79.46±5.61	88.85±4.55	89.56±5.00	80.37±6.29	81.66±6.63	81.76±6.15
LARCM + IS	81.49±5.41	87.88±4.78	80.12±6.65	87.34±5.25	89.20±4.05	79.80±4.99	81.63±5.48	81.73±5.34
LARCM + J-Measure	83.29±5.52	88.88±4.63	81.61±5.61	89.07±4.96	89.20±4.91	80.32±5.93	82.49±5.58	83.85±5.15
LARCM + Kappa	84.39±5.93	89.15±4.14	79.98±6.27	89.59±3.94	90.24±4.32	80.24±5.52	82.63±5.54	82.44±5.45
LARCM + Klosgen	84.12±5.36	87.90±4.99	81.20±5.39	87.32±4.77	89.78±5.10	80.68±5.98	80.51±5.83	79.71±5.79
LARCM + Lambda	82.17±6.03	86.68±4.85	79.17±6.47	87.07±4.90	86.46±5.19	78.12±6.28	79.66±5.90	80.02±6.40
LARCM + Laplace	83.22±6.23	87.02±4.97	79.61±6.10	86.83±5.29	87.61±5.08	79.66±6.23	80.68±6.23	80.85±6.40
LARCM + Lift	80.44±5.85	88.44±4.82	77.10±6.24	87.29±5.00	87.37±4.49	77.24±6.59	78.39±6.83	79.51±6.35
LARCM + Mutual Information LHS	82.66±5.40	87.44±4.78	81.51±5.25	86.73±4.77	87.49±4.95	79.05±5.79	81.20±5.37	81.85±5.31
LARCM + Novelty	82.66±6.07	87.46±4.87	84.39±5.54	86.63±5.10	88.32±4.48	79.15±6.01	81.32±5.81	81.68±5.48
LARCM + Odds Ratio	88.68±4.81	91.78±4.38	82.61±5.96	91.95±4.02	92.68±4.33	86.63±5.58	86.20±5.85	86.20±5.65

Tabela B.14: Acurácia dos classificadores para cada modelo de redução de dimensionalidade, com a quantidade de tópicos obtidos $k = 100$ na coleção de documentos ACM-2.

Método	NB	MNB	J48	SMO c=1	SMO c=10	IBk 3	IBk 5	IBk 7
LDA + bag-of-related-words	86.95±4.80	85.02±5.13	78.32±6.98	83.68±5.32	86.29±4.92	68.93±7.41	70.17±7.51	76.51±6.63
LDA + bag-of-words	88.05±4.38	88.15±4.70	79.15±5.75	86.15±4.95	86.66±4.98	79.46±5.95	79.83±6.22	77.41±5.88
LARCM + Added Value	85.17±4.65	89.24±3.83	82.27±6.17	88.93±3.63	89.88±4.11	77.46±5.97	80.12±5.57	81.56±5.26
LARCM + Certainty Factor	84.78±5.03	88.15±5.18	87.41±6.47	87.63±5.54	89.68±4.69	77.27±6.70	79.22±6.21	80.49±5.79
LARCM + Collective Strength	81.34±4.75	88.71±5.17	79.68±5.55	88.27±4.70	90.20±4.23	76.20±6.01	80.37±6.01	80.27±6.00
LARCM + Confiança	85.46±4.88	87.93±4.81	82.71±5.75	87.95±4.52	91.07±3.82	80.63±5.66	83.05±5.45	84.10±5.26
LARCM + Conviction	72.71±7.64	75.39±6.71	67.76±7.02	74.78±6.83	74.29±6.66	64.54±7.14	67.10±7.33	65.32±7.69
LARCM + ϕ -Coefficient	84.71±5.05	88.88±4.02	82.34±5.03	89.95±3.45	91.44±3.81	79.49±6.24	83.66±5.69	85.68±5.03
LARCM + Gini Index	85.63±5.07	88.80±4.19	82.34±5.46	88.66±3.99	89.90±4.23	77.27±6.18	79.90±5.26	82.39±4.92
LARCM + IS	84.68±5.45	89.02±4.92	83.29±5.41	88.24±5.19	88.71±4.30	78.46±6.31	79.54±6.49	81.24±6.01
LARCM + J-Measure	85.39±5.10	88.39±4.63	82.27±6.44	87.27±4.82	88.10±4.40	77.71±6.30	80.71±6.50	80.56±6.19
LARCM + Kappa	84.10±5.89	87.29±5.28	85.00±5.40	87.68±5.20	88.98±4.34	79.02±6.39	81.54±6.11	82.37±6.06
LARCM + Klosgen	81.83±6.30	88.76±4.17	87.24±6.21	87.17±4.47	88.78±4.03	79.93±5.52	81.05±5.82	80.29±5.47
LARCM + Lambda	83.24±4.89	87.17±4.77	79.80±5.61	86.54±5.02	88.22±5.02	77.34±5.65	77.85±6.37	78.41±6.23
LARCM + Laplace	82.78±5.48	88.10±4.77	82.78±5.76	86.66±5.05	88.02±4.99	78.34±6.54	79.85±6.36	81.54±6.45
LARCM + Lift	84.83±4.96	88.76±5.39	81.22±5.42	87.88±5.23	88.85±5.18	78.34±6.17	80.12±6.73	82.56±6.79
LARCM + Mutual Information LHS	83.76±5.70	88.88±4.64	82.49±5.47	87.78±5.19	89.71±4.53	79.15±5.45	81.39±5.22	81.56±4.76
LARCM + Novelty	84.61±5.09	86.59±5.61	81.07±5.84	86.56±5.71	88.49±4.92	79.17±5.53	81.46±5.56	81.44±5.57
LARCM + Odds Ratio	89.83±5.11	91.22±4.30	86.85±5.01	91.07±4.36	92.46±4.01	82.68±5.86	85.54±5.57	85.59±5.84

Tabela B.15: Acurácia dos classificadores para cada modelo de redução de dimensionalidade, com a quantidade de tópicos obtidos $k = 150$ na coleção de documentos ACM-2.

Método	NB	MNB	J48	SMO c=1	SMO c=10	IBk 3	IBk 5	IBk 7
malletborw-t01	79.51±5.95	81.27±5.55	65.00±7.31	80.02±5.55	81.51±5.55	70.05±6.63	66.46±7.72	61.63±7.07
LDA + <i>bag-of-words</i>	78.76±5.43	81.80±4.67	69.51±6.80	79.56±5.41	82.98±5.55	71.41±6.30	73.07±6.33	67.95±6.86
LARCM + Added Value	86.02±5.13	89.12±4.57	82.32±5.73	87.73±4.88	90.49±4.17	79.17±5.71	81.93±5.87	83.24±5.71
LARCM + Certainty Factor	86.61±5.24	88.17±5.15	87.22±4.81	86.78±5.25	88.80±4.47	79.24±5.91	82.29±5.47	83.05±5.58
LARCM + Collective Strength	85.39±5.58	87.07±5.27	83.07±5.87	84.78±5.60	87.07±5.70	76.93±6.12	78.76±6.22	80.27±6.02
LARCM + Confiança	83.66±6.23	87.80±4.84	83.66±5.53	86.22±5.09	90.10±4.71	76.22±5.60	78.20±6.69	79.12±5.99
LARCM + Conviction	67.59±6.93	74.29±6.36	65.85±7.23	72.66±5.91	72.95±6.73	62.93±7.65	65.10±6.67	66.32±6.71
LARCM + ϕ -Coefficient	87.10±4.71	88.98±5.11	83.66±5.73	88.34±5.14	90.07±4.36	78.41±5.59	80.56±5.73	81.51±5.18
LARCM + Gini Index	88.10±5.31	88.88±4.80	85.46±5.08	86.80±4.98	90.10±4.21	79.12±5.86	79.85±5.29	82.07±5.66
LARCM + IS	85.41±4.62	87.10±4.18	86.80±4.92	85.24±5.06	87.63±4.60	77.61±6.21	78.61±6.32	79.80±6.19
LARCM + J-Measure	85.37±5.63	88.05±5.25	84.98±5.84	86.24±5.15	88.44±4.68	77.07±6.28	79.49±6.32	81.22±6.42
LARCM + Kappa	83.56±5.46	89.05±4.04	83.12±5.72	87.71±4.76	90.02±4.41	77.46±5.19	81.39±5.59	82.66±5.52
LARCM + Klosgen	85.46±5.37	88.44±4.51	83.95±5.14	87.93±4.87	89.66±4.23	76.85±6.02	79.22±6.07	80.85±5.89
LARCM + Lambda	86.20±4.88	87.27±4.87	78.90±5.93	85.15±5.04	85.32±5.85	76.05±6.81	77.88±6.22	77.85±6.08
LARCM + Laplace	85.22±5.06	87.93±4.78	85.68±5.18	85.63±5.11	87.73±4.32	79.07±5.34	82.15±4.97	83.02±5.08
LARCM + Lift	84.61±6.07	88.85±4.94	80.95±6.32	86.44±4.96	88.78±4.40	76.00±6.63	78.95±5.93	79.78±6.23
LARCM + Mutual Information LHS	84.78±5.41	88.37±5.02	81.41±5.47	87.39±4.98	87.00±5.01	80.27±7.21	83.73±6.28	83.76±6.07
LARCM + Novelty	86.00±4.67	87.76±4.62	82.59±5.89	87.02±4.75	90.49±4.34	80.85±6.30	82.56±5.76	84.15±5.48
LARCM + Odds Ratio	90.93±4.44	91.20±4.20	86.07±5.41	90.76±4.14	93.00±3.78	85.49±5.39	87.12±4.77	88.39±4.53

Tabela B.16: Acurácia dos classificadores para cada modelo de redução de dimensionalidade, com a quantidade de tópicos obtidos $k = 50$ na coleção de documentos ACM-3.

Método	NB	MNB	J48	SMO c=1	SMO c=10	IBk 3	IBk 5	IBk 7
malletborw-t01	86.39±4.83	84.66±4.88	84.37±4.99	80.02±5.41	83.01±5.02	73.86±6.40	74.90±5.72	77.45±5.64
LDA + <i>bag-of-words</i>	91.19±4.17	92.07±4.04	89.20±4.69	91.73±4.22	91.20±4.29	87.12±4.47	89.02±4.73	88.48±4.26
LARCM + Added Value	81.06±6.13	82.70±5.42	71.64±5.79	83.66±5.07	84.19±5.24	76.80±6.39	77.84±5.91	77.78±6.00
LARCM + Certainty Factor	83.31±5.89	84.47±5.77	75.41±5.95	86.48±5.56	86.11±4.91	77.20±6.00	78.21±6.07	79.95±6.22
LARCM + Collective Strength	80.87±5.52	83.43±5.17	73.51±6.40	84.34±5.17	86.35±4.97	75.82±5.81	76.76±5.91	77.65±5.85
LARCM + Confiança	80.69±5.24	83.11±4.96	73.79±6.49	84.83±4.46	85.03±4.66	76.42±5.70	78.17±4.87	79.16±4.90
LARCM + Conviction	68.14±6.94	75.73±6.70	65.41±7.89	74.64±6.98	74.16±6.80	70.27±7.08	70.15±7.13	69.88±6.85
LARCM + ϕ -Coefficient	82.34±5.40	85.38±4.83	74.61±7.03	86.53±5.14	86.32±4.55	81.05±5.50	82.78±5.71	82.38±5.61
LARCM + Gini Index	82.22±5.68	85.75±5.07	74.78±6.69	87.29±4.59	86.92±4.45	77.06±5.66	79.00±4.97	78.66±5.63
LARCM + IS	82.33±5.49	83.61±5.86	75.26±5.44	85.22±5.47	85.46±5.52	76.00±6.20	77.47±6.26	78.78±6.25
LARCM + J-Measure	81.43±5.37	84.31±5.14	73.83±6.48	85.17±4.56	84.45±5.28	79.55±5.85	80.28±5.61	80.52±5.96
LARCM + Kappa	80.02±5.93	82.26±4.69	76.09±6.28	84.23±5.09	85.56±4.86	76.54±6.22	77.06±5.59	77.72±5.62
LARCM + Klosgen	82.31±5.53	85.07±5.59	73.23±6.70	86.58±5.13	86.59±4.52	80.08±5.80	80.38±5.20	81.06±5.27
LARCM + Lambda	81.52±5.58	83.66±5.42	72.26±7.07	85.39±4.58	83.47±5.06	74.52±5.57	76.51±6.04	76.72±5.74
LARCM + Laplace	83.57±5.53	86.41±4.71	74.87±6.62	85.89±5.12	86.81±4.97	79.13±5.65	80.05±5.32	80.84±5.18
LARCM + Lift	80.75±5.86	84.34±5.86	75.60±7.41	86.06±5.41	86.44±5.11	75.67±6.32	78.67±6.54	79.88±6.38
LARCM + Mutual Information LHS	82.76±5.92	85.49±4.82	74.21±5.87	85.54±5.13	85.90±5.81	77.83±5.72	78.67±5.21	79.52±5.39
LARCM + Novelty	83.57±6.30	84.33±5.46	75.14±6.73	85.26±5.49	87.30±5.06	79.06±6.34	79.76±6.93	80.60±6.57
LARCM + Odds Ratio	86.75±4.73	85.36±5.38	80.76±5.72	87.65±4.81	90.34±4.43	86.06±5.78	86.90±5.46	85.89±5.41

Tabela B.17: Acurácia dos classificadores para cada modelo de redução de dimensionalidade, com a quantidade de tópicos obtidos $k = 100$ na coleção de documentos ACM-3.

Método	NB	MNB	J48	SMO c=1	SMO c=10	IBk 3	IBk 5	IBk 7
malletborw-t01	84.19±4.62	82.29±4.79	80.73±5.99	81.66±4.92	84.15±4.51	70.36±5.30	72.04±5.96	74.18±6.06
LDA + <i>bag-of-words</i>	84.13±5.23	87.17±5.18	77.40±5.72	85.00±5.17	84.96±5.03	82.11±5.25	82.99±5.18	81.20±5.36
LARCM + Added Value	83.05±6.19	82.50±5.64	80.31±5.72	86.98±5.16	87.55±4.98	79.19±6.71	78.90±6.70	80.12±6.75
LARCM + Certainty Factor	84.71±5.81	83.81±5.37	74.32±6.31	87.05±4.71	85.60±5.32	77.29±6.05	79.84±5.69	80.33±5.97
LARCM + Collective Strength	85.45±5.04	84.40±4.65	76.80±6.29	87.14±4.60	87.61±4.17	78.74±6.15	80.65±5.11	82.06±4.79
LARCM + Confiança	84.24±5.42	82.72±6.00	79.30±5.66	85.03±5.04	85.32±5.04	76.76±6.39	78.18±6.45	78.28±6.29
LARCM + Conviction	66.17±6.64	73.51±7.37	63.61±6.69	71.79±7.11	71.81±7.49	68.65±7.87	70.47±7.51	70.64±7.16
LARCM + ϕ -Coefficient	84.45±4.82	85.37±4.79	79.47±6.33	87.43±4.92	87.65±4.67	80.05±6.24	81.40±6.04	81.74±6.28
LARCM + Gini Index	84.38±5.28	82.59±5.48	77.64±5.56	87.08±4.60	86.84±5.01	77.68±6.07	79.28±5.81	79.18±5.93
LARCM + IS	82.48±5.23	81.86±5.65	77.39±5.79	85.44±4.67	83.42±4.81	75.26±6.01	77.66±5.54	78.52±5.52
LARCM + J-Measure	84.16±4.53	81.88±5.36	78.16±6.26	86.01±4.74	86.51±4.18	74.92±6.46	76.60±6.11	79.15±5.60
LARCM + Kappa	86.87±4.97	82.53±5.35	76.76±6.67	85.61±5.07	86.69±4.94	77.27±6.45	78.54±5.71	77.77±5.59
LARCM + Klosgen	84.43±5.41	83.53±5.28	75.05±5.77	85.84±5.05	84.88±5.15	75.68±6.06	79.67±5.46	79.54±6.34
LARCM + Lambda	82.09±5.10	82.76±5.21	74.69±6.51	85.94±5.03	85.05±5.55	72.57±6.45	74.92±6.36	76.69±6.44
LARCM + Laplace	84.62±4.87	82.67±5.46	76.13±6.12	86.41±4.95	86.97±5.48	75.34±6.45	78.42±6.53	78.99±6.46
LARCM + Lift	83.85±5.28	81.11±5.04	75.41±6.61	86.18±4.94	85.72±4.55	73.39±6.03	76.25±6.28	76.37±6.11
LARCM + Mutual Information LHS	82.52±5.50	83.00±5.17	76.07±5.77	84.88±5.29	84.92±5.80	73.57±5.37	74.87±6.03	76.43±5.72
LARCM + Novelty	86.47±4.72	84.16±5.12	75.92±6.74	86.08±4.91	87.38±5.32	78.25±5.47	78.42±5.41	80.03±5.32
LARCM + Odds Ratio	88.99±4.25	86.09±5.07	82.15±5.06	89.19±4.63	91.11±4.43	83.23±5.02	86.07±4.97	86.33±4.73

Tabela B.18: Acurácia dos classificadores para cada modelo de redução de dimensionalidade, com a quantidade de tópicos obtidos $k = 150$ na coleção de documentos ACM-3.

Método	NB	MNB	J48	SMO c=1	SMO c=10	IBk 3	IBk 5	IBk 7
malletborw-t01	76.95 \pm 6.25	78.36 \pm 6.64	67.85 \pm 6.46	78.27 \pm 6.48	81.06 \pm 6.18	66.61 \pm 7.89	66.42 \pm 8.31	58.70 \pm 7.90
LDA + bag-of-words	77.33 \pm 5.61	80.68 \pm 5.56	72.79 \pm 5.42	65.29 \pm 6.53	81.32 \pm 5.78	71.02 \pm 6.92	76.68 \pm 5.22	73.98 \pm 6.21
LARCM + Added Value	83.73 \pm 5.56	80.51 \pm 6.12	80.27 \pm 6.44	83.92 \pm 5.07	85.41 \pm 5.38	74.21 \pm 6.70	74.95 \pm 6.48	76.40 \pm 6.09
LARCM + Certainty Factor	84.92 \pm 5.90	81.08 \pm 5.96	79.33 \pm 5.67	83.94 \pm 5.88	86.27 \pm 5.96	75.48 \pm 6.25	76.25 \pm 6.62	76.89 \pm 6.00
LARCM + Collective Strength	83.13 \pm 5.72	81.42 \pm 5.87	75.85 \pm 5.59	85.49 \pm 5.46	85.72 \pm 5.79	75.72 \pm 5.26	77.15 \pm 5.39	77.24 \pm 5.18
LARCM + Confiança	84.86 \pm 5.62	81.81 \pm 5.37	81.05 \pm 7.08	84.53 \pm 5.56	85.73 \pm 5.36	77.22 \pm 5.85	79.14 \pm 5.58	79.16 \pm 5.73
LARCM + Conviction	65.10 \pm 6.81	72.89 \pm 6.07	62.00 \pm 7.29	72.91 \pm 5.78	72.43 \pm 5.93	62.55 \pm 7.58	61.42 \pm 7.74	59.50 \pm 7.50
LARCM + ϕ -Coefficient	86.84 \pm 5.19	82.30 \pm 6.41	79.70 \pm 6.35	87.45 \pm 5.43	87.17 \pm 4.77	80.96 \pm 6.08	81.89 \pm 6.32	82.63 \pm 5.94
LARCM + Gini Index	84.68 \pm 5.99	80.35 \pm 6.20	79.22 \pm 6.96	85.57 \pm 5.88	86.87 \pm 5.19	74.60 \pm 6.11	76.86 \pm 6.66	77.34 \pm 6.96
LARCM + IS	84.67 \pm 5.42	79.69 \pm 5.61	78.44 \pm 6.29	84.98 \pm 4.99	85.07 \pm 4.71	75.94 \pm 6.78	79.16 \pm 5.52	80.05 \pm 5.70
LARCM + J-Measure	83.58 \pm 5.45	81.55 \pm 5.71	78.40 \pm 6.11	84.60 \pm 5.79	86.33 \pm 4.94	75.33 \pm 6.45	78.37 \pm 5.55	79.32 \pm 5.80
LARCM + Kappa	83.67 \pm 4.93	81.95 \pm 5.37	77.69 \pm 6.55	86.08 \pm 5.31	87.16 \pm 4.72	78.16 \pm 6.19	80.14 \pm 5.52	80.06 \pm 5.48
LARCM + Klogen	83.90 \pm 5.27	80.82 \pm 5.37	76.51 \pm 5.71	85.03 \pm 5.40	86.81 \pm 5.03	76.25 \pm 6.23	76.98 \pm 6.55	78.93 \pm 6.29
LARCM + Lambda	82.08 \pm 5.53	81.45 \pm 5.72	72.50 \pm 7.01	85.78 \pm 5.61	85.43 \pm 5.58	72.85 \pm 6.60	72.87 \pm 6.87	75.74 \pm 7.41
LARCM + Laplace	87.09 \pm 4.89	82.44 \pm 4.88	79.38 \pm 6.20	84.12 \pm 5.31	86.25 \pm 4.89	74.93 \pm 6.02	78.40 \pm 5.74	78.21 \pm 5.71
LARCM + Lift	84.14 \pm 4.65	80.77 \pm 5.25	76.36 \pm 6.86	85.86 \pm 4.81	86.51 \pm 4.85	76.54 \pm 6.54	78.53 \pm 6.05	79.21 \pm 5.91
LARCM + Mutual Information LHS	85.76 \pm 5.22	78.48 \pm 6.33	75.91 \pm 5.98	85.01 \pm 6.05	85.07 \pm 5.83	75.74 \pm 7.29	76.73 \pm 6.97	78.06 \pm 6.92
LARCM + Novelty	86.91 \pm 4.71	81.91 \pm 6.05	80.07 \pm 5.54	84.66 \pm 5.59	86.48 \pm 5.21	78.11 \pm 6.18	79.07 \pm 5.69	77.60 \pm 5.81
LARCM + Odds Ratio	91.11 \pm 4.26	83.81 \pm 5.69	85.69 \pm 5.48	88.02 \pm 4.90	92.64 \pm 4.09	84.03 \pm 6.44	85.96 \pm 5.51	85.57 \pm 5.36

Tabela B.19: Acurácia dos classificadores para cada modelo de redução de dimensionalidade, com a quantidade de tópicos obtidos $k = 50$ na coleção de documentos ACM-4.

Método	NB	MNB	J48	SMO c=1	SMO c=10	IBk 3	IBk 5	IBk 7
malletborw-t01	93.56 \pm 3.57	91.59 \pm 4.13	91.78 \pm 4.56	90.78 \pm 4.27	93.68 \pm 3.98	83.82 \pm 5.12	84.84 \pm 4.98	85.63 \pm 5.30
LDA + bag-of-words	94.85 \pm 3.69	91.87 \pm 3.79	92.15 \pm 4.40	95.58 \pm 3.13	94.82 \pm 3.16	92.41 \pm 4.00	92.26 \pm 4.14	93.45 \pm 3.81
LARCM + Added Value	89.19 \pm 4.15	84.61 \pm 4.12	85.20 \pm 5.20	88.22 \pm 3.96	91.80 \pm 4.13	84.95 \pm 4.70	86.42 \pm 4.58	87.14 \pm 4.72
LARCM + Certainty Factor	87.88 \pm 4.29	85.19 \pm 3.88	82.58 \pm 6.06	88.33 \pm 4.81	91.28 \pm 3.93	83.91 \pm 5.34	86.60 \pm 5.10	87.39 \pm 4.74
LARCM + Collective Strength	89.87 \pm 4.15	85.48 \pm 3.59	84.41 \pm 5.78	89.36 \pm 4.20	92.56 \pm 3.41	87.69 \pm 4.36	88.14 \pm 4.23	88.37 \pm 4.46
LARCM + Confiança	88.91 \pm 4.93	85.11 \pm 4.30	87.67 \pm 4.79	88.56 \pm 4.86	90.79 \pm 4.56	85.19 \pm 6.13	86.79 \pm 5.90	86.89 \pm 5.94
LARCM + Conviction	76.48 \pm 6.97	79.87 \pm 5.68	73.81 \pm 6.68	82.57 \pm 6.38	83.12 \pm 5.80	78.25 \pm 6.56	79.95 \pm 6.25	80.39 \pm 6.10
LARCM + ϕ -Coefficient	88.17 \pm 4.06	86.57 \pm 3.84	81.80 \pm 4.83	90.61 \pm 4.19	91.82 \pm 4.28	87.20 \pm 4.78	88.07 \pm 4.94	88.32 \pm 4.78
LARCM + Gini Index	86.13 \pm 5.11	85.91 \pm 3.93	83.31 \pm 4.80	89.54 \pm 4.22	91.69 \pm 4.24	86.21 \pm 5.25	86.59 \pm 4.92	87.12 \pm 5.05
LARCM + IS	87.52 \pm 5.70	83.83 \pm 4.61	83.75 \pm 5.54	87.61 \pm 4.84	90.69 \pm 4.34	85.10 \pm 5.23	85.43 \pm 5.38	85.10 \pm 5.46
LARCM + J-Measure	88.17 \pm 4.30	86.40 \pm 3.94	83.76 \pm 5.16	89.87 \pm 4.16	92.11 \pm 3.82	86.44 \pm 4.82	87.00 \pm 4.84	87.05 \pm 4.93
LARCM + Kappa	88.35 \pm 4.82	84.03 \pm 4.56	85.46 \pm 5.61	86.62 \pm 4.65	90.77 \pm 4.63	86.96 \pm 4.95	85.48 \pm 4.83	86.17 \pm 4.83
LARCM + Klogen	85.26 \pm 5.33	85.99 \pm 3.96	85.32 \pm 4.99	88.99 \pm 4.54	90.79 \pm 4.10	84.29 \pm 5.58	85.51 \pm 5.90	85.50 \pm 5.65
LARCM + Lambda	83.95 \pm 4.95	84.55 \pm 3.92	80.80 \pm 5.05	87.74 \pm 4.13	89.27 \pm 4.40	85.91 \pm 4.09	86.30 \pm 4.59	86.25 \pm 4.81
LARCM + Laplace	86.65 \pm 4.78	84.34 \pm 4.24	86.14 \pm 4.59	88.43 \pm 4.54	91.63 \pm 3.90	86.40 \pm 5.44	86.87 \pm 5.19	86.73 \pm 5.57
LARCM + Lift	88.28 \pm 4.25	84.11 \pm 4.60	82.31 \pm 4.94	88.74 \pm 4.46	91.55 \pm 3.84	85.76 \pm 4.76	86.54 \pm 4.86	86.98 \pm 4.76
LARCM + Mutual Information LHS	88.45 \pm 4.52	84.71 \pm 4.49	86.25 \pm 5.39	88.22 \pm 5.00	89.32 \pm 4.43	86.57 \pm 5.06	86.52 \pm 4.71	86.14 \pm 5.23
LARCM + Novelty	89.97 \pm 4.74	85.81 \pm 4.62	83.48 \pm 6.05	89.44 \pm 4.73	91.53 \pm 4.12	85.61 \pm 5.31	87.11 \pm 5.27	87.31 \pm 5.11
LARCM + Odds Ratio	92.18 \pm 4.39	89.62 \pm 3.83	89.60 \pm 4.30	94.26 \pm 3.32	94.74 \pm 3.19	92.23 \pm 3.76	92.74 \pm 3.69	92.96 \pm 3.72

Tabela B.20: Acurácia dos classificadores para cada modelo de redução de dimensionalidade, com a quantidade de tópicos obtidos $k = 100$ na coleção de documentos ACM-4.

Método	NB	MNB	J48	SMO c=1	SMO c=10	IBk 3	IBk 5	IBk 7
malletborw-t01	89.52 \pm 4.78	86.75 \pm 5.18	89.96 \pm 4.71	89.12 \pm 4.99	93.28 \pm 3.73	80.79 \pm 6.02	82.95 \pm 5.97	83.76 \pm 5.49
LDA + bag-of-words	90.68 \pm 4.39	81.92 \pm 3.94	83.66 \pm 5.36	87.78 \pm 3.99	91.40 \pm 3.76	86.52 \pm 4.53	87.73 \pm 5.25	86.46 \pm 5.62
LARCM + Added Value	88.35 \pm 4.67	82.28 \pm 3.86	85.36 \pm 5.40	84.21 \pm 4.63	90.10 \pm 3.86	82.07 \pm 5.79	84.81 \pm 5.68	84.99 \pm 5.11
LARCM + Certainty Factor	90.30 \pm 4.51	83.17 \pm 4.52	87.68 \pm 5.25	86.30 \pm 4.98	89.83 \pm 4.33	83.55 \pm 5.50	84.87 \pm 5.03	86.07 \pm 4.99
LARCM + Collective Strength	88.47 \pm 4.95	83.00 \pm 3.62	85.00 \pm 5.64	87.14 \pm 4.49	90.03 \pm 3.98	86.22 \pm 5.29	85.43 \pm 5.43	85.73 \pm 5.44
LARCM + Confiança	90.36 \pm 4.12	82.79 \pm 4.50	90.02 \pm 4.32	88.51 \pm 4.23	91.12 \pm 4.20	84.22 \pm 5.24	85.18 \pm 5.62	85.57 \pm 5.25
LARCM + Conviction	73.41 \pm 5.86	78.07 \pm 5.34	73.98 \pm 7.05	77.38 \pm 5.81	80.76 \pm 6.41	74.94 \pm 6.28	78.83 \pm 5.91	79.71 \pm 5.79
LARCM + ϕ -Coefficient	91.83 \pm 4.05	84.67 \pm 2.83	87.73 \pm 5.27	87.88 \pm 4.60	91.83 \pm 4.14	86.07 \pm 4.79	87.64 \pm 5.04	86.75 \pm 4.64
LARCM + Gini Index	91.02 \pm 4.36	83.72 \pm 3.95	85.30 \pm 4.91	87.41 \pm 4.27	91.09 \pm 4.18	85.71 \pm 5.08	85.66 \pm 4.46	85.66 \pm 5.01
LARCM + IS	90.73 \pm 4.17	84.06 \pm 3.58	86.04 \pm 4.67	86.35 \pm 4.60	90.46 \pm 4.52	84.09 \pm 5.30	83.83 \pm 5.06	84.06 \pm 4.99
LARCM + J-Measure	89.03 \pm 4.35	83.75 \pm 3.90	87.24 \pm 5.25	87.67 \pm 4.36	90.58 \pm 4.04	82.92 \pm 5.27	82.13 \pm 5.04	83.45 \pm 4.79
LARCM + Kappa	90.09 \pm 4.72	83.75 \pm 4.03	85.85 \pm 5.20	86.07 \pm 4.60	90.89 \pm 5.01	81.45 \pm 5.07	82.08 \pm 4.57	82.79 \pm 5.34
LARCM + Klogen	90.12 \pm 4.42	82.72 \pm 3.32	89.00 \pm 4.65	85.05 \pm 4.21	90.88 \pm 4.32	82.52 \pm 5.45	85.40 \pm 5.40	86.65 \pm 5.04
LARCM + Lambda	86.62 \pm 5.17	82.79 \pm 4.34	84.66 \pm 5.83	85.62 \pm 5.25	90.88 \pm 5.18	82.92 \pm 5.98	85.59 \pm 5.24	85.55 \pm 5.22
LARCM + Laplace	89.90 \pm 3.84	83.10 \pm 4.09	87.82 \pm 4.23	87.60 \pm 4.77	90.77 \pm 4.49	85.74 \pm 5.16	86.45 \pm 5.25	86.80 \pm 5.34
LARCM + Lift	89.54 \pm 4.38	82.26 \pm 4.28	86.81 \pm 5.02	87.79 \pm 4.38	90.94<math			

Tabela B.21: Acurácia dos classificadores para cada modelo de redução de dimensionalidade, com a quantidade de tópicos obtidos $k = 150$ na coleção de documentos ACM-4.

Método	NB	MNB	J48	SMO c=1	SMO c=10	IBk 3	IBk 5	IBk 7
malletborw-t01	85.26±5.60	79.36±4.93	80.09±5.71	85.07±5.35	88.02±4.83	76.61±6.54	70.73±6.09	66.22±6.56
LDA + <i>bag-of-words</i>	89.23±4.53	73.68±5.09	80.72±6.36	81.51±5.46	88.23±5.07	85.60±5.47	84.10±5.68	82.68±5.62
LARCM + Added Value	91.73±4.29	79.81±4.14	88.47±4.86	81.97±4.49	90.55±4.71	86.52±5.44	87.65±5.53	87.71±5.10
LARCM + Certainty Factor	89.90±5.01	81.98±4.15	84.52±6.20	85.89±4.92	91.38±3.87	86.14±5.18	86.09±5.59	87.10±4.69
LARCM + Collective Strength	92.06±4.35	81.34±3.64	85.16±4.75	81.98±4.16	92.54±3.57	86.27±4.89	87.11±5.05	86.71±5.25
LARCM + Confiança	90.38±3.80	80.92±3.78	87.94±4.88	84.50±4.79	90.71±4.44	83.73±4.67	84.92±4.59	84.75±4.65
LARCM + Conviction	71.14±6.78	76.73±4.68	72.01±6.84	76.80±5.20	80.57±6.29	74.96±6.35	73.52±6.82	71.44±7.65
LARCM + ϕ -Coefficient	91.27±4.55	81.93±3.77	87.64±4.63	86.14±4.61	91.01±4.08	84.04±5.82	84.41±5.05	84.61±4.77
LARCM + Gini Index	92.17±4.19	80.96±4.01	86.85±4.99	81.85±4.60	90.26±4.19	82.60±5.69	81.86±5.56	82.55±5.51
LARCM + IS	91.50±4.42	81.95±3.32	83.85±5.23	83.32±4.70	91.30±4.74	82.88±6.41	84.00±5.23	85.33±4.83
LARCM + J-Measure	90.51±4.00	80.86±3.53	86.60±5.03	81.55±3.85	91.25±4.13	78.24±5.61	80.51±5.18	82.33±5.11
LARCM + Kappa	90.53±4.29	81.34±3.59	87.82±4.79	81.37±3.97	90.23±4.26	84.24±4.98	84.38±5.24	84.56±4.73
LARCM + Klosgen	91.14±4.07	79.98±3.68	86.52±5.00	82.59±3.63	90.71±4.33	85.89±5.32	86.39±5.22	86.90±5.21
LARCM + Lambda	88.88±4.57	81.07±4.08	83.51±5.50	82.22±5.08	88.13±4.63	80.30±5.91	82.87±5.36	83.81±5.16
LARCM + Laplace	91.75±4.42	81.32±3.69	85.66±4.92	83.30±4.50	90.18±4.45	83.86±5.19	83.49±5.19	82.81±5.45
LARCM + Lift	91.06±4.18	79.92±4.16	82.57±5.52	81.16±4.14	89.36±4.86	79.46±5.30	81.95±5.29	83.19±4.65
LARCM + Mutual Information LHS	89.67±4.16	81.71±3.67	86.17±5.18	83.33±4.11	90.43±4.42	83.99±4.96	84.57±5.69	85.20±5.24
LARCM + Novelty	89.34±4.59	81.01±3.69	87.66±4.93	81.79±3.94	89.34±3.80	82.18±5.54	84.01±5.12	83.73±5.34
LARCM + Odds Ratio	94.26±3.29	83.57±3.53	91.07±4.39	88.39±4.32	94.49±3.40	91.85±4.39	91.93±4.46	91.30±4.64

Tabela B.22: Acurácia dos classificadores para cada modelo de redução de dimensionalidade, com a quantidade de tópicos obtidos $k = 50$ na coleção de documentos ACM-5.

Método	NB	MNB	J48	SMO c=1	SMO c=10	IBk 3	IBk 5	IBk 7
malletborw-t01	81.67±5.21	85.80±5.29	85.22±4.80	84.14±6.02	87.33±4.95	78.61±5.85	77.19±6.30	77.21±6.41
LDA + <i>bag-of-words</i>	86.62±4.64	90.13±4.33	88.09±4.28	89.72±4.15	90.59±3.94	83.31±4.79	83.86±5.05	84.66±4.52
LARCM + Added Value	72.18±6.61	76.45±4.80	69.62±5.06	77.25±5.08	77.89±5.17	67.91±6.60	69.85±6.26	71.65±5.81
LARCM + Certainty Factor	73.80±5.43	76.48±5.81	68.75±5.74	78.62±5.41	78.15±4.91	71.36±5.77	73.21±5.48	73.74±5.18
LARCM + Collective Strength	72.44±5.25	76.75±5.64	66.99±6.37	77.64±5.84	77.56±5.63	70.78±5.79	73.62±5.89	73.43±5.43
LARCM + Confiança	74.33±5.58	77.22±4.72	71.09±6.35	79.39±4.96	78.98±4.93	72.91±6.21	73.74±5.97	73.89±6.21
LARCM + Conviction	62.38±6.14	71.66±6.01	60.44±6.69	70.83±6.10	70.34±6.74	65.89±5.68	66.71±5.49	68.26±5.80
LARCM + ϕ -Coefficient	74.03±5.93	79.08±4.98	70.17±6.00	79.89±5.53	80.59±6.13	73.25±5.58	75.07±5.77	75.58±6.01
LARCM + Gini Index	72.23±6.15	77.41±5.50	68.94±6.28	79.36±5.82	78.68±5.62	73.23±6.91	74.60±6.47	75.39±6.64
LARCM + IS	72.46±5.83	77.84±5.74	66.57±6.56	77.86±5.51	77.97±5.49	71.02±6.02	73.12±5.15	73.70±5.60
LARCM + J-Measure	70.65±5.53	76.87±5.19	68.83±6.41	77.62±5.37	77.19±5.44	72.25±6.09	73.37±6.38	73.80±5.58
LARCM + Kappa	70.65±6.51	76.35±5.73	67.17±6.72	77.75±5.84	76.07±6.17	71.51±6.11	73.72±5.50	73.12±5.79
LARCM + Klosgen	72.72±5.18	76.73±5.05	67.83±6.87	77.64±5.54	77.90±6.36	69.79±6.44	71.25±5.89	71.82±6.10
LARCM + Lambda	72.66±6.25	78.22±5.50	68.66±6.59	78.54±5.62	78.20±5.66	71.47±6.10	72.77±6.23	74.10±6.51
LARCM + Laplace	73.13±6.23	77.08±4.92	70.24±5.86	77.71±4.88	76.65±5.74	70.72±5.51	72.21±5.47	74.38±5.36
LARCM + Lift	73.72±5.38	78.20±4.99	67.84±5.29	78.63±4.90	79.13±5.39	75.33±5.59	74.95±5.36	76.01±5.19
LARCM + Mutual Information LHS	72.06±6.08	76.90±5.38	67.62±5.64	76.80±5.86	78.11±5.65	72.13±6.16	72.27±6.14	73.50±6.36
LARCM + Novelty	72.93±5.73	76.82±5.32	72.99±6.30	77.78±5.28	78.27±5.58	73.78±5.62	74.02±5.68	74.36±6.08
LARCM + Odds Ratio	78.88±4.70	81.60±4.79	74.00±5.81	82.19±4.35	84.44±4.75	78.75±5.29	80.28±4.28	81.22±4.40

Tabela B.23: Acurácia dos classificadores para cada modelo de redução de dimensionalidade, com a quantidade de tópicos obtidos $k = 100$ na coleção de documentos ACM-5.

Método	NB	MNB	J48	SMO c=1	SMO c=10	IBk 3	IBk 5	IBk 7
malletborw-t01	80.68±5.28	77.81±4.97	80.59±5.78	79.94±5.53	82.18±5.58	64.56±6.07	62.06±6.64	58.53±5.93
LDA + <i>bag-of-words</i>	86.56±4.85	85.81±4.56	85.90±4.94	86.43±4.84	88.68±4.58	79.15±5.87	79.85±5.47	79.24±5.18
LARCM + Added Value	75.54±5.93	78.51±4.71	71.45±5.94	76.88±5.06	78.21±5.51	71.09±5.91	72.65±5.65	73.27±6.15
LARCM + Certainty Factor	76.90±4.90	78.71±4.84	71.96±6.25	78.89±5.53	80.06±5.20	73.97±6.01	75.12±6.39	76.56±5.58
LARCM + Collective Strength	73.90±6.59	77.43±4.95	68.58±6.50	77.98±5.35	77.92±6.02	70.17±6.00	73.59±6.18	73.89±6.87
LARCM + Confiança	74.72±6.47	78.18±5.33	71.66±5.82	78.15±5.28	79.30±5.47	71.78±5.67	72.14±5.82	71.91±5.85
LARCM + Conviction	59.51±6.18	70.87±5.67	60.78±6.31	70.74±6.66	71.08±6.28	64.97±6.63	65.90±6.87	67.62±6.21
LARCM + ϕ -Coefficient	77.66±6.11	77.35±5.23	71.05±5.40	78.28±5.11	79.32±5.57	71.57±6.09	74.01±5.30	75.16±5.41
LARCM + Gini Index	74.37±5.35	77.63±5.00	70.68±6.05	77.35±6.02	78.96±5.66	74.26±5.35	73.52±5.10	74.52±5.54
LARCM + IS	72.06±5.48	77.33±4.89	69.39±5.98	76.94±4.95	77.90±5.55	70.38±6.32	71.15±5.87	72.70±6.53
LARCM + J-Measure	74.37±5.61	78.32±4.96	70.91±6.95	77.26±5.46	78.43±5.23	70.48±6.17	72.56±6.12	72.80±5.49
LARCM + Kappa	75.20±5.32	78.12±5.28	68.83±6.19	77.73±5.55	77.65±6.14	68.30±5.88	70.47±5.31	71.41±5.72
LARCM + Klosgen	76.15±5.72	77.78±5.24	72.82±5.64	77.48±5.10	77.22±5.13	70.55±6.21	72.99±5.54	73.61±5.57
LARCM + Lambda	73.01±6.76	78.98±4.97	67.81±6.83	78.79±5.04	79.43±5.84	69.51±6.78	70.35±6.78	70.34±6.60
LARCM + Laplace	76.20±6.03	77.85±5.02	72.76±6.04	78.02±5.37	78.04±5.80	70.40±6.25	71.21±6.57	73.65±5.69
LARCM + Lift	73.57±6.27	77.50±4.97	67.60±6.91	79.07±5.23	79.26±5.30	71.07±6.24	72.82±5.64	73.23±5.28
LARCM + Mutual Information LHS	74.78±5.83	77.62±5.00	70.15±6.14	77.32±5.40	78.60±5.30	70.36±6.11	72.39±5.78	73.95±6.03
LARCM + Novelty	74.82±5.48	78.28±5.42	70.90±5.74	78.19±5.60	79.81±5.43	71.32±6.47	72.51±6.29	73.82±5.98
LARCM + Odds Ratio	83.38±5.10	80.49±5.30	77.73±5.41	82.06±5.39	84.82±5.10	78.84±5.68	80.91±5.21	81.79±5.68

Tabela B.24: Acurácia dos classificadores para cada modelo de redução de dimensionalidade, com a quantidade de tópicos obtidos $k = 150$ na coleção de documentos ACM-5.

Método	NB	MNB	J48	SMO c=1	SMO c=10	IBk 3	IBk 5	IBk 7
malletborw-t01	76.45 \pm 5.71	77.37 \pm 5.91	78.87 \pm 5.95	76.44 \pm 5.23	80.75 \pm 6.06	57.47 \pm 6.95	58.78 \pm 7.79	62.42 \pm 6.02
LDA + bag-of-words	84.25 \pm 4.72	80.90 \pm 4.72	76.75 \pm 5.53	71.30 \pm 5.81	82.66 \pm 5.08	67.83 \pm 5.50	70.15 \pm 6.30	67.49 \pm 6.44
LARCM + Added Value	77.43 \pm 5.50	78.01 \pm 5.43	74.75 \pm 6.04	76.82 \pm 5.76	80.43 \pm 6.03	70.73 \pm 5.70	72.57 \pm 5.83	73.33 \pm 6.19
LARCM + Certainty Factor	75.99 \pm 5.70	78.70 \pm 5.01	71.52 \pm 5.99	76.56 \pm 5.07	79.73 \pm 5.77	70.78 \pm 6.52	71.93 \pm 6.66	73.20 \pm 7.17
LARCM + Collective Strength	77.41 \pm 5.35	77.61 \pm 4.97	68.79 \pm 5.88	77.94 \pm 5.20	79.54 \pm 5.09	72.66 \pm 6.13	73.78 \pm 6.25	75.14 \pm 6.03
LARCM + Confiança	80.34 \pm 5.02	77.71 \pm 4.69	73.17 \pm 6.60	77.18 \pm 4.53	80.94 \pm 4.97	69.28 \pm 5.53	70.38 \pm 4.90	72.23 \pm 5.38
LARCM + Conviction	64.36 \pm 5.33	70.43 \pm 5.99	58.59 \pm 6.54	70.26 \pm 6.09	69.07 \pm 6.33	62.76 \pm 6.29	63.40 \pm 5.54	63.38 \pm 6.38
LARCM + ϕ -Coefficient	79.79 \pm 4.92	79.04 \pm 4.48	71.89 \pm 6.22	78.85 \pm 4.78	79.79 \pm 5.34	71.21 \pm 6.65	73.35 \pm 6.63	75.01 \pm 5.98
LARCM + Gini Index	76.47 \pm 5.64	77.98 \pm 5.47	74.55 \pm 6.04	76.56 \pm 5.94	80.21 \pm 5.83	70.17 \pm 6.50	71.48 \pm 7.25	73.25 \pm 6.47
LARCM + IS	76.33 \pm 5.51	77.39 \pm 5.22	69.81 \pm 6.05	75.37 \pm 5.80	78.55 \pm 5.29	68.42 \pm 6.43	70.97 \pm 5.97	72.27 \pm 5.40
LARCM + J-Measure	76.37 \pm 5.52	77.69 \pm 4.18	70.28 \pm 6.28	77.35 \pm 4.61	80.15 \pm 5.65	69.79 \pm 6.31	70.93 \pm 6.90	72.30 \pm 6.56
LARCM + Kappa	79.74 \pm 5.35	77.43 \pm 5.14	72.21 \pm 6.89	76.75 \pm 4.98	78.19 \pm 5.91	69.96 \pm 5.99	72.88 \pm 6.20	72.84 \pm 5.98
LARCM + Klogen	75.78 \pm 5.65	77.71 \pm 4.81	69.72 \pm 6.41	75.88 \pm 5.24	79.54 \pm 5.92	70.92 \pm 5.79	71.83 \pm 6.60	72.59 \pm 6.65
LARCM + Lambda	74.86 \pm 4.84	77.77 \pm 5.21	67.06 \pm 5.72	76.11 \pm 5.03	79.55 \pm 5.76	66.24 \pm 5.50	67.32 \pm 6.08	69.89 \pm 5.34
LARCM + Laplace	80.51 \pm 5.04	78.50 \pm 4.99	72.38 \pm 5.83	77.12 \pm 5.08	80.85 \pm 5.23	67.98 \pm 6.42	69.19 \pm 6.45	71.57 \pm 6.22
LARCM + Lift	78.03 \pm 6.13	77.50 \pm 5.30	66.47 \pm 7.35	76.27 \pm 5.43	78.60 \pm 6.41	67.28 \pm 6.16	70.05 \pm 6.10	70.64 \pm 6.03
LARCM + Mutual Information LHS	75.92 \pm 6.54	77.90 \pm 5.54	66.54 \pm 6.36	76.90 \pm 5.31	79.36 \pm 6.39	69.34 \pm 6.39	71.08 \pm 6.74	71.99 \pm 6.34
LARCM + Novelty	81.42 \pm 5.70	78.74 \pm 5.04	76.15 \pm 6.25	77.72 \pm 5.01	81.52 \pm 5.62	70.02 \pm 6.57	71.82 \pm 5.81	74.04 \pm 6.31
LARCM + Odds Ratio	84.16 \pm 5.29	80.42 \pm 5.54	77.45 \pm 5.52	79.66 \pm 5.21	84.66 \pm 5.09	78.08 \pm 5.93	79.53 \pm 6.18	80.68 \pm 5.61

Tabela B.25: Acurácia dos classificadores para cada modelo de redução de dimensionalidade, com a quantidade de tópicos obtidos $k = 50$ na coleção de documentos ACM-6.

Método	NB	MNB	J48	SMO c=1	SMO c=10	IBk 3	IBk 5	IBk 7
malletborw-t01	88.95 \pm 4.94	87.12 \pm 4.96	87.99 \pm 4.82	83.82 \pm 5.33	86.86 \pm 5.31	80.83 \pm 5.75	78.38 \pm 5.76	78.95 \pm 6.24
LDA + bag-of-words	93.09 \pm 4.23	94.62 \pm 3.59	92.06 \pm 4.69	92.99 \pm 3.77	94.12 \pm 3.33	90.76 \pm 4.44	90.35 \pm 4.24	90.78 \pm 3.95
LARCM + Added Value	85.49 \pm 4.89	88.60 \pm 4.20	76.50 \pm 5.65	88.23 \pm 4.39	88.16 \pm 4.62	83.89 \pm 4.58	83.29 \pm 4.65	83.84 \pm 5.00
LARCM + Certainty Factor	87.53 \pm 4.57	90.17 \pm 4.20	77.07 \pm 6.47	89.81 \pm 4.13	90.98 \pm 3.90	83.41 \pm 5.11	84.92 \pm 4.91	85.70 \pm 4.57
LARCM + Collective Strength	86.43 \pm 5.06	88.37 \pm 4.59	77.93 \pm 5.05	88.47 \pm 4.32	89.29 \pm 4.30	81.73 \pm 5.54	82.96 \pm 5.10	83.92 \pm 5.28
LARCM + Confiança	86.68 \pm 4.56	86.91 \pm 4.59	81.57 \pm 6.36	85.67 \pm 4.77	89.02 \pm 4.22	81.25 \pm 5.36	82.60 \pm 5.14	83.89 \pm 4.88
LARCM + Conviction	76.58 \pm 6.12	79.12 \pm 4.99	72.85 \pm 5.60	77.97 \pm 5.19	77.73 \pm 5.42	73.70 \pm 5.38	73.85 \pm 5.57	75.18 \pm 5.57
LARCM + ϕ -Coefficient	85.88 \pm 4.77	90.25 \pm 3.95	79.29 \pm 5.55	88.83 \pm 4.58	90.55 \pm 4.04	84.60 \pm 5.33	85.68 \pm 4.57	85.99 \pm 4.86
LARCM + Gini Index	86.96 \pm 4.81	88.35 \pm 4.28	77.79 \pm 6.09	87.51 \pm 4.30	88.36 \pm 4.08	82.87 \pm 5.32	84.44 \pm 4.71	83.58 \pm 4.74
LARCM + IS	85.76 \pm 4.75	87.32 \pm 5.04	80.60 \pm 4.93	86.59 \pm 5.06	87.21 \pm 4.89	81.81 \pm 5.87	81.38 \pm 5.73	81.68 \pm 5.80
LARCM + J-Measure	87.58 \pm 4.86	86.94 \pm 4.75	75.42 \pm 6.19	87.94 \pm 4.59	89.12 \pm 4.36	83.49 \pm 4.56	83.48 \pm 4.62	84.35 \pm 4.66
LARCM + Kappa	87.62 \pm 4.33	87.92 \pm 4.61	76.91 \pm 6.01	88.56 \pm 4.57	88.98 \pm 4.46	81.95 \pm 5.62	82.91 \pm 4.81	83.87 \pm 5.03
LARCM + Klogen	85.47 \pm 5.10	86.96 \pm 4.02	77.76 \pm 6.02	86.96 \pm 4.02	88.68 \pm 4.71	80.58 \pm 5.71	81.37 \pm 6.02	81.97 \pm 5.28
LARCM + Lambda	82.52 \pm 4.81	84.57 \pm 5.21	75.58 \pm 5.92	85.88 \pm 5.30	88.31 \pm 4.40	78.99 \pm 6.17	80.55 \pm 5.94	81.55 \pm 5.84
LARCM + Laplace	86.00 \pm 5.33	88.29 \pm 5.06	77.68 \pm 5.83	88.95 \pm 4.70	90.35 \pm 4.18	82.20 \pm 4.81	83.71 \pm 5.25	84.81 \pm 4.83
LARCM + Lift	84.10 \pm 5.16	87.02 \pm 5.55	78.37 \pm 5.44	87.99 \pm 4.91	88.95 \pm 5.20	79.13 \pm 5.83	82.33 \pm 5.22	82.59 \pm 5.56
LARCM + Mutual Information LHS	87.57 \pm 4.14	86.78 \pm 4.79	78.60 \pm 5.28	88.68 \pm 4.73	90.58 \pm 4.22	82.46 \pm 5.58	84.31 \pm 5.31	84.81 \pm 5.37
LARCM + Novelty	88.61 \pm 4.80	88.76 \pm 4.49	78.49 \pm 5.85	89.86 \pm 4.25	90.93 \pm 4.13	84.07 \pm 4.77	85.85 \pm 5.43	86.34 \pm 5.04
LARCM + Odds Ratio	89.56 \pm 4.78	91.00 \pm 3.66	84.50 \pm 5.01	91.37 \pm 3.74	93.22 \pm 3.40	87.55 \pm 4.07	87.62 \pm 4.24	88.39 \pm 4.78

Tabela B.26: Acurácia dos classificadores para cada modelo de redução de dimensionalidade, com a quantidade de tópicos obtidos $k = 100$ na coleção de documentos ACM-6.

Método	NB	MNB	J48	SMO c=1	SMO c=10	IBk 3	IBk 5	IBk 7
malletborw-t01	87.52 \pm 4.55	86.13 \pm 4.19	79.95 \pm 5.43	84.83 \pm 4.76	88.50 \pm 4.58	74.77 \pm 6.87	78.25 \pm 6.71	79.34 \pm 6.49
LDA + bag-of-words	92.27 \pm 3.84	92.12 \pm 3.71	79.71 \pm 5.73	92.08 \pm 3.99	92.31 \pm 4.11	86.68 \pm 4.79	90.00 \pm 4.40	90.20 \pm 4.56
LARCM + Added Value	86.96 \pm 5.52	87.44 \pm 4.27	78.42 \pm 5.63	89.43 \pm 4.37	89.92 \pm 4.47	82.07 \pm 5.41	83.29 \pm 5.26	83.51 \pm 4.89
LARCM + Certainty Factor	88.45 \pm 4.25	86.81 \pm 4.01	83.13 \pm 5.79	86.74 \pm 4.97	89.53 \pm 4.55	82.54 \pm 5.76	83.11 \pm 5.58	84.32 \pm 5.53
LARCM + Collective Strength	87.81 \pm 4.93	85.89 \pm 4.75	76.87 \pm 6.31	86.75 \pm 4.99	87.90 \pm 4.88	80.12 \pm 5.69	80.92 \pm 5.55	81.04 \pm 4.98
LARCM + Confiança	86.47 \pm 4.63	85.74 \pm 4.47	80.57 \pm 5.98	85.63 \pm 5.12	88.28 \pm 5.05	79.42 \pm 5.64	82.05 \pm 5.29	83.13 \pm 5.28
LARCM + Conviction	73.45 \pm 6.13	74.87 \pm 5.97	69.72 \pm 6.25	74.26 \pm 5.84	75.72 \pm 6.51	67.16 \pm 6.64	68.83 \pm 6.32	70.16 \pm 6.15
LARCM + ϕ -Coefficient	87.10 \pm 5.36	86.79 \pm 4.76	78.65 \pm 6.05	85.93 \pm 5.20	88.05 \pm 4.91	81.76 \pm 5.81	82.51 \pm 5.18	83.25 \pm 5.12
LARCM + Gini Index	87.27 \pm 4.64	86.12 \pm 4.87	79.59 \pm 5.60	85.48 \pm 4.75	87.45 \pm 4.84	79.54 \pm 5.41	80.89 \pm 5.86	81.98 \pm 5.30
LARCM + IS	87.05 \pm 4.61	87.45 \pm 4.51	78.17 \pm 6.22	85.97 \pm 4.94	86.63 \pm 5.37	80.46 \pm 6.20	81.66 \pm 5.44	81.70 \pm 5.38
LARCM + J-Measure	88.13 \pm 4.65	85.54 \pm 4.26	81.32 \pm 6.36	85.31 \pm 4.55	88.45 \pm 4.64	77.30 \pm 5.61	78.22 \pm 5.14	79.50 \pm 4.86
LARCM + Kappa	86.05 \pm 5.19	87.58 \pm 4.41	79.68 \pm 6.05	87.46 \pm 4.56	87.42 \pm 4.88	79.79 \pm 4.84	80.52 \pm 4.81	81.25 \pm 5.85
LARCM + Klogen	86.44 \pm 4.51	84.85 \pm 4.67	79.85 \pm 4.85	83.87 \pm 4.66	86.96 \pm 4.62	77.20 \pm 5.95	79.28 \pm 6.46	79.05 \pm 6.24
LARCM + Lambda	86.89 \pm 4.99	84.92 \pm 4.09	75.78 \pm 6.19	85.89 \pm 4.82	88.40 \pm 4.93	77.22 \pm 5.92	79.11 \pm 6.22	79.30 \pm 5.45
LARCM + Laplace	87.72 \pm 4.43	86.33 \pm 4.33	78.29 \pm 5.83	86.15 \pm 4.38	87.50 \pm 4.78	79.61 \pm 5.50	80.79 \pm 5.77	79.72 \pm 5.80
LARCM + Lift	86.10 \pm 4.71	85.79 \pm 4.34	79.75 \pm 5.75	87.66 \pm 4.15	88.15<math			

Tabela B.27: Acurácia dos classificadores para cada modelo de redução de dimensionalidade, com a quantidade de tópicos obtidos $k = 150$ na coleção de documentos ACM-6.

Método	NB	MNB	J48	SMO c=1	SMO c=10	IBk 3	IBk 5	IBk 7
malletborw-t01	83.33 \pm 4.97	82.52 \pm 5.40	72.54 \pm 6.25	81.26 \pm 5.36	84.35 \pm 4.81	69.96 \pm 7.40	69.20 \pm 7.74	66.57 \pm 7.83
LDA + <i>bag-of-words</i>	86.54 \pm 4.18	86.26 \pm 4.31	75.10 \pm 5.80	81.23 \pm 5.22	87.68 \pm 4.42	77.68 \pm 5.27	75.24 \pm 5.50	70.69 \pm 6.36
LARCM + Added Value	87.07 \pm 4.38	84.17 \pm 4.91	80.19 \pm 6.26	83.78 \pm 5.74	87.99 \pm 4.38	78.95 \pm 5.71	79.79 \pm 5.80	79.59 \pm 5.86
LARCM + Certainty Factor	90.03 \pm 4.34	86.50 \pm 4.17	82.06 \pm 5.16	85.67 \pm 5.03	88.92 \pm 4.28	79.92 \pm 5.51	81.12 \pm 6.08	81.81 \pm 5.79
LARCM + Collective Strength	88.61 \pm 4.45	83.93 \pm 4.68	79.61 \pm 4.90	85.89 \pm 4.98	87.86 \pm 4.48	80.50 \pm 5.94	81.57 \pm 6.13	82.25 \pm 5.50
LARCM + Confiança	89.14 \pm 4.16	84.70 \pm 3.83	83.08 \pm 6.19	84.21 \pm 4.78	88.72 \pm 4.25	80.72 \pm 4.86	80.74 \pm 5.47	80.99 \pm 4.90
LARCM + Conviction	71.57 \pm 6.54	77.11 \pm 6.32	68.53 \pm 7.87	73.75 \pm 6.10	77.05 \pm 6.14	66.30 \pm 6.22	68.82 \pm 6.35	68.32 \pm 6.14
LARCM + ϕ -Coefficient	89.21 \pm 4.87	84.88 \pm 5.02	78.81 \pm 6.12	85.45 \pm 4.91	88.72 \pm 4.48	78.29 \pm 6.21	80.01 \pm 5.90	80.78 \pm 5.74
LARCM + Gini Index	88.90 \pm 4.74	84.69 \pm 4.55	80.07 \pm 5.90	85.29 \pm 5.04	87.11 \pm 4.65	79.08 \pm 6.14	80.20 \pm 5.81	81.07 \pm 5.59
LARCM + IS	88.57 \pm 4.28	84.89 \pm 4.55	74.60 \pm 6.25	86.13 \pm 5.08	88.11 \pm 4.54	82.17 \pm 4.68	83.22 \pm 5.11	83.02 \pm 4.93
LARCM + J-Measure	87.25 \pm 4.92	84.33 \pm 4.58	81.10 \pm 5.19	85.39 \pm 5.22	89.98 \pm 4.35	79.27 \pm 5.17	81.05 \pm 5.16	81.58 \pm 5.09
LARCM + Kappa	88.94 \pm 4.32	83.82 \pm 4.69	77.87 \pm 5.50	86.06 \pm 4.86	87.76 \pm 4.84	77.80 \pm 6.03	80.32 \pm 5.48	80.48 \pm 5.06
LARCM + Klosgen	89.15 \pm 5.09	85.82 \pm 4.68	83.08 \pm 5.82	85.01 \pm 5.36	88.63 \pm 4.79	79.52 \pm 6.39	81.28 \pm 5.85	81.71 \pm 5.50
LARCM + Lambda	86.84 \pm 4.69	84.42 \pm 4.68	73.96 \pm 6.88	83.80 \pm 4.60	87.19 \pm 4.67	73.59 \pm 6.02	76.22 \pm 5.43	75.58 \pm 6.17
LARCM + Laplace	88.15 \pm 4.46	86.38 \pm 4.56	80.19 \pm 5.18	87.46 \pm 4.80	90.16 \pm 3.99	80.99 \pm 5.63	81.77 \pm 5.21	82.39 \pm 4.77
LARCM + Lift	87.62 \pm 5.18	83.06 \pm 4.99	81.06 \pm 5.35	83.70 \pm 5.24	87.83 \pm 4.45	76.11 \pm 6.77	76.32 \pm 6.14	76.43 \pm 5.80
LARCM + Mutual Information LHS	87.91 \pm 4.56	85.69 \pm 5.34	80.50 \pm 5.82	85.64 \pm 5.63	88.07 \pm 4.68	78.55 \pm 6.65	79.94 \pm 6.07	80.95 \pm 5.74
LARCM + Novelty	89.33 \pm 4.29	87.26 \pm 4.58	82.22 \pm 5.41	87.89 \pm 4.51	88.92 \pm 4.34	82.96 \pm 5.21	85.45 \pm 5.15	85.56 \pm 4.85
LARCM + Odds Ratio	90.84 \pm 3.52	88.01 \pm 4.21	85.93 \pm 5.05	89.50 \pm 4.30	91.17 \pm 3.66	85.31 \pm 5.23	86.50 \pm 5.22	86.43 \pm 5.12

Tabela B.28: Acurácia dos classificadores para cada modelo de redução de dimensionalidade, com a quantidade de tópicos obtidos $k = 50$ na coleção de documentos ACM-7.

Método	NB	MNB	J48	SMO c=1	SMO c=10	IBk 3	IBk 5	IBk 7
malletborw-t01	86.05 \pm 4.85	84.46 \pm 5.26	85.93 \pm 5.32	81.51 \pm 5.58	83.88 \pm 5.43	75.93 \pm 5.93	75.03 \pm 5.54	74.65 \pm 5.21
LDA + <i>bag-of-words</i>	91.11 \pm 4.00	91.40 \pm 4.03	87.76 \pm 4.52	90.32 \pm 4.36	92.13 \pm 3.82	87.16 \pm 4.59	85.71 \pm 4.75	86.61 \pm 4.90
LARCM + Added Value	85.22 \pm 5.23	88.05 \pm 4.50	84.22 \pm 5.86	87.37 \pm 4.68	88.76 \pm 4.33	81.74 \pm 5.59	83.02 \pm 5.31	82.91 \pm 5.26
LARCM + Certainty Factor	87.19 \pm 4.52	88.91 \pm 3.58	85.21 \pm 5.67	87.93 \pm 3.73	89.81 \pm 3.53	81.22 \pm 5.65	81.60 \pm 5.41	82.88 \pm 4.78
LARCM + Collective Strength	84.75 \pm 5.03	86.82 \pm 4.19	85.37 \pm 4.40	86.71 \pm 4.24	87.63 \pm 4.27	81.59 \pm 5.35	83.52 \pm 5.03	84.26 \pm 4.64
LARCM + Confiança	85.17 \pm 5.36	87.49 \pm 4.30	84.95 \pm 4.76	86.96 \pm 4.42	88.19 \pm 4.63	79.41 \pm 5.57	82.24 \pm 4.94	82.71 \pm 5.28
LARCM + Conviction	69.27 \pm 6.50	76.80 \pm 5.74	73.29 \pm 5.94	74.52 \pm 5.48	73.95 \pm 5.92	70.02 \pm 6.15	70.43 \pm 6.01	71.15 \pm 5.98
LARCM + ϕ -Coefficient	86.74 \pm 4.78	88.59 \pm 4.40	86.93 \pm 5.02	87.99 \pm 4.61	88.91 \pm 3.99	82.09 \pm 5.23	83.96 \pm 5.18	84.60 \pm 4.63
LARCM + Gini Index	84.76 \pm 5.11	87.21 \pm 4.35	84.99 \pm 4.88	87.02 \pm 4.57	88.46 \pm 4.17	81.17 \pm 4.74	82.41 \pm 4.95	82.81 \pm 5.03
LARCM + IS	85.48 \pm 4.77	87.21 \pm 4.45	86.13 \pm 4.52	86.42 \pm 4.47	88.43 \pm 4.70	80.26 \pm 5.33	81.58 \pm 5.34	81.71 \pm 5.66
LARCM + J-Measure	84.55 \pm 4.70	87.53 \pm 4.63	86.96 \pm 5.04	87.10 \pm 4.57	88.51 \pm 4.36	81.07 \pm 5.82	81.30 \pm 5.25	82.14 \pm 5.40
LARCM + Kappa	82.83 \pm 5.00	87.57 \pm 4.58	84.11 \pm 5.19	86.90 \pm 4.92	88.50 \pm 4.43	81.28 \pm 4.80	83.30 \pm 5.11	83.86 \pm 5.20
LARCM + Klosgen	87.53 \pm 4.50	87.67 \pm 4.15	85.48 \pm 4.68	86.50 \pm 4.49	89.44 \pm 4.41	81.94 \pm 5.44	81.92 \pm 5.29	81.94 \pm 5.24
LARCM + Lambda	82.82 \pm 5.26	86.12 \pm 5.00	81.75 \pm 5.98	84.67 \pm 5.46	85.87 \pm 5.42	77.49 \pm 6.18	79.74 \pm 5.82	80.23 \pm 5.77
LARCM + Laplace	84.69 \pm 5.22	87.93 \pm 4.98	85.50 \pm 5.05	87.50 \pm 5.18	88.31 \pm 4.55	81.04 \pm 5.96	82.84 \pm 5.24	83.62 \pm 5.60
LARCM + Lift	85.59 \pm 5.36	87.31 \pm 4.62	84.46 \pm 5.51	86.42 \pm 4.75	87.74 \pm 4.78	80.55 \pm 5.68	80.68 \pm 5.62	81.85 \pm 5.55
LARCM + Mutual Information LHS	84.77 \pm 5.94	87.41 \pm 5.17	83.98 \pm 5.29	87.54 \pm 4.90	88.42 \pm 5.04	79.16 \pm 5.85	80.12 \pm 5.78	82.42 \pm 5.50
LARCM + Novelty	84.14 \pm 5.22	87.25 \pm 4.05	85.68 \pm 5.08	86.76 \pm 4.20	90.04 \pm 4.42	81.77 \pm 4.93	83.20 \pm 4.80	82.99 \pm 5.07
LARCM + Odds Ratio	88.12 \pm 4.36	89.53 \pm 3.96	88.29 \pm 5.17	90.11 \pm 4.25	91.47 \pm 4.01	87.74 \pm 4.99	87.55 \pm 4.74	88.23 \pm 4.97

Tabela B.29: Acurácia dos classificadores para cada modelo de redução de dimensionalidade, com a quantidade de tópicos obtidos $k = 100$ na coleção de documentos ACM-7.

Método	NB	MNB	J48	SMO c=1	SMO c=10	IBk 3	IBk 5	IBk 7
malletborw-t01	81.88 \pm 5.89	83.24 \pm 5.85	83.56 \pm 5.69	83.07 \pm 5.86	83.22 \pm 5.59	72.79 \pm 6.74	72.54 \pm 6.85	73.75 \pm 6.40
LDA + <i>bag-of-words</i>	84.73 \pm 4.73	87.23 \pm 4.42	82.19 \pm 6.00	85.44 \pm 5.02	87.87 \pm 4.61	80.05 \pm 5.35	82.01 \pm 5.12	80.47 \pm 5.44
LARCM + Added Value	84.39 \pm 4.58	88.25 \pm 4.12	84.28 \pm 4.27	86.63 \pm 3.95	88.38 \pm 3.65	79.00 \pm 5.17	79.87 \pm 5.66	80.87 \pm 5.44
LARCM + Certainty Factor	84.31 \pm 5.86	88.68 \pm 3.97	86.55 \pm 4.04	86.18 \pm 4.11	89.04 \pm 4.11	79.15 \pm 5.86	80.73 \pm 6.02	81.99 \pm 5.60
LARCM + Collective Strength	85.52 \pm 5.02	89.23 \pm 4.21	84.22 \pm 4.71	87.49 \pm 4.61	88.66 \pm 4.36	80.21 \pm 5.48	81.88 \pm 5.12	83.03 \pm 5.52
LARCM + Confiança	82.80 \pm 5.10	88.31 \pm 4.43	86.46 \pm 5.32	87.10 \pm 4.59	90.02 \pm 4.61	79.70 \pm 5.60	82.07 \pm 5.62	82.56 \pm 5.73
LARCM + Conviction	67.86 \pm 5.32	76.67 \pm 6.04	69.52 \pm 6.07	75.16 \pm 6.03	75.84 \pm 6.05	69.46 \pm 6.25	70.59 \pm 6.51	72.73 \pm 6.40
LARCM + ϕ -Coefficient	86.14 \pm 4.62	88.98 \pm 3.73	87.49 \pm 4.50	88.04 \pm 4.13	90.66 \pm 3.87	81.52 \pm 5.07	83.07 \pm 4.83	84.05 \pm 4.81
LARCM + Gini Index	86.04 \pm 4.84	87.89 \pm 4.35	87.31 \pm 4.60	87.10 \pm 3.89	89.47 \pm 3.71	80.41 \pm 5.52	81.83 \pm 5.16	83.01 \pm 5.24
LARCM + IS	85.29 \pm 4.75	88.70 \pm 4.55	84.20 \pm 4.57	86.34 \pm 4.98	88.89 \pm 4.37	78.38 \pm 5.62	79.45 \pm 4.56	81.67 \pm 5.39
LARCM + J-Measure	83.94 \pm 5.23	87.80 \pm 4.53	86.99 \pm 4.95	86.16 \pm 4.58	88.34 \pm 3.95	80.00 \pm 5.60	81.62 \pm 6.02	81.70 \pm 5.71
LARCM + Kappa	84.86 \pm 5.43	88.19 \pm 4.41	88.20 \pm 4.30	87.53 \pm 4.61	88.96 \pm 4.70	80.66 \pm 6.09	81.80 \pm 5.70	83.69 \pm 5.39
LARCM + Klosgen	86.46 \pm 4.79	88.98 \pm 4.53	86.03 \pm 4.48	87.40 \pm 4.91	90.49 \pm 3.85	79.00 \pm 4.95	81.00 \pm 4.56	81.92 \pm 4.48
LARCM + Lambda	81.94 \pm 6.13	87.55 \pm 4.79	82.62 \pm 5.41	85.25 \pm 5.46	86.91 \pm 4.90	75.76 \pm 6.11	77.69 \pm 6.22	79.52 \pm 5.86
LARCM + Laplace	84.52 \pm 4.62	89.02 \pm 4.32	87.59 \pm 5.54	87.70 \pm 4.08	89.02 \pm 4.15	78.21 \pm 5.98	80.83 \pm 5.56	82.22 \pm 5.22
LARCM + Lift	86.31 \pm 5.00	88.27 \pm 4.72	84.88 \pm 4.88	88.0				

Tabela B.30: Acurácia dos classificadores para cada modelo de redução de dimensionalidade, com a quantidade de tópicos obtidos $k = 150$ na coleção de documentos ACM-7.

Método	NB	MNB	J48	SMO c=1	SMO c=10	IBk 3	IBk 5	IBk 7
malletborw-t01	82.37±5.21	78.22±5.52	75.59±6.13	78.81±5.77	81.58±5.37	68.55±5.90	69.34±6.17	66.74±5.76
LDA + bag-of-words	83.41±4.69	82.09±4.86	72.50±6.35	81.75±5.06	82.99±5.30	75.78±6.03	72.07±5.64	67.76±6.27
LARCM + Added Value	87.95±4.43	87.71±4.21	85.24±5.38	85.26±4.58	88.27±4.21	79.40±5.72	81.30±5.42	81.53±5.51
LARCM + Certainty Factor	85.39±5.14	87.68±4.29	82.77±5.57	84.03±4.61	88.64±4.02	76.46±4.94	80.90±4.38	82.13±5.10
LARCM + Collective Strength	86.35±5.20	87.33±4.59	85.46±5.10	85.93±5.19	87.87±4.69	79.70±5.09	79.38±6.02	81.19±5.67
LARCM + Confiança	85.02±4.65	87.55±4.36	84.30±4.87	85.33±4.97	89.32±4.12	79.19±5.48	82.13±4.75	83.62±4.85
LARCM + Conviction	66.99±5.94	77.17±5.73	69.78±7.37	75.51±5.38	75.46±5.87	65.74±6.21	68.19±6.53	69.19±6.03
LARCM + ϕ -Coefficient	87.49±4.21	88.17±4.33	87.36±5.22	85.10±4.83	89.30±4.67	80.71±6.32	82.18±5.74	82.94±6.03
LARCM + Gini Index	85.50±5.03	88.46±4.41	84.99±4.30	85.99±4.83	89.15±4.12	81.37±5.11	82.00±5.28	82.73±5.49
LARCM + IS	84.71±5.30	85.46±4.94	82.56±5.61	83.04±5.58	86.79±4.96	78.28±5.33	80.60±4.88	80.52±5.23
LARCM + J-Measure	86.37±4.62	87.12±4.27	84.84±4.95	85.60±4.89	87.82±4.48	79.08±5.87	81.64±5.33	82.39±5.00
LARCM + Kappa	85.31±5.12	87.02±4.59	82.41±4.89	84.89±4.98	87.81±3.90	79.76±5.57	81.56±5.29	81.39±5.20
LARCM + Klogen	86.50±5.13	88.29±3.52	85.45±5.20	85.01±4.21	88.55±3.68	79.15±5.54	81.03±5.57	82.14±5.22
LARCM + Lambda	84.44±5.40	86.10±4.29	80.79±5.87	83.22±5.30	87.72±4.49	77.55±5.45	78.59±4.44	79.89±5.15
LARCM + Laplace	84.99±4.30	88.55±3.89	84.09±5.26	85.63±4.62	88.46±3.58	79.04±5.18	81.68±4.91	82.33±4.80
LARCM + Lift	86.42±4.78	87.57±5.06	80.88±5.81	85.61±4.78	88.06±4.58	80.13±5.92	82.11±6.27	82.07±6.12
LARCM + Mutual Information LHS	86.14±4.33	87.16±4.45	82.32±5.00	84.58±4.54	88.80±4.45	78.50±5.88	81.64±6.00	83.17±5.22
LARCM + Novelty	84.58±4.57	88.51±4.50	88.08±4.62	85.20±4.54	89.00±4.11	80.34±5.07	82.88±4.54	83.39±4.63
LARCM + Odds Ratio	88.52±4.57	90.98±3.99	89.62±4.61	89.53±4.04	91.90±3.69	87.08±4.77	88.43±4.55	89.53±4.25

Tabela B.31: Acurácia dos classificadores para cada modelo de redução de dimensionalidade, com a quantidade de tópicos obtidos $k = 50$ na coleção de documentos ACM-8.

Método	NB	MNB	J48	SMO c=1	SMO c=10	IBk 3	IBk 5	IBk 7
malletborw-t01	81.63±5.52	81.03±6.03	79.70±5.88	77.42±6.46	81.94±5.89	72.91±6.83	75.37±6.81	73.90±6.47
LDA + bag-of-words	86.22±4.80	87.35±4.64	83.32±4.99	86.52±4.57	86.04±4.89	83.28±5.09	83.74±5.10	84.17±5.02
LARCM + Added Value	69.81±6.15	74.49±5.82	62.93±5.84	74.21±6.22	77.12±5.60	66.65±6.27	67.91±6.22	68.37±6.29
LARCM + Certainty Factor	72.26±6.48	74.60±5.91	64.16±7.32	74.26±5.64	76.10±6.02	66.77±5.96	69.60±5.58	70.21±6.06
LARCM + Collective Strength	69.66±6.71	72.36±6.01	60.10±6.25	71.35±6.05	73.31±6.08	64.49±6.87	65.92±7.17	67.35±6.91
LARCM + Confiança	70.86±6.30	74.74±5.53	65.75±6.70	73.63±5.62	76.44±5.98	66.79±6.00	67.74±6.03	67.70±6.33
LARCM + Conviction	55.54±7.05	63.28±7.16	53.80±7.65	61.06±7.10	62.17±6.97	54.72±7.36	56.59±7.24	56.73±6.86
LARCM + ϕ -Coefficient	70.75±6.15	73.01±6.20	64.04±7.07	74.04±6.44	74.34±5.86	65.30±6.57	66.71±7.06	66.49±7.22
LARCM + Gini Index	70.68±5.96	77.75±5.32	65.82±6.72	75.24±5.58	78.07±5.44	67.19±7.70	69.89±6.95	70.66±6.74
LARCM + IS	73.26±5.20	74.04±5.45	63.34±6.57	72.77±5.42	75.10±5.93	66.08±6.31	66.24±5.78	66.29±6.14
LARCM + J-Measure	70.24±6.52	74.83±5.00	62.04±6.45	73.03±6.06	75.05±5.46	65.08±6.31	67.59±6.46	67.84±6.02
LARCM + Kappa	72.08±6.16	75.93±5.48	66.36±6.53	74.30±5.88	77.75±5.72	70.66±5.80	69.31±5.45	68.96±5.63
LARCM + Klogen	69.22±5.49	74.22±5.99	63.64±6.39	72.84±6.10	76.29±6.11	66.07±6.20	67.11±6.07	66.82±6.14
LARCM + Lambda	68.97±5.63	72.80±6.26	64.41±6.93	70.78±6.23	73.73±6.48	60.22±6.35	61.25±6.79	64.20±6.53
LARCM + Laplace	70.03±5.77	74.34±6.24	59.80±7.09	72.86±5.65	74.34±5.19	62.43±6.36	67.08±6.08	67.45±5.95
LARCM + Lift	71.30±6.24	74.52±5.21	66.89±6.77	73.90±5.09	75.77±5.58	66.76±6.01	70.30±5.39	69.45±5.76
LARCM + Mutual Information LHS	67.67±6.46	73.77±5.23	61.58±6.03	72.77±5.90	75.75±5.84	64.30±6.60	66.57±6.60	66.89±6.65
LARCM + Novelty	72.19±6.50	75.02±6.38	65.72±6.92	74.88±6.74	75.48±5.74	66.76±6.29	67.48±6.70	67.50±6.80
LARCM + Odds Ratio	76.01±5.86	77.70±5.30	65.41±6.95	79.36±4.97	80.34±4.93	72.55±5.74	74.79±5.31	74.71±5.31

Tabela B.32: Acurácia dos classificadores para cada modelo de redução de dimensionalidade, com a quantidade de tópicos obtidos $k = 100$ na coleção de documentos ACM-8.

Método	NB	MNB	J48	SMO c=1	SMO c=10	IBk 3	IBk 5	IBk 7
LDA + bag-of-related-words	77.82±5.58	77.24±6.04	72.38±6.54	75.50±6.73	80.23±6.38	63.46±6.95	65.94±7.13	69.35±6.49
LDA + bag-of-words	84.16±5.04	84.99±5.11	71.92±7.58	82.83±5.49	85.51±4.86	76.20±5.70	76.95±5.74	76.97±5.09
LARCM + Added Value	76.57±5.72	77.31±6.33	67.25±6.71	73.89±5.99	79.33±5.65	65.10±5.88	66.63±6.01	67.07±5.94
LARCM + Certainty Factor	74.86±5.70	75.44±5.51	65.90±6.82	72.62±5.81	78.31±5.36	64.79±5.92	66.72±5.83	68.12±6.00
LARCM + Collective Strength	73.26±5.96	76.56±6.08	65.90±7.27	72.79±6.46	77.58±5.72	65.47±6.86	65.72±6.43	66.86±5.97
LARCM + Confiança	74.07±6.45	76.85±5.48	67.64±6.22	74.87±5.97	80.11±5.11	67.88±6.78	70.89±6.69	71.36±6.20
LARCM + Conviction	51.74±7.16	60.98±6.80	49.63±6.44	58.98±7.04	58.91±6.91	47.82±6.97	50.34±7.46	52.86±7.10
LARCM + ϕ -Coefficient	76.37±5.84	78.58±5.03	69.39±5.62	76.73±5.35	79.40±5.23	65.86±6.21	66.97±6.61	66.99±6.93
LARCM + Gini Index	76.21±6.22	76.94±6.03	65.17±6.52	74.02±6.02	78.36±5.97	64.39±7.15	67.16±7.05	67.46±7.01
LARCM + IS	73.30±5.66	74.61±5.92	65.46±6.73	72.35±6.17	75.71±5.70	66.03±6.61	67.18±6.24	69.12±5.91
LARCM + J-Measure	74.79±6.86	75.29±5.62	67.91±7.07	72.08±5.51	75.34±5.40	62.26±6.51	64.42±6.20	65.14±6.08
LARCM + Kappa	71.70±5.62	76.79±5.50	67.70±6.31	72.98±5.75	77.58±5.83	64.00±7.03	66.69±6.33	67.92±5.53
LARCM + Klogen	75.05±5.49	78.00±4.90	68.37±6.10	74.16±4.72	77.30±5.42	63.41±6.04	66.27±6.10	67.26±6.01
LARCM + Lambda	71.36±6.05	74.16±5.13	64.02±6.35	71.01±5.97	76.01±5.00	61.16±6.32	63.36±6.24	64.41±6.38
LARCM + Laplace	75.19±5.72	75.70±5.83	67.93±6.31	73.55±5.90	78.04±6.15	65.65±6.28	66.62±7.09	68.01±6.46
LARCM + Lift	73.28±5.68	76.11±5.73	69.08±6.49	73.03±6.29	78.81±4.90	67.00±5.99	66.51±6.10	67.98±6.50
LARCM + Mutual Information LHS	73.16±6.16	75.40±5.42	68.20±5.82	74.04±5.75	78.98±5.88	65.72±6.67	65.91±6.55	67.69±5.81
LARCM + Novelty	73.31±6.36	75.00±6.04	68.29±6.23	71.85±6.38	76.55±6.22	64.09±6.89	67.47±7.92	68.16±7.62
LARCM + Odds Ratio	78.11±5.13	79.52±4.83	72.36±5.97	77.62±4.90	80.71±5.12	69.28±6.67	70.96±5.57	72.03±5.68

Tabela B.33: Acurácia dos classificadores para cada modelo de redução de dimensionalidade, com a quantidade de tópicos obtidos $k = 150$ na coleção de documentos ACM-8.

Método	NB	MNB	J48	SMO c=1	SMO c=10	IBk 3	IBk 5	IBk 7
LDA + <i>bag-of-related-words</i>	76.16±5.82	76.03±5.69	70.17±7.05	76.05±5.74	76.71±6.07	63.91±5.95	68.09±5.91	68.08±6.69
LDA + <i>bag-of-words</i>	74.36±5.69	80.74±4.60	66.37±5.84	75.88±4.58	78.49±5.21	73.74±5.80	75.74±5.71	74.25±5.76
LARCM + Added Value	76.48±5.77	77.37±5.62	69.96±6.22	73.27±6.18	79.07±5.19	66.25±6.20	66.36±6.63	69.01±6.19
LARCM + Certainty Factor	76.00±5.89	77.85±6.02	68.56±6.15	74.92±6.01	79.96±5.54	66.77±6.64	67.76±6.97	67.94±6.70
LARCM + Collective Strength	75.37±6.00	76.67±5.99	66.12±5.41	71.84±6.81	79.37±5.08	65.04±7.16	65.82±6.64	68.28±6.83
LARCM + Confiança	76.22±5.04	78.18±6.09	66.33±6.84	74.39±6.30	81.14±6.06	67.91±6.83	68.87±6.81	69.32±6.70
LARCM + Conviction	52.87±6.31	66.54±6.96	51.43±7.01	64.02±6.74	62.22±7.31	54.96±7.33	56.31±7.06	57.88±7.12
LARCM + ϕ -Coefficient	76.70±5.49	78.25±6.10	71.13±6.14	75.68±6.66	80.68±5.82	68.84±5.58	70.86±5.92	71.63±6.12
LARCM + Gini Index	76.11±5.44	78.62±6.11	68.74±5.64	75.05±6.28	79.57±5.69	66.68±6.38	68.02±6.20	68.10±6.01
LARCM + IS	74.69±6.01	76.42±5.35	69.59±5.46	73.35±5.31	80.14±5.69	68.18±6.31	68.57±6.06	69.68±6.05
LARCM + J-Measure	77.64±6.08	77.23±5.14	67.47±5.97	73.36±5.78	79.81±5.54	61.36±6.79	65.03±6.35	66.58±6.76
LARCM + Kappa	76.18±6.33	76.84±5.64	68.90±5.87	73.71±5.39	79.57±5.31	66.98±6.83	68.31±6.21	69.28±6.27
LARCM + Klosgen	76.03±5.20	75.94±5.18	68.45±5.75	72.24±5.96	80.08±5.07	65.75±5.64	69.01±5.95	68.27±6.04
LARCM + Lambda	72.76±5.38	73.87±5.61	61.32±6.27	69.48±5.80	75.06±5.60	61.73±7.05	64.07±6.71	64.07±6.46
LARCM + Laplace	76.93±5.29	77.49±5.51	69.92±5.74	74.14±5.80	80.32±4.97	66.63±6.66	68.25±6.34	67.52±6.21
LARCM + Lift	77.35±5.68	75.81±5.10	67.08±6.43	72.66±5.79	79.26±4.95	63.63±6.66	65.53±6.89	66.35±6.50
LARCM + Mutual Information LHS	74.97±6.08	76.84±5.62	68.60±7.11	74.01±5.65	78.62±5.33	65.63±5.89	66.02±6.00	67.22±6.05
LARCM + Novelty	77.21±5.76	78.74±5.47	69.84±6.39	75.40±5.63	79.67±5.05	67.35±7.18	66.20±6.44	67.43±5.78
LARCM + Odds Ratio	82.00±4.70	82.04±5.51	72.47±6.34	78.65±5.59	84.61±5.59	74.48±6.56	75.08±6.54	74.98±5.96

Tabela B.34: Acurácia dos classificadores para cada modelo de redução de dimensionalidade, com a quantidade de tópicos obtidos $k = 50$ na coleção de documentos Re8.

Método	NB	MNB	J48	SMO c=1	SMO c=10	IBk 3	IBk 5	IBk 7
LDA + <i>bag-of-related-words</i>	75.54±1.46	86.60±1.12	87.59±1.16	88.44±1.23	89.17±1.19	87.07±1.12	87.24±1.19	86.82±1.22
LDA + <i>bag-of-words</i>	82.83±1.31	90.10±0.80	90.89±1.01	93.29±0.78	93.56±0.79	92.90±0.75	92.89±0.80	92.73±0.84
LARCM + Added Value	71.28±1.57	82.79±1.16	82.43±1.26	83.88±1.29	84.40±1.22	82.44±1.30	82.81±1.09	82.96±1.12
LARCM + Certainty Factor	69.24±1.58	82.88±1.16	83.00±1.18	83.85±1.36	84.37±1.37	82.51±1.39	82.82±1.33	83.07±1.31
LARCM + Collective Strength	69.82±1.60	82.11±1.09	81.56±1.15	82.66±1.19	83.20±1.21	81.30±1.16	81.84±1.17	81.98±1.19
LARCM + Confiança	68.11±1.71	83.83±1.22	83.39±1.40	84.31±1.31	85.09±1.29	82.48±1.34	83.04±1.37	83.12±1.45
LARCM + Conviction	68.12±1.73	81.53±1.07	81.87±1.20	81.64±1.11	82.59±1.09	81.11±1.17	81.42±1.13	81.36±1.25
LARCM + ϕ -Coefficient	71.49±1.42	83.63±1.18	83.78±1.32	84.97±1.25	85.54±1.20	83.57±1.26	84.10±1.33	84.06±1.35
LARCM + Gini Index	70.05±1.75	83.49±1.19	82.96±1.19	84.09±1.25	84.75±1.21	82.80±1.25	83.29±1.15	83.35±1.23
LARCM + IS	70.76±1.57	82.51±1.26	82.74±1.25	83.22±1.27	83.72±1.21	82.17±1.38	82.72±1.40	82.72±1.43
LARCM + J-Measure	70.89±1.45	83.09±1.18	82.62±1.34	83.80±1.33	84.18±1.29	82.58±1.36	82.88±1.24	83.01±1.22
LARCM + Kappa	70.27±1.56	83.07±1.03	82.98±1.08	83.88±1.19	84.65±1.17	82.38±1.21	82.89±1.26	83.21±1.29
LARCM + Klosgen	69.85±1.53	83.35±1.11	82.96±1.20	84.25±1.18	84.82±1.15	82.85±1.07	83.34±1.18	83.50±1.11
LARCM + Lambda	67.60±1.55	82.24±1.09	82.51±1.35	82.95±1.21	83.40±1.26	82.12±1.34	82.57±1.25	82.66±1.19
LARCM + Laplace	69.66±1.61	83.42±1.19	83.91±1.16	84.46±1.23	85.06±1.17	83.39±1.23	83.56±1.26	83.63±1.23
LARCM + Lift	69.39±1.84	82.85±1.43	82.68±1.36	83.47±1.49	83.96±1.41	82.21±1.24	82.74±1.36	83.01±1.31
LARCM + Mutual Information LHS	70.27±1.54	83.14±1.15	82.77±1.18	83.45±1.13	83.92±1.19	82.51±1.25	82.84±1.35	82.98±1.24
LARCM + Novelty	70.47±1.60	83.29±1.29	83.22±1.26	83.92±1.28	84.32±1.37	82.61±1.44	83.06±1.49	83.29±1.44
LARCM + Odds Ratio	73.07±2.39	85.89±1.01	85.49±1.12	86.56±1.01	87.61±1.03	85.58±1.21	85.97±1.05	86.12±1.01

Tabela B.35: Acurácia dos classificadores para cada modelo de redução de dimensionalidade, com a quantidade de tópicos obtidos $k = 100$ na coleção de documentos Re8.

Método	NB	MNB	J48	SMO c=1	SMO c=10	IBk 3	IBk 5	IBk 7
LDA + <i>bag-of-related-words</i>	72.30±1.52	83.56±1.11	84.44±1.20	87.30±1.10	88.55±1.00	84.38±1.16	84.47±1.19	84.24±1.20
LDA + <i>bag-of-words</i>	81.34±1.37	90.30±0.85	90.17±1.06	94.44±0.82	94.77±0.71	92.69±0.83	92.95±0.73	92.97±0.75
LARCM + Added Value	71.47±1.45	84.76±0.93	83.47±1.22	85.52±1.06	85.88±1.04	83.51±1.11	83.68±1.20	83.95±1.21
LARCM + Certainty Factor	73.65±1.61	85.13±1.29	83.62±1.52	85.71±1.32	86.30±1.29	83.86±1.31	84.27±1.21	84.14±1.14
LARCM + Collective Strength	72.26±1.59	84.10±1.26	83.21±1.26	84.72±1.32	85.45±1.24	83.09±1.18	83.40±1.12	83.28±1.29
LARCM + Confiança	74.66±1.49	85.45±1.20	83.96±1.26	86.03±1.18	86.86±1.11	84.25±1.09	84.82±1.19	84.81±1.14
LARCM + Conviction	70.66±1.65	83.27±1.05	82.18±1.32	83.11±1.30	84.23±1.29	81.68±1.30	81.92±1.30	81.98±1.32
LARCM + ϕ -Coefficient	76.83±1.34	86.06±1.10	84.27±1.29	86.86±1.04	87.80±0.96	84.58±1.12	85.05±1.05	85.01±1.12
LARCM + Gini Index	72.15±1.56	85.22±1.14	83.59±1.35	85.55±1.13	86.09±1.13	83.51±1.25	83.88±1.25	84.22±1.24
LARCM + IS	72.21±1.32	84.76±1.22	83.17±1.16	85.12±1.28	85.68±1.32	83.61±1.29	83.56±1.22	83.52±1.20
LARCM + J-Measure	71.48±1.63	84.84±1.16	83.19±1.31	85.28±1.28	86.01±1.39	83.52±1.45	84.07±1.36	84.25±1.34
LARCM + Kappa	71.99±1.59	84.76±0.99	83.08±1.24	85.17±1.07	85.65±1.21	83.34±1.22	83.48±1.20	83.68±1.14
LARCM + Klosgen	72.56±1.46	84.54±1.08	83.48±1.18	85.47±1.21	85.97±1.32	83.22±1.33	83.78±1.32	83.78±1.26
LARCM + Lambda	72.64±1.72	83.96±0.97	82.01±1.28	84.66±1.08	84.98±1.10	82.62±1.20	82.88±1.21	83.01±1.23
LARCM + Laplace	74.28±1.71	85.05±1.11	83.94±1.27	85.52±1.26	86.14±1.32	83.85±1.27	84.06±1.25	84.00±1.22
LARCM + Lift	72.90±1.51	85.14±1.15	83.45±1.30	85.43±1.19	85.91±1.26	83.54±1.29	84.11±1.34	84.30±1.24
LARCM + Mutual Information LHS	72.37±1.52	84.42±0.98	82.88±1.39	84.90±1.06	85.48±1.09	83.09±1.22	83.40±1.15	83.60±1.19
LARCM + Novelty	73.92±1.59	85.36±1.12	83.21±1.34	85.62±1.20	86.20±1.18	83.90±1.26	84.18±1.28	84.17±1.29
LARCM + Odds Ratio	78.50±1.42	87.47±1.06	86.27±1.23	88.72±1.09	89.86±1.05	87.23±1.13	87.40±1.11	87.68±1.12

Tabela B.36: Acurácia dos classificadores para cada modelo de redução de dimensionalidade, com a quantidade de tópicos obtidos $k = 150$ na coleção de documentos Re8.

Método	NB	MNB	J48	SMO c=1	SMO c=10	IBk 3	IBk 5	IBk 7
LDA + <i>bag-of-related-words</i>	68.19±1.70	81.13±1.14	83.74±1.20	87.65±1.15	89.00±1.12	83.12±1.26	83.90±1.18	83.96±1.25
LDA + <i>bag-of-words</i>	77.00±1.26	89.10±0.81	90.24±0.97	94.02±0.75	94.92±0.75	91.44±0.77	91.60±0.74	91.78±0.80
LARCM + Added Value	72.22±1.50	85.80±1.17	83.96±1.42	87.03±1.23	87.51±1.18	84.68±1.32	85.11±1.34	85.01±1.40
LARCM + Certainty Factor	72.09±1.59	85.84±1.12	83.62±1.26	86.60±1.13	87.32±1.14	84.36±1.04	84.53±1.12	84.59±1.11
LARCM + Collective Strength	70.47±1.79	84.51±0.98	82.67±1.24	85.41±1.09	86.27±1.08	82.98±1.31	83.44±1.23	83.64±1.17
LARCM + Confiança	74.42±1.42	86.21±1.06	83.97±1.29	87.37±1.12	87.96±0.97	85.02±1.23	85.24±1.27	85.36±1.26
LARCM + Conviction	69.45±1.66	84.81±1.07	82.95±1.30	85.20±1.15	85.51±1.16	83.19±1.24	83.62±1.15	83.55±1.16
LARCM + ϕ -Coefficient	76.43±1.45	86.17±0.95	84.83±1.36	87.39±0.96	88.23±0.94	84.99±1.12	85.38±1.08	85.31±1.08
LARCM + Gini Index	72.77±1.57	86.32±1.09	83.98±1.19	86.79±1.05	87.58±1.08	84.60±1.18	84.96±1.19	84.96±1.24
LARCM + IS	70.42±1.52	85.11±1.28	83.07±1.37	85.81±1.20	86.30±1.17	83.51±1.37	83.92±1.40	83.91±1.34
LARCM + J-Measure	73.14±1.75	85.67±1.08	83.68±1.39	86.41±1.20	86.83±1.15	84.21±1.18	84.52±1.20	84.53±1.19
LARCM + Kappa	71.50±1.62	85.79±1.18	83.47±1.21	86.58±1.14	87.14±1.04	84.12±1.19	84.63±1.17	84.76±1.13
LARCM + Klosgen	72.68±1.55	86.34±1.08	84.10±1.16	86.98±1.11	87.37±1.11	84.65±1.11	85.16±1.06	85.36±1.11
LARCM + Lambda	69.93±1.67	85.07±1.15	83.52±1.42	85.84±1.11	86.47±1.15	83.58±1.18	84.18±1.15	84.55±1.21
LARCM + Laplace	75.45±1.41	86.28±1.15	84.42±1.18	87.32±1.13	88.12±1.10	85.19±1.14	85.54±1.19	85.44±1.18
LARCM + Lift	70.76±1.66	85.36±1.16	83.63±1.22	85.74±1.27	86.41±1.18	83.98±1.34	84.57±1.39	84.49±1.38
LARCM + Mutual Information LHS	71.14±1.58	85.39±1.19	83.57±1.30	85.99±1.16	86.43±1.03	83.70±1.27	84.11±1.29	84.64±1.26
LARCM + Novelty	72.78±1.55	86.07±1.08	83.68±1.20	86.65±1.12	87.05±1.07	84.32±1.22	84.72±1.22	84.73±1.23
LARCM + Odds Ratio	77.93±1.38	88.03±0.85	85.98±1.00	89.50±0.85	90.20±0.97	87.43±1.06	87.56±1.03	87.76±1.04

Tabelas com os Resultados da Avaliação da Interpretabilidade

Neste trabalho de doutorado, foi proposta uma metodologia de avaliação (Seção 4.3) que está dividida em duas partes. A segunda parte avalia se o modelo LARCM produz tópicos com melhor interpretabilidade quando comparados aos obtidos com o LDA. Neste apêndice, são apresentados os valores de Coerência Observada para o melhor tópico e o menor valor da medida para o tópico do quartil superior tanto do modelo LARCM combinado com todas as medidas objetivas apresentadas na Tabela 5.2 quanto do modelo LDA combinado com a *bag-of-words* e *bag-of-related-words*, detalhados para cada coleção de documentos e para cada quantidade de tópicos extraídos k . As Tabelas 5.4 e 5.5, apresentadas na Seção 5.3, correspondem a uma compilação dos valores apresentados a seguir, selecionando os 3 melhores resultados para o LARCM considerando o melhor valor de CO e os resultados dos 2 modelos obtidos com o LDA.

Tabela C.1: Resultados para a coleção ACM-1 do valor de Coerência Observada (CO) do tópico com melhor avaliação e do tópico com menor valor de medida entre os 25% melhores tópicos. Os resultados estão ordenados pelo melhor valor de CO obtido.

ACM-1					
K=50		K=100			
Configuration	maximum value	upper quartile	Configuration	maximum value	upper quartile
LDA + <i>bag-of-words</i>	0,35	0,27	LDA + <i>bag-of-words</i>	0,36	0,26
LDA + <i>bag-of-related-words</i>	0,32	0,26	LDA + <i>bag-of-related-words</i>	0,35	0,26
LARCM + IS	0,27	0,19	LARCM + Lift	0,33	0,12
LARCM + Mutual Information LHS	0,26	0,21	LARCM + Kappa	0,31	0,09
LARCM + J-Measure	0,25	0,20	LARCM + Collective Strength	0,30	0,11
LARCM + Collective Strength	0,22	0,08	LARCM + J-Measure	0,29	0,19
LARCM + Novelty	0,21	0,08	LARCM + Mutual Information LHS	0,27	0,19
LARCM + Conviction	0,21	0,07	LARCM + IS	0,25	0,15
LARCM + Lift	0,20	0,11	LARCM + Conviction	0,25	0,11
LARCM + Gini Index	0,20	0,08	LARCM + Gini Index	0,25	0,11
LARCM + Certainty Factor	0,20	0,05	LARCM + Laplace	0,25	0,07
LARCM + Confiança	0,20	0,04	LARCM + Added Value	0,24	0,09
LARCM + ϕ -Coefficient	0,18	0,06	LARCM + Novelty	0,23	0,09
LARCM + Kappa	0,16	0,09	LARCM + ϕ -Coefficient	0,23	0,06
LARCM + Laplace	0,16	0,04	LARCM + Certainty Factor	0,22	0,09
LARCM + Added Value	0,15	0,07	LARCM + Klosgen	0,21	0,09
LARCM + Klosgen	0,15	0,05	LARCM + Odds Ratio	0,21	0,06
LARCM + Odds Ratio	0,14	0,06	LARCM + Confiança	0,18	0,07
LARCM + Lambda	0,11	0,05	LARCM + Lambda	0,18	0,05
K=150					
Configuration	maximum value	upper quartile	Configuration	maximum value	upper quartile
LDA + <i>bag-of-words</i>	0,36	0,26	LDA + <i>bag-of-words</i>	0,34	0,25
LDA + <i>bag-of-related-words</i>	0,34	0,25	LARCM + Conviction	0,33	0,10
LARCM + Conviction	0,33	0,10	LARCM + Lift	0,32	0,13
LARCM + Lift	0,32	0,13	LARCM + ϕ -Coefficient	0,32	0,06
LARCM + ϕ -Coefficient	0,32	0,06	LARCM + IS	0,30	0,15
LARCM + IS	0,30	0,15	LARCM + Novelty	0,29	0,10
LARCM + Novelty	0,29	0,10	LARCM + Certainty Factor	0,28	0,11
LARCM + Certainty Factor	0,28	0,11	LARCM + Collective Strength	0,27	0,11
LARCM + Collective Strength	0,27	0,11	LARCM + Confiança	0,27	0,07
LARCM + Mutual Information LHS	0,26	0,17	LARCM + Mutual Information LHS	0,26	0,17
LARCM + J-Measure	0,24	0,17	LARCM + J-Measure	0,24	0,17
LARCM + Kappa	0,23	0,12	LARCM + Kappa	0,23	0,12
LARCM + Added Value	0,23	0,10	LARCM + Laplace	0,23	0,08
LARCM + Laplace	0,23	0,08	LARCM + Odds Ratio	0,23	0,06
LARCM + Odds Ratio	0,23	0,06	LARCM + Gini Index	0,22	0,11
LARCM + Gini Index	0,22	0,11	LARCM + Klosgen	0,21	0,09
LARCM + Klosgen	0,21	0,09	LARCM + Lambda	0,19	0,06

Tabela C.2: Resultados para a coleção ACM-2 do valor de Coerência Observada (CO) do tópico com melhor avaliação e do tópico com menor valor de medida entre os 25% melhores tópicos. Os resultados estão ordenados pelo melhor valor de CO obtido.

ACM-2					
K=50		K=100			
Configuration	maximum value	upper quartile	Configuration	maximum value	upper quartile
LARCM + Kappa	0,38	0,10	LARCM + Added Value	0,48	0,11
LDA + <i>bag-of-related-words</i>	0,34	0,28	LARCM + Lift	0,44	0,15
LDA + <i>bag-of-words</i>	0,34	0,28	LARCM + Certainty Factor	0,43	0,12
LARCM + J-Measure	0,29	0,22	LDA + <i>bag-of-words</i>	0,36	0,27
LARCM + Klogen	0,29	0,09	LDA + <i>bag-of-related-words</i>	0,34	0,26
LARCM + Collective Strength	0,28	0,12	LARCM + Klogen	0,34	0,12
LARCM + Lift	0,28	0,10	LARCM + Mutual Information LHS	0,32	0,19
LARCM + Mutual Information LHS	0,27	0,21	LARCM + Novelty	0,32	0,13
LARCM + Confiança	0,27	0,09	LARCM + Gini Index	0,31	0,14
LARCM + Laplace	0,27	0,08	LARCM + Kappa	0,28	0,12
LARCM + IS	0,26	0,21	LARCM + Confiança	0,27	0,09
LARCM + Gini Index	0,26	0,11	LARCM + J-Measure	0,26	0,19
LARCM + Added Value	0,23	0,08	LARCM + Conviction	0,26	0,13
LARCM + Novelty	0,22	0,09	LARCM + ϕ -Coefficient	0,26	0,10
LARCM + Conviction	0,20	0,11	LARCM + Laplace	0,26	0,09
LARCM + ϕ -Coefficient	0,17	0,10	LARCM + IS	0,25	0,18
LARCM + Certainty Factor	0,16	0,07	LARCM + Collective Strength	0,25	0,14
LARCM + Odds Ratio	0,14	0,06	LARCM + Odds Ratio	0,23	0,07
LARCM + Lambda	0,12	0,06	LARCM + Lambda	0,14	0,07
K=150					
		Configuration	maximum value	upper quartile	
		LARCM + Lift	0,48	0,14	
		LARCM + Gini Index	0,43	0,13	
		LARCM + Added Value	0,42	0,11	
		LARCM + Klogen	0,42	0,11	
		LARCM + Novelty	0,39	0,13	
		LARCM + Kappa	0,38	0,13	
		LDA + <i>bag-of-related-words</i>	0,35	0,26	
		LARCM + Collective Strength	0,35	0,12	
		LARCM + Confiança	0,35	0,10	
		LARCM + Laplace	0,35	0,09	
		LDA + <i>bag-of-words</i>	0,34	0,26	
		LARCM + Mutual Information LHS	0,33	0,18	
		LARCM + J-Measure	0,32	0,18	
		LARCM + Certainty Factor	0,31	0,12	
		LARCM + IS	0,26	0,16	
		LARCM + ϕ -Coefficient	0,26	0,09	
		LARCM + Conviction	0,25	0,12	
		LARCM + Odds Ratio	0,25	0,07	
		LARCM + Lambda	0,21	0,07	

Tabela C.3: Resultados para a coleção ACM-3 do valor de Coerência Observada (CO) do tópico com melhor avaliação e do tópico com menor valor de medida entre os 25% melhores tópicos. Os resultados estão ordenados pelo melhor valor de CO obtido.

ACM-3					
K=50		K=100			
Configuration	maximum value	upper quartile	Configuration	maximum value	upper quartile
LDA + <i>bag-of-words</i>	0,34	0,27	LARCM + IS	0,37	0,17
LARCM + Mutual Information LHS	0,34	0,20	LARCM + Kappa	0,36	0,11
LARCM + Lift	0,31	0,08	LARCM + Klosgen	0,36	0,10
LDA + <i>bag-of-related-words</i>	0,30	0,27	LDA + <i>bag-of-words</i>	0,34	0,26
LARCM + J-Measure	0,29	0,19	LARCM + Conviction	0,34	0,14
LARCM + ϕ -Coefficient	0,27	0,09	LDA + <i>bag-of-related-words</i>	0,33	0,26
LARCM + Kappa	0,26	0,05	LARCM + J-Measure	0,30	0,18
LARCM + IS	0,25	0,19	LARCM + Mutual Information LHS	0,29	0,19
LARCM + Collective Strength	0,23	0,09	LARCM + Odds Ratio	0,29	0,07
LARCM + Conviction	0,22	0,13	LARCM + Novelty	0,26	0,11
LARCM + Laplace	0,21	0,06	LARCM + Lift	0,25	0,12
LARCM + Gini Index	0,18	0,09	LARCM + Collective Strength	0,25	0,10
LARCM + Novelty	0,18	0,08	LARCM + Gini Index	0,25	0,09
LARCM + Klosgen	0,16	0,07	LARCM + Certainty Factor	0,22	0,09
LARCM + Certainty Factor	0,16	0,06	LARCM + ϕ -Coefficient	0,22	0,09
LARCM + Added Value	0,16	0,05	LARCM + Added Value	0,21	0,10
LARCM + Confiança	0,15	0,06	LARCM + Laplace	0,20	0,07
LARCM + Odds Ratio	0,12	0,07	LARCM + Lambda	0,19	0,06
LARCM + Lambda	0,12	0,05	LARCM + Confiança	0,18	0,07
K=150					
Configuration	maximum value	upper quartile	Configuration	maximum value	upper quartile
LARCM + Mutual Information LHS	0,40	0,17	LARCM + <i>bag-of-related-words</i>	0,35	0,26
LDA + <i>bag-of-related-words</i>	0,35	0,26	LDA + <i>bag-of-words</i>	0,35	0,25
LARCM + Certainty Factor	0,35	0,10	LARCM + Novelty	0,33	0,11
LARCM + Novelty	0,33	0,11	LARCM + Odds Ratio	0,32	0,06
LARCM + Lift	0,31	0,11	LARCM + Kappa	0,31	0,11
LARCM + Collective Strength	0,28	0,11	LARCM + J-Measure	0,27	0,16
LARCM + J-Measure	0,27	0,16	LARCM + Lift	0,27	0,13
LARCM + Added Value	0,27	0,10	LARCM + Added Value	0,27	0,10
LARCM + ϕ -Coefficient	0,27	0,07	LARCM + IS	0,26	0,15
LARCM + IS	0,26	0,15	LARCM + Laplace	0,26	0,08
LARCM + Laplace	0,25	0,08	LARCM + Confiança	0,25	0,08
LARCM + Conviction	0,24	0,12	LARCM + Klosgen	0,24	0,11
LARCM + Gini Index	0,24	0,10	LARCM + Lambda	0,24	0,10
LARCM + Lambda	0,18	0,06			

Tabela C.4: Resultados para a coleção ACM-4 do valor de Coerência Observada (CO) do tópico com melhor avaliação e do tópico com menor valor de medida entre os 25% melhores tópicos. Os resultados estão ordenados pelo melhor valor de CO obtido.

ACM-4					
K=50		K=100			
Configuration	maximum value	upper quartile	Configuration	maximum value	upper quartile
LARCM + Conviction	0,38	0,12	LARCM + Conviction	0,37	0,13
LARCM + Kappa	0,36	0,08	LARCM + Kappa	0,36	0,10
LDA + <i>bag-of-words</i>	0,33	0,26	LDA + <i>bag-of-words</i>	0,33	0,25
LARCM + Collective Strength	0,33	0,10	LARCM + Collective Strength	0,33	0,11
LDA + <i>bag-of-related-words</i>	0,31	0,26	LDA + <i>bag-of-related-words</i>	0,31	0,25
LARCM + IS	0,29	0,17	LARCM + Klosgen	0,30	0,08
LARCM + Klosgen	0,27	0,07	LARCM + Certainty Factor	0,29	0,10
LARCM + J-Measure	0,25	0,20	LARCM + ϕ -Coefficient	0,29	0,08
LARCM + Mutual Information LHS	0,25	0,20	LARCM + Lift	0,28	0,13
LARCM + Lift	0,25	0,12	LARCM + J-Measure	0,27	0,17
LARCM + Novelty	0,24	0,09	LARCM + Novelty	0,27	0,10
LARCM + Gini Index	0,22	0,09	LARCM + Mutual Information LHS	0,26	0,19
LARCM + Added Value	0,21	0,07	LARCM + IS	0,26	0,15
LARCM + ϕ -Coefficient	0,20	0,05	LARCM + Gini Index	0,26	0,09
LARCM + Odds Ratio	0,20	0,05	LARCM + Laplace	0,24	0,08
LARCM + Certainty Factor	0,19	0,06	LARCM + Confiança	0,23	0,07
LARCM + Confiança	0,18	0,06	LARCM + Added Value	0,21	0,09
LARCM + Laplace	0,16	0,06	LARCM + Odds Ratio	0,21	0,06
LARCM + Lambda	0,14	0,04	LARCM + Lambda	0,21	0,05
K=150					
Configuration	maximum value	upper quartile	Configuration	maximum value	upper quartile
LARCM + Lambda	0,46	0,07	LARCM + Lambda	0,46	0,07
LARCM + Kappa	0,36	0,11	LARCM + Kappa	0,36	0,11
LDA + <i>bag-of-words</i>	0,34	0,25	LDA + <i>bag-of-words</i>	0,34	0,25
LARCM + Collective Strength	0,33	0,11	LARCM + Collective Strength	0,33	0,11
LDA + <i>bag-of-related-words</i>	0,32	0,25	LDA + <i>bag-of-related-words</i>	0,32	0,25
LARCM + Conviction	0,32	0,12	LARCM + Conviction	0,32	0,12
LARCM + ϕ -Coefficient	0,32	0,07	LARCM + ϕ -Coefficient	0,32	0,07
LARCM + Mutual Information LHS	0,30	0,18	LARCM + Mutual Information LHS	0,30	0,18
LARCM + IS	0,30	0,13	LARCM + IS	0,30	0,13
LARCM + Odds Ratio	0,29	0,07	LARCM + Odds Ratio	0,29	0,07
LARCM + Certainty Factor	0,27	0,11	LARCM + Certainty Factor	0,27	0,11
LARCM + Novelty	0,27	0,11	LARCM + Novelty	0,27	0,11
LARCM + J-Measure	0,26	0,15	LARCM + J-Measure	0,26	0,15
LARCM + Gini Index	0,25	0,11	LARCM + Gini Index	0,25	0,11
LARCM + Klosgen	0,25	0,10	LARCM + Klosgen	0,25	0,10
LARCM + Lift	0,24	0,13	LARCM + Lift	0,24	0,13
LARCM + Added Value	0,24	0,11	LARCM + Added Value	0,24	0,11
LARCM + Laplace	0,23	0,08	LARCM + Laplace	0,23	0,08
LARCM + Confiança	0,23	0,07	LARCM + Confiança	0,23	0,07

Tabela C.5: Resultados para a coleção ACM-5 do valor de Coerência Observada (CO) do tópico com melhor avaliação e do tópico com menor valor de medida entre os 25% melhores tópicos. Os resultados estão ordenados pelo melhor valor de CO obtido.

ACM-5					
K=50		K=100			
Configuration	maximum value	upper quartile	Configuration	maximum value	upper quartile
LARCM + Gini Index	0,39	0,09	LARCM + Kappa	0,43	0,11
LDA + <i>bag-of-words</i>	0,34	0,27	LARCM + Added Value	0,41	0,09
LARCM + J-Measure	0,34	0,20	LARCM + Collective Strength	0,39	0,09
LDA + <i>bag-of-related-words</i>	0,33	0,28	LDA + <i>bag-of-related-words</i>	0,35	0,26
LARCM + Novelty	0,33	0,08	LDA + <i>bag-of-words</i>	0,34	0,27
LARCM + ϕ -Coefficient	0,28	0,08	LARCM + Certainty Factor	0,33	0,10
LARCM + Collective Strength	0,27	0,08	LARCM + Laplace	0,32	0,07
LARCM + Conviction	0,27	0,08	LARCM + IS	0,29	0,17
LARCM + Klosgen	0,26	0,07	LARCM + Novelty	0,29	0,12
LARCM + Odds Ratio	0,26	0,07	LARCM + Conviction	0,29	0,11
LARCM + Mutual Information LHS	0,25	0,20	LARCM + J-Measure	0,28	0,18
LARCM + IS	0,25	0,18	LARCM + Klosgen	0,28	0,09
LARCM + Kappa	0,23	0,08	LARCM + Lift	0,27	0,13
LARCM + Added Value	0,21	0,08	LARCM + Mutual Information LHS	0,26	0,20
LARCM + Lift	0,21	0,08	LARCM + Gini Index	0,26	0,13
LARCM + Certainty Factor	0,21	0,07	LARCM + ϕ -Coefficient	0,26	0,09
LARCM + Confiança	0,19	0,08	LARCM + Odds Ratio	0,26	0,06
LARCM + Laplace	0,16	0,07	LARCM + Confiança	0,24	0,08
LARCM + Lambda	0,12	0,04	LARCM + Lambda	0,21	0,05
K=150					
Configuration	maximum value	upper quartile	Configuration	maximum value	upper quartile
LARCM + Added Value	0,41	0,10	LARCM + ϕ -Coefficient	0,40	0,08
LARCM + ϕ -Coefficient	0,40	0,08	LARCM + Klosgen	0,39	0,10
LARCM + Klosgen	0,39	0,10	LARCM + Gini Index	0,38	0,12
LARCM + Gini Index	0,38	0,12	LARCM + Novelty	0,38	0,11
LARCM + Novelty	0,38	0,11	LDA + <i>bag-of-related-words</i>	0,36	0,25
LDA + <i>bag-of-related-words</i>	0,36	0,25	LDA + <i>bag-of-words</i>	0,35	0,26
LDA + <i>bag-of-words</i>	0,35	0,26	LARCM + Kappa	0,34	0,10
LARCM + Kappa	0,34	0,10	LARCM + Collective Strength	0,33	0,11
LARCM + Collective Strength	0,33	0,11	LARCM + Confiança	0,33	0,08
LARCM + Confiança	0,33	0,08	LARCM + Conviction	0,32	0,12
LARCM + Conviction	0,32	0,12	LARCM + Laplace	0,32	0,08
LARCM + Laplace	0,32	0,08	LARCM + Odds Ratio	0,32	0,06
LARCM + Odds Ratio	0,32	0,06	LARCM + J-Measure	0,30	0,17
LARCM + J-Measure	0,30	0,17	LARCM + Mutual Information LHS	0,29	0,17
LARCM + Mutual Information LHS	0,29	0,17	LARCM + IS	0,27	0,15
LARCM + IS	0,27	0,15	LARCM + Lift	0,25	0,13
LARCM + Lift	0,25	0,13	LARCM + Certainty Factor	0,24	0,10
LARCM + Certainty Factor	0,24	0,10	LARCM + Lambda	0,21	0,06

Tabela C.6: Resultados para a coleção ACM-6 do valor de Coerência Observada (CO) do tópico com melhor avaliação e do tópico com menor valor de medida entre os 25% melhores tópicos. Os resultados estão ordenados pelo melhor valor de CO obtido.

ACM-6					
K=50		K=100			
Configuration	maximum value	upper quartile	Configuration	maximum value	upper quartile
LARCM + Confiança	0,38	0,05	LARCM + Certainty Factor	0,40	0,09
LDA + <i>bag-of-words</i>	0,36	0,28	LARCM + Laplace	0,38	0,06
LDA + <i>bag-of-related-words</i>	0,34	0,27	LARCM + Lift	0,36	0,12
LARCM + IS	0,33	0,21	LARCM + Added Value	0,36	0,08
LARCM + Gini Index	0,33	0,08	LARCM + Conviction	0,35	0,13
LARCM + Novelty	0,33	0,08	LARCM + IS	0,34	0,18
LARCM + Kappa	0,33	0,08	LDA + <i>bag-of-words</i>	0,33	0,26
LARCM + Laplace	0,33	0,05	LARCM + Gini Index	0,33	0,10
LARCM + Conviction	0,29	0,09	LARCM + Kappa	0,33	0,09
LARCM + Certainty Factor	0,28	0,05	LARCM + Klosgen	0,33	0,09
LARCM + J-Measure	0,27	0,21	LARCM + Confiança	0,33	0,07
LARCM + Mutual Information LHS	0,26	0,21	LDA + <i>bag-of-related-words</i>	0,32	0,26
LARCM + Klosgen	0,25	0,06	LARCM + J-Measure	0,30	0,18
LARCM + Lift	0,24	0,09	LARCM + Novelty	0,30	0,08
LARCM + Added Value	0,20	0,04	LARCM + Mutual Information LHS	0,29	0,19
LARCM + Collective Strength	0,19	0,07	LARCM + Collective Strength	0,25	0,11
LARCM + Odds Ratio	0,19	0,05	LARCM + Lambda	0,23	0,05
LARCM + ϕ -Coefficient	0,18	0,07	LARCM + ϕ -Coefficient	0,18	0,08
LARCM + Lambda	0,10	0,04	LARCM + Odds Ratio	0,18	0,06
K=150					
Configuration	maximum value	upper quartile	Configuration	maximum value	upper quartile
LARCM + IS	0,42	0,15	LARCM + Collective Strength	0,42	0,11
LARCM + Collective Strength	0,42	0,11	LARCM + Gini Index	0,42	0,11
LARCM + Gini Index	0,42	0,11	LARCM + Kappa	0,39	0,10
LARCM + Kappa	0,39	0,10	LARCM + Confiança	0,38	0,06
LARCM + Mutual Information LHS	0,37	0,18	LARCM + Mutual Information LHS	0,37	0,18
LARCM + Klosgen	0,36	0,09	LARCM + Klosgen	0,36	0,09
LDA + <i>bag-of-words</i>	0,35	0,26	LDA + <i>bag-of-words</i>	0,35	0,25
LDA + <i>bag-of-related-words</i>	0,35	0,25	LARCM + Lift	0,35	0,11
LARCM + Lift	0,35	0,11	LARCM + Certainty Factor	0,35	0,10
LARCM + Certainty Factor	0,35	0,10	LARCM + J-Measure	0,33	0,17
LARCM + J-Measure	0,33	0,17	LARCM + Novelty	0,33	0,11
LARCM + Novelty	0,33	0,11	LARCM + Added Value	0,33	0,10
LARCM + Added Value	0,33	0,10	LARCM + ϕ -Coefficient	0,33	0,08
LARCM + ϕ -Coefficient	0,33	0,08	LARCM + Laplace	0,33	0,06
LARCM + Laplace	0,33	0,06	LARCM + Conviction	0,28	0,12
LARCM + Conviction	0,28	0,12	LARCM + Lambda	0,27	0,05
LARCM + Lambda	0,27	0,05	LARCM + Odds Ratio	0,21	0,06

Tabela C.7: Resultados para a coleção ACM-7 do valor de Coerência Observada (CO) do tópico com melhor avaliação e do tópico com menor valor de medida entre os 25% melhores tópicos. Os resultados estão ordenados pelo melhor valor de CO obtido.

ACM-7					
K=50		K=100			
Configuration	maximum value	upper quartile	Configuration	maximum value	upper quartile
LDA + <i>bag-of-related-words</i>	0,34	0,27	LARCM + Odds Ratio	0,41	0,06
LDA + <i>bag-of-words</i>	0,33	0,27		0,36	0,10
LARCM + IS	0,29	0,19		0,33	0,26
LARCM + Added Value	0,28	0,05		0,31	0,26
LARCM + Gini Index	0,27	0,08		0,30	0,08
LARCM + Collective Strength	0,27	0,06		0,28	0,09
LARCM + Mutual Information LHS	0,26	0,21		0,27	0,20
LARCM + J-Measure	0,26	0,20		0,26	0,18
LARCM + ϕ -Coefficient	0,26	0,07		0,26	0,07
LARCM + Klosgen	0,26	0,06		0,25	0,09
LARCM + Certainty Factor	0,24	0,06		0,24	0,18
LARCM + Confiança	0,24	0,06		0,24	0,08
LARCM + Kappa	0,24	0,06		0,24	0,08
LARCM + Laplace	0,24	0,06		0,24	0,07
LARCM + Novelty	0,22	0,08		0,22	0,09
LARCM + Lift	0,20	0,07		0,22	0,08
LARCM + Conviction	0,18	0,09		0,21	0,11
LARCM + Odds Ratio	0,18	0,07		0,20	0,07
LARCM + Lambda	0,08	0,03		0,12	0,05
K=150					
Configuration	maximum value	upper quartile			
LARCM + Conviction	0,36	0,10			
LARCM + Novelty	0,36	0,10			
LARCM + Klosgen	0,35	0,09			
LDA + <i>bag-of-words</i>	0,33	0,25			
LARCM + IS	0,33	0,15			
LARCM + ϕ -Coefficient	0,33	0,07			
LDA + <i>bag-of-related-words</i>	0,31	0,25			
LARCM + Odds Ratio	0,31	0,06			
LARCM + Added Value	0,28	0,10			
LARCM + Mutual Information LHS	0,27	0,17			
LARCM + Certainty Factor	0,27	0,08			
LARCM + J-Measure	0,26	0,17			
LARCM + Gini Index	0,25	0,11			
LARCM + Kappa	0,25	0,10			
LARCM + Confiança	0,24	0,07			
LARCM + Laplace	0,24	0,07			
LARCM + Collective Strength	0,23	0,10			
LARCM + Lift	0,22	0,11			
LARCM + Lambda	0,19	0,05			

Tabela C.8: Resultados para a coleção ACM-8 do valor de Coerência Observada (CO) do tópico com melhor avaliação e do tópico com menor valor de medida entre os 25% melhores tópicos. Os resultados estão ordenados pelo melhor valor de CO obtido.

ACM-8					
K=50		K=100			
Configuration	maximum value	upper quartile	Configuration	maximum value	upper quartile
LDA + <i>bag-of-words</i>	0,33	0,27	LDA + <i>bag-of-words</i>	0,34	0,26
LDA + <i>bag-of-related-words</i>	0,31	0,26	LDA + <i>bag-of-related-words</i>	0,33	0,25
LARCM + J-Measure	0,29	0,19	LARCM + Mutual Information LHS	0,31	0,19
LARCM + Collective Strength	0,28	0,07	LARCM + Collective Strength	0,28	0,07
LARCM + IS	0,27	0,18	LARCM + ϕ -Coefficient	0,28	0,06
LARCM + Mutual Information LHS	0,26	0,20	LARCM + J-Measure	0,27	0,17
LARCM + Gini Index	0,26	0,06	LARCM + Conviction	0,27	0,10
LARCM + Conviction	0,24	0,09	LARCM + Lift	0,26	0,10
LARCM + Novelty	0,23	0,08	LARCM + IS	0,25	0,16
LARCM + Klosgen	0,23	0,06	LARCM + Gini Index	0,24	0,09
LARCM + Lift	0,21	0,07	LARCM + Novelty	0,24	0,06
LARCM + Kappa	0,20	0,07	LARCM + Klosgen	0,23	0,07
LARCM + ϕ -Coefficient	0,20	0,06	LARCM + Kappa	0,22	0,08
LARCM + Odds Ratio	0,19	0,06	LARCM + Odds Ratio	0,21	0,07
LARCM + Confiança	0,18	0,05	LARCM + Certainty Factor	0,20	0,07
LARCM + Laplace	0,18	0,05	LARCM + Added Value	0,19	0,06
LARCM + Lambda	0,18	0,03	LARCM + Laplace	0,18	0,06
LARCM + Certainty Factor	0,17	0,05	LARCM + Confiança	0,18	0,05
LARCM + Added Value	0,17	0,04	LARCM + Lambda	0,17	0,04
K=150					
Configuration	maximum value	upper quartile	Configuration	maximum value	upper quartile
LDA + <i>bag-of-words</i>	0,32	0,25	LDA + <i>bag-of-words</i>	0,32	0,25
LDA + <i>bag-of-related-words</i>	0,32	0,25	LARCM + ϕ -Coefficient	0,31	0,06
LARCM + ϕ -Coefficient	0,31	0,06	LARCM + J-Measure	0,28	0,16
LARCM + Lift	0,28	0,11	LARCM + Lift	0,28	0,11
LARCM + Collective Strength	0,28	0,09	LARCM + Odds Ratio	0,28	0,06
LARCM + Odds Ratio	0,28	0,06	LARCM + IS	0,27	0,14
LARCM + IS	0,27	0,14	LARCM + Mutual Information LHS	0,25	0,18
LARCM + Mutual Information LHS	0,25	0,18	LARCM + Gini Index	0,25	0,09
LARCM + Gini Index	0,25	0,09	LARCM + Conviction	0,24	0,10
LARCM + Conviction	0,24	0,10	LARCM + Kappa	0,24	0,08
LARCM + Kappa	0,24	0,08	LARCM + Added Value	0,24	0,07
LARCM + Added Value	0,24	0,07	LARCM + Novelty	0,23	0,08
LARCM + Novelty	0,23	0,08	LARCM + Klosgen	0,21	0,08
LARCM + Klosgen	0,21	0,08	LARCM + Certainty Factor	0,21	0,07
LARCM + Certainty Factor	0,21	0,07	LARCM + Lambda	0,21	0,06
LARCM + Lambda	0,21	0,06	LARCM + Laplace	0,20	0,06
LARCM + Laplace	0,20	0,06	LARCM + Confiança	0,18	0,05
LARCM + Confiança	0,18	0,06			

Tabela C.9: Resultados para a coleção Re8 do valor de Coerência Observada (CO) do tópico com melhor avaliação e do tópico com menor valor de medida entre os 25% melhores tópicos. Os resultados estão ordenados pelo melhor valor de CO obtido.

Re8					
K=50		K=100			
Configuration	maximum value	upper quartile	Configuration	maximum value	upper quartile
LDA + <i>bag-of-related-words</i>	0,30	0,24	LDA + <i>bag-of-related-words</i>	0,31	0,23
LARCM + Collective Strength	0,29	0,15	LDA + <i>bag-of-words</i>	0,30	0,25
LARCM + Novelty	0,29	0,15	LARCM + Added Value	0,28	0,16
LARCM + Odds Ratio	0,29	0,09	LARCM + Collective Strength	0,28	0,15
LDA + <i>bag-of-words</i>	0,28	0,22	LARCM + Certainty Factor	0,26	0,17
LARCM + J-Measure	0,25	0,18	LARCM + Laplace	0,26	0,11
LARCM + Mutual Information LHS	0,25	0,18	LARCM + IS	0,25	0,18
LARCM + IS	0,25	0,17	LARCM + Lift	0,25	0,17
LARCM + Added Value	0,25	0,16	LARCM + Gini Index	0,25	0,16
LARCM + Certainty Factor	0,25	0,16	LARCM + Kappa	0,25	0,16
LARCM + Gini Index	0,25	0,16	LARCM + Odds Ratio	0,25	0,13
LARCM + Laplace	0,24	0,10	LARCM + J-Measure	0,24	0,18
LARCM + Conviction	0,23	0,16	LARCM + Mutual Information LHS	0,24	0,18
LARCM + Kappa	0,23	0,16	LARCM + Conviction	0,24	0,16
LARCM + Lift	0,23	0,16	LARCM + Confiança	0,24	0,11
LARCM + Confiança	0,23	0,09	LARCM + Klosgen	0,23	0,17
LARCM + Klosgen	0,22	0,14	LARCM + Novelty	0,23	0,16
LARCM + ϕ -Coefficient	0,21	0,10	LARCM + ϕ -Coefficient	0,23	0,10
LARCM + Lambda	0,20	0,12	LARCM + Lambda	0,20	0,13
K=150					
Configuration	maximum value	upper quartile	Configuration	maximum value	upper quartile
LARCM + Collective Strength	0,44	0,16	LARCM + Collective Strength	0,44	0,16
LARCM + Added Value	0,35	0,16	LARCM + Novelty	0,34	0,16
LARCM + Novelty	0,34	0,16	LARCM + Klosgen	0,34	0,16
LARCM + Klosgen	0,34	0,16	LARCM + ϕ -Coefficient	0,34	0,12
LARCM + ϕ -Coefficient	0,34	0,12	LDA + <i>bag-of-words</i>	0,32	0,24
LDA + <i>bag-of-words</i>	0,32	0,17	LARCM + IS	0,32	0,17
LARCM + Mutual Information LHS	0,31	0,17	LDA + <i>bag-of-related-words</i>	0,31	0,22
LARCM + Lambda	0,31	0,13	LARCM + Lambda	0,31	0,17
LARCM + Lift	0,30	0,17	LARCM + Conviction	0,29	0,17
LARCM + Conviction	0,29	0,17	LARCM + J-Measure	0,29	0,17
LARCM + J-Measure	0,29	0,17	LARCM + Gini Index	0,29	0,16
LARCM + Gini Index	0,29	0,16	LARCM + Kappa	0,28	0,16
LARCM + Kappa	0,28	0,16	LARCM + Certainty Factor	0,27	0,16
LARCM + Certainty Factor	0,27	0,16	LARCM + Odds Ratio	0,27	0,13
LARCM + Odds Ratio	0,27	0,13	LARCM + Confiança	0,27	0,10
LARCM + Confiança	0,27	0,10	LARCM + Laplace	0,27	0,10

Sistemas de Recomendação Sensíveis ao Contexto

Um sistema de recomendação é uma tecnologia de filtragem de informação que pode ser utilizada para prever avaliações de itens e/ou gerar um *ranking* de itens que podem ser de interesse do usuário. Em sites de comércio eletrônico, em geral, os itens são produtos recomendados para compra. Já em uma biblioteca digital eles podem ser textos ou outras mídias que sejam relevantes às preferências dos usuários.

Um sistema de recomendação sensível ao contexto é uma tecnologia de filtragem de informação que, além do comportamento e interesse do usuário, utiliza também informação contextual (contexto), como tempo ou local, para recomendar a este uma lista ordenada de itens/recomendações que lhe é de interesse (Sundermann, 2015). Neste capítulo, serão apresentados os conceitos básicos dos algoritmos utilizados no estudo de caso apresentado na Seção 5.4. Uma revisão mais detalhada da literatura da área e dos algoritmos aplicados nesses sistemas pode ser encontrada no trabalho de Sundermann (2015).

Neste trabalho de doutorado, foi adotado o processo de avaliação proposto por Sundermann et al. (2014) para o estudo de caso. Em sua proposta, foram avaliados 3 algoritmos de recomendação sensíveis ao contexto, *cReduction*, *weightPoF* e *filterPoF*, que foram comparados com o algoritmo de recomendação *Item-Based Collaborative Filtering* (IBCF). Os 3 algoritmos sensíveis ao contexto são baseados no algoritmo IBCF (Sundermann et al., 2014), e este último não utiliza informação de contexto para recomendação. Seja m o número de usuários do conjunto $U = u_1, u_2, \dots, u_m$ e n o número de itens que podem ser recomendados do conjunto $I = i_1, i_2, \dots, i_n$. Um modelo M de filtragem colaborativa baseada no item (IBCF) é uma matriz que representa as similaridades entre todos os pares de itens, de acordo com uma métrica de similaridade. Na Tabela D.1 é ilustrada a matriz de similaridade, na qual cada item $i \in I$ é um item acessado, por exemplo, uma página da Web.

Tabela D.1: Exemplo de matriz de similaridade de itens para algoritmos de recomendação baseados no IBCF.

	i_1	i_2	\dots	i_n
i_1	1	$sim(i_1, i_2)$	\dots	$sim(i_1, i_n)$
i_2	$sim(i_2, i_1)$	1	\dots	$sim(i_2, i_n)$
\vdots	\vdots	\ddots	1	\vdots
i_n	$sim(i_n, i_1)$	$sim(i_n, i_2)$	\dots	1

A similaridade $sim(i_1, i_2)$ entre os itens i_1 e i_2 é calculada pela similaridade cosseno entre \vec{i}_1 e \vec{i}_2 , tal que:

$$sim(i_1, i_2) = \cos(\vec{i}_1, \vec{i}_2) = \frac{\vec{i}_1 \cdot \vec{i}_2}{\|\vec{i}_1\| * \|\vec{i}_2\|} \quad (\text{D.1})$$

em que \vec{i}_1 e \vec{i}_2 são vetores de avaliações com m posições, e cada posição correspondendo ao valor de uma avaliação do usuário sobre o item. Na proposta de Sundermann et al. (2014), são consideradas avaliações implícitas e seus valores são binários: o valor 1 indica que o usuário acessou o item, enquanto que o valor 0 indica que o usuário não acessou o item. Para um usuário u_a , com um conjunto de itens previamente avaliados (ou acessados) conhecido $O \in I$, as recomendações são geradas da seguinte forma: o conjunto de itens candidatos a recomendação R é selecionado tal que $R = I - O$. Então, para cada item candidato $r \in R$, a nota de recomendação é calculada como:

$$nota(u_a, O, r) = \frac{\sum_{i \in K_r \cap O} sim(r, i)}{\sum_{i \in K_r} sim(r, i)} \quad (\text{D.2})$$

em que K_r é o conjunto com os k itens mais similares ao item candidato r . Os N itens candidatos com maiores valores de score são recomendados ao usuário u_a .

O algoritmo de recomendação *cReduction* (Adomavicius et al., 2005) aplica a informação de contexto na etapa de pré-filtragem contextual. Nesse processo, os dados são separados para cada um dos possíveis contextos. Por exemplo, as páginas Web analisadas são separadas em segmentos de acordo com o dia da semana que foram acessadas. Em seguida, um modelo de recomendação do IBCF é construído para cada um dos segmentos. Em seguida, de acordo com o contexto da sessão ativa, seleciona-se o melhor modelo obtido dos diferentes segmentos para realizar a recomendação.

Na categoria de pós-filtragem contextual encontram-se os algoritmos *weightPoF* e *filterPoF* (Panniello and Gorgoglione, 2012). Nessa categoria, a informação contextual é usada após a construção de um modelo de recomendação tradicional para filtrar ou reordenar as recomendações. Para os dois algoritmos utilizados, primeiro é calculada a probabilidade de o usuário escolher um dado item em um determinado contexto, que será utilizada para reordenar as recomendações obtidas pelo IBCF. A probabilidade do usuário u_a acessar um item $i \in I$ no contexto c é dada por:

$$P_c(u_a, i) = \frac{Num_c(u, i)}{Num_c(u)} \quad (\text{D.3})$$

em que $Num_c(u, i)$ é o total de usuários em U que acessaram o item i no contexto c e $Num_c(u)$ é o total de usuários que acessaram qualquer item no contexto c . Em seguida, a nota de recomendação de um item $i \in I$ para um usuário u_a no contexto c para o algoritmo *weightPoF* é calculada por:

$$nota_c(u_a, O, i) = nota(u_a, O, i) \times P_c(u_a, i) \quad (\text{D.4})$$

Para o algoritmo *filterPof*, a probabilidade calculada é utilizada como um filtro no qual, dado um limiar P^* , itens i cujo valor de $P_c(u_a, i)$ seja menor que P^* recebem nota zero. Caso contrário, a nota $nota_c(u_a, O, i)$ é dada pelo valor da nota obtida pelo algoritmo IBCF.