Avaliação de métodos não-supervisionados de seleção de atributos para Mineração de Textos

Bruno Magalhães Nogueira

Avaliação de métodos não-supervisionados de seleção de atributos para Mineração de Textos

Bruno Magalhães Nogueira

Orientadora: *Profª Drª Solange Oliveira Rezende*

Dissertação apresentada ao Instituto de Ciências Matemáticas e de Computação - ICMC-USP como parte dos requisitos para obtenção do título de Mestre em Ciências de Computação e Matemática Computacional.

USP - São Carlos
Fevereiro/2009

Dedicatória

Aos meus pais,
Rita e José Geraldo.

Agradecimentos

A Deus pelas bênçãos alcançadas ao longo dessa jornada.

Aos meus pais, José Geraldo e Rita, pelo amor, apoio e compreensão irrestritos com que pude contar durante todo o tempo, em todas as etapas da minha vida. Se consegui chegar até aqui, foi porque me espelhei e me apoiei em vocês, exemplos de determinação. Agradeço também aos meus irmãos, Túlio e Thayse, pelo amor e camaradagem. Ter de deixar o convívio diário com todos vocês foi a decisão mais difícil que já tomei. Mas saibam que, mesmo a 600 quilômetros de distância, não tem um único dia que não penso em vocês e que nossa casa continua sendo meu porto seguro.

À professora Solange Rezende, pela confiança e orientação sem as quais eu não chegaria a esse ponto. Além de uma educadora exemplar, você é uma pessoa ímpar, um exemplo a ser seguido por todos. Obrigado pela oportunidade que me deu, pelos conhecimentos transmitidos, pela paciência sempre demonstrada e pela amizade conquistada.

À Vanessa, minha namorada, pelo companheirismo tão importante com que pude contar nesse tempo em que estamos juntos. Obrigado por ser essa pessoa tão especial, carinhosa e compreensiva, trazendo-me conforto nessa árdua caminhada.

À minha família, em especial às minhas avós Eva e Izabel (*in memoriam*), pelo apoio em toda a minha vida e pela compreensão nesses tempos de ausência.

Ao pessoal do grupo de pesquisa em Mineração de Textos, Fabiano Santos, Maria Fernanda Moura, Merley Conrado, Rafael Rossi, Ricardo Marcacini e Tatiane Nogueira, pelas dicas e risadas das nossas reuniões. Em especial, agradeço à Maria Fernanda pelas inúmeras contribuições para a realização deste trabalho.

Aos amigos e professores do LABIC, pelo companheirismo e pelos bons momentos de convivência, dentro e fora do laboratório: André Maletzke, André Rossi, Carlos Ferrero, Edson Matsubara, Gustavo Batista, Ígor Braga, Leonardo Almeida, Márcio Basgalupp, Maria Carolina Monard, Merley Conrado, Rafael Giusti, Renato Silva, Robson Motta, Ronaldo Prati e Victor Laguna.

Aos amigos de Belo Horizonte, pelo companheirismo mesmo à distância: Bruno Marcos, Felipe Gonçalves, Leandro Cardoso, Maria Fernanda Fonseca e Rodrigo Costa.

Ao professor Eduardo Hruschka, pela contribuição ao processo de avaliação adotado neste trabalho.

A Anandsing Dwarkasing, pelas revisões de inglês de artigos e do *abstract* desta dissertação.

Às funcionárias do setor de Pós-Graduação do ICMC/USP, Ana Paula, Beth, Laura e Lívia, pelos excelentes serviços prestados à comunidade acadêmica dessa unidade.

À Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) pelo auxílio financeiro essencial para a realização deste trabalho.

A todos que, direta ou indiretamente, contribuíram para a realização deste trabalho.

Resumo

Selecionar atributos é, por vezes, uma atividade necessária para o correto desenvolvimento de tarefas de aprendizado de máquina. Em Mineração de Textos, reduzir o número de atributos em uma base de textos é essencial para a eficácia do processo e a compreensibilidade do conhecimento extraído, uma vez que se lida com espaços de alta dimensionalidade e esparsos. Quando se lida com contextos nos quais a coleção de textos é não-rotulada, métodos não-supervisionados de redução de atributos são utilizados. No entanto, não existe forma geral predefinida para a obtenção de medidas de utilidade de atributos em métodos não-supervisionados, demandando um esforço maior em sua realização. Assim, este trabalho aborda a seleção não-supervisionada de atributos por meio de um estudo exploratório de métodos dessa natureza, comparando a eficácia de cada um deles na redução do número de atributos em aplicações de Mineração de Textos. Dez métodos são comparados - *Ranking* por *Term Frequency*, *Ranking* por *Document Frequency*, *Term Frequency-Inverse Document Frequency*, *Term Contribution*, *Term Variance*, *Term Variance Quality*, Método de Luhn, Método LuhnDF, Método de Salton e *Zone-Scored Term Frequency* - sendo dois deles aqui propostos - Método LuhnDF e *Zone-Scored Term Frequency*. A avaliação se dá em dois focos, supervisionado, pelo medida de acurácia de quatro classificadores (C4.5, SVM, KNN e Naïve Bayes), e não-supervisionado, por meio da medida estatística de *Expected Mutual Information Measure*. Aos resultados de avaliação, aplica-se o teste estatístico de Kruskal-Wallis para determinação de significância estatística na diferença de desempenho dos diferentes métodos de seleção de atributos comparados. Seis bases de textos são utilizadas nas avaliações experimentais, cada uma relativa a um grande domínio e contendo subdomínios, os quais correspondiam às classes usadas para avaliação supervisionada. Com esse estudo, este trabalho visa contribuir com uma aplicação de Mineração de Textos que visa extrair taxonomias de tópicos a partir de bases textuais não-rotuladas, selecionando os atributos mais representativos em uma coleção de textos. Os resultados das avaliações mostram que não há diferença estatística significativa entre os métodos não-supervisionados de seleção de atributos comparados. Além disso, comparações desses métodos não-supervisionados com outros supervisionados (Razão de Ganho e Ganho de Informação) apontam que é possível utilizar os métodos não-supervisionados em atividades supervisionadas de Mineração de Textos, obtendo eficiência compatível com os métodos supervisionados, dado que não detectou-se diferença estatística nessas comparações, e com um custo computacional menor.

Abstract

Feature selection is an activity sometimes necessary to obtain good results in machine learning tasks. In Text Mining, reducing the number of features in a text base is essential for the effectiveness of the process and the comprehensibility of the extracted knowledge, since it deals with high dimensionalities and sparse contexts. When dealing with contexts in which the text collection is not labeled, unsupervised methods for feature reduction have to be used. However, there aren't any general predefined feature quality measures for unsupervised methods, therefore demanding a higher effort for its execution. So, this work broaches the unsupervised feature selection through an exploratory study of methods of this kind, comparing their efficacies in the reduction of the number of features in the Text Mining process. Ten methods are compared - Ranking by Term Frequency, Ranking by Document Frequency, Term Frequency-Inverse Document Frequency, Term Contribution, Term Variance, Term Variance Quality, Luhn's Method, LuhnDF Method, Salton's Method and Zone-Scored Term Frequency - and two of them are proposed in this work - LuhnDF Method and Zone-Scored Term Frequency. The evaluation process is done in two ways, supervised, through the accuracy measure of four classifiers (C4.5, SVM, KNN and Naïve Bayes), and unsupervised, using the Expected Mutual Information Measure. The evaluation results are submitted to the statistical test of Kruskal-Wallis in order to determine the statistical significance of the performance difference of the different feature selection methods. Six text bases are used in the experimental evaluation, each one related to one domain and containing sub domains, which correspond to the classes used for supervised evaluation. Through this study, this work aims to contribute with a Text Mining application that extracts topic taxonomies from unlabeled text collections, through the selection of the most representative features in a text collection. The evaluation results show that there is no statistical difference between the unsupervised feature selection methods compared. Moreover, comparisons of these unsupervised methods with other supervised ones (Gain Ratio and Information Gain) show that it is possible to use unsupervised methods in supervised Text Mining activities, obtaining an efficiency compatible with supervised methods, since there isn't any statistical difference the statistical test detected in these comparisons, and with a lower computational effort.

Esta dissertação foi preparada com o formatador de textos LaTeX. Foi utilizado um estilo (*style*) desenvolvido por Ronaldo Cristiano Prati. O sistema de citações de referências bibliográficas utiliza o padrão *Apalike* do sistema BibTeX.

Algumas palavras utilizadas neste trabalho não foram traduzidas da língua inglesa para a portuguesa por serem amplamente conhecidas e difundidas na comunidade acadêmica.

Sumário

Resumo ix
Abstract xi
Sumário xvi
Lista de Figuras xvii
Lista de Tabelas xx

Lista de Abreviaturas **xxi**

1 Introdução **1**

2 Mineração de Textos **5**
2.1 Processo de Mineração de Textos 5
2.1.1 Identificação do Problema 6
2.1.2 Pré-processamento 7
2.1.3 Extração de Padrões 8
2.1.4 Pós-processamento e Utilização do Conhecimento 9
2.2 Metodologia para Extração de Taxonomias de Tópicos 10
2.3 Considerações Finais 14

3 Redução de Dimensionalidade de Atributos em Mineração de Textos **15**
3.1 Representação de Documentos 16
3.2 Seleção e Extração de Atributos 18
3.2.1 Abordagens para Seleção de Atributos: Filtros, *Wrappers* e Híbridas 19
3.3 Métodos de Seleção Não-Supervisionada de Atributos 21
3.3.1 Baseados em Freqüência de Termos ou Freqüência de Documentos . 21
3.3.2 Baseados em Variância de Termos 24
3.3.3 Baseados em contexto 26
3.4 Considerações Finais 27

4 Metodologia para Avaliação Experimental **29**
4.1 Metodologia de Avaliação 29
4.1.1 Seleção do Corpus 29
4.1.2 Geração de Atributos 31

4.1.3 Montagem da Matriz Atributo-Valor . . . 32
4.1.4 Seleção de Atributos . . . 33
4.1.5 Avaliação dos Subconjuntos de Atributos . . . 34
4.2 Implementações Realizadas para Viabilização da Metodologia de Avaliação 37
4.3 Bases de Textos Selecionadas . . . 38
4.4 Considerações Finais . . . 40

5 Avaliação Experimental 41
5.1 Seleção dos Corpus, Geração de Termos e Montagem da Matriz Atributo-Valor . . . 41
5.2 Aplicação dos Métodos de Seleção de Atributos . . . 42
5.3 Avaliação dos Subconjuntos de Atributos . . . 43
5.3.1 Avaliação Não-Supervisionada . . . 43
5.3.2 Avaliação Supervisionada . . . 48
5.4 Considerações Finais . . . 54

6 Conclusões e Trabalhos Futuros 57
6.1 Principais Contribuições . . . 59
6.2 Limitações . . . 59
6.3 Trabalhos Futuros . . . 60

Referências 68

A Tabelas com os Resultados da Avaliação Não-Supervisionada 69

B Tabelas com os Resultados da Avaliação Supervisionada 73

Lista de Figuras

2.1 Etapas do processo de Mineração de Textos 6
2.2 Etapas da metodologia de extração de taxonomias de tópicos 11

3.1 Estados de busca de subconjuntos para quatro atributos 19
3.2 *Framework* de filtros para seleção de atributos 20
3.3 *Framework wrapper* para seleção de atributos 20
3.4 Ilustração dos Cortes de Luhn 22

4.1 Metodologia para avaliação de métodos não-supervisionados de seleção de atributos 30
4.2 Atividades para seleção do corpus trabalhado pela metodologia 30
4.3 Geração de atributos na metodologia de avaliação 31
4.4 Montagem de uma matriz atributo-valor a partir de uma coleção padronizada 32
4.5 Aplicação dos métodos de seleção não-supervisionada de atributos 33
4.6 Avaliação dos subconjuntos de atributos obtidos pelos diferentes métodos . 35

5.1 Valores de EMIM para os subconjuntos de atributos na base de textos *CIIS* 44
5.2 Valores de EMIM para os subconjuntos de atributos na base de textos *IA* . 44
5.3 Valores de EMIM para os subconjuntos de atributos na base de textos *IFM* 44
5.4 Valores de EMIM para os subconjuntos de atributos na base de textos *Chemistry* 45
5.5 Valores de EMIM para os subconjuntos de atributos na base de textos *CS* 45
5.6 Valores de EMIM para os subconjuntos de atributos na base de textos *Physics* 45
5.7 Valores de acurácia para os subconjuntos de atributos na base de textos *CIIS* 49
5.8 Valores de acurácia para os subconjuntos de atributos na base de textos *IA* 49
5.9 Valores de acurácia para os subconjuntos de atributos na base de textos *IFM* 50
5.10 Valores de acurácia para os subconjuntos de atributos na base de textos *Chemistry* 50
5.11 Valores de acurácia para os subconjuntos de atributos na base de textos *CS* 51
5.12 Valores de acurácia para os subconjuntos de atributos na base de textos *Physics* 51

Lista de Tabelas

3.1 Padrão de matriz atributo-valor 17

4.1 Descrição das bases de textos utilizadas neste trabalho 39

5.1 Resultados da geração de termos nas bases de textos utilizadas 42
5.2 Número de *stems* contido nos subconjuntos gerados 42
5.3 Pesos adotados para os termos pertencentes a cada zona dos documentos das bases *Chemistry*, *CS* e *Physics* 43
5.4 p-valores obtidos para o grupo 1 de comparações não-supervisionadas . . . 46
5.5 Valor de EMIM para os métodos Luhn, LuhnDF e Salton, obtidos para todas as bases 46
5.6 Valores de EMIM para as diferentes técnicas com subconjuntos de cardinalidade igual ao método de Luhn 47
5.7 Valores de EMIM para as diferentes técnicas com subconjuntos de cardinalidade igual ao método de Luhn-DF 47
5.8 Valores de EMIM para as diferentes técnicas com subconjuntos de cardinalidade igual ao método de Salton 47
5.9 p-valores obtidos para o grupo 1 de comparações supervisionadas 52
5.10 Resultados do pós-teste de Dunn para as comparações em que houve diferença estatística na base CIIS 53
5.11 P-valores obtidos para o grupo 3 de comparações supervisionadas 54

A.1 Valores de EMIM obtidos na variação de porcentagem de atributos selecionados para as bases CIIS, IA e IFM 70
A.2 Valores de EMIM obtidos na variação de porcentagem de atributos selecionados para as bases Chemistry, CS e Physics 71

B.1 Acurácias preditivas e desvios padrões obtidos com variação de porcentagem de atributos na base de textos CIIS 74
B.2 Acurácias preditivas e desvios padrões obtidos com variação de porcentagem de atributos na base de textos IA 75
B.3 Acurácias preditivas e desvios padrões obtidos com variação de porcentagem de atributos na base de textos IFM 76

B.4 Acurácias preditivas e desvios padrões obtidos com variação de porcentagem de atributos na base de textos Chemsitry 77
B.5 Acurácias preditivas e desvios padrões obtidos com variação de porcentagem de atributos na base de textos CS 78
B.6 Acurácias preditivas e desvios padrões obtidos com variação de porcentagem de atributos na base de textos Physics 79
B.7 Valores de acurácia obtidos com os diferentes classificadores para os subconjuntos gerados pelos métodos Luhn, LuhnDF e Salton 80
B.8 Acurácia preditiva obtida pelos diferentes métodos de seleção de atributos com subconjuntos de cardinalidade igual ao método de Luhn 81
B.9 Acurácia preditiva obtida pelos diferentes métodos de seleção de atributos com subconjuntos de cardinalidade igual ao método de LuhnDF 81
B.10 Acurácia preditiva obtida pelos diferentes métodos de seleção de atributos com subconjuntos de cardinalidade igual ao método de Salton 82

Lista de Abreviaturas

CIIS	Base de textos sobre *Case Based Reasoning, Inductive Logic Programming, Information Retrieval e Sonification*
CS	Base de textos sobre Ciência da Computação (*Computer Science*)
DF	*Document Frequency*
EMIM	*Expected Mutual Information Measure*
GR	*Gain Ratio*
IA	Base de textos sobre Inteligência Artificial
IDF	*Inverse Document Frequency*
IFM	Base de textos do Instituto Fábrica do Milênio
IESystem	*Information Extraction System*
IG	*Information Gain*
KNN	*k-Nearest Neighbor*
LABIC	Laboratório de Inteligência Computacional
LSA	*Latent Semantic Analysis*
LuhnDF	Adaptação do método de Luhn para *Document Frequency*
MT	Mineração de Textos
PCA	*Principal Component Analysis*
PreText	Ferramenta para Pré-Processamento de Textos
RDF	*Ranking* por *Document Frequency*
RTF	*Ranking* por *Term Frequency*
SVM	*Support Vector Machines*

TC *Term Contribution*

TF *Term Frequency*

TFIDF *Term Frequency-Inverse Document Frequency*

TV *Term Variance*

TVQ *Term Variance Quality*

XML *Extensible Markup Language*

ZSTF *Zone-Scored Term Frequency*

Capítulo

1

Introdução

A quantidade de informação no universo digital era de 161 hexabytes, de acordo com estimativas feitas no ano de 2006, o que correspondia a 3 milhões de vezes a quantidade de informação contida em todos os livros já escritos. No período de 2006 a 2010, estima-se que o crescimento de informações no universo digital será de 161 hexabytes para 988 hexabytes (Gantz et al., 2007). Ante a esse cenário, a capacidade humana de analisar e compreender esses dados é excedida. Dessa forma, grande parte da informação e do conhecimento potencialmente útil que poderia ser retirado dessas bases é perdida. Como conseqüência, o desenvolvimento de ferramentas e técnicas de extração de conhecimento de repositórios de dados tem se tornado cada vez mais indispensável (Ebecken et al., 2003).

O processo de Mineração de Textos (MT), por meio de técnicas computacionais de extração de conhecimento, atua nesse cenário como um agente transformador, extraindo dessa enorme quantidade de dados textuais conhecimento útil que pode ser utilizado como vantagem competitiva e suporte à tomada de decisão. Esse processo pode ser visto como um caso particular do processo de Mineração de Dados, sendo composto por cinco etapas: Identificação do Problema, Pré-Processamento, Extração de Padrões, Pós-Processamento e Uso do Conhecimento. Essas etapas podem ser instanciadas de acordo com os objetivos do processo (Rezende et al., 2003).

A etapa de Pré-Processamento em Mineração de Textos é uma etapa crucial para a viabilidade do processo. Essa etapa visa transformar a coleção de documentos em um formato manipulável por algoritmos de aprendizado, envolvendo tarefas como tratamento, limpeza e redução dos dados. Além disso, as atividades desenvolvidas nessa etapa determinam a qualidade dos dados analisados e, conseqüentemente, a eficiência dos algoritmos de aprendizado e a confiabilidade dos resultados obtidos. Ao fim do Pré-Processamento, obtém-se uma matriz atributo-valor, na qual cada linha corresponde a um documento e cada coluna corresponde a um termo presente na coleção de documentos, de maneira que as posições internas da tabela representam uma medida correspondente a um termo em

um determinado documento. Nesse tipo de transformação, o número de palavras candidatas a atributos excede o número de documentos em mais de uma ordem de magnitude (Forman, 2003), gerando matrizes esparsas e de alta dimensionalidade.

Em geral, ao lidar com matrizes que apresentam essas características, o desempenho de algoritmos de aprendizado cai drasticamente. Algoritmos de agrupamento de dados, por exemplo, são, geralmente, muito sensíveis à alta dimensionalidade e a dados esparsos. Além disso, a grande dimensionalidade exige um alto tempo de processamento, tornando, por vezes, inviável a solução do problema. A fim de reduzir o tamanho da matriz atributo-valor e viabilizar o uso de algoritmos de aprendizado de máquina, faz-se necessário um processo de redução da dimensionalidade de atributos, de maneira que permaneçam na matriz atributo-valor apenas os termos[1] mais representativos da coleção de textos.

A atividade de redução de dimensionalidade de atributos é adotada na etapa de Pré-Processamento da Mineração de Textos. Por exemplo, a redução do número de atributos é uma das atividades da etapa de Pré-Processamento na metodologia de Moura et al. (2008a), na qual uma taxonomia de tópicos é extraída a partir de uma coleção de documentos[2] não-rotulados relativa a um domínio específico, possibilitando a organização hierárquica desses documentos e da informação neles contida. Em contextos não-supervisionados, como é o caso dessa metodologia, a redução de dimensionalidade de atributos é uma tarefa especialmente difícil, sendo, hoje, um dos principais problemas defrontados em Mineração de Textos (Guyon e Elisseeff, 2003). Ao contrário dos contextos supervisionados, nos quais a importância de um atributo pode ser medida em função de sua utilidade na predição de valores do atributo classe, a qualidade de um atributo em atividades não-supervisionadas varia de acordo do objetivo da tarefa a ser realizada.

Nesse sentido, o presente trabalho tem por objetivo o estudo exploratório de técnicas não-supervisionadas de seleção de atributos para aplicações de Mineração de Textos. Com essa análise, este trabalho pretende colaborar com a metodologia proposta por Moura et al. (2008a), atuando na tarefa de redução de atributos para a obtenção de taxonomias de tópicos. Entretanto, é necessário ressaltar que as análises aqui realizadas e os respectivos resultados obtidos são genéricos para tarefas de aprendizado não-supervisionado com bases textuais.

Portanto, a hipótese que rege este trabalho é que é possível, por meio de técnicas de seleção não-supervisionada de atributos, "otimizar" conjuntos de atributos oriundos de bases não-rotuladas em tarefas não-supervisionadas de Mineração de Textos. Por "otimizar" entende-se a obtenção de um conjunto de atributos de menor dimensionalidade, mas não menos representativo ao objetivo do processo, isto é, que não tenham perda de informação importante relativa ao domínio do problema.

Embora grande parte dos estudos presentes na literatura da área apresentem comparações de técnicas de seleção de atributos supervisionadas, o foco deste trabalho está no estudo comparativo de técnicas de seleção de atributos não-supervisionadas, pois tarefas de aprendizado não-supervisionado, de modo geral, lidam exclusivamente com bases

[1] Neste trabalho, *termo*, *atributo* e *palavra* são, por vezes, utilizados com o mesmo significado, podendo se referir a elementos simples ou compostos, como, por exemplo, "engenharia", "engenharia civil" ou "engenharia de software".

[2] Os termos *texto* e *documento* são utilizados indistintamente neste trabalho.

de textos não-rotulados. Dez métodos de seleção de atributos são comparados, a saber: *Ranking* por *Term Frequency* (RTF), *Ranking* por *Document Frequency* (RDF), *Term Frequency Inverse Document Frequency* (TFIDF) (Salton e Buckley, 1987), Método de Luhn (Luhn, 1958), Método de Salton (Salton et al., 1975), Método LuhnDF, *Term Contribution* (TC) (Liu et al., 2003), *Term Variance* (TV) (Liu et al., 2005), *Term Variance Quality* (TVQ) (Dhillon et al., 2003) e *Zone-Scored Term Frequency* (ZSTF). Desses métodos, Luhn-DF e ZSTF são aqui propostos.

Os subgrupos de termos obtidos por essas diferentes técnicas de seleção de atributos são avaliados por medidas objetivas que mensuram a qualidade de cada um desses subgrupos. Neste trabalho, optou-se, como critério de qualidade dos diferentes métodos de seleção de atributos comparados, avaliar a perda de informação acarretada pela eliminação de atributos ao utilizar cada um desses métodos, no intuito de detectar quais métodos são mais propensos a eliminar termos importantes ao domínio considerado. Com essa finalidade, foram aplicados dois métodos de avaliação aos subconjuntos gerados pelas diferentes técnicas de seleção de atributos. Uma primeira medida de avaliação, não-supervisionada, se deu por meio da medida estatística de *Expected Mutual Information Measure* (EMIM). Uma segunda forma de avaliação, supervisionada, foi realizada usando medidas de acurácia preditiva de quatro diferentes classificadores muito utilizados para tarefas de classificação de textos (C4.5, SVM, KNN e *Naïve* Bayes). Para que fosse possível realizar essa segunda categoria de avaliação, bases de textos rotuladas foram utilizadas ao longo dos experimentos. Os valores de classe dos documentos, entretanto, foram utilizados apenas para o processo de avaliação dos subconjuntos, sendo desconsiderados no processo de seleção de atributos.

Este trabalho está dividido da seguinte forma: no Capítulo 2 são apresentados o processo de Mineração de Textos e a metodologia de extração de taxonomias de tópicos com a qual este trabalho pretende colaborar. No Capítulo 3, aspectos de redução de dimensões de atributos são apresentados, bem como são discutidos métodos para seleção não-supervisionada de atributos. Já no Capítulo 4, apresenta-se a metodologia de avaliação experimental aqui adotada, bem como as bases de textos aplicadas. Em seguida, no Capítulo 5 são mostrados e discutidos os respectivos resultados obtidos na avaliação experimental desenvolvida neste trabalho. Por fim, no Capítulo 6, apresenta-se as conclusões obtidas com este trabalho, bem como são levantados pontos a serem desenvolvidos em trabalhos futuros.

Capítulo

2

Mineração de Textos

Em um contexto em que grande parte dos dados corporativos encontra-se disponível em forma textual, o processo de Mineração de Textos surge como uma poderosa ferramenta de apoio à gestão de conhecimento. Ebecken et al. (2003) definem a Mineração de Textos como "um conjunto de técnicas e processos que descobrem conhecimento inovador nos textos". Nesse sentido, o objetivo da prática de Mineração de Textos centra-se na busca por padrões, tendências e regularidades em documentos escritos em linguagem natural.

O processo de Mineração de Textos pode ser dividido em uma seqüência de etapas genéricas, formando um ciclo no qual, ao final, obtém-se o conhecimento acerca dos dados analisados. Esse processo pode ser instanciado de acordo com a necessidade de cada aplicação, como é o caso da metodologia apresentada por Moura et al. (2008a), em que o processo de Mineração de Textos é instanciado com o objetivo de extrair taxonomias de tópicos a partir de bases textuais. Neste capítulo, são descritas as etapas constituintes do processo de Mineração de Textos, destacando, posteriormente, como cada uma delas é instanciada na metodologia de extração de taxonomias de tópicos com a qual este trabalho colabora.

2.1 Processo de Mineração de Textos

O processo de Mineração de Textos, como uma especialização do processo de Mineração de Dados, pode ser dividido em cinco grandes fases: Identificação do Problema, Pré-processamento, Extração de Padrões, Pós-Processamento e Utilização do Conhecimento. Na Figura 2.1 é ilustrado o ciclo formado por essas etapas, as quais são discutidas nas próximas seções deste trabalho.

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

2.1.1 Identificação do Problema

Nessa etapa, delimita-se o escopo do problema com o qual irá trabalhar-se, definindo o objetivo da aplicação do processo de Mineração de Textos. Basicamente, o analista define a base de textos com a qual irá trabalhar, o que se espera obter com a análise dos dados e como o resultado da análise pode ser utilizado no contexto de sua organização. Nesse sentido, Rezende et al. (2003) definem quatro questões essenciais a serem respondidas nessa etapa:

- Quais são as principais metas do processo?
- Quais critérios de desempenho são importantes?
- O conhecimento extraído deve ser compreensível a seres humanos ou um modelo do tipo caixa-preta é apropriado?
- Qual deve ser a relação entre simplicidade e precisão do modelo extraído?

As decisões tomadas nesse ponto guiarão os passos consecutivos e poderão ter reflexo no desempenho da aplicação. Nesse sentido, com o auxílio de especialistas no domínio, uma atividade imprescindível a ser realizada nessa fase é um estudo acerca do domínio do problema para se adquirir um conhecimento inicial, visando o auxílio na tomada de decisão nas etapas subseqüentes do processo (Fayyad et al., 1996).

Outro aspecto de grande importância é a definição da coleção de documentos a ser utilizada, devendo-se selecionar textos que sejam mais relevantes ao domínio e aplicação do conhecimento a ser extraído, podendo esses ser obtidos de diversas fontes, como livros e artigos disponíveis na forma eletrônica e documentos da *Web* (Imamura, 2001).

Essa é uma atividade crítica, uma vez que os dados podem não estar disponíveis no formato adequado, como documentos não-digitalizados ou mesmo não-rotulados, em casos de atividades de classificação.

2.1.2 Pré-processamento

A etapa de Pré-Processamento é uma das etapas que mais tempo consome durante todo o ciclo. Nesse ponto do processo, busca-se colocar os dados em um formato adequado para a extração de conhecimento, sendo necessárias atividades de tratamento, limpeza e redução do volume de dados contido na base.

Em um aspecto mais geral da coleção de documentos, o analista deve verificar nesse ponto características que assegurem a confiabilidade, não-redundância e o balanceamento de sua coleção. Nesse sentido, Moura (2006) cita uma série de ações assistidas por especialistas que podem ser tomadas quando necessárias, como:

- Eliminação de repetições de documentos;
- Balanceamento da coleção por reamostragem;
- Redução da quantidade de documentos, quando o objetivo assim permitir;
- Verificação da existência de uma estrutura prévia nos documentos, a fim de utilizar essa informação na estruturação final da coleção;
- Análise do tamanho dos documentos na coleção, verificando a necessidade de uma normalização dos pesos atribuídos aos termos em função do tamanho dos textos.

Após a filtragem da base de documentos, no Pré-Processamento de textos deve-se buscar a estruturação dos documentos de maneira a torná-los processáveis por parte dos algoritmos de extração de padrões. A transformação mais comum para dados textuais é a representação desses em tabelas atributo-valor, de maneira que cada linha corresponda a um documento da coleção e cada coluna corresponda a um termo presente nos documentos da coleção. A cada célula da tabela formada é associada uma medida que relacione um termo a um documento. Uma definição mais detalhada das matrizes atributo-valor utilizadas em processos de Mineração de Textos, bem como das medidas de relação entre termo e documento são apresentadas na Seção 3.1 deste trabalho.

De modo geral, os termos presentes nas coleções textuais são previamente analisados e tratados, de maneira a gerar atributos mais consistentes para a formação das tabelas atributo-valor. Em um esforço inicial, busca-se desconsiderar dos textos os termos que não representem conhecimento útil, via eliminação de *stopwords*. Essas palavras não são relevantes para a análise dos textos, sendo geralmente constituídas por preposições, pronomes, artigos, interjeições, dentre outras. Para alguns domínios, também é comum existirem *stoplists* de domínio, que são conjuntos de palavras que, ao lidar com aquele domínio, de acordo com o objetivo da aplicação de Mineração de Textos, podem ser desconsideradas. Posteriormente, busca-se identificar similaridades de significados entre palavras, como em casos de variações morfológicas ou de palavras sinônimas (Ebecken

et al., 2003). Para tal, pode-se, por exemplo, reduzir uma palavra à sua raiz por meio de processos de *stemming*, reduzir as palavras ao seu lema (lematização), substantivar os termos (substantivação) ou mesmo usar dicionários ou *thesaurus*. Além disso, é possível buscar na coleção a formação de termos compostos, ou n-gramas, que são termos formados por mais de um elemento, porém apresentando um único significado semântico (Manning e Schtze, 1999; Conrado e Rezende, 2008).

Uma vez formada, a matriz atributo-valor referente à coleção de documentos é inerentemente esparsa e de alta dimensionalidade, o que, por vezes, pode tornar o processo de análise computacionalmente muito custoso ou até mesmo inviável, além de afetar negativamente o resultado de alguns algoritmos de extração de conhecimento. É vital para o processo de análise, portanto, selecionar os termos mais relevantes da coleção de documentos, tornando o conjunto de termos com o qual se trabalha mais conciso, porém não menos representativo em relação ao conjunto original. Esse processo é conhecido como redução de dimensionalidade de atributos e, conforme explicitado no Capítulo 1, é o foco principal deste trabalho e será explicado com maior riqueza de detalhes no Capítulo 3.

2.1.3 Extração de Padrões

Delimitado o problema e devidamente representados os textos, o processo avança, então, para a fase de extração de padrões. As tarefas a serem realizadas nesse ponto são definidas de acordo com o objetivo final do processo de extração de conhecimento. Assim como na Mineração de Dados, pode-se resumir as principais atividades de extração de padrões em textos em duas grandes tarefas: **preditivas** e **descritivas**.

Atividades **preditivas** são aquelas que produzem um modelo descrito pelo conjunto de dados para predizer o valor de uma ou mais variáveis de interesse (rótulos). Atividades **descritivas**, por sua vez, produzem conhecimento novo baseado nos dados textuais por meio de padrões que podem ser interpretados por humanos (Kantardzic, 2003).

Para as atividades **preditivas** são utilizados algoritmos 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 e Baranauskas, 2003). Essa classe de algoritmos se divide em suas 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 reais. 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 não-supervisionados, os quais extraem padrões de dados não-rotulados. As principais tarefas desse ramo da Mineração de Textos são a obtenção de regras de associação, o agrupamento e a sumarização de documentos.

As **regras de associação** são relações lógicas inferidas entre dados correlacionados de uma ou mais bases de dados, analisados conjuntamente (Agrawal e Srikant, 1994). Em outras palavras, as regras de associação indicam relações entre dois conjuntos disjuntos de atributos L e R, representando tais relações na forma $L \rightarrow R$, na qual L é conhecido

como antecedente e R como o conseqüente da regra. Em Mineração de Textos, o número de regras de associação obtidas tende a ser muito grande, uma vez que cada vocábulo de um texto pode ser considerado um atributo e, como existem diversos vocábulos diferentes em um documento, várias relações podem ser encontradas. Portanto, as regras obtidas devem, também, ser mineradas até que delas seja possível obter um conhecimento útil e conclusivo, cabendo essa tarefa à etapa de Pós-processamento do processo de Mineração de Textos.

O **agrupamento de dados**, também conhecido como *clustering*, visa agrupar dados de acordo com alguma medida de similaridade entre eles. Basicamente, esse processo tem por objetivo formar grupos nos dados de maneira que dados dentro de um mesmo grupo sejam o mais similares possível, enquanto objetos de grupos diferentes apresentem máxima dissimilaridade. Como resultado do processo, pode-se obter agrupamentos *flat*, nos quais os grupos são isolados, ou hierárquicos, nos quais os grupos apresentam alguma estrutura hierárquica. A utilização de agrupamento de dados pode ser interessante para fins de mineração, exploração e sumarização de coleções de textos (Peltonen et al., 2002). Além disso, em operações de recuperação de informação, o agrupamento de dados é bastante aplicado, uma vez que um usuário, ao recuperar um documento, pode estar interessado em outros documentos similares a ele, os quais estariam agrupados no mesmo *cluster* (Chakrabarti, 2003).

Sumarização, por sua vez, é uma atividade que visa obter uma descrição compacta para um conjunto de dados. Em aplicações de Mineração de Textos, a aplicação principal de sumarização é na obtenção automática de resumos. Para tal, os documentos são agrupados por tópicos, extraindo a informação desses grupos em um único documento representativo. Assim, consegue-se reduzir o conteúdo do repositório sem, no entanto, perder as informações importantes. A extração do resumo é feita sentença por sentença, as quais são identificadas e etiquetadas (Rino e Pardo, 2003). Após isso, são atribuídas ponderações às sentenças em função de alguma medida, como a freqüência de um conjunto de interesse de palavras nas mesmas. Ao final, obtém-se um conjunto de sentenças no qual é mantida a ordem em que elas ocorreram no documento. É importante ressaltar que, embora a sumarização seja uma técnica passível de ser utilizada para redução de dimensionalidade em bases de dados textuais, essa utiliza aspectos lingüísticos referentes aos textos, o que foge ao escopo deste trabalho.

2.1.4 Pós-processamento e Utilização do Conhecimento

Uma vez obtidos os padrões acerca dos dados, esses devem ser analisados e interpretados. Deve-se avaliar, junto a um especialista, aspectos do conhecimento extraído de acordo com o contexto do problema, como a representatividade do conhecimento obtido, a novidade contida nos resultados e a maneira que o conhecimento deverá ser utilizado.

No que tange à validade do conhecimento, deve-se verificar se os padrões obtidos condizem com a realidade e são aplicáveis ao que se propôs no objetivo inicial. Por exemplo, em aplicações que envolvam tarefas preditivas, pode-se avaliar os modelos com medidas relativas à precisão na predição de dados não vistos, como taxa de erro, precisão e

cobertura. A avaliação de modelos descritivos é um pouco mais complicada, uma vez que o critério de avaliação varia de acordo com o objetivo do processo. Dessa forma, algumas vezes podem não existir medidas objetivas para avaliar o resultado do processo, ficando a avaliação a cargo de um especialista do domínio (avaliação subjetiva).

Outro aspecto importante refere-se à compreensibilidade do conhecimento. Os algoritmos de extração de padrões podem gerar uma quantidade de padrões muito grande, dificultando a compreensão por parte do usuário e fazendo necessária a aplicação de mecanismos que forneçam aos usuários apenas os padrões mais interessantes (Silberschatz e Tuzhilin, 1995). Em tarefas de Mineração de Textos, devido à presença de altas dimensionalidades, esse é um problema recorrente, demandando especial atenção (Carvalho et al., 2007).

Uma tarefa que tem um importante papel no auxílio ao especialista nesse ponto do processo, mas que também pode ser utilizada ao longo das demais etapas, é a **visualização** dos dados. Card et al. (1999) definem essa tarefa como "uma representação visual de dados interativa baseada no computador para ampliar a cognição". Diversos tipos de gráficos e diagramas podem ser utilizados ao longo de todo o processo pelo analista, facilitando a compreensão dos resultados das diversas etapas e auxiliando a tomada de decisão. A análise visual pode, por exemplo, indicar falhas no pré-processamento dos dados e a necessidade de refazê-lo. Além disso, diversas formas finais de conhecimento são facilmente associáveis a formas de visualização de informação, como é o caso do *clustering* hierárquico, cujo resultado final é uma árvore que pode ser visualizada, por exemplo, por meio de árvores de diretórios ou árvores hiperbólicas (Lamping et al., 1995; Marcacini, 2008).

Após ter percorrido todas as etapas do processo, o conhecimento encontra-se apto a ser utilizado pelo usuário. Com o correto desenvolvimento das etapas anteriores, pode-se garantir que o conhecimento, nesse ponto, é válido e útil, podendo ser aplicado no apoio à tomada de decisão para os aspectos estabelecidos nos objetivos do processo.

2.2 Metodologia para Extração de Taxonomias de Tópicos

Nesta seção, a metodologia de extração de taxonomias de tópicos com a qual o presente trabalho colabora é descrita. Embora o foco deste trabalho não se restrinja ao contexto dessa metodologia, a descrição da mesma faz-se necessária para exemplificar um processo instanciado de Mineração de Textos, com atividades e tarefas bem definidas. Além disso, a exposição de algumas especificidades desse processo permitem clarear alguns dos aspectos que nortearam as decisões tomadas ao longo deste trabalho. É necessário ressaltar que embora tenham sido adotadas decisões a fim de atender às necessidades específicas da metodologia, essas não impossibilitam a seleção de atributos, como sugerido neste trabalho, a outras tarefas de Mineração de Textos.

Conforme mencionado anteriormente, o processo de Mineração de Textos é um processo genérico, podendo ser instanciado de acordo com a necessidade de aplicação por uma adaptação das diferentes etapas. Nesse sentido, Moura (2006) propôs uma metodologia baseada no processo de Mineração de Textos para auxiliar a organização e o gerenciamento

da informação disponível na forma de documentos acerca de um domínio de conhecimento restrito. Dessa forma, o objetivo principal da metodologia é auxiliar o especialista do domínio a organizar os documentos sob uma taxonomia de tópicos, bem como auxiliá-lo a tomar decisões acerca da geração da mesma.

Considera-se taxonomia como uma classificação hierárquica de itens, na qual os itens inferiores são especializações do itens superiores (Adamo, 2001). Assim, uma taxonomia de tópicos pode ser aqui encarada como uma classificação hierárquica formada por tópicos extraídos de uma coleção de textos. Dessa forma, a cada nível da taxonomia pode-se associar recursos da base textual do domínio relativos ao tópico ali contido, facilitando a gerência das informações. Na Figura 2.2 são apresentados os passos constituintes da metodologia proposta para a extração dessas taxonomias.

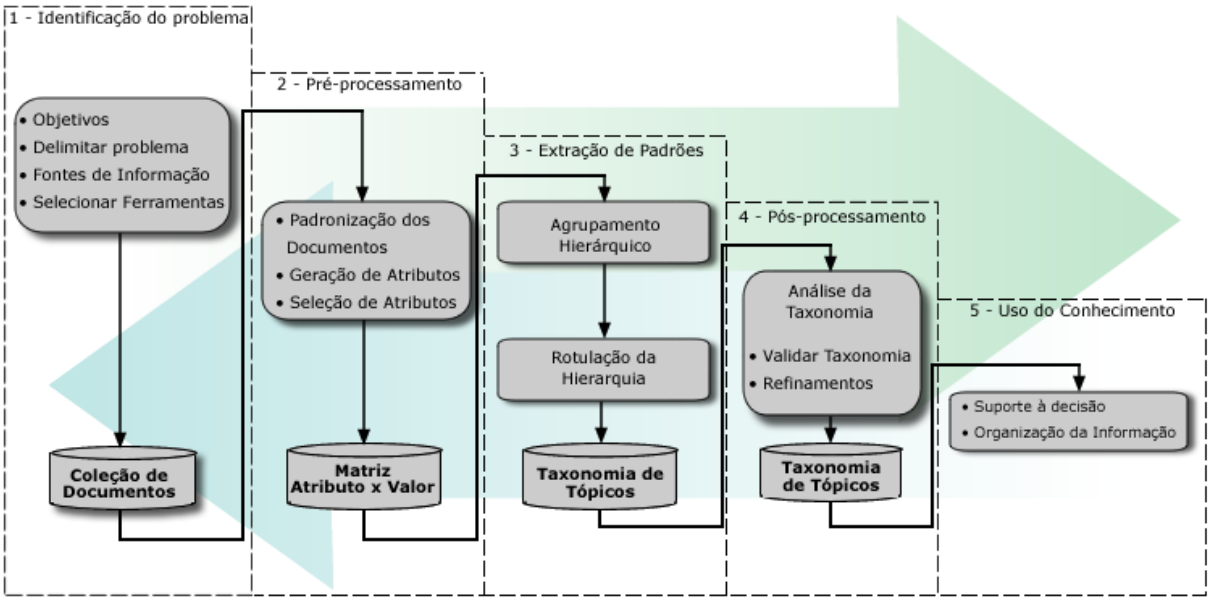


Figura 2.2: Etapas da metodologia de extração de taxonomias de tópicos

Na etapa de **Identificação do Problema**, o analista deve, juntamente com um especialista do domínio, delimitar a coleção de textos a ser utilizada. Deve-se selecionar documentos potencialmente úteis ao domínio específico com o qual se deseja trabalhar. Se a coleção de informações acerca do domínio contiver elementos não-textuais, como filmes e imagens, deve-se considerar os metadados acerca dos mesmos. Reunidos os dados de interesse, são criados documentos marcados com *tags* XML, a fim de separar informações como título do documento, autores, palavras-chave e corpo do texto. A partir dessas marcações, é possível ao especialista definir partes do documento que são do seu interesse, desprezando as demais (por exemplo, para um determinado fim pode ser interessante considerar apenas o título e o resumo dos documentos). Pode-se, então, armazenar as partes de interesse dos documentos em novos documentos reduzidos, formando, assim, a base de dados com a qual se irá trabalhar.

Já no passo de **Pré-Processamento**, o analista deve verificar a necessidade de alterações na coleção de textos, como o descarte de documentos desnecessários. Feito isso, passa-se à construção da matriz atributo-valor, sob a qual é estruturada a coleção de documentos de maneira que essa coleção seja manipulada pelos algoritmos de extração de padrões.

Inicialmente, uma filtragem de termos é efetuada na coleção de documentos. Para facilitar a identificação de termos similares, essa metodologia permite, de acordo com a necessidade do problema ou a preferência do usuário, a escolha de um dos seguintes métodos para a normalização dos termos:

- Redução do termo à sua raiz (*stemming*) (Krovetz, 1993);
- Redução do termo ao seu lema (lematização) (Arampatzis et al., 2000);
- Substantivação dos termos (Conrado e Rezende, 2008);
- Mapeamento de termos variantes - sinônimos, abreviações, acrônimos e ortografias alternativas - para um único termo por meio de *thesaurus* (Ebecken et al., 2003).

Feito isso, passa-se, então, à remoção de *stopwords*, removendo palavras pouco ou nada representativas para a extração de conhecimento, como preposições, pronomes e artigos. Em alguns casos, é possível haver uma *stoplist* (conjunto de *stopwords*), contendo palavras que semanticamente não trazem informação importante ao domínio da aplicação, devendo também ser descartadas nesse momento.

Logo a seguir, a metodologia utiliza a busca de termos compostos por uma ou mais palavras, conhecidos como *n-gramas*, onde n indica o número de elementos que compõem esses termos (Mladenic, 1998; Imamura, 2001). Deve-se escolher nesse ponto o número máximo de *gramas* que um termo poderá possuir, uma vez que, quanto maior o número de *gramas* procuradas, maiores serão o tempo de processamento e os recursos computacionais necessários para essa busca. Além disso, quanto maior for o número de *gramas* que um termo composto puder assumir, maiores serão as possibilidades de combinações possíveis entre termos simples. Conseqüentemente, um maior número de atributos na base será gerado, aumentando sua dimensionalidade.

Desse modo, forma-se a matriz atributo-valor que representa a coleção de textos. Nessa metodologia, os atributos da matriz correspondem aos termos selecionados no passo anterior, enquanto os objetos são os documentos presentes na coleção. Em cada célula interna da matriz, essa metodologia adota a contabilização da freqüência absoluta de um termo em um documento.

A matriz formada, entretanto, é inerentemente esparsa e de alta dimensionalidade, o que pode ser extremamente prejudicial ao processo de extração de conhecimento. Nesse sentido, o próximo passo da metodologia sugere a aplicação de alguns métodos para eliminar atributos, visando diminuir a dimensionalidade da matriz. A fim de manter a configuração original dos atributos presentes na base, a utilização de métodos de extração de atributos, como a combinação linear de atributos, para redução de dimensionalidade não é indicada, sendo adotada na metodologia a utilização de métodos de seleção de atributos. Essa metodologia, por lidar com bases textuais de grande dimensionalidade, sugere a utilização da abordagem de filtros para seleção de atributos devido ao seu menor custo computacional, embora seja, também, possível a utilização de abordagens *wrappers* (Mladenic, 1998). Uma vez que os textos são não-rotulados, somente processos de seleção não-supervisionada de atributos são aplicáveis. Como o objetivo deste trabalho é o estudo

exploratório de técnicas para esse fim, o processo de seleção será alvo de explicação no Capítulo 3.

Terminada a etapa de pré-processamento, passa-se à etapa de **Extração de Padrões**. Dado que o objetivo principal na metodologia é fornecer ao especialista uma taxonomia de tópicos, a etapa de extração de padrões efetua um processo de agrupamento hierárquico de documentos, de maneira a agrupar documentos com conteúdos semelhantes. A metodologia considerada utiliza métodos que trabalham sobre uma matriz de similaridade entre os objetos. Por exemplo, podem ser aplicados os métodos de agrupamento hierárquico *complete linkage*, *average likage* ou *single linkage* (Mardia et al., 1979).

Uma vez gerada a hierarquia, os agrupamentos obtidos refletem tópicos ou sub tópicos aos quais os documentos se referem. Passa-se, então, à geração dos rótulos para cada grupo encontrado pela obtenção dos termos mais significativos. Essa atividade pode ser considerada um processo de seleção supervisionada de atributos, considerando que os agrupamentos seriam as classes presentes na base de textos (Weiss et al., 2005), existindo diversos métodos para tal fim (Moura e Rezende, 2007).

Na etapa de **Pós-Processamento**, realiza-se a visualização e a validação da taxonomia de tópicos obtida. No processo de **visualização**, a taxonomia resultante do processo é exibida. Uma vez que o conhecimento extraído é de natureza hierárquica, uma maneira intuitiva de exibí-la ao usuário é por meio de árvores. Assim, a metodologia considerada representa o conhecimento extraído em árvores do conhecimento (Souza et al., 2005). Uma árvore do conhecimento é uma estrutura hierárquica em formato de árvore que representa o conhecimento acerca de um domínio, de maneira que os elementos mais genéricos se encontram na raiz da estrutura e os mais específicos se distribuem nos níveis mais baixos. Cada um dos nós da estrutura é identificado com termos acerca de seu conteúdo, caracterizando uma taxonomia. É possível, ainda, adicionar a cada nó recursos de informação de tópicos ali representados, como documentos e imagens associadas.

Essa metodologia sugere a exibição dessas árvores para o usuário por meio de técnicas como as árvores de diretórios (*folder trees*), cuja principal vantagem é a fácil adição de descrição dos nodos, ou árvores hiperbólicas (Lamping et al., 1995), as quais são muito indicadas por facilitarem a visualização de detalhes da hierarquia sem, no entanto, perder a noção de contexto global. Entretanto, por não estar a metodologia fixa a uma determinada ferramenta, outras formas de visualização podem ser utilizadas, variando de acordo com a necessidade da aplicação.

Com a visualização dos resultados, inicia-se o processo de **validação** dos dados, no qual, com a intervenção do especialista, busca-se avaliar os resultados e estabelecer critérios para ajustes na taxonomia obtida, como a poda da árvore de *cluster* ou mesmo edição da taxonomia obtida. Dado que esse é um processo iterativo, caso o especialista julgue necessário efetuar mudanças mais profundas na taxonomia, volta-se à etapa anterior do processo na qual deseja-se fazer a alteração, como as etapas de pré-processamento e geração dos rótulos. Caso contrário, o conhecimento, expresso na forma de taxonomia de tópicos, está pronto para ser utilizado (etapa de **Utilização do Conhecimento**).

2.3 Considerações Finais

A extração de conhecimento a partir de bases de dados textuais tornou-se um importante aliado às mais diversas corporações, trazendo dos documentos textuais conhecimento útil que auxilia o processo de tomada de decisão. A Mineração de Textos é o elemento intermediário entre os dados textuais e o conhecimento, extraindo dos documentos informações novas e potencialmente úteis por meio de padrões detectados ao longo das coleções.

O processo de Mineração de Textos é constituído de uma seqüência de cinco passos: identificação do problema, pré-processamento, extração de padrões, pós-processamento e utilização do conhecimento. Cada um desses é constituído de atividades genéricas, as quais podem ser instanciadas de acordo com a especificidade de cada aplicação por meio de, por exemplo, seleção do algoritmo de aprendizado que melhor cumpra o objetivo final, ou de alguma medida que selecione o subconjunto de atributos mais representativo daquela coleção.

Como exemplo de um processo instanciado, cita-se na Seção 2.2 deste trabalho uma aplicação de Mineração de Textos voltada para a organização de conhecimento textual por meio da extração de uma taxonomia de tópicos de uma coleção de documentos. A característica esparsa e de alta dimensionalidade das matrizes atributo-valor geradas no Pré-Processamento de Mineração de Textos, prejudica, especialmente, o desempenho de tarefas de agrupamento, como é o caso da referida metodologia para extração de taxonomias de tópicos. Faz-se necessário, então, adotar processos de eliminação de atributos não-necessários, visando a redução de dimensionalidade e tornar o espaço o menos esparso possível. Dessa forma, este trabalho visa colaborar com essa metodologia, atuando na etapa de pré-processamento da mesma, por meio de uma avaliação de técnicas de seleção não-supervisionada de atributos potencialmente aplicáveis para uma eficiente redução de atributos, descrita com maior riqueza de detalhes no Capítulo 3.

Capítulo

3

Redução de Dimensionalidade de Atributos em Mineração de Textos

Reduzir o número de atributos é uma tarefa importante em processos de aquisição automática de conhecimento, não somente no sentido de determinar as melhores variáveis para modelagem, mas também para aspectos de compreensibilidade e escalabilidade dos modelos resultantes (Kim et al., 2000). Yu e Liu (2004) destacam que, considerando conjuntos limitados de dados para algoritmos de aprendizado de máquina, o excesso de atributos causa lentidão no processo de treinamento, bem como decremento na qualidade do conhecimento extraído (*overfitting*, por exemplo). Em processos de Mineração de Textos, a redução de atributos assume um papel essencial para o sucesso do processo na medida em que os textos apresentam grande dimensionalidade e variabilidade de termos.

Basicamente, a idéia central da redução do número de atributos é reduzir o espaço de busca pela solução, obtendo um conjunto de atributos de menor dimensionalidade e consistente sem, no entanto, afetar a qualidade da solução final. Quanto à importância dos atributos para um domínio de problema, John et al. (1994) propõem a separação desses em três classes: fortemente relevantes, fracamente relevantes e irrelevantes. Atributos fortemente relevantes são sempre necessários em um subconjunto ótimo para a correta caracterização de um domínio de problema. Atributos fracamente relevantes são necessários apenas em alguns casos. Já os atributos irrelevantes não são necessários, podendo ser descartados. Dentre os atributos fracamente relevantes, encontram-se os atributos redundantes, que são aqueles cuja informação que carregam a respeito do domínio do problema pode ser predita por um ou mais atributos. É importante ressaltar que, embora tal definição seja formalizada em função da predição de classes em aprendizado supervisionado e não seja diretamente aplicável a contextos não-supervisionados, faz-se, aqui, um relaxamento da mesma para o aprendizado não-supervisionado, considerando a importância dos atributos na determinação de, por exemplo, grupos de documentos.

Encontram-se na literatura trabalhos relacionados à redução de atributos para textos. A grande maioria desses, porém, é voltada para dados rotulados (Yang e Pedersen, 1997; Forman, 2003; Saleh e El-Sonbaty, 2007; Chen et al., 2009; Dasgupta et al., 2007; Aghdam et al., 2009). Estudos comparativos em redução não-supervisionada de atributos, especialmente voltadas para dados textuais (Liu et al., 2005; Wiratunga et al., 2006; Gonzaga et al., 2007; Junling et al., 2007; Shafiei et al., 2007), são mais difíceis de serem efetuados e avaliados e, talvez por isso, mais raros na literatura. Mesmo nesses trabalhos, não há uma extensa comparação de métodos dessa natureza. A dificuldade de avaliação, por exemplo, emana do fato de que embora existam algumas medidas para avaliação da redução de atributos em aprendizado não-supervisionado, muitas vezes tais medidas não se encaixam ao contexto em que se está realizando a redução de atributos. Assim, é comum utilizar da subjetividade de especialistas para avaliar, ou mesmo transformar a avaliação para um contexto supervisionado.

Considerando que a metodologia de extração de taxonomias de tópicos com a qual este trabalho colabora lida com coleções de documentos não-rotuladas, efetua-se, aqui, uma análise exploratória de métodos não-supervisionados para seleção de atributos em Mineração de Textos. Nas seções seguintes deste capítulo, serão apresentados aspectos relativos à seleção de atributos em documentos, bem como serão discutidas algumas medidas encontradas na literatura para tal fim.

3.1 Representação de Documentos

Devido à natureza não-estruturada de documentos textuais, uma atividade essencial na etapa de pré-processamento em todo processo de Mineração de Textos se refere à estruturação desses, de maneira a torná-los manipuláveis por algoritmos de extração de conhecimento. Embora alguns algoritmos exijam informações mais sofisticadas, como os baseados em conhecimentos lingüísticos, a grande maioria dos algoritmos de extração de padrões necessita apenas que os documentos estejam em um formato de planilha, no qual as linhas correspondam aos documentos e as colunas aos termos presentes na coleção. Esse formato é denominado "*bag of words*", na qual os termos são considerados independentes, formando um conjunto desordenado em que a ordem de ocorrência das palavras não importa.

Uma forma de representar a *bag of words* é por meio de tabelas atributo-valor. Um exemplo dessa representação pode ser observado na Tabela 3.1 (Martins, 2003), na qual d_i corresponde ao i-ésimo documento, t_j representa o j-ésimo atributo (termo), a_{ij} é a medida que relaciona o i-ésimo documento com o j-ésimo atributo e y_i representa a classe do i-ésimo documento. É importante ressaltar que esse último elemento não aparece em casos nos quais a coleção de textos seja não-rotulada.

Dessa forma, cada documento pode ser representado como um vetor $\vec{d_i} = (\vec{a}_i, y_i)$, no qual $\vec{a}_i = (a_{i1}, a_{i2}, \ldots, a_{iM})$. Segundo Salton e Buckley (1987), o cálculo do valor de a_{ij} depende de três componentes:

- o componente do *documento*, relacionado à estatística de um determinado termo

	t_1	t_2	$\ldots$	t_M	Y
d_1	a_{11}	a_{12}	$\ldots$	a_{1M}	y_1
d_2	a_{21}	a_{22}	$\ldots$	a_{2M}	y_2
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
d_N	a_{N1}	a_{N2}	$\ldots$	a_{NM}	y_N

Tabela 3.1: Padrão de matriz atributo-valor

estar presente em um dado documento;

- o componente da *coleção*, relacionado à distribuição da presença do termo ao longo da coleção de documentos;

- o componente da *normalização*, o qual possibilita que documentos de diferentes tamanhos possam ser comparados segundo uma mesma escala.

Diversas medidas foram propostas para calcular os valores dos pesos a_{ij}. Essas medidas podem ser classificadas em dois tipos: **binárias** e **baseadas em freqüência**. **Pesos binários** indicam a ocorrência ou não de um termo em um determinado documento, podendo ser utilizados, por exemplo, para extrair informações relativas à semelhança de documentos a partir do número de termos em comum. Já os **pesos baseados em freqüência** visam contabilizar o número de ocorrências de um determinado termo em um dado documento, servindo como base para a extração de diversas medidas estatísticas na extração de padrões, sendo esses os mais comumente aplicados.

Dentre os pesos baseados em medidas de freqüência, pode-se distinguir duas classes: **pesos supervisionados**, que dependem de dados com valor de classe conhecido, medindo a importância de um dado atributo na determinação do valor de classe; e **pesos não-supervisionados**, aplicáveis a dados não-rotulados.

Dentre os pesos supervisionados, pode-se citar medidas como a *ConfWeight* (Soucy e Mineau, 2005), a Entropia, a Entropia Conjunta (Haykin, 2001) e a Informação Mútua (Yao, 2003). Como exemplos de pesos não-supervisionados, pode-se citar a *TF* (*term frequency*), a qual considera a freqüência absoluta dos termos nos documentos (van Rijsbergen, 1979); a *IDF* (*Inverse Document Frequency*) (Salton et al., 1975), que calcula a freqüência inversa do termo, favorecendo termos que aparecem em poucos documentos da coleção; e a *TFIDF* (Salton e Buckley, 1987), que consiste em uma combinação das duas anteriores (TF e IDF).

Como cada termo contido na coleção de documentos é candidato a tornar-se um atributo, pode-se afirmar que a representação da coleção tem por características a alta dimensionalidade e o fato de ser esparsa. A alta dimensionalidade é causada pelo grande número de termos contidos em uma coleção de textos. Já a natureza esparsa da matriz é ocasionada pelo fato de que grande parte dos termos ocorre somente em uma pequena parte dos documentos, de maneira que a medida de relação com todos os outros documentos da coleção será igual a zero. Dessa forma, é de extrema importância a redução do espaço de atributos para obter uma otimização dos resultados do processo de Mineração de Textos e melhorar a representatividade dos dados (Martins, 2003).

3.2 Seleção e Extração de Atributos

No sentido de reduzir a dimensionalidade dos atributos, existem duas técnicas comumente usadas: **extração de atributos** e **seleção 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 (Wyse et al., 1980). O novo conjunto gerado é, geralmente, de menor dimensão que o conjunto original, uma vez que os novos atributos são gerados a partir de combinações dos anteriores, substituindo-os. A principal desvantagem desse método é que os atributos gerados não mantêm uma correlação explícita com a configuração real do problema, tornando os modelos gerados de difícil interpretação (Dash e Liu, 1997). Como exemplos de técnicas, pode-se citar a *Principal Component Analysis* (PCA) (Jolliffe, 2002), *Latent Semantic Analysis* (LSA) (Landauer et al., 1998) e o agrupamento de palavras (Slonim e Tishby, 2000).

Selecionar atributos, por sua vez, remete à seleção de um subconjunto de atributos a partir do conjunto original, seguindo alguns critérios. Os atributos selecionados mantêm a interpretação física original, facilitando a compreensão do modelo gerado (Liu et al., 2005). Em um conjunto de dados que possui n atributos, existem $2^n - 1$ subconjuntos candidatos (excluindo-se o subconjunto vazio), devendo-se selecionar aquele que apresenta melhor desempenho segundo algum critério de avaliação. Um exemplo de busca nos possíveis subconjuntos de um conjunto de quatro atributos é ilustrado na Figura 3.1, na qual os atributos selecionados são representados pelos círculos escuros. Entretanto, a busca exaustiva pelo melhor subconjunto, em contextos de dimensionalidade muito alta, pode tornar-se inviável. Dessa forma, Dash e Liu (1997) sugerem a adoção de um critério de parada no processo de busca, o qual se constituiria de quatro passos básicos:

1. Um procedimento de geração do próximo subconjunto a ser analisado, podendo ser segundo alguma heurística, por meio de busca completa (ou exaustiva) ou aleatória;
2. Uma função de avaliação para os subconjuntos gerados;
3. Um critério de parada; e
4. Um procedimento de validação dos subconjuntos.

Os métodos de busca mais comuns são aqueles que efetuam uma busca gulosa. Esses podem ser divididos em dois tipos: *forward selection* e *backward elimination*. Os métodos *forward selection* iniciam a busca com um subconjunto vazio de atributos, colocando progressivamente um atributo, escolhido segundo algum critério de avaliação, a cada iteração. Já os métodos *backward elimination* iniciam o processo com todos os atributos, eliminando, a cada passo, o menos interessante deles. Essas abordagens podem gerar diferentes subconjuntos e, dependendo da aplicação e dos objetivos, uma pode ser mais indicada que a outra (Guyon et al., 2006). Existem, ainda, métodos que não efetuam busca para selecionar atributos, os quais, geralmente, fornecem *ranking* de atributos, sem apontar combinações de atributos que formariam subconjuntos "ótimos" locais.

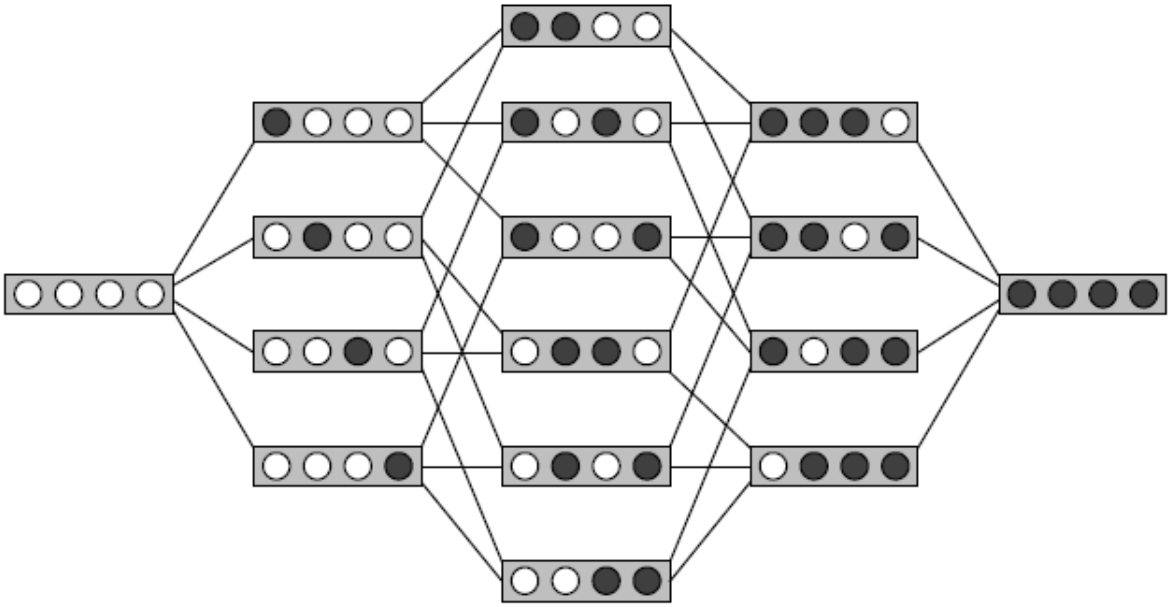


Figura 3.1: Estados de busca de subconjuntos para quatro atributos. Fonte: Langley (1994)

.

A escolha dos processos de seleção de atributos depende, também, da existência ou não de rótulo nos textos utilizados. Caso existam, pode-se usar métodos supervisionados de seleção de atributos, que avaliam a importância do atributo na determinação do valor de classe, como Ganho de Informação, χ^2, *odds ratio* e taxa de probabilidade (Forman, 2003). Por outro lado, métodos não-supervisionados devem ser utilizados quando se lida com dados não-rotulados. Mitra et al. (2002) dividem os métodos não-supervisionados em duas classes: aqueles que buscam maximizar o desempenho em problemas de agrupamento e aqueles que consideram a seleção de atributos baseada na relevância e dependência dos atributos.

Considerando que o objetivo principal deste trabalho é colaborar com uma metodologia de extração de uma taxonomia de tópicos a partir de bases textuais, é essencial que o conhecimento extraído ao final do processo seja compreensível ao usuário. Dessa forma, foram utilizados métodos de seleção de atributos, a fim de manter a relação física dos atributos selecionados com os atributos iniciais. Como essa metodologia aborda coleções de textos não-rotuladas, os métodos comparados são não-supervisionados, selecionando aqueles atributos que sejam mais relevantes na representação dos documentos. Uma vez que tarefas de Mineração de Textos lidam com um número muito grande de atributos, métodos que efetuam busca, em geral, não se mostram eficientes devido ao alto custo computacional. Assim, optou-se por métodos que não efetuam busca no processo de seleção dos atributos mais relevantes ao domínio.

3.2.1 Abordagens para Seleção de Atributos: Filtros, *Wrappers* e Híbridas

Existem três *frameworks* básicos para processos de seleção de atributos: filtros, *wrappers* e híbridos (Liu e Yu, 2005). Na abordagem dos filtros, os atributos são selecionados de maneira independente, considerando as características intrínsecas dos atributos, sem

qualquer *feedback* de informações do algoritmo de aprendizado. Basicamente, as variáveis são avaliadas de acordo com medidas de interesse utilizadas, de maneira a conservar na base de dados os atributos com melhor classificação. O esquema desse *framework* pode ser observado na Figura 3.2. Esse tipo de abordagem tem como principais vantagens o fato de que não herda nenhum tipo de bias de algoritmos de aprendizado e o seu baixo custo computacional, muito atrativo para tarefas de Mineração de Textos, em que lida-se com contextos de alta dimensionalidade.

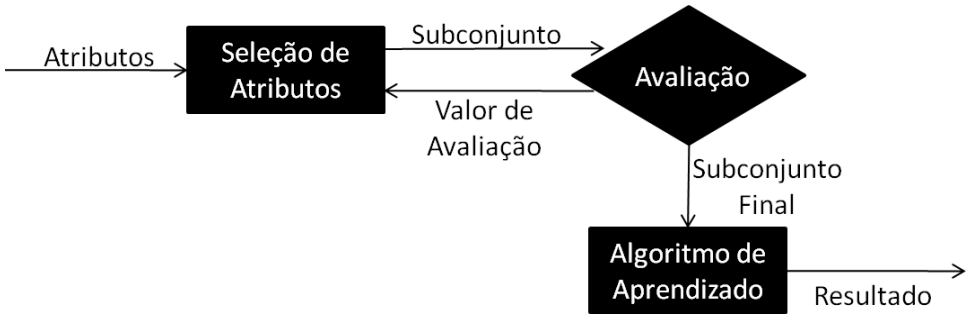


Figura 3.2: *Framework* de filtros para seleção de atributos

Já no *framework* de *wrappers*, o algoritmo de aprendizado é utilizado no processo de busca por um subconjunto quasi-ótimo de atributos. Cada subconjunto de atributos selecionado é aplicado ao algoritmo de aprendizado, que é considerado como uma "caixa-preta". A avaliação de cada subconjunto é relacionada ao desempenho do algoritmo de aprendizado, selecionando-se aquele que apresentar o melhor resultado. Um *framework wrapper* é exemplificado na Figura 3.3. Liu e Yu (2005) destacam que, uma vez que essa abordagem utiliza um algoritmo de aprendizado para controlar a qualidade dos subconjuntos obtidos, os *wrappers* tendem a apresentar melhores resultados na fase de extração de padrões. Isso se deve ao fato de que, geralmente, os algoritmos de aprendizado utilizados nessa etapa são os mesmos utilizados para controlar a seleção dos subconjuntos de atributos. Entretanto, os *wrappers* apresentam um custo computacional maior que o apresentado pelos métodos de filtro de atributos (John et al., 1994).

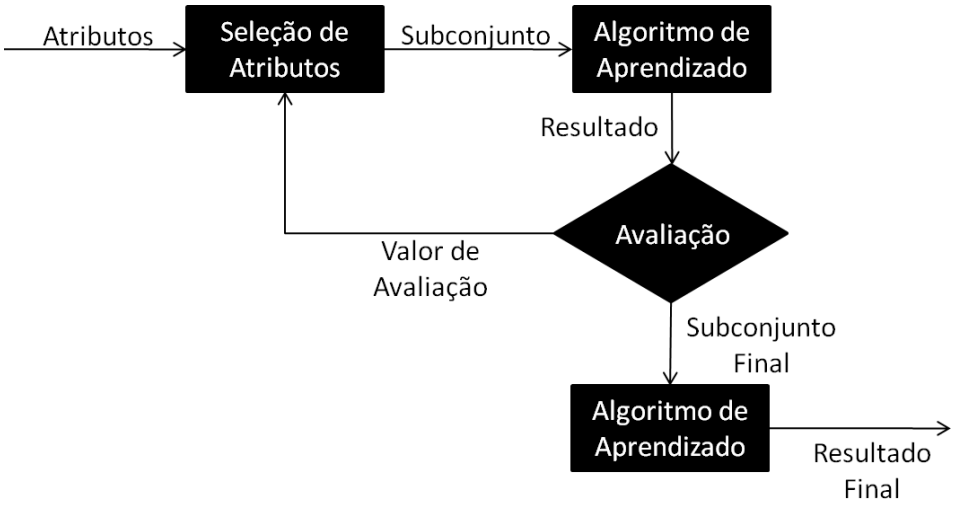


Figura 3.3: *Framework wrapper* para seleção de atributos

A abordagem híbrida combina os benefícios de ambas as abordagens anteriormente descritas, utilizando tanto uma medida independente quanto um algoritmo de aprendizado

para avaliar subconjuntos. Usualmente, a medida independente é utilizada para avaliar subconjuntos de atributos de uma determinada cardinalidade. Em seguida, o algoritmo de aprendizado é utilizado para escolher o melhor dentre os melhores subconjuntos de todas as cardinalidades.

Uma das grandes preocupações em processos de Mineração de Textos é o custo computacional dos algoritmos empregados, devido à alta dimensionalidade enfrentada em processos dessa natureza. Assim, optou-se, neste trabalho, pela adoção de métodos de menor custo, a fim de viabilizar um eficiente pré-processamento de grandes bases de textos. Embora os métodos *wrapper* e híbridos tendam a apresentar melhores resultados do que métodos de filtro, *wrappers* e híbridos demandam um custo de processamento maior que os filtros. Dessa forma, adotou-se neste trabalho a abordagem de filtro de atributos para selecionar aqueles que melhor representam o conhecimento contido em uma base textual.

3.3 Métodos de Seleção Não-Supervisionada de Atributos

Nesta seção, são apresentados os métodos de seleção de atributos comparados durante a realização deste trabalho. Para tal, os métodos foram divididos em três categorias, de acordo com o tipo de medida de relação entre documento e atributo no qual o método se baseia: freqüência de termos ou documentos, variância de termo e contexto. Cada uma dessas medidas, bem como os métodos a elas pertencentes, são apresentados nas seções a seguir. Para efeitos de padronização de notação, assume-se, daqui em diante, que os conjuntos originais de atributos possuem N documentos e M atributos, sendo o índice i utilizado para o i-ésimo documento e o índice j para o j-ésimo atributo.

3.3.1 Baseados em Freqüência de Termos ou Freqüência de Documentos

Freqüência de termo, conhecida como TF (do inglês *term frequency*), e freqüência de documentos, conhecida como DF (do inglês *document frequency*), são duas medidas de relação entre atributos e documentos facilmente calculadas e com custo computacional de $O(MN)$. A freqüência de termo contabiliza a freqüência absoluta de um determinado termo ao longo da coleção de documentos. Como métodos baseados nessa medida, tem-se o *ranking* pela freqüência de termos, o corte de Luhn e a *Term Frequency Inverse Document Frequency*.

A freqüência de documentos, por sua vez, contabiliza o número de documentos que um determinado termo aparece. Das medidas aqui apresentadas, o *ranking* pela freqüência de documentos, o método de Salton, o método LuhnDF e também a *Term Frequency Inverse Document Frequency* utilizam a DF em seu cálculo.

Ranking pela Freqüência de Termos (RTF)

O *ranking* pela freqüência de termos utiliza a TF como medida de *score* para um determinado atributo, dando maior nota àquele que apresenta maior freqüência ao longo

da coleção. O *score* de um atributo da coleção pode ser calculado conforme a Equação 3.1

$$TF_j = \sum_{i=1}^{N} f_{ij} \tag{3.1}$$

na qual TF_j é a TF para o j-ésimo atributo e f_{ij} é a freqüência do j-ésimo atributo no i-ésimo documento.

Método de Luhn

Luhn (1958) propôs em seu trabalho a classificação dos termos segundo a sua significância, medida em função da TF desse termo. Esse é um método tradicional, sendo extensamente utilizado em trabalhos da área (Losee, 2001; Martins, 2003; Matsubara et al., 2003; Cummins e Oriordan, 2005). Nesse método, o autor baseou-se na Lei de Zipf (Zipf, 1949), também conhecida como **Princípio do Menor Esforço**, a qual afirma que a freqüência de ocorrência de alguns eventos está relacionada com uma função de ordenação. Em textos, ao contabilizar a freqüência dos termos e ordenar o histograma resultante em ordem decrescente, forma-se a chamada *Curva de Zipf*, na qual o k-ésimo termo mais comum ocorre com freqüência inversamente proporcional a k.

Nesse sentido, Luhn (1958) propôs pontos de corte superior e inferior da Curva de Zipf, de maneira que termos com alta e baixa freqüência são descartados, considerando termos de alta significância os de freqüência intermediária. Termos de alta freqüência são julgados não relevantes por geralmente aparecerem na grande maioria dos textos, não trazendo, em geral, informações úteis. Já os termos de baixa freqüência são considerados muito raros e não possuem, de maneira geral, caráter discriminatório.

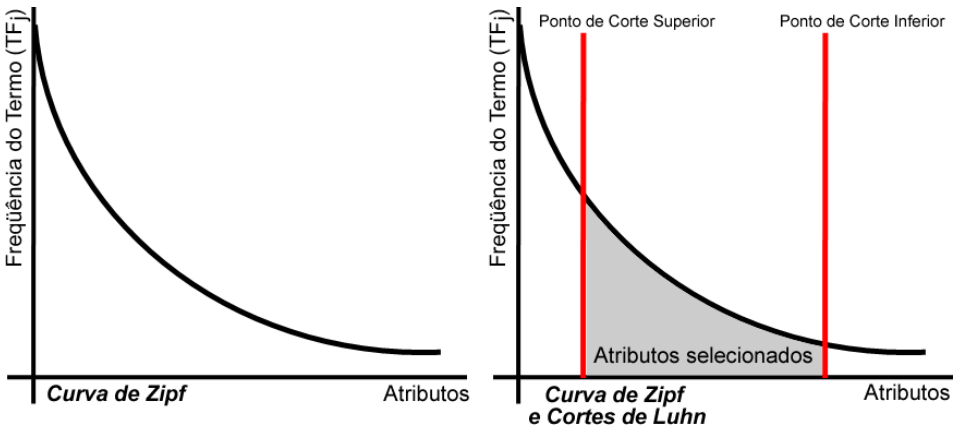


Figura 3.4: Cortes de Luhn - Adaptado de Matsubara et al. (2003)

Dado o baixo processamento demandado por esse método, que necessita apenas do cálculo da TF dos termos presentes na base e a ordenação dessas freqüências para a geração de um gráfico, esse método pode ser facilmente escalável para bases de textos muito grandes, apresentando custo de $O(M * log(M))$. Além disso, esse é um dos poucos métodos cuja aplicação já retorna para o usuário exatamente quantos atributos considerar e quais são esses atributos. Entretanto, os pontos de corte superior e inferior sugeridos pelo método não são exatos, sendo a subjetividade da escolha desses pontos o principal

ponto negativo de aplicabilidade do método de Luhn. Neste trabalho, para amenizar esse problema, a determinação dos pontos de corte se dá nos pontos de corte aos pontos de inflexão da curva de Zipf, os quais são, também, determinados subjetivamente.

Ranking pela Freqüência de Documentos (RDF)

Esse método é baseado na freqüência de documentos. Basicamente, o método computa o número de documentos nos quais os termos aparecem, assumindo que termos que aparecem em poucos documentos não são importantes na coleção, podendo ser descartados. Essa é uma medida que apresenta eficiência na seleção de atributos em processos de categorização de textos (Yang e Pedersen, 1997), podendo ser obtida conforme mostrado na Equação 3.2:

$$DF_j = \sum_{i=1}^{N}(1|f_{ij} \neq 0) \tag{3.2}$$

na qual f_{ij} é a freqüência do j-ésimo atributo no i-ésimo documento.

Método de Salton

Salton et al. (1975) propõem em seu trabalho uma medida de obtenção do poder de discriminação de um termo, ou seja, quão bem um termo é capaz de discriminar um documento da coleção de outro documento. O valor de discriminação de um termo reflete, portanto, o quanto a separação média entre documentos muda quando esse termo é considerado para representação de um documento, de maneira que os melhores termos são aqueles que atingem maior grau de separação.

Dessa forma, o trabalho sugere considerar, na coleção, termos que não apresentem freqüência muito alta ou muito baixa, seguindo a distribuição mostrada na Equação 3.3:

$$\frac{N}{100} \leq DF_j \leq \frac{N}{10} \tag{3.3}$$

na qual DF_j é a freqüência de documentos do j-ésimo termo na coleção.

Esse é um trabalho tradicional na área de Mineração de Textos, utilizado por diversos autores (Modha e Spangler, 2003; Kang, 2004; Price e Thelwall, 2005) e apresentando, geralmente, grandes contribuições. Assim como o método de Luhn, esse método sugere pontos exatos de corte, obtendo-se, após a aplicação do mesmo, o conjunto exato de atributos a ser considerado. Seu custo linear em função do número de atributos ($O(MN)$) o torna aplicável para grandes bases de documentos. Entretanto, o intervalo de DF proposto para os termos a serem considerados é bastante restrito, resultando em um grande corte de termos da base.

Método LuhnDF

Com o objetivo de aproveitar as idéias dos métodos de Luhn e Salton, foi proposta, neste trabalho, uma adaptação do método de Luhn para a freqüência de documentos, resultando em um método que foi denominado "LuhnDF". Nesse, gera-se os histogramas para a DF_j dos termos de forma descendente, adotando, assim como o método de Luhn

tradicional, um ponto de corte superior e outro inferior para os atributos, como apresentado na Figura 3.4, trocando-se, no eixo vertical, a TF pela DF de um termo. O custo computacional do método é o mesmo do método orignal ($O(M * log(M))$). Esse método seleciona, assim como o de Salton, termos cuja DF não é tão grande, nem tão pequena, sendo, porém, menos agressivo e mais maleável quanto aos pontos de corte. Assim como no método de Luhn, a seleção dos pontos de corte superior e inferior é empírica, ficando a cargo do usuário a determinação desses pontos. Neste trabalho, os pontos de corte para LuhnDF são determinados, também, em regiões de inflexão da curva de tendência.

Term Frequency Inverse Document Frequency(TFIDF)

A TFIDF é uma das medidas de relação entre atributos e documentos, especialmente utilizada em tarefas de recuperação de informação (Soucy e Mineau, 2005). A idéia básica dessa medida, apresentada por Salton e Buckley (1987), é ponderar a TF dos termos em função de suas distribuições ao longo da coleção, dando menor peso àqueles termos que aparecem em muitos documentos. Assim, termos muito comuns na coleção são considerados não discriminativos, uma vez que não trariam informação necessária para distinguir um documento de outro. Para isso, é introduzido o valor do inverso da freqüência de documentos (IDF - do inglês *Inverse Document Frequency*) para um atributo, o qual pode ser obtido conforme apresentado na Equação 3.4.

$$IDF_j = log\left(\frac{N}{DF_j}\right) \tag{3.4}$$

na qual DF_j é a freqüência de documento do j-ésimo atributo. A TFIDF é, portanto, o resultado da multiplicação da TF pela IDF, podendo ser calculada conforme mostrado na Equação 3.5:

$$TFIDF_j = \sum_{i=1}^{N} f_{ij} * IDF_j \tag{3.5}$$

sendo f_{ij} a freqüência do j-ésimo atributo no i-ésimo documento e IDF_j o inverso da freqüência de documentos do j-ésimo atributo.

O baixo custo computacional ($O(MN)$) e a ponderação dos pesos de um termo pela distribuição do mesmo ao longo da coleção de documentos tornam essa medida relevante, sendo, portanto, objeto de pesquisa na realização deste trabalho.

3.3.2 Baseados em Variância de Termos

Um segundo grupo de métodos avaliados neste trabalho consiste de modelos baseados na variância dos termos. A variância é uma medida estatística que mede a dispersão de uma variável em relação ao seu valor esperado. Dessa forma, pode ser utilizada para avaliar a distribuição das freqüências de um termo ao longo da coleção. Os métodos que se valem dessa medida para selecionar atributos estão dispostos a seguir.

Contribuição do Termo (TC)

A contribuição do termo (TC - do inglês *Term Contribution*), proposta por Liu et al. (2003), tem como principal motivação o fato de que métodos mais simples como a DF assumem que cada termo tem a mesma importância em todos os documentos da coleção. Considerando uma tarefa de agrupamento de dados, tais medidas são facilmente distorcidas por termos que apresentam alta freqüência de documentos mas distribuição uniforme sobre diferentes grupos de documentos.

O conceito que rege esse método é baseado no conceito de valor discriminativo de um termo (Salton e McGill, 1986), segundo o qual a importância de um termo pode ser vista como a contribuição desse para a similaridade de documentos. A similaridade entre dois documentos pode ser calculada utilizando a TFIDF (ver Equação 3.5, considerando apenas um documento), segundo a Equação 3.6:

$$Sim_{x,y} = \sum_{j=1}^{M} f_{j,x} * IDF_j * f_{j,y} * IDF_j \tag{3.6}$$

considerando x e y dois documentos da coleção, $f_{j,i}$ a freqüência do j-ésimo atributo no i-ésimo documento e IDF_j o inverso da freqüência de documentos do j-ésimo atributo.

De maneira similar, a contribuição do termo pode ser obtida utilizando a Equação 3.7:

$$TC_j = \sum_{x=1}^{N} \sum_{y=1}^{N} f_{j,x} * IDF_j * f_{j,y} * IDF_j \tag{3.7}$$

A contribuição do termo provê maior *score* àqueles termos que aparecem em poucos documentos, ignorando atributos muito raros ou muito freqüentes. Esse método apresenta um custo computacional de $O(MN)$. Além disso, tem sido um método adotado em outros trabalhos na área (Wiratunga et al., 2006; Liu et al., 2005; Gonzaga et al., 2007), obtendo bons resultados.

Variância do Termo

Esse método, proposto por Liu et al. (2005), afirma que a qualidade de um termo é calculada a partir de sua variância pura. Essa medida considera que os termos importantes são aqueles que não apresentam baixa freqüência de documentos e mantêm uma distribuição não-uniforme ao longo da coleção (maior variância). Assim como a TC, esse método visa evitar a distorção de termos comuns que ocorrem de maneira uniforme ao longo da coleção, independente do grupo ao qual o documento pertença. Para a aplicação desse método, calcula-se a variância de todos os atributos do domínio, conforme apresentado na Equação 3.8:

$$TV_j = \sum_{i=1}^{N} \left[f_{ij} - \bar{f}_j \right]^2 \tag{3.8}$$

na qual f_{ij} é a freqüência absoluta do j-ésimo atributo no i-ésimo documento e $\bar{f}_j$ é a média das freqüências do j-ésimo termo nos documentos da coleção.

Esse método, no trabalho em que foi proposto, apresenta bons resultados, sendo tam-

bém utilizado no trabalho de Gonzaga et al. (2007). O custo computacional apresentado por esse método é linear em função do número de atributos ($O(MN)$), sendo aplicável para grandes bases textuais.

Qualidade da variância do termo

Esse método, proposto por Dhillon et al. (2003), faz uma adaptação da medida estatística de variância, a fim de quantificar a qualidade da variância para os diferentes termos. Seu cálculo é demonstrado na Equação 3.9:

$$TVQ_j = \sum_{i=1}^{N} f_{ij}^2 - \frac{1}{N}\left[\sum_{i=1}^{N} f_{ij}\right]^2 \tag{3.9}$$

considerando f_{ij} como a freqüência do j-ésimo termo no i-ésimo documento.

É possível perceber que, assim como no método TV, o valor de avaliação de um atributo, segundo essa medida, é proporcional à variância de sua freqüência. Assim, assume-se que termos com pequena variação apresentam pequeno poder discriminante, uma vez que ocorrem de maneira uniforme ao longo de toda a coleção. Essa medida também já foi utilizada em vários outros trabalhos (Liu et al., 2005; Archetti et al., 2006; Messina et al., 2006; Xu et al., 2007), apresentando bons resultados, o que motivou sua inserção na comparação apresentada neste trabalho. Seu custo, também linear em função do número de atributos ($O(MN)$), permite a sua utilização em bases com muitos documentos.

3.3.3 Baseados em contexto

A terceira classe de métodos aqui apresentada é a dos baseados em contexto. A idéia de contexto de um termo é oriunda da indexação por zonas (Manning et al., 2008), muito utilizada em tarefas de recuperação de informação. Zonas são partes bem delimitadas de documentos de tamanho arbitrário, constituídas por textos escritos. Por exemplo, em artigos científicos, o título, o resumo e a conclusão são zonas do documento. A partir desta separação do documento por partes, pode-se dar pesos diferentes a termos, de acordo com a zona em que esses ocorram. Dessa forma, pondera-se a freqüência de um termo pela importância da seção em que ele aparece, obtendo *rankings* de atributos que melhor representam a qualidade da informação trazida pelos mesmos. Seguindo essa idéia, no presente trabalho é apresentado um método baseado em zonas do documento para seleção de termos, o qual é apresentado a seguir.

Zone-Scored Term Frequency (ZSTF)

Esse método aqui proposto visa ponderar a freqüência de um termo de acordo com a posição no documento (ou zona) que esse aparece. A grande motivação para a sua proposta é o fato de que grande parte dos documentos envolvidos em tarefas de gerência de conhecimento, contexto no qual este trabalho se insere, são claramente separados em zonas distintas. Nesse tipo de documentos, é fácil perceber que algumas zonas trazem mais informação relevante sobre o conteúdo do documento do que outras zonas. Em

artigos científicos, por exemplo, os termos do título condensam, intuitivamente, muito mais informação do que termos presentes em sua introdução.

Baseado nesses fatos, o método *Zone-Scored Term Frequency* aqui proposto aborda a idéia de que termos que ocorrem em zonas de maior impacto, ou que nelas se concentre maior informação relativa ao conteúdo do documento, tenham maior peso associado a eles no processo de seleção de atributos, atribuindo-os um maior grau de importância. Considerando-se uma coleção de documentos, cada um desses compostos por L zonas, a ZSTF de um termo pode ser calculada conforme mostrado na Equação 3.10

$$ZSTF_j = \sum_{i=1}^{N} \sum_{l=1}^{L} f_{ij} * P_l \tag{3.10}$$

na qual f_{ij} é a freqüência do j-ésimo termo no i-ésimo documento e P_l é o peso aferido à l-ésima zona dos documentos da coleção. O valor de P_l é subjetivo, devendo obedecer às restrições apresentadas na Equação 3.11.

$$\sum_{l=1}^{L} P_l = 1 | 0 < P_l \leq 1 \tag{3.11}$$

Uma vez que o peso de uma zona é proporcional à importância dessa zona na descrição do conteúdo de todo o documento, a ZSTF dará maior peso àqueles termos que aparecem em zonas mais discriminativas. Dessa forma, termos com maior *score* ZSTF tendem a serem mais representativos ao longo da coleção. Com um custo computacional de $O(MN)$, a ZSTF é plenamente aplicável a bases de documentos de grandes dimensionalidades. Sua principal limitação, entretanto, é a restrição do uso em documentos que possuam zonas bem delimitadas, impossibilitando sua aplicação em algumas bases de dados textuais. Além disso, é preciso conhecer o domínio com o qual se trabalha para melhor atribuir os pesos às zonas.

3.4 Considerações Finais

A representação de documentos estruturada na forma de matrizes atributo-valor é a mais comumente encontrada nos diversos trabalhos da área. Nessa representação, cada termo presente na coleção de textos é candidato a atributo, enquanto as instâncias correspondem aos documentos da base. Entretanto, como o número de atributos geralmente excede o número de instâncias em mais de um grau de magnitude, a matriz formada é inerentemente esparsa e de alta dimensionalidade. Para processos eficazes de extração de conhecimento, faz-se necessário selecionar aqueles termos que melhor representem a coleção de documentos, reduzindo o número de atributos na base.

Para tal, existem duas estratégias básicas: selecionar ou extrair atributos. Uma vez que o processo de extração gera novos atributos e não mantém explícita a relação física entre atributos da base e a configuração real do problema, a seleção de atributos é a mais adequada para contextos em que, assim como no contexto com o qual este trabalho visa colaborar, é de suma importância a compreensibilidade dos modelos extraídos, mantendo na base o subconjunto mais representativo dentre os atributos originais.

Selecionar atributos em um contexto não-supervisionado é uma atividade difícil. Ao contrário dos contextos supervisionados, nos quais selecionam-se atributos em função de seu poder preditivo em relação aos valores do atributo classe, a seleção não-supervisionada trabalha com aspectos intrínsecos aos atributos, valendo-se de medidas estatísticas para tal. Outro ponto de dificuldade remete-se à forma de avaliação dos subconjuntos gerados. Poucos critérios objetivos de avaliação dos subconjuntos nas tarefas de extração de conhecimento são disponíveis e adequados ao problema, ficando, freqüentemente, a avaliação a critério da subjetividade de um especialista no domínio do problema ou levando-se a avaliação para um contexto supervisionado.

Uma vez que este trabalho é voltado para tarefas que lidam com coleções de texto não-rotuladas, realiza-se uma análise exploratória de métodos não-supervisionados para seleção de atributos, como os mostrados na Seção 3.3 deste capítulo. Para a realização dessas comparações, os métodos de seleção de atributos foram aplicados a bases de documentos coletadas e foram avaliados seguindo um processo proposto neste trabalho. Tanto as bases de documentos selecionadas para o desenvolvimento deste trabalho quanto o processo de avaliação adotado são apresentados no capítulo a seguir.

Capítulo

4

Metodologia para Avaliação Experimental

Neste capítulo, apresenta-se a metodologia adotada no presente trabalho para a avaliação de métodos de seleção não-supervisionada de atributos. São descritos os passos adotados no processo de avaliação e as bases de dados utilizadas neste processo, detalhando as especificidades de cada uma delas.

4.1 Metodologia de Avaliação

A avaliação de métodos não-supervisionados de seleção de atributos realizada neste trabalho, é baseada nas atividades de um processo de Mineração de Textos, desde a seleção dos documentos até a atividade de seleção de atributos na fase de Pré-Processamento. Dessa maneira, para avaliar os resultados, cinco passos básicos são seguidos: Seleção do Corpus, Geração de Atributos, Montagem da Matriz Atributo-Valor, Seleção de Atributos e Avaliação dos Subconjuntos de Atributos. A disposição dos passos e uma ilustração dos mesmos em alto nível pode ser obtida na Figura 4.1. Nas seções a seguir, detalha-se as atividades realizadas em cada um desses passos.

4.1.1 Seleção do Corpus

Nessa etapa, seleciona-se os dados com os quais se deseja trabalhar, tal como na etapa de Identificação do Problema em tarefas de Mineração de Textos. Para tal, a metodologia aqui adotada efetua três atividades, conforme descrito na Figura 4.2.

Todos os documentos pertencentes à base devem ser transformados para o formato de texto plano, a fim de tornar o conteúdo dos mesmos mais facilmente processável. Uma vez nesse formato, é possível identificar diferentes partes constituintes em alguns tipos de documentos como, por exemplo, título, autoria e conclusão. Isso permite escolher com quais partes dos documentos deseja-se trabalhar, eliminando da análise aquelas partes cuja informação for considerada não necessária ao objetivo do processo. Como este tra-

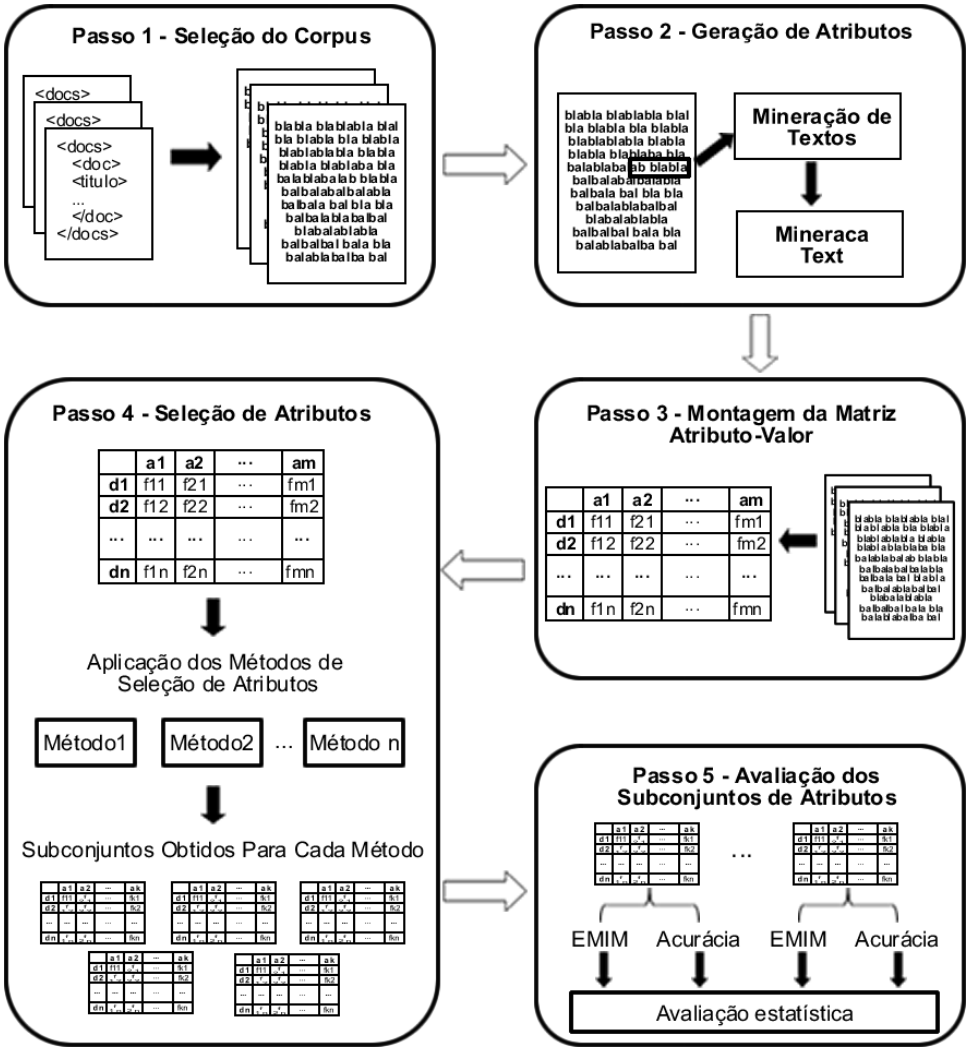


Figura 4.1: Metodologia para avaliação de métodos não-supervisionados de seleção de atributos

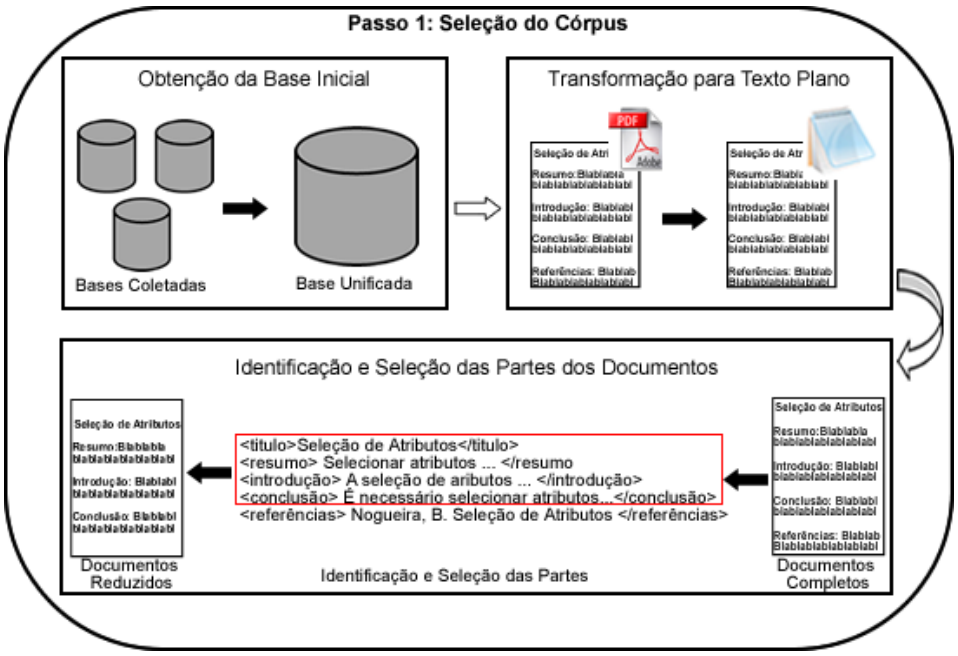


Figura 4.2: Atividades para seleção do corpus trabalhado pela metodologia

balho tem por meta colaborar com a metodologia de extração de taxonomias de tópicos de Moura et al. (2008a), utilizou-se documentos científicos, como artigos, tal qual na re-

ferida metodologia. Nesse tipo de documentos, a divisão em diferentes partes é nítida, sendo possível construir *parsers* de documentos para separá-las. Essa é uma atividade extremamente necessária para a aplicação da classe de métodos descrita na Seção 3.3.3 deste trabalho.

Para essa etapa do processo, é possível utilizar a ferramenta *IESystem*, atualmente em desenvolvimento no Laboratório de Inteligência Computacional (LABIC - ICMC/USP)[1], a qual transforma documentos do formato *PDF* para o formato de texto plano e, para artigos científicos, identifica e marca as diferentes partes que o constituem.

4.1.2 Geração de Atributos

Esse passo tem por objetivo garantir que todos os termos da coleção considerada estejam em um formato padrão, tal como feito na fase de pré-processamento da Mineração de Textos, mostrada na Seção 2.1.2. Em um primeiro momento, faz-se uma padronização dos caracteres, removendo os caracteres considerados desnecessários ao objetivo da avaliação por não apresentarem informação relevante, como acentos, pontuação, cedilhas, números e *underlines*. Todos os caracteres restantes são convertidos para a sua forma minúscula. Logo após, são eliminadas as *stopwords*, como uma maneira preliminar de filtragem de atributos não-representativos ao domínio, definindo o conjunto de palavras candidatas a atributos. Em uma etapa subseqüente, todas as palavras restantes são reduzidas ao seu radical, por meio de um processo de *stemming*. Embora existam outras formas de redução do termo e extinção de variações morfológicas de um mesmo termo (Conrado e Rezende, 2008), a adoção do processo de *stemming* foi adotado devido ao seu baixo custo computacional, quando comparado às demais técnicas. Assim, tem-se o conjunto final de termos da coleção. Um exemplo das atividades realizadas nesse passo é mostrado na Figura 4.3.

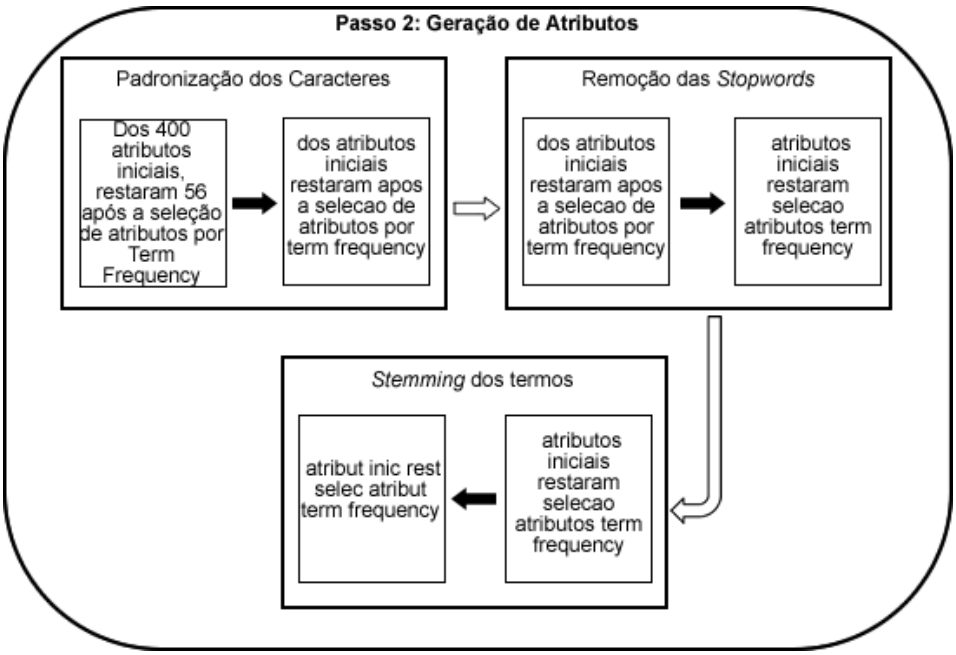


Figura 4.3: Geração de atributos na metodologia de avaliação

É importante ressaltar que, embora a metodologia de Moura et al. (2008a) adote o

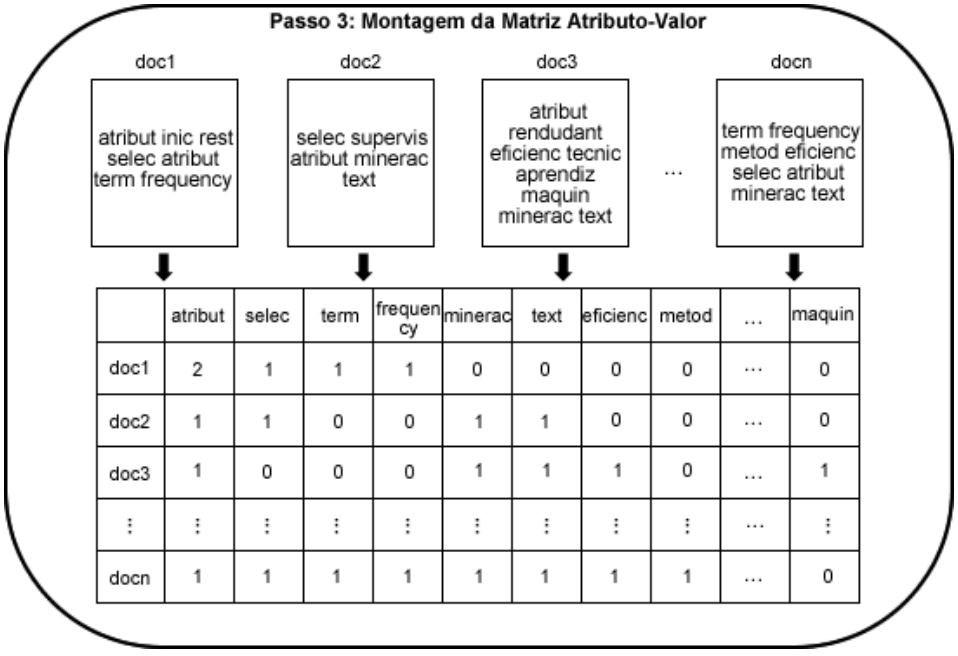
[1]http://labic.icmc.usp.br

uso de termos compostos (n-gramas) para fins de maior compreensibilidade da taxonomia de tópicos extraída, a avaliação aqui apresentada sugere a utilização de apenas termos simples. Essa decisão se deve ao fato de que, conforme afirmado em Moura et al. (2008c), métodos de seleção de atributos que analisam estatisticamente a freqüência dos termos, como os aqui mostrados, são mais confiáveis para conjuntos de termos simples, pois apresentam tendência em favorecer termos que não são muito freqüentes nem muito raros. Experimentalmente, é fácil perceber que os termos compostos mais interessantes são, geralmente, os termos com maiores freqüência, embora essa não seja condição suficiente para considerá-los termos relevantes ao domínio. Ainda em Moura et al. (2008c) é possível encontrar um eficiente método para seleção de termos compostos, subseqüente à seleção de termos simples, baseado na não-redundância de termos.

Esse passo pode ser efetuado com o suporte da ferramenta *PreText* (Matsubara et al., 2003), a qual remove as *stopwords*, identifica os termos da coleção e aplica a eles o processo de *stemming*. Para esse processo de *stemming*, essa ferramenta baseia-se no algoritmo de Porter (Porter, 1997), o qual foi adaptado para três idiomas: Português, Espanhol e Inglês.

4.1.3 Montagem da Matriz Atributo-Valor

Uma vez definido o conjunto inicial de atributos que relacione a coleção, a metodologia de avaliação sugere a transformação dos dados para a forma de matrizes atributo-valor, como exemplificado na Figura 4.4, a qual contém uma amostra da matriz atributo-valor obtida na conversão de documentos fictícios.



	atribut	selec	term	frequency	minerac	text	eficienc	metod	...	maquin
doc1	2	1	1	1	0	0	0	0	...	0
doc2	1	1	0	0	1	1	0	0	...	0
doc3	1	0	0	0	1	1	1	0	...	1
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	...	⋮
docn	1	1	1	1	1	1	1	1	...	0

Figura 4.4: Montagem de uma matriz atributo-valor a partir de uma coleção padronizada

Na matriz atributo-valor gerada para avaliação dos métodos de seleção de atributos, a medida de relação entre atributos e instâncias utilizada é a freqüência absoluta de um termo em um documento (TF, definida anteriormente na Seção 3.1). Todos os termos são inseridos na matriz considerando cada texto como uma *bag-of-words*, sem considerar informações a respeito do contexto em que o termo se encontra. Esse passo pode, também,

ser realizado com o auxílio da ferramenta *PreText*.

4.1.4 Seleção de Atributos

A partir da matriz atributo-valor, é possível aplicar os métodos não-supervisionados de seleção de atributos comparados neste trabalho, enumerados na Seção 3.3. Para a aplicação do método *Zone Scored Term Frequency*, efetua-se inicialmente, nesse passo, a ponderação dos atributos pela zona em que eles ocorrem. Esse processo se dá pela multiplicação da TF do termo em uma zona pelo peso definido para aquela zona.

Aos demais métodos, a aplicação se dá diretamente na matriz atributo-valor original. Conforme ilustrado na Figura 4.5, a aplicação dos métodos que sugerem pontos exatos de corte (Luhn, Salton e LuhnDF) gera, para cada base de textos, um único subconjunto de dados. Já para os métodos que fornecem apenas um *ranking* de atributos, sem fornecer o número exato a ser considerado, diversos subconjuntos eram gerados, variando a porcentagem de atributos selecionados em relação ao conjunto original (5%, 10%, 20%, 30%, ... , 90%). Os subconjuntos de dados gerados reduzem apenas o número de atributos, mantendo o número de documentos.

Além disso, para efeito de comparação com as técnicas que sugerem pontos exatos de corte, são, também, gerados subconjuntos de cardinalidade de atributos idênticas aos subconjuntos gerados pelas técnicas que sugerem pontos exatos de corte. Dessa forma, para cada subconjunto gerado, selecionava-se os p atributos com melhor *ranking*, considerando p a cardinalidade de atributos desejada para o subconjunto. Uma outra tarefa aqui efetuada, essa unicamente para fins de comparação estatística na validação supervisionada de atributos, é a aplicação de dois métodos supervisionados de seleção de atributos bastante conhecidos na literatura que geralmente apresentam bons resultados: Ganho de Informação (IG, do inglês *Information Gain*) e Razão de Ganho (GR, do inglês *Gain Ratio*) (Yang e Pedersen, 1997).

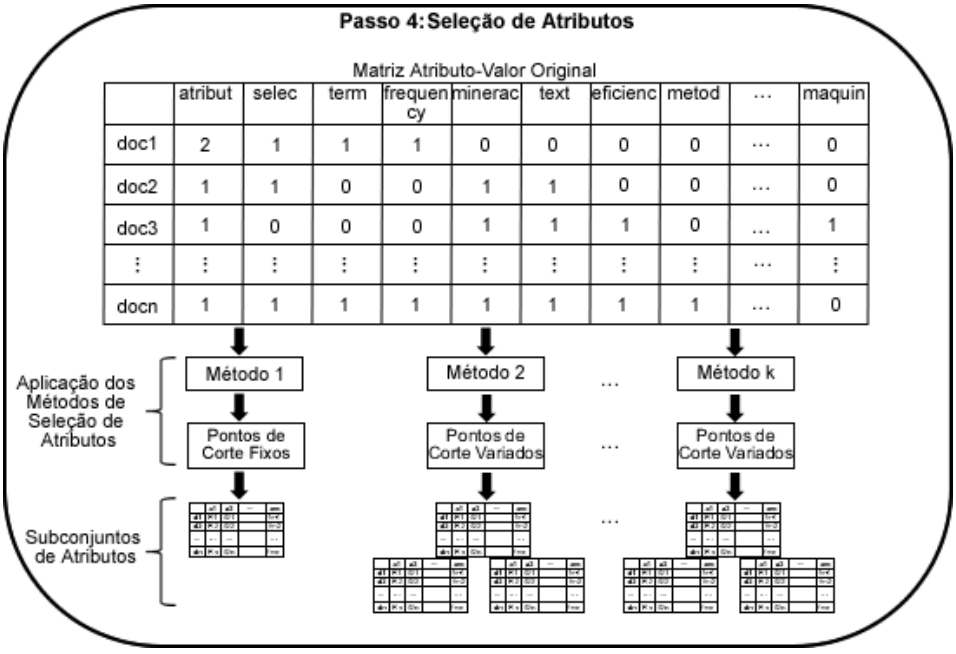


Figura 4.5: Aplicação dos métodos de seleção não-supervisionada de atributos

4.1.5 Avaliação dos Subconjuntos de Atributos

O último passo da metodologia seguida neste trabalho se refere à avaliação dos subconjuntos de atributos. Após selecionar subconjuntos de atributos com as diferentes técnicas, é necessário avaliá-los a fim de verificar aquele que melhor representa a coleção de documentos.

Considerando que o presente trabalho visa colaborar com um contexto cujo objetivo é a obtenção de uma taxonomia de tópicos a partir de uma coleção de documentos, pode-se definir como interesse principal da seleção de atributos a manutenção na base de dados daqueles termos mais representativos. Em outras palavras, busca-se, neste trabalho, uma redução de dimensões de atributos sem, no entanto, ocasionar perda de informações relevantes acerca do conteúdo dos documentos, sendo capaz de manter a distinção entre os mesmos no processo de agrupamento de documentos.

Dessa forma, para a avaliação dos diferentes subconjuntos gerados, adotou-se, conforme mostrado na Figura 4.6, duas diferentes formas, sendo uma não-supervisionada, por meio da aplicação de *Expected Mutual Information Measure*, e outra supervisionada, pela análise da acurácia de diferentes classificadores utilizados na área de categorização de textos.

A avaliação dos subconjuntos de atributos gerados pelos diferentes métodos de seleção de atributos foi dividida em várias comparações estatísticas a fim de melhor verificar o impacto da aplicação dos métodos de seleção de atributos. Assim, tanto para a avaliação não-supervisionada quanto para a avaliação supervisionada, três grupos de comparações foram estabelecidos:

- Grupo 1 - Variação da porcentagem de atributos selecionados, comparando métodos sem ponto exato de corte: visa verificar o impacto desse tipo de método, computacionalmente mais custosos, buscando por diferenças significativas no desempenho desses;

- Grupo 2 - Comparação de métodos que fornecem pontos exatos de corte, para todas as bases de textos: compara o impacto desses métodos quanto à qualidade dos pontos de corte que esses fornecem;

- Grupo 3 - Comparação de métodos que não fornecem pontos exatos de corte e métodos com pontos exatos de corte, gerando com o primeiro tipo de métodos subconjuntos de cardinalidade igual à dos métodos com ponto exato de corte: tem por objetivo defrontar os métodos com ponto exato de corte, computacionalmente mais simples, com métodos mais complexos, a fim de verificar se a aplicação do segundo grupo de métodos é realmente mais eficaz, fazendo valer o esforço computacional adicional.

Com os resultados de cada um desses grupos, é possível comparar objetivamente o desempenho dos diferentes métodos de seleção de atributos. Para tal, a fim de identificar diferença significativa nos desempenhos dos diferentes métodos de seleção de atributos, neste trabalho utiliza-se o teste estatístico não-paramétrico de Kruskal-Wallis (Kruskal e

Wallis, 1952) para amostras não-emparelhadas, aplicando o pós-teste de múltiplas comparações de Dunn.

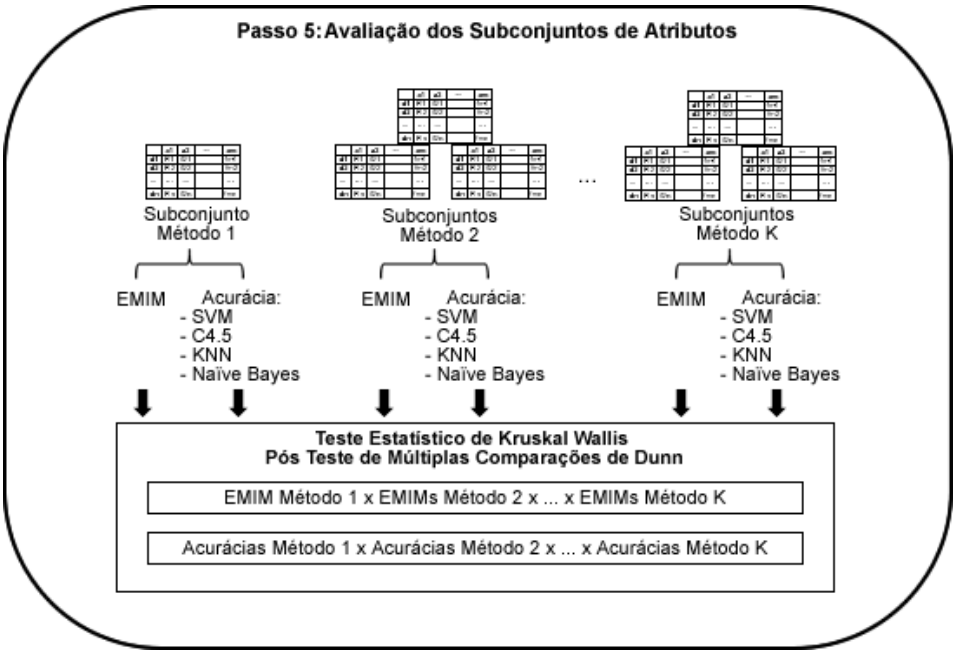


Figura 4.6: Avaliação dos subconjuntos de atributos obtidos pelos diferentes métodos

Expected Mutual Information Measure (EMIM)

A *Expected Mutual Information Measure* (van Rijsbergen, 1979) é uma medida estatística que representa, dados dois conjuntos de objetos, o quanto a distribuição desses dois conjuntos desvia da independência estocástica. Em termos de avaliação de subconjuntos de atributos, a aplicação da EMIM visa mensurar quão bem os termos selecionados em um determinado subconjunto de atributos consegue predizer o restante do vocabulário da coleção de documentos (Lawrie et al., 2001). O cálculo da EMIM para um subconjunto de atributos com S termos selecionados a partir de um conjunto inicial de M termos pode ser feito de acordo com a Equação 4.1:

$$EMIM_S = \sum_{j=1}^{M}\sum_{s=1}^{S} P(j,s) * log\frac{P(j,s)}{P(j) * P(s)} \tag{4.1}$$

na qual s é um atributo selecionado e j é um atributo contido no conjunto original de atributos. A probabilidade de um atributo $P(j)$ é apresentada na Equação 4.2:

$$P(j) = \frac{TF_{.,j}}{TF_{.,.}} \tag{4.2}$$

$$TF_{.,j} = \sum_{i=1}^{N} f_{i,j} \tag{4.3}$$

$$TF_{.,.} = \sum_{i=1}^{N} TF_{.,j} \tag{4.4}$$

na qual $TF_{i,j}$ é a freqüência do j-ésimo atributo no i-ésimo documento, $TF_{.,j}$ é a freqüência total de um atributo em na coleção de documentos, cujo cálculo é mostrado na Equação

4.3, e $TF_{.,.}$ é a freqüência total de atributos na coleção de documentos, calculado conforme Equação 4.4. De maneira análoga, a probabilidade conjunta $p(j, s)$ pode ser calculada conforme demonstrado na Equação 4.5:

$$P(j,s) = \sum_{i=1}^{N} \frac{TF_{i,j} * TFi, s}{TF_{.,.} * TF_{i,.}} \tag{4.5}$$

$$TF_{i,.} = \sum_{j=1}^{M} f_{i,j} \tag{4.6}$$

$$TF_{.,.} = \sum_{i=1}^{N} TF_{i,.} \tag{4.7}$$

considerando $TF_{i,.}$ como a freqüência total de atributos em um documento, cujo cálculo é mostrado na Equação 4.6 e $TF_{.,.}$ é a freqüência total de atributos na coleção de documentos, calculado conforme Equação 4.7.

É possível perceber que o valor de EMIM para dois termos independentes é zero. Assim, consegue-se, com essa medida, detectar se houve eliminação de algum termo cuja informação não possa ser obtida por meio de outros termos da base. Dessa forma, avalia-se os métodos de seleção de atributos mensurando a quantidade de informação perdida quando eliminando os atributos não selecionados por esses métodos. Assume-se, portanto, que os melhores métodos de seleção de atributos são aqueles que apresentam maior valor de EMIM quando comparado a outros métodos, considerando subconjuntos de mesma cardinalidade.

Acurácia Preditiva de Classificadores

Uma segunda forma de avaliação de métodos de seleção de atributos adotada neste trabalho é a avaliação da acurácia preditiva de classificadores. Algoritmos de classificação conseguem aprender hipóteses a partir de dados de treinamento, gerando modelos de classificação que atribuem, a exemplos de um domínio, uma classe contida em um conjunto de classes predefinidas. A acurácia preditiva é uma maneira de testar uma hipótese aprendida por um algoritmo de classificação, verificando quão bem os dados são aprendidos e generalizados pelo modelo de classificação (Mitchell, 1997).

Para estimar uma acurácia não-distorcida, utiliza-se o conceito de conjuntos de treinamento e teste. Um conjunto de treinamento consiste dos dados mostrados ao algoritmo de classificação para o aprendizado do modelo de classificação. O conjunto de teste, por sua vez, é um conjunto independente do conjunto de treinamento utilizado para avaliar a acurácia do modelo gerado ao lidar com novos dados, testando a generalização da hipótese aprendida. Para maior aproximação dos resultados obtidos dos valores reais, efetua-se uma validação cruzada, por meio do processo de *k-fold cross validation*, adotando valor de k igual a dez.

No presente trabalho, para avaliar os métodos de seleção de atributos, adota-se a utilização de quatro classificadores muito utilizados em tarefas de classificação textual: Árvores de Decisão C4.5 (Quinlan, 1993), K-Vizinhos Mais Próximos (KNN, do inglês

K-Nearest Neighbor, Máquinas de Vetores de Suporte (SVM, do inglês *Support Vector Machines*) (Vapnik, 1998; Burges, 1998) e Classificadores Bayesianos Simples, também conhecidos como *Naïve Bayesian Classifiers* (Mitchell, 1997). Para a sua aplicação, as bases de dados utilizadas são supervisionadas, sendo o rótulo das mesmas desconsiderado para fins de seleção de atributos, sendo utilizados apenas para o processo de validação dos subconjuntos obtidos. Embora essa seja uma análise simplificada se contextualizada em tarefas de agrupamentos, a análise de acurácia de classificadores pode ser utilizada para verificar a capacidade do subconjunto gerado em preservar a caracterização das classes nos domínios. Isso pode ser também útil em processos de agrupamento, dado que elementos de uma mesma classe tendem a ser similares entre si e, portanto, a serem reunidos em um mesmo grupo, sendo, inclusive, utilizado em outros trabalhos da área (Dash e Liu, 1997; Wiratunga et al., 2006).

Para efeitos de comparação de métodos de seleção de atributos, tem-se por hipótese que utilizando apenas atributos relevantes ao domínio do problema, descartando os atributos irrelevantes, os classificadores conseguem obter maior acurácia, ou pelo menos a mantêm, porém com menor custo computacional. Assim, os melhores métodos de seleção de atributos são aqueles cujos subconjuntos de atributos por ele gerados apresentaram maiores acurácias que os subconjuntos gerados por outros métodos quando submetidos a um mesmo algoritmo de classificação. Para um contexto não-supervisionado, a aplicação da avaliação por acurácia preditiva visa medir a capacidade dos subconjuntos de atributos gerados pelas diferentes técnicas em manter a estrutura dos dados, medida por meio da preservação da caracterização das classes. Em outras palavras, se um subconjunto de atributos obtém maior acurácia de classificação do que outro, assume-se que esse manteve na base atributos mais importantes à caracterização do domínio do problema considerado, o que é de extrema importância em tarefas de organização do conhecimento textual.

Além disso, a avaliação desses métodos não-supervisionados de seleção de atributos em contextos supervisionados pode, também, indicar a aplicabilidade desses métodos em tarefas de aprendizado supervisionado, dado que os métodos de seleção de atributos aqui considerados apresentam baixo custo computacional se comparados a métodos tradicionais de seleção supervisionada de atributos. Os métodos Chi-Quadrado, Informação Mútua e Ganho de Informação, por exemplo, apresentam custo computacional quadrático em função do número de atributos ($O(M^2)$), enquanto a maioria dos métodos não-supervisionados previamente apresentados tem custo linear ($O(M)$). Para averiguar a aplicabilidade de métodos não-supervisionados em contextos supervisionados, compara-se estatisticamente esses métodos não-supervisionados com os métodos Ganho de Informação e Razão de Ganho, aplicados conforme disposto na Seção 4.1.4.

4.2 Implementações Realizadas para Viabilização da Metodologia de Avaliação

Para a viabilização da avaliação de métodos de seleção de atributos proposta neste trabalho, diversas implementações foram realizadas, relativas aos diferentes passos da

metodologia proposta. Dado que o trabalho foi desenvolvido no âmbito do Laboratório de Inteligência Computacional - LABIC / ICMC-USP, no qual criou-se a ferramenta de pré-processamento de textos *PreText* e o sistema para descoberta de conhecimento *Discover* (Prati et al., 2003), as implementações realizadas neste trabalho foram projetadas de modo a serem totalmente integráveis a essas ferramentas, seguindo o formato de entrada e saída de dados por elas utilizado, conhecido como formato *"discover"*.

Em um primeiro momento, desenvolveu-se um *script* em linguagem de programação Java[2] para a transformação da matriz atributo-valor, fornecida pela *PreText*, para um formato esparso, eliminando os valores iguais a zero, a fim de reduzir o consumo de memória quando ocorre o carregamento da mesma para a memória principal e facilitar a manipulação da mesma. Essa matriz esparsa é a entrada para a maioria dos métodos de seleção de atributos implementados.

A implementação dos métodos de seleção de atributos foi dividida em duas linguagens de programação: R[3] e Java. Métodos que utilizam geração de gráficos para análise de um especialista para a aplicação de cortes com pontos fixos em relação à DF ou TF de um termo, como os métodos de Luhn, LuhnDF e Salton, foram implementados na linguagem R. A escolha por essa linguagem é devido ao fato dessa possuir uma ampla biblioteca facilmente manipulável para geração de gráficos, além de facilidade para lidar com matrizes e selecionar seus intervalos. Para esses métodos, a entrada é a matriz atributo-valor gerada pela ferramenta *PreText*. Os demais métodos de seleção de atributos utilizam a matriz no formato esparso e foram implementados na linguagem Java, a qual possui artifícios que possibilitam manipular mais facilmente matrizes nesse formato. A saída de cada um desses métodos é uma lista com o valor de *score* atribuído pelo método a cada atributo. A partir dessas medidas de score, é feito um *ranking* de atributos, de maneira a facilitar a seleção dos subconjuntos de atributos.

Para fins de integração com as ferramentas *PreText* e *Discover*, o resultado de todos os métodos de seleção de atributos é convertido, após a seleção, para o formato atributo-valor utilizado por essas ferramentas. Basicamente, reduz-se a matriz atributo-valor inicial, selecionando apenas as colunas relativas aos atributos selecionados pelo método. Foram desenvolvidos, ainda, *scripts* de conversão dessas matrizes no formato *"discover"* para formatos de entrada de ferramentas que podem ser utilizadas para a avaliação dos subconjuntos de atributos. Dessa forma, geram-se matrizes no formato *"arff"*, utilizado pelo ambiente Weka (Witten e Frank, 2005), e matrizes planas utilizadas pelo ambiente Matlab[4].

4.3 Bases de Textos Selecionadas

Para a avaliação dos métodos de seleção de atributos, seis bases de textos contendo artigos científicos foram utilizadas. Optou-se por trabalhar apenas com bases de textos científicos por duas razões. A primeira razão é relativa à confiabilidade dos resultados

[2]http://java.sun.com
[3]http://www.r-project.org
[4]http://www.mathworks.com

obtidos, no que tange à qualidade do vocabulário utilizado nos documentos. Documentos científicos são escritos utilizando termos padrões relativos a um domínio, enquanto que documentos não formais não são escritos nem revistos seguindo um mesmo critério, o que pode acarretar no uso de vocabulário menos específico. Com isso, a eficiência de métodos de seleção de atributos pode ser afetada, tal como relatado em Nogueira et al. (2008b). Uma segunda motivação para a adoção de textos científicos foi possibilitar o uso de métodos de seleção de atributos baseados em contexto (ver Seção 3.3.3), dado que documentos científicos possuem zonas bem delimitadas e de fácil extração.

Cada uma das bases aqui utilizadas é relativa a um grande domínio de conhecimento, tal como sugerido na metodologia de Moura et al. (2008a). A fim de possibilitar a aplicação de avaliação supervisionada, os documentos para um mesmo domínio foram divididos em subdomínios, de acordo com o tema do mesmo, de maneira que cada subdomínio corresponda a uma classe. Os subdomínios são bem delimitados e foram escolhidos de maneira a serem o mais distintos possível, sem interseção de assuntos. A Tabela 4.1 contém uma descrição inicial das bases de textos utilizadas, seus domínios, subdomínios e número de documentos por classe.

Base	**Domínio**	**Subdomínios**	**# Docs**
CIIS	Inteligência Artificial	*Case Based Reasoning*	276
		Inductive Logic Programming	119
		Information Retrieval	179
		Sonification	101
IA	Inteligência Artificial	*Agents & Multiagents*	100
		Fuzzy Logic	100
		Machine Learning	100
		Planning & Schedule	100
		Robotics	100
IFM	Instituto Fábrica do Milênio	*Work Package* 01	80
		Work Package 02	123
		Work Package 03	291
		Work Package 04	97
CS	Ciência da Computação	*Computer Hardware*	89
		Human-Computer Interaction	98
		Artificial Intelligence	94
		Security & Criptology	127
Chemistry	Química	*Analytical Chemistry*	100
		Inorganic Chemistry	97
		Organic Chemistry	100
		Polymer Science	100
Physics	Física	*Biophysics*	95
		Geophysics	97
		Mechanics	117
		Quantum Physics	82

Tabela 4.1: Descrição das bases de textos utilizadas neste trabalho

Das seis bases de textos aqui consideradas, cinco são escritas na língua inglesa, enquanto a base *IFM* é escrita em língua portuguesa. Todas as bases de documentos aqui citadas foram montadas a partir de documentos disponíveis em repositórios de artigos

científicos na internet exclusivamente para a realização deste trabalho, à exceção da base *CIIS*[5], a qual havia sido previamente montada com resumos de artigos científicos, e da base *IFM*. Essa última base contém documentos gerados pelo projeto Instituto Fábrica do Milênio[6], uma instituição brasileira cujo foco é a busca por soluções manufatureiras para as necessidades das indústrias. Cada um dos subdomínios dessa base de textos é formada por documentos relativos a um *Work Package*, que são áreas de pesquisa desenvolvidas pela instituição. A montagem dessas bases se fez necessária devido à ausência de bases de textos compostas apenas de artigos científicos disponíveis para o uso.

Considerando que durante o desenvolvimento deste trabalho ainda não havia ferramenta capaz de selecionar partes dos documentos de maneira eficiente, as bases *CS*, *Chemistry* e *Physics* foram coletadas de maneira que cada subdomínio de cada uma delas possuísse artigos científicos escritos segundo um mesmo modelo (por exemplo, o modelo *Lecture Notes in Computer Science* para artigos do subdomínio *Artificial Intelligence* da base *CS*), de maneira a facilitar a identificação manual das zonas dos documentos. Dessa forma, somente essas três bases foram utilizadas para a avaliação do método ZSTF, apresentado na Seção 3.3, enquanto que para os demais métodos, todas as bases de textos foram utilizadas.

4.4 Considerações Finais

A metodologia para avaliação experimental adotada neste trabalho, constituída de cinco passos, foi criada para a padronização e confiabilidade dos resultados obtidos. Os dois primeiros passos da metodologia, Seleção do Corpus e Padronização dos Documentos, são utilizados para assegurar a qualidade da coleção de documentos trabalhada, bem como garantir que todos se encontrassem em um padrão único antes de serem manipulados pelos algoritmos de seleção de atributos. O terceiro passo, Montagem da Matriz Atributo-Valor, tem por objetivo estruturar os documentos em um formato adequado para o uso de algoritmos de seleção de atributos e, posteriormente, de extração de conhecimento. Por fim, nos passos de Seleção de Atributos e Avaliação dos Subconjuntos de Atributos é onde, respectivamente, se desenvolve a aplicação e avaliação dos métodos de seleção não-supervisionada de atributos comparados neste trabalho.

Para o processo de avaliação, seis bases de dados textuais contendo artigos científicos foram utilizadas. Cada uma dessas bases é relativa a um único grande domínio, contendo, dentro das mesmas, diversos subdomínios que são considerados classes para processo de avaliação supervisionada. Os resultados obtidos com a aplicação de cada método de seleção de atributos comparado e as avaliações da eficiência dos mesmos são apresentados no próximo capítulo deste trabalho.

[5]http://infoserver.lcad.icmc.usp.br/infovis2/PEx
[6]http://www.ifm.org.br

Capítulo

5

Avaliação Experimental

Conforme apresentado anteriormente, o presente trabalho utiliza, para avaliação dos resultados, seis bases de textos e, sobre elas, aplica-se os métodos de seleção não-supervisionada de atributos dispostos na Seção 3.3 deste trabalho. A avaliação da eficiência desses métodos foi dividida em dois focos: avaliação não-supervisionada, utilizando a medida de *Expected Mutual Information Measure*, e avaliação supervisionada, por meio de medida de acurácia de classificadores. Os resultados são apresentados nas seções a seguir.

5.1 Seleção dos Corpus, Geração de Termos e Montagem da Matriz Atributo-Valor

Após serem coletados das diversas fontes de origem, os documentos de cada base de documentos foram unificados em um único repositório. Todos os documentos estavam originalmente no formato *pdf* e foram convertidos para texto plano. Quanto à identificação de zonas no documento, apenas as bases de texto *CS*, *Chemistry* e *Physics* passaram por esse processo, conforme afirmado anteriormente. Os documentos dessas bases foram divididos em cinco zonas: título, resumo, corpo do documento, conclusão e referências. As referências dos artigos foram descartadas, dado que a informação contida nessa zona dos documentos pode não ser estritamente relativa ao conteúdo dos mesmos, podendo prejudicar o processo de avaliação.

Após isso, passou-se para o processo de geração de atributos, conforme sugerido pela metodologia descrita no capítulo anterior. Na Tabela 5.1 são apresentados os resultados da geração de atributos nas bases de textos aqui adotadas, por meio da ferramenta PreText. É possível perceber que há uma diferença significativa do número de *stems* gerados nas bases *CIIS* e *IA* para as demais bases. A base *CIIS* é constituída apenas por resumos de artigos, tendo, por conseqüência, bem menos termos que as demais bases, constituídas

por textos completos. Já a base IA contém documentos de grande extensão sobre uma variedade de assuntos bem distintos, justificando o alto número de termos contidos.

Uma vez gerados os atributos, construiu-se, para cada base de textos, uma matriz atributo-valor. Em todas elas, uma última coluna correspondia à classe do documento, artifício esse necessário para a avaliação supervisionada dos métodos. Com isso, os documentos encontravam-se dispostos em uma representação cujo formato é apto a ser aplicado aos métodos de seleção de atributos.

Base	**# Docs**	**# *Stems* Gerados**
CIIS	675	4101
IA	500	72974
IFM	582	34747
CS	398	23295
Chemistry	372	28194
Physics	383	22195

Tabela 5.1: Resultados da geração de termos nas bases de textos utilizadas

5.2 Aplicação dos Métodos de Seleção de Atributos

Os métodos de seleção não-supervisionada de atributos foram aplicados sobre as diferentes bases de textos, gerando os subconjuntos para cada método (um único subconjunto para cada método com ponto exato de corte e vários subconjuntos para os métodos com *ranking* de atributos). Foram, também, aplicados os métodos de seleção supervisionada de atributos. Para esses métodos, gerou-se, por meio da variação de porcentagem de atributos selecionados, subconjuntos de atributos melhores classificados pelos mesmos. A Tabela 5.2 contém o número de *stems* mantidos em cada um dos subconjuntos gerados para cada base de textos utilizada.

Subconjunto	**CIIS**	**IA**	**IFM**	**CS**	**Chemistry**	**Physics**
Luhn	1181	14710	12551	6937	9067	8638
Luhn-DF	823	7037	8881	3840	4016	4788
Salton	810	7539	6553	3898	5390	4578
5%	206	3649	1738	1165	1410	1110
10%	411	7298	3475	2330	2820	2220
20%	821	14595	6950	4659	5639	4439
30%	1231	21893	10425	6989	8459	6659
40%	1641	29190	13899	9318	11278	8878
50%	2051	36487	17374	11648	14097	11098
60%	2461	43785	20849	13977	16917	13317
70%	2871	51082	24323	16307	19736	15537
80%	3281	58380	27798	18636	22556	17756
90%	3691	65677	31273	20966	25375	19976

Tabela 5.2: Número de *stems* contido nos subconjuntos gerados

Conforme mostrado na Seção 4.3, o método ZSTF, baseado em contexto, foi aplicado somente às bases de textos *Physics*, *Chemistry* e *Computer Science*. Um aspecto a ser

Zona	Peso
Título	0,4
Resumo	0,3
Corpo	0,1
Conclusão	0,2

Tabela 5.3: Pesos adotados para os termos pertencentes a cada zona dos documentos das bases *Chemistry*, *CS* e *Physics*

definido no momento da aplicação desse método é relativo aos pesos atribuídos a cada uma das zonas dos documentos. Neste trabalho, os pesos aplicados aos termos foram decididos a priori, de acordo com o que se esperava subjetivamente da importância da informação de um termo pertencente àquela zona. Os pesos por zona adotados são descritos na Tabela 5.3.

5.3 Avaliação dos Subconjuntos de Atributos

A avaliação dos subconjuntos de atributos gerados pelos diferentes métodos de seleção de atributos comparados neste trabalho foi, conforme sugerido pela metodologia de avaliação adotada, dividida em dois focos, avaliação supervisionada e não-supervisionada. Em cada uma dessas avaliações, três grupos de comparações foram realizados, aplicando a cada um deles o teste estatístico de Kruskal-Wallis. As seções a seguir apresentam e discutem os resultados obtidos. Por questões de espaço e melhor legibilidade deste trabalho, algumas tabelas contendo dados gerados e aqui referenciadas foram colocadas nos Apêndices A e B.

5.3.1 Avaliação Não-Supervisionada

Na avaliação não-supervisionada, calculou-se o valor de EMIM obtido pelos diferentes subconjuntos a fim de verificar a quantidade de informação perdida, em cada base de textos, quando da seleção de atributos por meio da aplicação das diversas técnicas.

Nas Figuras 5.1, 5.2, 5.3, 5.4, 5.5 e 5.6 é possível observar os valores de EMIM obtidos com a aplicação dos diferentes métodos de seleção de atributos em função do número de atributos considerados em cada subconjunto gerado. É necessário frisar que as linhas nos gráficos são apenas para fins de indicação de tendência, não sendo calculados os valores de EMIM nos intervalos entre os subconjuntos gerados. Em uma análise primária dessas figuras, é possível perceber que há grande proximidade de desempenho entre as técnicas que não fornecem pontos exatos de corte (TC, TV, TVQ, RTF, RDF, TFIDF e ZSTF). Outro ponto que chama a atenção é que as técnicas que fornecem um ponto exato de corte (Luhn, LuhnDF e Salton) geralmente apresentam um valor de EMIM abaixo do apresentado pelas técnicas sem ponto exato de corte. Isso não permite, no entanto, afirmar que existe diferença estatística entre os métodos, a qual é averiguada nos grupos de comparações dispostos a seguir.

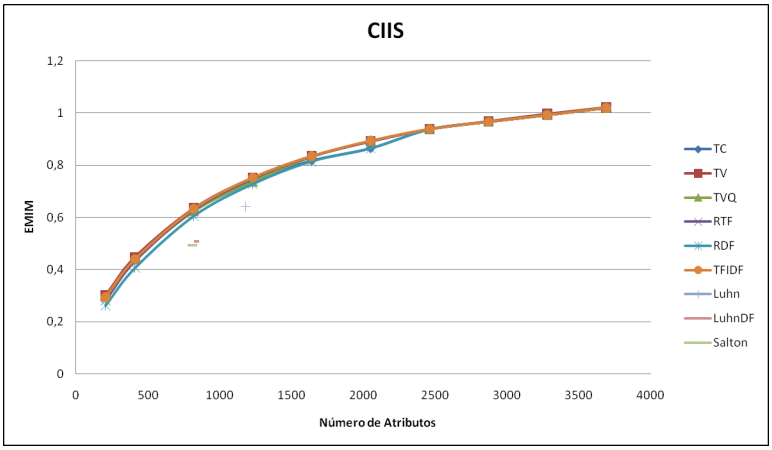


Figura 5.1: Valor de EMIM x Número de atributos em cada subconjunto - Base de textos *CIIS*

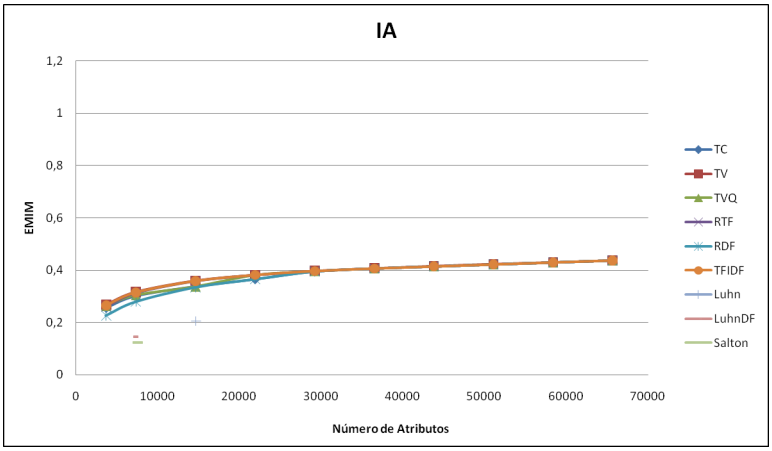


Figura 5.2: Valor de EMIM x Número de atributos em cada subconjunto - Base de textos *IA*

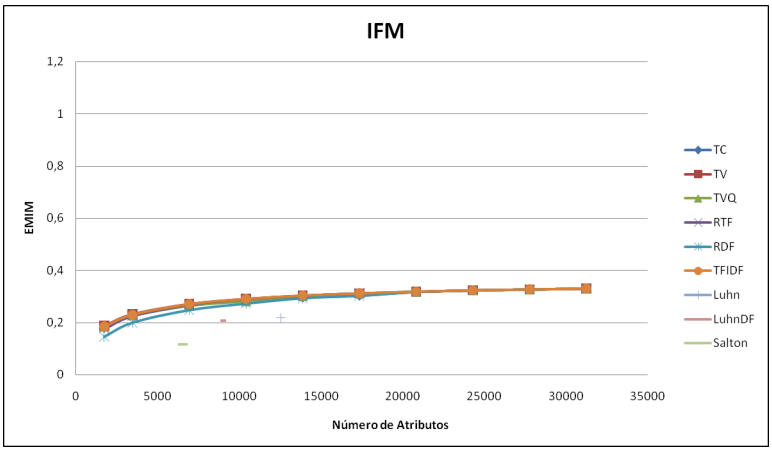


Figura 5.3: Valor de EMIM x Número de atributos em cada subconjunto - Base de textos *IFM*

Grupo 1: Comparação dos Métodos Sem Pontos Exatos de Corte

Os métodos sem ponto exato de corte foram comparados entre si, a fim de verificar, dentre esses métodos computacionalmente mais custosos, quais conseguem preservar a maior quantidade de informação dentro da base para subconjuntos de atributos de mesma cardinalidade. Os valores de EMIM obtidos com a variação de atributos, para os métodos TC, TV, TVQ, RTF, RDF, TFIDF e ZSTF (quando aplicável) em todas as bases de

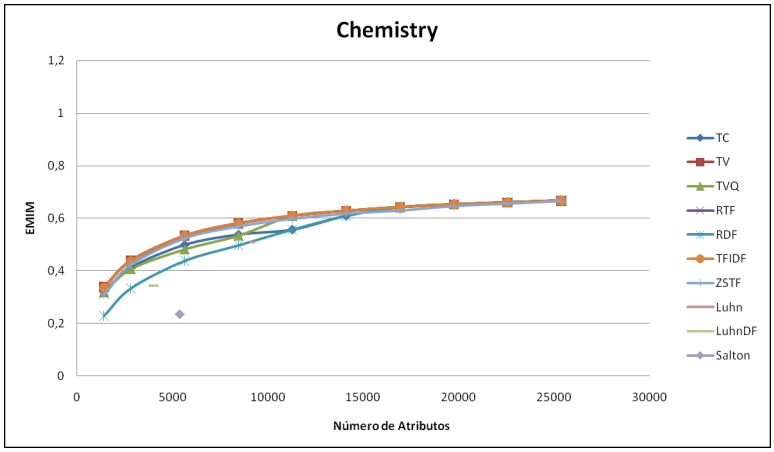


Figura 5.4: Valor de EMIM x Número de atributos em cada subconjunto - Base de textos *Chemistry*

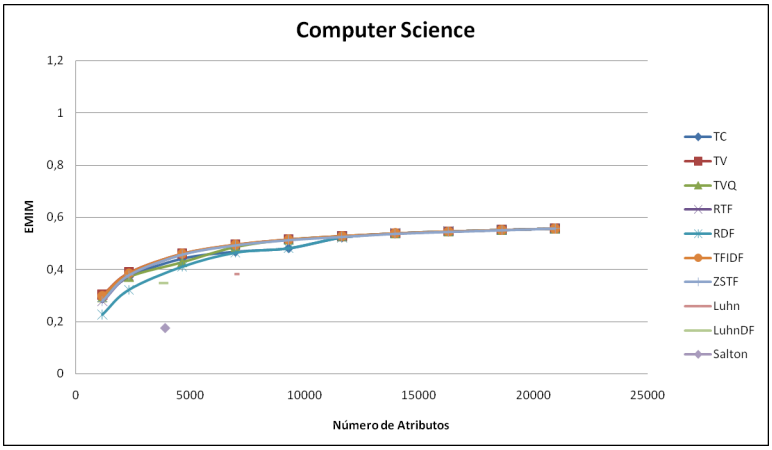


Figura 5.5: Valor de EMIM x Número de atributos em cada subconjunto - Base de textos *CS*

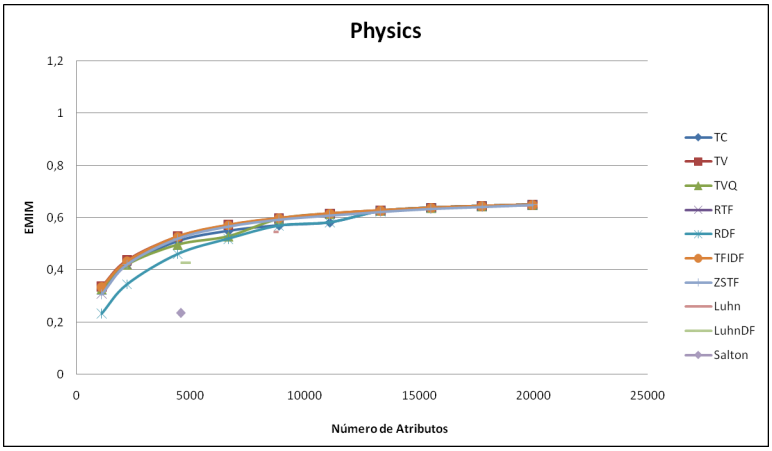


Figura 5.6: Valor de EMIM x Número de atributos em cada subconjunto - Base de textos *Physics*

textos aqui consideradas podem ser verificados nas Tabelas A.1 e A.2.

Com os valores de EMIM obtidos, é possível perceber, como era esperado, que quanto maior a quantidade de atributos, maior a quantidade de informação preservada. Outro aspecto importante refere-se ao alto desempenho apresentado pelas técnicas na base CIIS. Esse desempenho ocorre devido ao fato de que essa base possui sensivelmente menos atributos que as demais bases, por se tratar de documentos de menor tamanho, com menor

número de *stems*, como pode ser observado na Tabela 5.1. Dessa forma, o poder de um atributo em predizer o valor de outro se torna maior. Outro ponto importante refere-se à constatação de que o método ZSTF, aqui proposto, possui desempenho semelhante aos demais métodos, apresentando valores de EMIM bastante similares aos melhores métodos em quase todos os casos.

Para comparar o impacto dos métodos de seleção de atributos, o teste de Kruskal-Wallis (Kruskal e Wallis, 1952) foi aplicado sobre os valores obtidos. Foram feitos testes estatísticos com os valores separados por base, a fim de verificar o impacto dos métodos ante as especificidades de cada base de textos, bem como testes considerando todas as bases de textos juntas. Nesse último caso, o teste foi dividido em dois outros testes: sem considerar o método ZSTF (todas as seis bases) e considerando o método ZSTF (somente as bases Physics, Chemistry e CS).

Os p-valores obtidos nos testes estatísticos desse grupo de comparações são exibidos na Tabela 5.4. Analisando o resultado das comparações estatísticas, é possível perceber que não há evidências estatísticas que indiquem diferenças entre os métodos comparados, uma vez que o desempenho dos mesmos é bastante similar e positivo. Isso nos mostra que não há, quanto à preservação de informações na base de textos, diferença significativa de desempenho quando é utilizado algum dentre esses métodos.

Base	**p-valor**
CIIS	0,9966
IA	0,9989
IFM	0,9989
Chemistry	0,9964
CS	0,998
Physics	0,9962
Todas sem ZSTF	0,9679
Todas com ZSTF	0,9685

Tabela 5.4: p-valores obtidos para o grupo 1 de comparações não-supervisionadas

Grupo 2: Comparação dos Métodos Com Pontos Exatos de Corte

O segundo grupo de comparações aqui utilizado contém a comparação de desempenho dos métodos Luhn, LuhnDF e Salton, a fim de verificar a qualidade dos pontos de corte sugeridos pelos mesmos na preservação de informação na base de documentos. Os resultados obtidos com esses métodos são mostrados na Tabela 5.5.

Base	**Luhn**	**LuhnDF**	**Salton**
CIIS	0,64129	0,50841	0,49452
IA	0,2052	0,14607	0,12199
IFM	0,21677	0,2081	0,11738
Chemistry	0,50994	0,34416	0,23424
CS	0,38107	0,34695	0,17612
Physics	0,54446	0,42816	0,23531

Tabela 5.5: Valor de EMIM para os métodos Luhn, LuhnDF e Salton, obtidos para todas as bases

Nesse grupo de comparações não foi possível efetuar teste estatístico de Kruskal-Wallis, devido ao fato de que o número de amostras era seis (correspondente às bases de dados), enquanto esse teste exige que, no mínimo, oito amostras estejam disponíveis. Assim, não se pode afirmar que um método é significativamente melhor ou pior que outro. Entretanto, é possível perceber que em todos os casos o método de Luhn obtém melhor desempenho, seguido, em ordem, pelos métodos de LuhnDF e Salton, tendendo, portanto, a apresentar melhores resultados.

Grupo 3: Comparação Entre Métodos Com e Sem Pontos Exatos de Corte

Esse terceiro grupo visa comparar, em subconjuntos de igual cardinalidade, o desempenho dos métodos que não fornecem pontos exatos de corte e os métodos de Luhn, LuhnDF e Salton, analisando a eficiência dos mesmos na preservação de informação nas bases de textos. As Tabelas 5.6, 5.7 e 5.8 contêm os resultados das análises efetuadas. Como o método ZSTF só pôde ser aplicado às bases *Chemistry*, *CS* e *Physics*, os valores para esse método são nulos para as bases *CIIS*, *IA* e *IFM*.

Base	**TC**	**TV**	**TVQ**	**RTF**	**RDF**	**TFIDF**	**ZSTF**	**Luhn**
CIIS	0,73016	0,74119	0,72587	0,7359	0,71792	0,74016	-	0,64129
IA	0,33832	0,36014	0,33748	0,35855	0,33544	0,35951	-	0,2052
IFM	0,29252	0,29935	0,29662	0,29921	0,28416	0,29955	-	0,21677
Chemistry	0,54197	0,58803	0,5732	0,58745	0,53567	0,58869	0,57614	0,50994
CS	0,46776	0,49471	0,48539	0,49365	0,46331	0,49461	0,49123	0,38107
Physics	0,56941	0,59661	0,59234	0,59606	0,56727	0,59646	0,5896	0,54446

Tabela 5.6: Valores de EMIM para as diferentes técnicas com subconjuntos de cardinalidade igual ao método de Luhn

Base	**TC**	**TV**	**TVQ**	**RTF**	**RDF**	**TFIDF**	**ZSTF**	**LuhnDF**
CIIS	0,62949	0,63606	0,62443	0,6297	0,60693	0,63364	-	0,50841
IA	0,29916	0,31414	0,30252	0,30938	0,27811	0,31245	-	0,14607
IFM	0,27781	0,28398	0,27623	0,28255	0,26367	0,28346	-	0,2081
Chemistry	0,45785	0,48989	0,44737	0,48102	0,38444	0,48797	0,47866	0,34416
CS	0,42413	0,44201	0,41462	0,43754	0,38579	0,44102	0,437	0,34695
Physics	0,51858	0,53765	0,50341	0,53411	0,46711	0,53701	0,52852	0,42816

Tabela 5.7: Valores de EMIM para as diferentes técnicas com subconjuntos de cardinalidade igual ao método de Luhn-DF

Base	**TC**	**TV**	**TVQ**	**RTF**	**RDF**	**TFIDF**	**ZSTF**	**Salton**
CIIS	0,62417	0,63132	0,61937	0,62483	0,5982	0,62926	-	0,49452
IA	0,30331	0,31867	0,30612	0,31451	0,28409	0,31723	-	0,12199
IFM	0,26223	0,2683	0,26253	0,26578	0,24339	0,26737	-	0,11738
Chemistry	0,49317	0,52842	0,47724	0,52374	0,43497	0,52789	0,51834	0,23424
CS	0,42582	0,44342	0,41579	0,43906	0,38643	0,44265	0,43861	0,17612
Physics	0,51364	0,53267	0,49984	0,52863	0,46444	0,5317	0,52327	0,23531

Tabela 5.8: Valores de EMIM para as diferentes técnicas com subconjuntos de cardinalidade igual ao método de Salton

Novamente, a limitação de amostras disponíveis impossibilitou a aplicação do teste estatístico de Krskal-Wallis aos resultados. Efetuando uma análise manual nos resultados obtidos, é possível perceber que os métodos sem ponto exato de corte apresentam sempre um valor de EMIM superior aos dos métodos com pontos exatos de corte. As maiores diferenças se dão nas comparações com o método de Salton, nas quais é possível perceber, em alguns casos, métodos de seleção com valores de EMIM acima do dobro do valor apresentado pelo método de Salton. Isso não permite, porém, que se delineie significância estatística no desempenho dos métodos.

Entre os métodos sem pontos de corte, embora o método TV tenha apresentado maior valor de EMIM para a maioria dos casos, não há uma nítida ordenação de desempenhos entre os métodos, havendo grande similaridade entre os mesmos. É importante ressaltar que o método aqui proposto, ZSTF, conseguiu desempenho bastante similar aos melhores métodos nas análises.

5.3.2 Avaliação Supervisionada

No processo de avaliação supervisionada, quatro algoritmos de classificação foram utilizados para comparação da acurácia preditiva: C4.5, Naïve Bayes, SVM e KNN. Para todos esses algoritmos, bem como para os métodos de seleção de atributos supervisionados (Ganho de Informação e Razão de Ganho), utilizou-se as implementações disponíveis no ambiente Weka (Witten e Frank, 2005). Dessa forma, o algoritmo C4.5 será, doravante, denominado J48, uma vez que esse é o nome da implementação das árvores de decisão C4.5 no ambiente Weka. Uma vez que o objetivo dessa avaliação não é a comparação do desempenho dos algoritmos de aprendizado, mas sim a análise do comportamento dos diferentes métodos de seleção de atributos ante um mesmo processo de avaliação, foram utilizados os parâmetros padrões dos algoritmos sugeridos pelo ambiente. Para o processo de k-vizinhos mais próximos, adotou-se k igual a 10. Para o treinamento dos algoritmos, utilizou-se 10-*fold cross validation*.

Nas figuras 5.7, 5.8, 5.9, 5.10, 5.11, 5.12 é possível observar as acurácias de predição obtidas pelos subconjuntos de atributos gerados pelos diferentes métodos de seleção de atributos aqui considerados. Assim como nos gráficos apresentados para a avaliação não-supervisionada, as linhas dos gráficos apresentados nessa seção são apenas um artifício para indicar tendência de evolução das medidas de acurácia ao longo dos subconjuntos, não sendo calculados os valores de acurácia nos intervalos entre os subconjuntos de atributos gerados. As tabelas contendo todos os valores de acurácia médios obtidas pelos diferentes classificadores nos diferentes subconjuntos de atributos gerados para cada base de textos, os quais deram origem aos gráficos supra-citados, bem como os respectivos desvios padrões, estão dispostas no Apêndice B.

Uma análise inicial dos gráficos permite observar dois aspectos que, embora não relevantes ao processo de comparação aqui proposto, merecem ser discutidos. Um primeiro fato é que, para a maioria dos casos, o algoritmo J48, a partir de uma dada cardinalidade de atributos, não sofre efeito pela adição de novos atributos à base. Isso se deve ao fato de que o algoritmo de indução de uma árvore de decisão possui um viés que busca selecionar

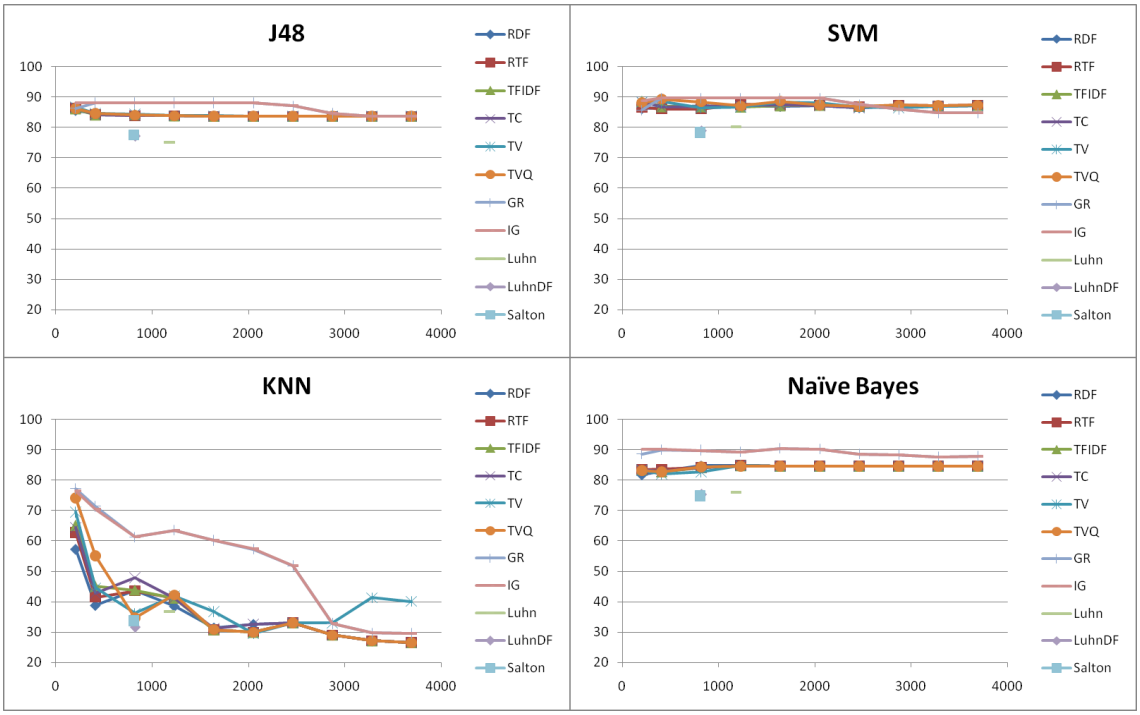


Figura 5.7: Valor de acurácia x Número de atributos em cada subconjunto - Base de textos *CIIS*

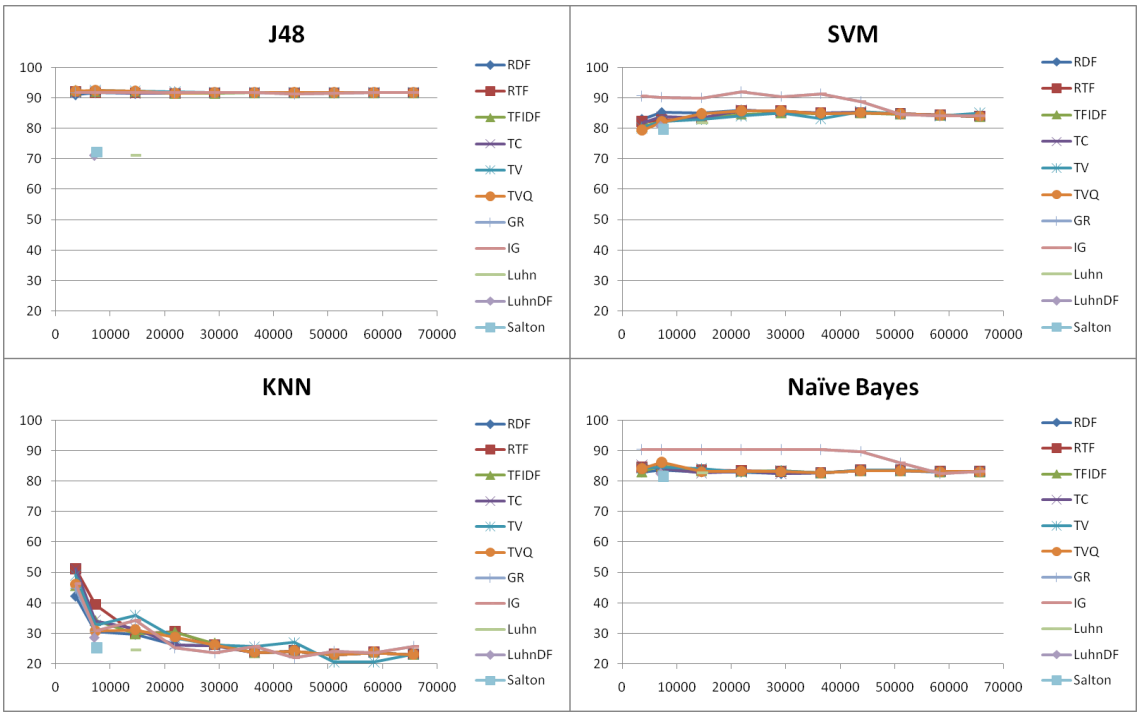


Figura 5.8: Valor de acurácia x Número de atributos em cada subconjunto - Base de textos *IA*

a melhor divisão da base de dados utilizando o menor número de atributos possível, o que pode ser encarado como um processo embutido de seleção de atributos. Dessa forma, uma vez que o algoritmo encontra um subconjunto ótimo de atributos para decidir sobre a classe de um objeto, ele desconsidera todos os demais.

Um segundo aspecto se remete ao desempenho do algoritmo KNN de acordo com o aumento da cardinalidade dos subconjuntos de atributos gerados. Em todas bases de textos, quanto maior é o tamanho do subconjunto de atributos obtido, pior é a acurácia preditiva

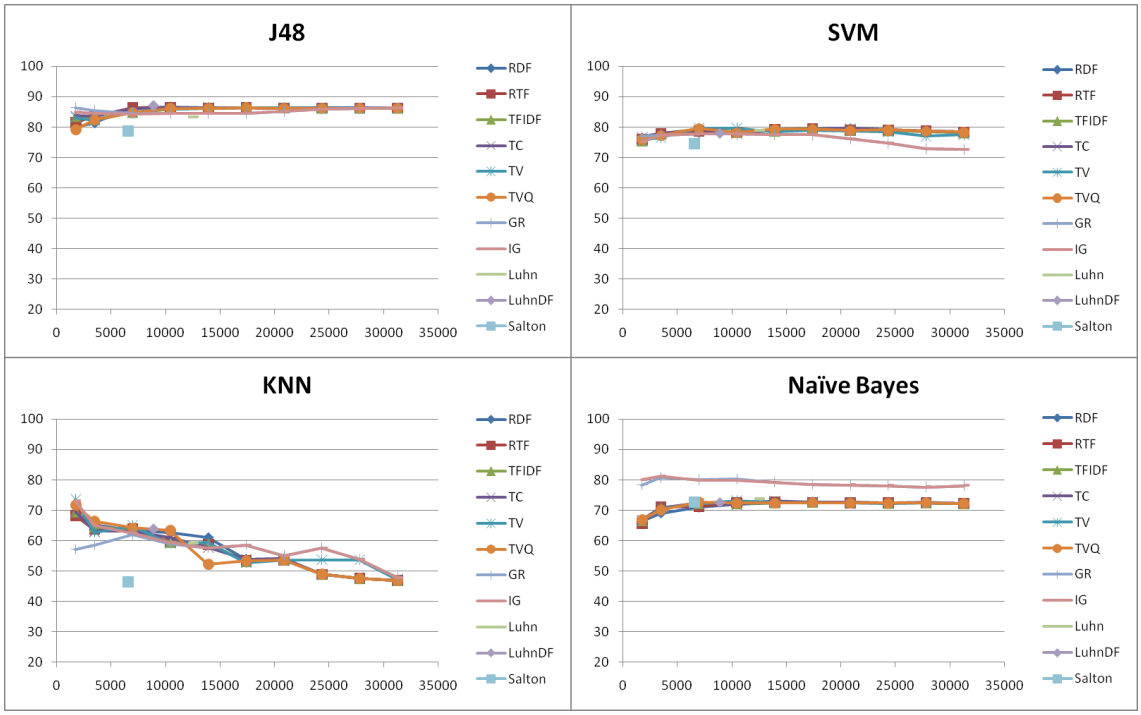


Figura 5.9: Valor de acurácia x Número de atributos em cada subconjunto - Base de textos *IFM*

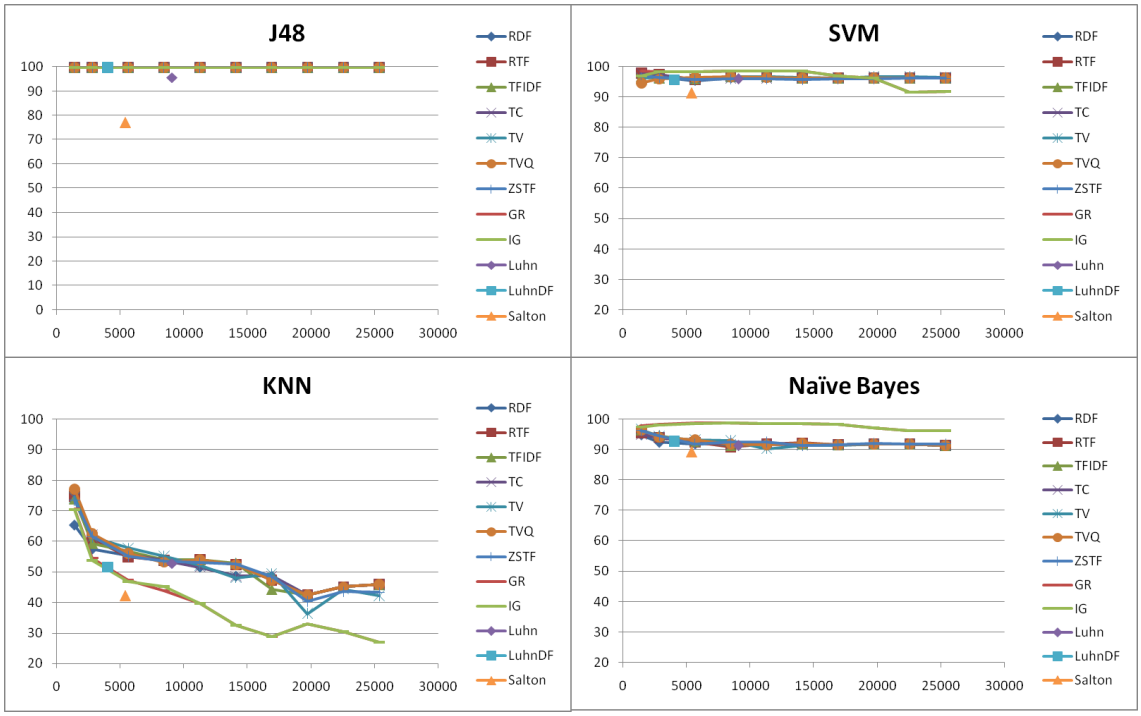


Figura 5.10: Valor de acurácia x Número de atributos em cada subconjunto - Base de Textos *Chemistry*

obtida. Esse método se mostra, portanto, altamente sensível à alta dimensionalidade de atributos, reforçando a importância de selecionar os melhores atributos da base em uma etapa prévia à aplicação desse algoritmo de aprendizado.

Por fim, é possível perceber que o método de Salton obtém, quase sempre, as menores acurácias dentre os métodos aqui comparados quando é considerada a mesma cardinalidade de atributos. Além disso, pode-se perceber que o desempenho dos métodos não-supervisionados de seleção de atributos que não fornecem pontos exatos de corte (RDF,

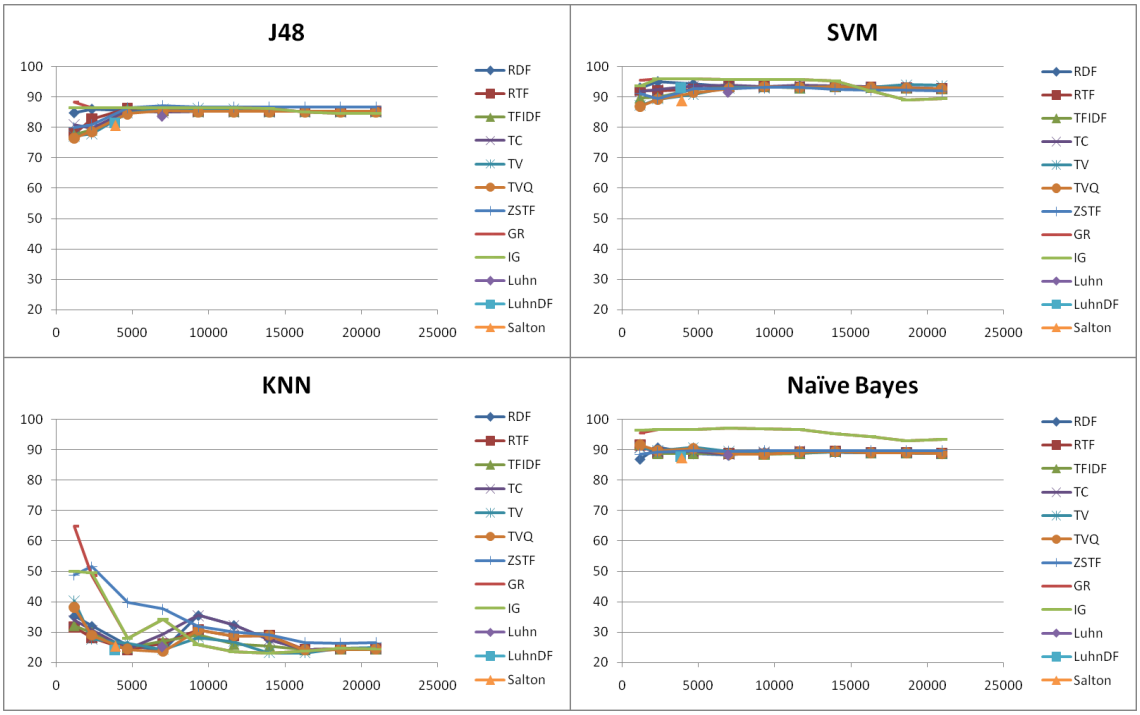


Figura 5.11: Valor de acurácia x Número de atributos em cada subconjunto - Base de textos *CS*

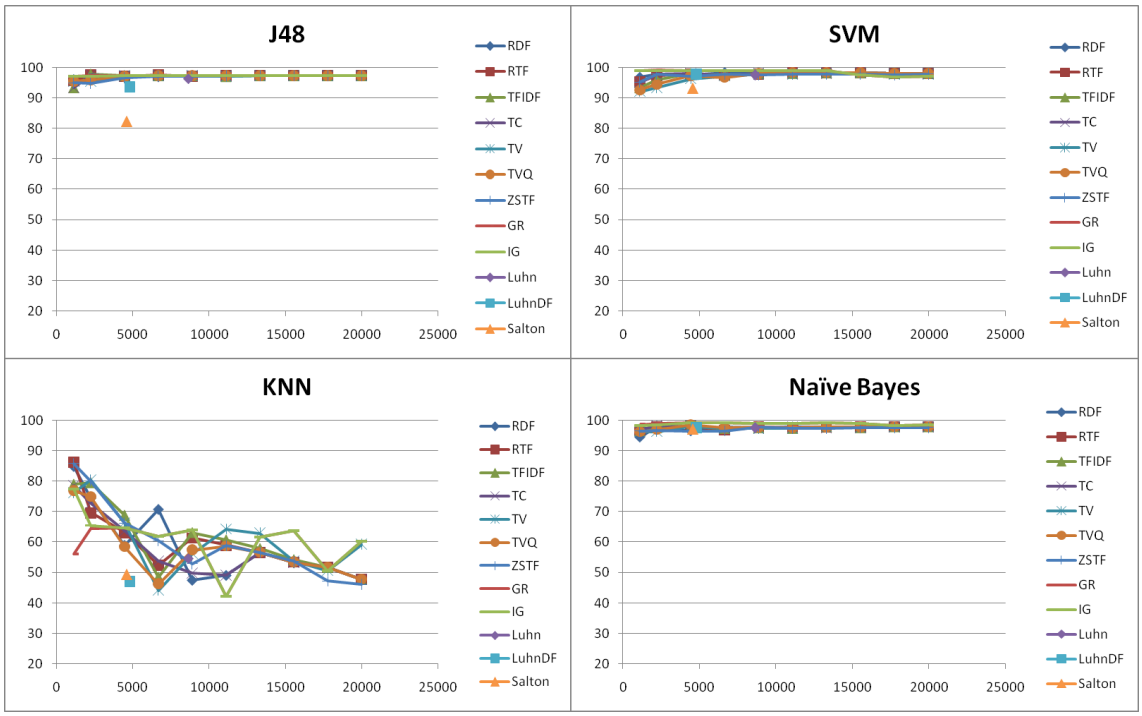


Figura 5.12: Valor de acurácia x Número de atributos em cada subconjunto - Base de textos *Physics*

RTF, TFIDF, TC, TV, TVQ e ZSTF) é bastante similar, com todos apresentando uma mesma tendência de desempenho com a variação da cardinalidade dos subconjuntos de atributos. Não se pode afirmar, entretanto, somente com essa análise primária, se há ou não, estatisticamente, significância na diferença de desempenho entre os métodos de seleção de atributos. Nos grupos de comparação sugeridos pela metodologia de avaliação aqui adotada, os três grupos de comparações efetuados buscam comprovar fatos como esse. Os resultados da realização desses grupos de comparações estão dispostos a seguir.

Grupo 1: Comparação dos Métodos Sem pontos exatos de Corte

Para a comparação dos métodos sem pontos exatos de corte, utilizou-se do artifício da variação da porcentagem dos atributos selecionados para cada método, analisando a acurácia preditiva obtida em cada um desses subconjuntos pelos diferentes classificadores aqui utilizados.

Sob os valores de acurácia obtidos durante os experimentos foi aplicado o teste estatístico de Kruskal-Wallis, com pós-teste de Dunn. Nesses testes, comparou-se a diferença estatística do desempenho dos métodos não-supervisionados entre si, bem como do desempenho desses frente aos métodos supervisionados - Razão de Ganho (GR) e Ganho de Informação (IG). Dessa forma, os testes desse grupo foram divididos em seis subgrupos: considerando todas bases de textos, com e sem métodos supervisionados; considerando apenas as bases cuja aplicação do ZSTF é possível, com e sem métodos supervisionados; e considerando cada base de textos separadamente, com e sem métodos supervisionados. Os p-valores obtidos com essa análise podem ser encontrados na Tabela 5.9. Os valores de acurácia obtidos, os quais deram origem a essa análise estatística, bem como os respectivos desvios padrões, podem ser observados nas Tabelas B.1, B.2, B.3, B.4, B.5 e B.6.

Base	**p-valor**	
	Com GR e IG	**Sem GR e IG**
CIIS	$< 0{,}0001$	0,9997
IA	0,5736	$> 0{,}9999$
IFM	0,9911	$> 0{,}9999$
Chemistry	0,8373	$> 0{,}9999$
CS	0,2811	0,9873
Physics	0,0538	0,6515
Todas sem ZSTF	0,0821	$> 0{,}9999$
Todas com ZSTF	0,1177	0,9997

Tabela 5.9: p-valores obtidos para o grupo 1 de comparações supervisionadas

De acordo com os resultados obtidos, para a grande maioria dos casos, não há diferença significativa no desempenho dos métodos de seleção, seja analisando os métodos não-supervisionados isoladamente ou em conjunto com os métodos supervisionados, com a exceção da base de textos *CIIS*. Isso nos permite afirmar que, para a maioria dos casos, métodos não supervisionados de seleção de atributos conseguem preservar a estrutura de classes (ou agrupamentos bem definidos) tão bem quanto métodos supervisionados. É necessário, entretanto, verificar que as análises que consideram apenas os métodos não-supervisionados apresentam p-valores maiores, indicando a maior igualdade entre esses métodos, dado que a hipótese nula é de que existe igualdade de desempenho entre os métodos. Os valores do pós-teste de Dunn para as comparações de métodos que houve diferença estatística de desempenho na base *CIIS* pode ser verificado na Tabela 5.10. Pode-se verificar que, nessa base, os métodos IG e GR ganham de todos os métodos não-supervisionados de atributo, sendo que o método de Ganho de Informação, para a maioria das comparações, a hipótese nula é rejeitada com 99% de confiança.

Métodos	p-valor
RDF vs. GR	P<0.05
RDF vs. IG	P<0.01
RTF vs. GR	P<0.05
RTF vs. IG	P<0.01
TFIDF vs. GR	P<0.05
TFIDF vs. IG	P<0.01
TC vs. GR	P<0.05
TC vs. IG	P<0.01
TV vs. GR	P<0.05
TV vs. IG	P<0.05
TVQ vs. GR	P<0.05
TVQ vs. IG	P<0.05

Tabela 5.10: Resultados do pós-teste de Dunn para as comparações em que houve diferença estatística na base CIIS

Grupo 2: Comparação dos Métodos Com Pontos Exatos de Corte

Com o intuito de avaliar a qualidade dos pontos de corte fornecidos pelas técnicas Luhn, LuhnDF e Salton, o presente trabalho efetuou uma comparação estatística da acurácia obtida pelos diferentes classificadores utilizando subconjuntos de atributos fornecidos pelos três métodos em todas as bases de documentos aqui utilizadas.

As análises estatísticas para as bases de textos isoladamente não foram possíveis de serem efetuadas devido à restrição do número de amostras (apenas quatro, uma para cada algoritmo de classificação). A aplicação do teste estatístico de Kruskal-Wallis para a comparação dos valores obtidos em todas as bases apontou um p-valor de 0,5748, o que indica que não há diferença estatística significativa entre os métodos. Dessa forma, não há diferença na aplicação de qualquer desses três métodos no que se refere à preservação da estrutura de classes ou grupos, dado que a qualidade do ponto de corte sugerido por todos para a caracterização das classes é muito similar. A acurácia decorrente da utilização dos subconjuntos de atributos gerados por esses três métodos, os quais deram origem a essa análise estatística, bem como os respectivos desvios padrões, podem ser verificados na Tabela B.7.

Grupo 3: Comparação Entre Métodos Com e Sem Pontos Exatos de Corte

O terceiro grupo de medidas supervisionadas visa comparar, com subconjuntos de mesma cardinalidade de atributos, a eficiência na preservação de estrutura de classe dos diferentes métodos de seleção de atributos que não fornecem ponto exato de corte e os método que fornecem ponto exato de corte.

Na avaliação estatística dessa comparação, buscou-se, novamente, comparar os métodos não-supervisionados entre si, bem como com os métodos supervisionados. Para isso, os testes estatísticos envolvendo todas as bases foram divididos em quatro subgrupos: considerando todas as bases de textos, com e sem métodos supervisionados; e considerando apenas as bases de textos cuja aplicação do método ZSTF é possível, com e sem métodos supervisionados. A aplicação para bases de textos separadamente também não foi possível devido à restrição de apenas quatro amostras por base. Os resultados obtidos

podem ser verificados na Tabela 5.11. Nas Tabelas B.8, B.9 e B.10 é possível observar as acurácias obtidas pelos diferentes métodos com subconjuntos de cardinalidade igual aos métodos de Luhn, LuhnDF e Salton, respectivamente, as quais foram utilizadas para a obtenção desse teste estatístico.

Cardinalidade Igual Luhn		
Base	**Com GR e IG**	**Sem GR e IG**
Todas sem ZSTF	0,792	0,9523
Todas com ZSTF	0,8943	0,9989
Cardinalidade Igual LuhnDF		
Base	**Com GR e IG**	**Sem GR e IG**
Todas sem ZSTF	0,8865	0,9791
Todas com ZSTF	0,7679	0,924
Cardinalidade Igual Salton		
Base	**Com GR e IG**	**Sem GR e IG**
Todas sem ZSTF	0,3123	0,4386
Todas com ZSTF	0,4403	0,6166

Tabela 5.11: P-valores obtidos para o grupo 3 de comparações supervisionadas

É possível verificar que não há significância estatística no diferença entre os métodos aqui comparados. Analisando os p-valores considerando os métodos GR e IG, é possível perceber que os métodos não-supervisionados apresentaram eficiência bastante similar aos métodos supervisionados, dado que os p-valores foram bastante altos. Quando não se considera esses métodos supervisionados, os p-valores assumem valores ainda maiores, indicando grande semelhança da eficiência dos métodos não-supervisionados entre si.

Como esses valores permanecem altos quando comparados com o método ZSTF, pode-se assumir que esse é um método cuja eficiência na preservação de estrutura de classes é bastante similar à apresentada pelos demais métodos não supervisionados sem, no entanto, superá-los. O mesmo ocorre quando comparando o método LuhnDF com os demais métodos. Os únicos valores que destoam negativamente se referem às comparações com cardinalidade de atributos igual à do método de Salton. Embora não exista diferença estatística significante entre esse método e os demais, os baixos p-valores podem ter sido puxados para baixo devido ao desempenho um pouco inferior apresentado por esse método, fazendo-o destacar dos demais.

5.4 Considerações Finais

Neste capítulo, foram apresentadas as avaliações experimentais realizadas para comparação objetiva dos métodos de seleção não-supervisionada de atributos. Essa comparação segue dois focos, visando analisar a capacidade dos métodos em preservar as informações presentes na base de documentos (avaliação não-supervisionada), bem como verificar a capacidade dos métodos de preservar a estrutura de classes ou grupos (avaliação supervisionada).

A maioria dos resultados não indica diferenças estatísticas entre os métodos não-supervisionados comparados, tampouco com métodos supervisionados de seleção de atributos. Isso indica, em primeira instância, que qualquer um dos métodos aqui comparados

se torna uma opção eficiente para seleção de atributos em um processo não-supervisionado de Mineração de Textos. As conclusões que podem ser feitas a partir desses resultados, bem como as contribuições e limitações deste trabalho e as propostas de trabalhos futuros serão apresentadas no próximo capítulo deste trabalho.

Capítulo

6

Conclusões e Trabalhos Futuros

Tarefas de Mineração de Textos lidam, geralmente, com espaços inerentemente esparsos e de alta dimensionalidade de atributos. Nesse contexto, a eficiência de métodos de aprendizado de máquina, como os algoritmos de agrupamento de dados, é muito prejudicada. Além disso, a alta dimensionalidade acarreta em maiores custos computacionais de memória e processamento para a realização das tarefas de aprendizado automático. Assim, faz-se necessária a utilização de métodos que consigam reduzir eficientemente os conjuntos de atributos, de maneira que não haja perda significativa de informações na base de textos. Quando lidando com contextos não-supervisionados, a redução de dimensionalidade de atributos é uma tarefa especialmente difícil, dado que a qualidade de um atributo em atividades não-supervisionadas varia de acordo com o objetivo da tarefa a ser realizada.

Neste trabalho, contribuiu-se com a seleção não-supervisionada de atributos para Mineração de Textos, cumprindo o objetivo principal proposto. Dez métodos não supervisionados foram comparados - *Ranking* por *Term Frequency* (RTF), *Ranking* por *Document Frequency* (RDF), *Term Frequency Inverse Document Frequency* (TFIDF), Método de Luhn, Método de Salton, Método LuhnDF, *Term Contribution* (TC), *Term Variance* (TV), *Term Variance Quality* (TVQ) e *Zone-Scored Term Frequency* (ZSTF). Destes métodos, LuhnDF e ZSTF foram aqui propostos. A avaliação desses métodos seguiu uma metodologia aqui apresentada, dividindo-se em dois tipos: não-supervisionada, por meio da medida de EMIM (*Expected Mutual Information Measure*), e supervisionada, pela obtenção da acurácia preditiva de quatro classificadores (C4.5, SVM, KNN e *Naïve Bayes*). Em cada um desses tipos de avaliação, três grupos de comparações foram efetuadas: confrontando apenas os métodos que não fornecem ponto exato de corte (RTF, RDF, TFIDF, TC, TV, TVQ e ZSTF); confrontando apenas os métodos que fornecem pontos exatos de corte (Luhn, LuhnDF e Salton); e confrontando todos os métodos.

De maneira geral, ao longo dos três grupos de comparações, não se apontou diferença estatística no desempenhos dos métodos não-supervisionados de atributos aqui compara-

dos. Analisando os resultados do primeiro grupo de comparações, pode-se perceber que os métodos que não apresentam pontos exatos de corte (RTF, RDF, TFIDF, TC, TV, TVQ e ZSTF) obtiveram resultados muito semelhantes entre si. Embora seja possível perceber que o método TV apresentou, de modo geral, medidas superiores às dos demais métodos na avaliação não-supervisionada, a diferença não foi grande o suficiente para que houvesse significância estatística e fosse possível afirmar que esse é o melhor método para preservação da informação em uma base textual. Já na avaliação supervisionada, houve grande alternância de métodos com melhor desempenho ao longo dos testes, o que mostra que não se pode afirmar que há diferença, dentre esses métodos, na preservação de estrutura de classes ou grupos. Nesse contexto, pode-se perceber, quanto ao método ZSTF aqui proposto, que esse é um método competitivo quanto aos demais que não fornecem ponto exato de corte, obtendo avaliações que apontam uma eficiência tão boa quanto à dos demais métodos.

O segundo grupo de comparações mostrou que os métodos Luhn, LuhnDF e Salton são bastante eficientes, não havendo diferenças significativas entre os mesmos. Embora ao longo das análises, tanto não-supervisionadas como supervisionadas, o método de Salton tenha geralmente apresentado eficiências abaixo das apresentadas pelos dois outros métodos, essa diferença não foi estatisticamente significativa. Outro fato que se pode notar é que o método LuhnDF, aqui proposto a partir de um relaxamento flexível do método de Salton no que tange aos intervalos de DF a serem apresentados pelos termos que são selecionados, apresentou, geralmente, valores de avaliação superiores à desse método, embora não superior o suficiente para que se aponte diferença estatística.

Por fim, o terceiro grupo de comparações nos permite afirmar que não existe diferença entre os métodos que fornecem ponto exato de corte (Luhn, LuhnDF e Salton) e os demais métodos. A avaliação de subconjuntos de mesma dimensionalidade mostrou que os pontos de cortes fornecidos por esses métodos são suficientes para que se preserve na base de textos a informação que a base de textos pode fornecer, bem como a estrutura das classes ou grupos que se delineiam na base.

Em suma, de acordo com as avaliações estatísticas efetuadas, os testes aqui realizados apontam que os métodos mais simples e de menor custo computacional, como os métodos de Luhn, LuhnDF, Salton, RTF e RDF, no que se refere à preservação de informações na base de texto e conservação da estrutura de classes ou grupos, são tão eficientes quanto métodos mais complexos. Há, ainda, a vantagem de que, aplicados os métodos Luhn, LuhnDF e Salton, já se sabe exatamente quantos e quais atributos devem ser mantidos. A alta eficiência demonstrada por esses métodos comprova a hipótese que rege este trabalho.

Uma outra conclusão, decorrente da avaliação supervisionada aqui realizada, é que métodos não-supervisionados podem ser utilizados em processos supervisionados de Mineração de Textos, com eficiência compatível à de dois métodos supervisionados bastante utilizados (Razão de Ganho e Ganho de Informação), tendo como vantagem o fato de apresentarem um custo computacional menor. Embora os métodos supervisionados tenham, geralmente, superado a eficiência de métodos não-supervisionados, a diferença das avaliações é pequena, não sendo suficiente, na maioria dos casos, para apontar-se significância estatística no desempenho dos métodos.

6.1 Principais Contribuições

A principal contribuição deste trabalho se refere às comparações objetivas entre métodos não-supervisionados de seleção de atributos para Mineração de Textos, devido à existência de poucos trabalhos na literatura envolvendo tal assunto. As análises aqui realizadas podem ser utilizadas como base para trabalhos que utilizam aprendizado não-supervisionado em Mineração de Textos. Como decorrência dessa comparação, contribuiu-se, com uma metodologia de extração de taxonomias de tópicos a partir de bases textuais realizada no âmbito do laboratório de pesquisa, na medida em que os métodos aqui comparados serão utilizados para a seleção de atributos na etapa de pré-processamento da mesma.

Outra contribuição decorrente deste trabalho é a proposta dos métodos LuhnDF e ZSTF para seleção não-supervisionada de atributos. Embora esses métodos não superem os métodos já existentes na literatura, quanto à preservação de informações na base textual e conservação de estrutura de classes ou grupos, esses métodos se mostraram competitivos aos demais, se tornando opções viáveis para usuários de processos de Mineração de Textos. O método LuhnDF tem como principal vantagem o fato de que é menos restritivo que o método de Salton, bem como apresenta um baixo custo computacional. Já o método ZSTF tem como principal vantagem a possibilidade de ponderar zonas dos documentos, sendo útil para casos em que existam zonas bem definidas e deseja-se efetuar uma distinção da informação trazida pelas mesmas.

Algumas dessas contribuições parciais tiveram seus resultados formalizados por meio de publicações em congressos científicos da área de Inteligência Artificial e um relatório técnico, disponíveis nos trabalhos de Nogueira et al. (2008b), Nogueira et al. (2008a), Moura et al. (2008a), Moura et al. (2008b) e Moura et al. (2008c).

6.2 Limitações

Algumas limitações podem ser observadas quanto à realização e aos resultados aqui apresentados. Uma primeira limitação refere-se à utilização de apenas seis bases de textos. Uma vez que decidiu-se trabalhar apenas com documentos científicos, para maior confiabilidade dos resultados e possibilitar a aplicação do método ZSTF, havia pouca disponibilidade de bases prontas para serem utilizadas. Além disso, a montagem de uma nova base de textos dessa natureza demanda grande esforço manual e bastante tempo para uma montagem adequada.

Outra limitação apresentada é quanto ao método de avaliação estatística aqui utilizado, uma vez que o teste estatístico de Kruskal-Wallis é extremamente conservador no que diz respeito à rejeição da hipótese nula e exige mais que sete amostras para que se possa apontar possíveis diferenças estatísticas. Entretanto, a natureza das comparações aqui executadas exigiam a utilização de métodos não-paramétricos e não-emparelhados. Nesse contexto, após uma consulta com especialistas da área de Estatística, verificou-se que o método de Kruskal-Wallis seria o mais indicado, apesar de suas limitações, devido à sua robustez e por não exigir nenhum pressuposto com relação à distribuição dos dados.

Por fim, uma vez que um dos objetivos é de colaborar com uma metodologia de extração de taxonomias de tópicos, uma limitação aqui apresentada se refere à não avaliação do impacto dos diferentes métodos no resultado final desse processo de extração de taxonomias de tópicos. A decisão por não avaliar esse impacto se deve ao grande número de variáveis no decorrer do processo que poderiam adicionar ruído às avaliações, como o tipo do algoritmo de agrupamento dos dados e o algoritmo de rotulação dos grupos obtidos.

6.3 Trabalhos Futuros

Como uma primeira proposta de trabalho futuro, pretende-se efetuar uma análise subjetiva formal da representatividade dos termos selecionados por cada método, o que pode ser um indicativo do impacto da utilização desses métodos nas taxonomias de tópicos geradas pela metodologia com a qual se colabora. Para isso, especialistas de domínio terão de ser consultados para manualmente analisar os termos eliminados por cada método e, a partir dessa análise, atribuir notas aos diferentes subconjuntos gerados.

Outra experimentação que se pretende investigar futuramente refere-se à variação dos pesos das zonas no ZSTF. Neste trabalho, considerou-se para as diferentes zonas apenas o peso que, subjetivamente, se esperava que um termo a ele pertencente possuísse. Entretanto, um processo de calibragem desses pesos e uma análise do impacto dos mesmos no processo pode ser interessante e aprimorar os resultados apresentados por essa técnica.

Um aspecto relevante a ser explorado também futuramente diz respeito à comparação de abordagens de seleção local para problemas de agrupamento de textos, o que pode trazer resultados interessantes ao contexto de geração de taxonomias de tópicos com o qual este trabalho colabora.

Por fim, pretende-se, ainda, implementar uma ferramenta que integre os diferentes métodos aqui comparados em um único ambiente, a fim de tornar mais fácil a sua utilização em conjunto ao ambiente *Discover* e à ferramenta *PreText*.

Referências Bibliográficas

Adamo, J.-M. (2001). *Data Mining for Association Rules and Sequential Patterns: Sequential and Parallel Algorithms*. Springer Verlag, New York, EUA. Citado na página 11.

Aghdam, M. H., Ghasem-Aghaee, N., e Basiri, M. E. (2009). Text feature selection using ant colony optimization. *Expert Systems with Applications*, 36(3):6843–6853. Citado na página 16.

Agrawal, R. e Srikant, R. (1994). Fast algorithms for mining association rules in large databases. Em *VLDB '94: Proceedings of the XX International Conference on Very Large Data Bases*, páginas 487–499, San Francisco, CA, EUA. Morgan Kaufmann Publishers Inc. Citado na página 8.

Arampatzis, A., van der Weide, T., Koster, C., e van Bommel, P. (2000). Linguistically-motivated information retrieval. Em *Encyclopedia of Library and Information Science*. Marcel Dekker, Inc., New York, Basel. Citado na página 12.

Archetti, F., Campanelli, P., Fersini, E., e Messina, E. (2006). A hierarchical document clustering environment based on the induced bisecting k-means. Em *FQAS '06: Proceedings of VII International Conference on Flexible Query Answering Systems*, volume 4027 of *Lecture Notes in Computer Science*, páginas 257–269. Springer. Citado na página 26.

Burges, C. J. C. (1998). A tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discovery*, 2:121–167. Citado na página 37.

Card, S., Mackinlay, J., e Schneiderman, B. (1999). Information visualization. Em Card, S., Mackinlay, J., e Schneiderman, B., editores, *Readings in Information Visualization: Using Vision to Think*, capítulo 1, páginas 1–34. Morgan Kaufmann Publishers, 1 edição. Citado na página 10.

Carvalho, V. O., Rezende, S. O., e Castro, M. (2007). An analytical evaluation of objective measures behavior for generalized association rules. Em *CIDM '07: I IEEE Symposium on Computational Intelligence and Data Mining*, páginas 43–50. IEEE. Citado na página 10.

Chakrabarti, S. (2003). *Mining the Web: Discovering knowledge from hypertext data*. Morgan Kaufmann Publishers. Citado na página 9.

Chen, J., Huang, H., Tian, S., e Qu, Y. (2009). Feature selection for text classification with naïve bayes. *Expert Systems with Applications*, 36(3):5432–5435. Citado na página 16.

Conrado, M. S. e Rezende, S. O. (2008). Avaliando a geração de termos a partir de coleções textuais. Em *WTDIA '08: Anais do IV Workshop de Teses e Dissertações em Inteligência Artificial - SBIA '08: XIX Simpósio Brasileiro de Inteligência Artificial*, páginas 1–10. São Carlos : ICMC/USP. Citado nas páginas 8, 12, e 31.

Cummins, R. e Oriordan, C. (2005). Evolving general term-weighting schemes for information retrieval: Tests on larger collections. *Artificial Intelligence Review*, 24(3-4):277–299. Citado na página 22.

Dasgupta, A., Drineas, P., Harb, B., Josifovski, V., e Mahoney, M. W. (2007). Feature selection methods for text classification. Em *KDD '07: Proceedings of the XIII ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, páginas 230–239, New York, NY, EUA. ACM. Citado na página 16.

Dash, M. e Liu, H. (1997). Feature selection for classification. *Intelligent Data Analysis*, 1(3):131–156. Citado nas páginas 18 e 37.

Dhillon, I., Kogan, J., e Nicholas, C. (2003). Feature selection and document clustering. Em Berry, M. W., editor, *Survey of Text Mining*, páginas 73–100. Springer. Citado nas páginas 3 e 26.

Ebecken, N. F. F., Lopes, M. C. S., e de Aragão Costa, M. C. (2003). Mineração de textos. Em Rezende, S. O., editor, *Sistemas Inteligentes: Fundamentos e Aplicações*, capítulo 13, páginas 337–370. Manole, 1 edição. Citado nas páginas 1, 5, 7, e 12.

Fayyad, U. M., Piatetsky-Shapiro, G., e Smyth, P. (1996). Knowledge discovery and data mining: Towards a unifying framework. Em *KDD '96: Proceedings of II International Conference on Knowledge Discovery and Data Mining*, páginas 82–88, EUA. AAAI Press. Citado na página 6.

Forman, G. (2003). An extensive empirical study of feature selection metrics for text classification. *The Journal of Machine Learning Research*, 3:1289–1305. Citado nas páginas 2, 16, e 19.

Gantz, J. F., Reinsel, D., Chute, C., Schlichting, W., McArthur, J., Minton, S., Xheneti, I., Toncheva, A., e Manfrediz, A. (2007). The expanding digital universe: A forecast of worldwide information growth through 2010. *External Publication of IDC (Analyse the Future) Information and Data*, páginas 1–21. Citado na página 1.

Gonzaga, L., Grivet, M., e Vasconcelos, A. T. (2007). A simple and fast term selection procedure for text clustering. Em *ISDA '07: Proceedings of the VII International Conference on Intelligent Systems Design and Applications*, páginas 777–781, Washington, DC, EUA. IEEE Computer Society. Citado nas páginas 16, 25, e 26.

Guyon, I. e Elisseeff, A. (2003). An introduction to variable and feature selection. *Journal of Machine Learning Research*, 3:1157–1182. Citado na página 2.

Guyon, I., Gunn, S., Nikravesh, M., e Zadeh, L. A. (2006). *Feature Extraction: Foundations and Applications (Studies in Fuzziness and Soft Computing)*. Springer, EUA. Citado na página 18.

Haykin, S. (2001). *Redes Neurais: Princípios e Práticas*. Bookman, Porto Alegre, RS, Brasil. Citado na página 17.

Imamura, C. Y.-M. (2001). Pré-processamento para extração de conhecimento de bases textuais. Dissertação de Mestrado, Instituto de Ciências Matemáticas e de Computação – USP – São Carlos. Citado nas páginas 6 e 12.

John, G. H., Kohavi, R., e Pfleger, K. (1994). Irrelevant features and the subset selection problem. Em *ICML '94: Proceedings of XI International Conference on Machine Learning*, páginas 121–129. Citado nas páginas 15 e 20.

Jolliffe, I. T. (2002). *Principal Component Analysis*. Springer, EUA. Citado na página 18.

Junling, X., Baowen, X., Weifeng, Z., Zifeng, C., e Wei, Z. (2007). A new feature selection method for text clustering. *Wuhan University Journal of Natural Sciences*, 12:912–916. Citado na página 16.

Kang, S. S. (2004). Selecting features by term importance for text categorization. Em *AIA '04: Proceedings of the IV IASTED International Conference on Artificial Intelligence and Aplications*, páginas 342–346, Innsbruck, Austria. ACTA Press. Citado na página 23.

Kantardzic, M. (2003). *Data Mining: Concepts, Models, Methods, and Algorithms*. John Wiley & Sons, Piscataway, NJ, EUA. Citado na página 8.

Kim, Y., Street, W. N., e Menczer, F. (2000). Feature selection in unsupervised learning via evolutionary search. Em *KDD '00: Proceedings of the VI ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, páginas 365–369, New York, NY, EUA. ACM. Citado na página 15.

Krovetz, R. (1993). Viewing morphology as an inference process. Em *SIGIR '93: Proceedings of the XVI Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, páginas 191–202, New York, NY, EUA. ACM. Citado na página 12.

Kruskal, W. H. e Wallis, W. A. (1952). Use of ranks in one-criterion variance analysis. *Journal of the American Statistical Association*, 47(260):583–621. Citado nas páginas 34 e 46.

Lamping, J., Rao, R., e Pirolli, P. (1995). A focus+context technique based on hyperbolic geometry for visualizing large hierarchies. Em *CHI '95: Proceedings of the XIII SIGCHI*

Conference on Human Factors in Computing Systems, páginas 401–408, New York, NY, EUA. ACM Press/Addison-Wesley Publishing Co. Citado nas páginas 10 e 13.

Landauer, T. K., Foltz, P. W., e Laham, D. (1998). An introduction to latent semantic analysis. *Discourse Processes*, 25(2-3):259–284. Citado na página 18.

Langley, P. (1994). Selection of relevant features in machine learning. Em *Proceedings of AAAI Fall Symposium on Relevance*, páginas 140–144. AAAI Press. Citado na página 19.

Lawrie, D., Croft, W. B., e Rosenberg, A. (2001). Finding topic words for hierarchical summarization. Em *SIGIR '01: Proceedings of the XXIV Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, páginas 349–357, New York, NY, EUA. ACM. Citado na página 35.

Liu, H. e Yu, L. (2005). Toward integrating feature selection algorithms for classification and clustering. *IEEE Transactions on Knowledge and Data Engineering*, 17(4):491–502. Citado nas páginas 19 e 20.

Liu, L., Kang, J., Yu, J., e Wang, Z. (2005). A comparative study on unsupervised feature selection methods for text clustering. Em *NLP-KE '05: Proceedings of 2005 IEEE International Conference on Natural Language Processing and Knowledge Engineering*, páginas 597–601. Citado nas páginas 3, 16, 18, 25, e 26.

Liu, T., Liu, S., e Chen, Z. (2003). An evaluation on feature selection for text clustering. Em *ICML '03: Proceedings of the XX International Conference on Machine Learning*, páginas 488–495, San Francisco, CA. Morgan Kaufmann. Citado nas páginas 3 e 25.

Losee, R. M. (2001). Term dependence: A basis for luhn and zipf models. *Journal of the American Society for Information Science and Technology*, 52(12):1019–1025. Citado na página 22.

Luhn, H. P. (1958). The automatic creation of literature abstracts. *IBM Journal os Research and Development*, 2(2):159–165. Citado nas páginas 3 e 22.

Manning, C. D., Raghavan, P., e Schtze, H. (2008). *Introduction to Information Retrieval*. Cambridge University Press, New York, NY, EUA. Citado na página 26.

Manning, C. D. e Schtze, H. (1999). *Foundations of Statistical Natural Language Processing*. MIT Press, Cambridge, MA, EUA. Citado na página 8.

Marcacini, R. M. (2008). Um ambiente interativo para análise visual de agrupamentos hierárquicos. Monografia conclusão de curso de graduação, Instituto de Ciências Matemáticas e de Computação – USP – São Carlos. Citado na página 10.

Mardia, K. V., Kent, J. T., e Bibby, J. M. (1979). *Multivariate Analysis*. Academic Press. Citado na página 13.

Martins, C. A. (2003). Uma abordagem para pré-processamento de dados textuais em algoritmos de aprendizado. Tese de Doutorado em Ciências da Computação e Matemática Computacional, Instituto de Ciências Matemáticas e de Computação – USP – São Carlos. Citado nas páginas 16, 17, e 22.

Matsubara, E. T., Martins, C. A., e Monard, M. C. (2003). Pre-text: uma ferramenta para pré-processamento de textos utilizando a abordagem bag-of-words. Relatório Técnico 209, Instituto de Ciências Matemáticas e de Computação – USP – São Carlos. Citado nas páginas 22 e 32.

Messina, E., Toscani, D., e Archetti, F. (2006). Up-dres user profiling for a dynamic recommendation system. Em Perner, P., editor, *ICDM '06: Proceedings of the VI Industrial Conference on Data Mining*, volume 4065 of *Lecture Notes in Computer Science*, páginas 146–160. Springer. Citado na página 26.

Mitchell, T. M. (1997). *Machine Learning*. McGraw-Hill Higher Education. Citado nas páginas 8, 36, e 37.

Mitra, P., Murthy, C. A., e Pal, S. K. (2002). Unsupervised feature selection using feature similarity. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(3):301–312. Citado na página 19.

Mladenic, D. (1998). Machine learning on non-homogeneous, distributed text data. PhD Thesis, University of Ljubljana – Faculty of Computer and Information Science – Eslovênia. Citado na página 12.

Modha, D. S. e Spangler, W. S. (2003). Feature weighting in k-means clustering. *Machine Learning*, 52(3):217–237. Citado na página 23.

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

Moura, M. F. (2006). Uma abordagem para a construção e atualização de taxonomias de tópicos a partir de coleções de textos dinâmicas. Monografia de Qualificação de Doutorado, Instituto de Ciências Matemáticas e de Computação – USP – São Carlos. Citado nas páginas 7 e 10.

Moura, M. F., Marcacini, R. M., Nogueira, B. M., da Silva Conrado, M., e Rezende, S. O. (2008a). A proposal for building domain topic taxonomies. Em *WTI '08: Proceedings of I Workshop on Web and Text Intelligence - SBIA '08: XIX Simpósio Brasileiro de Inteligência Artificial*, páginas 83–84. São Carlos: ICMC/USP. Citado nas páginas 2, 5, 30, 31, 39, e 59.

Moura, M. F., Marcacini, R. M., Nogueira, B. M., da Silva Conrado, M., e Rezende, S. O. (2008b). Uma abordagem completa para a construção de taxonomias de tópicos em um domínio. Relatório técnico 329, ICMC-USP, São Carlos - SP. Citado na página 59.

Moura, M. F., Nogueira, B. M., da Silva Conrado, M., dos Santos, F. F., e Rezende, S. O. (2008c). Making good choices of non-redundant n-gramwords. Em Library, I. X. D., editor, *DMAI '08: Proceedings of I International Workshop on Data Mining and Artificial Intelligence - ICCIT '08: XI IEEE International Conference on Computer and Information Technology*, páginas 64–71. Citado nas páginas 32 e 59.

Moura, M. F. e Rezende, S. O. (2007). Choosing a hierarchical cluster labelling method for a specific domain document collection. Em Neves, J., Santos, M. F., e Machado, J. M., editores, *New Trends in Artificial Intelligence*, capítulo 11, páginas 812–823. Lisboa, Portugal: APPIA - Associação Portuguesa para Inteligência Artificial. EPIA-Encontro Português de Inteligência Artificial, 2007, Guimarães, Portugal, 1 edição. Citado na página 13.

Nogueira, B. M., Moura, M. F., Conrado, M. S., e Rezende, S. O. (2008a). Avaliação de métodos não-supervisionados de seleção de atributos para mineração de textos. Em *WTI '08: Proceedings of I Workshop on Web and Text Intelligence - SBIA '08: XIX Simpósio Brasileiro de Inteligência Artificial*, páginas 59–66. São Carlos: ICMC/USP. Citado na página 59.

Nogueira, B. M., Moura, M. F., Conrado, M. S., Rossi, R. G., Marcacini, R. M., e Rezende, S. O. (2008b). Winning some of the document preprocessing challenges in a text mining process. Em *WAAMD '08: Anais do IV Workshop em Algoritmos e Aplicações de Mineração de Dados - SBBD '08: XXIII Simpósio Brasileiro de Banco de Dados*, páginas 10–18. Porto Alegre : SBC. Citado nas páginas 39 e 59.

Peltonen, J., Sinkhonen, J., e Kaski, S. (2002). Discriminative clustering of text documents. Em *ICONIP '02: Proceedings of IEEE IX International Conference on Neural Information Processing*, volume 4, páginas 1956–1960. Citado na página 9.

Porter, M. F. (1997). An algorithm for suffix stripping. *Readings in Information Retrieval*, páginas 313–316. Citado na página 32.

Prati, R. C., Geromini, M. R., e Monard, M. C. (2003). An integrated environment for data mining. Em *LAPTEC '03: Proceedings of IV Congress of Logic Applied to Technology*, volume 2, páginas 55–62, Marília - SP. Pléiade. Citado na página 38.

Price, L. e Thelwall, M. (2005). The clustering power of low frequency words in academic webs. *Journal of the American Society for Information Science and Technology*, 56(8):883–888. Citado na página 23.

Quinlan, J. R. (1993). *C4.5: programs for machine learning*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA. Citado na página 36.

Rezende, S. O., Pugliesi, J. B., Melanda, E. A., e Paula, M. F. (2003). Mineração de dados. Em Rezende, S. O., editor, *Sistemas Inteligentes: Fundamentos e Aplicações*, capítulo 12, páginas 307–335. Manole, 1 edição. Citado nas páginas 1 e 6.

Rino, L. H. M. e Pardo, T. A. S. (2003). A sumarização automática de textos: Principais características e metodologias. Em Vieira, R., editor, *Jornada de Atualização em Inteligência Artificial*, páginas 203–245. SBC, viii edição. Citado na página 9.

Saleh, S. N. e El-Sonbaty, Y. (2007). A feature selection algorithm with redundancy reduction for text classification. Em *ISCIS '07: XXII International International Symposium on Computer and Information Sciences*, páginas 1–6. Citado na página 16.

Salton, G. e Buckley, C. (1987). Term weighting approaches in automatic text retrieval. Relatório técnico, Ithaca, NY, EUA. Citado nas páginas 3, 16, 17, e 24.

Salton, G. e McGill, M. J. (1986). *Introduction to Modern Information Retrieval*. McGraw-Hill, Inc., New York, NY, EUA. Citado na página 25.

Salton, G., Yang, C. S., e Yu, C. T. (1975). A theory of term importance in automatic text analysis. *Journal of the American Association Science*, 1(26):33–44. Citado nas páginas 3, 17, e 23.

Shafiei, M., Wang, S., Zhang, R., Milios, E. E., Tang, B., Tougas, J., e Spiteri, R. J. (2007). Document representation and dimension reduction for text clustering. Em *TDMM '07: Workshop on Text Data Mining and Management - XXIII IEEE International Conference on Data Engineering*, páginas 770–779, Istanbul, Turkey. Citado na página 16.

Silberschatz, A. e Tuzhilin, A. (1995). On subjective measures of interestingness in knowledge discovery. Em *KDD '95: Proceedings of I International Conference on Knowledge Discovery and Data Mining*, páginas 275–281. Citado na página 10.

Slonim, N. e Tishby, N. (2000). Document clustering using word clusters via the information bottleneck method. Em *SIGIR '00: Proceedings of the XXIII Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, páginas 208–215, New York, NY, EUA. ACM. Citado na página 18.

Soucy, P. e Mineau, G. W. (2005). Beyond tfidf weighting for text categorization in the vector space model. Em *IJCAI '05: Proceedings of the XIX International Joint Conferences on Artificial Inteligence*, páginas 1130–1135. Citado nas páginas 17 e 24.

Souza, K. X. S., Davis, J., Evangelista, S. R. M., Souza, M. I. F., Santos, A. D., e Moura, M. F. (2005). The evolution of knowledge representation within Embrapa´s information agency. Em *EFITA '05: Proceedings of V Conference of the European Federation for Information Technology in Agriculture, Food and Environment - WCCA '05: III World Congress on Computers in Agriculture and Natural Recources*, páginas 464–470, Vila Real, Portugal. Universidade Trás-os-Montes e Alto Douro. Citado na página 13.

van Rijsbergen, C. J. (1979). *Information Retrieval*. Butterworth. Citado nas páginas 17 e 35.

Vapnik, V. N. (1998). *Statistical Learning Theory*. Wiley-Interscience. Citado na página 37.

Weiss, S. M., Indurkhya, N., Zhang, T., e Damerau, F. J. (2005). *Text Mining - Predictive Methods for Analizing Unstructured Information*. Springer Science+Business Media, Inc. Citado na página 13.

Wiratunga, N., Lothian, R., e Massie, S. (2006). Unsupervised feature selection for text data. Em *ECCBR '06: Proceedings of VIII European Conference on Case-Based Reasoning*, páginas 340–354, Heidelberg. Springer Berlin. Citado nas páginas 16, 25, e 37.

Witten, I. H. e Frank, E. (2005). *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann Series in Data Management Systems. Morgan Kaufmann, second edição. Citado nas páginas 38 e 48.

Wyse, N., Dubes, R., e Jain, A. (1980). A critical evaluation of intrinsic dimensionality algorithms. Em Gelsema, E. e Kanal, L., editores, *Pattern Recognition in Practice*, páginas 415–425. North-Holland. Citado na página 18.

Xu, Z., Akella, R., Ching, M., e Tang, R. (2007). Semi-supervised clustering using bayesian regularization. Em *ICDMW '07: Proceedings of the VII IEEE International Conference on Data Mining Workshops*, páginas 361–366, Washington, DC, EUA. IEEE Computer Society. Citado na página 26.

Yang, Y. e Pedersen, J. O. (1997). A comparative study on feature selection in text categorization. Em *ICML '97: Proceedings of XIV International Conference on Machine Learning*, páginas 412–420, Nashville, US. Morgan Kaufmann Publishers, San Francisco, US. Citado nas páginas 16, 23, e 33.

Yao, Y. Y. (2003). Information-theoretic measures for knowledge discovery and data mining. Em *Entropy Measures, Maximum Entropy Principle and Emerging Applications*, páginas 115–136. Springer, Berlim, Alemanha. Citado na página 17.

Yu, L. e Liu, H. (2004). Efficient feature selection via analysis of relevance and redundancy. *Journal of Machine Learning Research*, 5:1205–1224. Citado na página 15.

Zipf, G. K. (1949). *Human Behavior and the Principle of Least Effort*. Addison-Wesley. Citado na página 22.

Apêndice

A

Tabelas com os Resultados da Avaliação Não-Supervisionada

Neste apêndice podem ser observadas as tabelas com os valores de EMIM obtidos pelos diferentes métodos de seleção de atributos nas diferentes bases de textos utilizadas. A Tabela A.1 contém os valores obtidos pela variação de porcentagem de atributos selecionados para as bases *CIIS*, *IA* e *IFM*. Já na Tabela A.2, são apresentados os valores com variação de porcentagem de atributos para as bases *Chemistry*, *CS* e *Physics*. Todas os valores de EMIM contidos nessas tabelas foram utilizadas para as análises estatísticas e para a construção dos gráficos apresentados na Seção 5.3.1 deste trabalho.

	%	TC	TV	TVQ	RTF	RDF	TFIDF
CIIS	5	0,29327	0,30108	0,29662	0,28227	0,26259	0,29349
	10	0,43677	0,44806	0,43973	0,43315	0,40659	0,43998
	20	0,62882	0,63517	0,62373	0,62878	0,60574	0,6328
	30	0,74145	0,75244	0,73707	0,74881	0,72831	0,75212
	40	0,81722	0,83513	0,8323	0,83319	0,815	0,83448
	50	0,86464	0,89302	0,89122	0,89122	0,86464	0,8934
	60	0,93743	0,93873	0,93806	0,93806	0,93743	0,93834
	70	0,96615	0,96863	0,96615	0,96615	0,96615	0,96615
	80	0,99229	0,99691	0,99229	0,99229	0,99229	0,99229
	90	1,0191	1,0221	1,0191	1,0191	1,0191	1,0191
IA	5	0,25583	0,26804	0,26249	0,25968	0,22608	0,26546
	10	0,3015	0,3165	0,30435	0,31209	0,2791	0,31494
	20	0,33794	0,35965	0,33719	0,35807	0,33457	0,35904
	30	0,36557	0,3819	0,3813	0,38136	0,36557	0,382
	40	0,39546	0,39711	0,39613	0,39613	0,39546	0,39706
	50	0,40673	0,40687	0,40673	0,40673	0,40673	0,40673
	60	0,41442	0,4155	0,41442	0,41442	0,41442	0,41442
	70	0,42211	0,42332	0,42211	0,42211	0,42211	0,42211
	80	0,42979	0,42983	0,42979	0,42979	0,42979	0,42979
	90	0,43747	0,43704	0,43747	0,43747	0,43747	0,43747
IFM	5	0,1809	0,18742	0,18535	0,17634	0,14564	0,1845
	10	0,22611	0,23209	0,22875	0,22631	0,19996	0,22993
	20	0,26547	0,27148	0,26533	0,26903	0,24802	0,27062
	30	0,28501	0,2915	0,28246	0,29053	0,27325	0,29109
	40	0,2966	0,30383	0,30296	0,30349	0,29392	0,3039
	50	0,30352	0,31195	0,31243	0,31242	0,3036	0,31241
	60	0,31799	0,31828	0,31875	0,31875	0,31799	0,31881
	70	0,3241	0,32412	0,3241	0,3241	0,3241	0,32402
	80	0,3275	0,32724	0,3275	0,3275	0,3275	0,32743
	90	0,33088	0,33101	0,33088	0,33088	0,33088	0,33081

Tabela A.1: Valores de EMIM obtidos na variação de porcentagem de atributos selecionados para as bases CIIS, IA e IFM

	%	TC	TV	TVQ	RTF	RDF	TFIDF	ZSTF
Chemistry	5	0,32037	0,33887	0,31902	0,30975	0,22871	0,33551	0,30976
	10	0,41217	0,43983	0,40669	0,42351	0,33341	0,43705	0,42351
	20	0,49839	0,53421	0,48134	0,52991	0,43764	0,53357	0,52372
	30	0,53736	0,58138	0,53363	0,5796	0,49658	0,58082	0,56853
	40	0,55619	0,60935	0,60783	0,60917	0,5563	0,61003	0,59761
	50	0,60877	0,62745	0,62838	0,62838	0,60877	0,6284	0,61741
	60	0,64169	0,64253	0,64215	0,64215	0,64169	0,64234	0,62891
	70	0,65291	0,65313	0,65291	0,65291	0,65291	0,6528	0,64578
	80	0,66022	0,66023	0,66022	0,66022	0,66022	0,66011	0,65514
	90	0,66753	0,66699	0,66753	0,66753	0,66753	0,66742	0,66502
CS	5	0,28864	0,30187	0,29347	0,27699	0,22628	0,29811	0,27826
	10	0,37198	0,38823	0,37148	0,37683	0,32197	0,38573	0,37737
	20	0,44148	0,46069	0,42781	0,45783	0,41029	0,46023	0,45611
	30	0,46824	0,49526	0,48655	0,49425	0,46429	0,49522	0,49177
	40	0,48126	0,51546	0,51481	0,51512	0,48126	0,51525	0,51194
	50	0,52252	0,52765	0,52798	0,52798	0,52252	0,52823	0,52404
	60	0,53835	0,53852	0,53852	0,53852	0,53835	0,53872	0,53596
	70	0,54584	0,54597	0,54584	0,54584	0,54584	0,54584	0,54325
	80	0,55137	0,55197	0,55137	0,55137	0,55137	0,55137	0,54954
	90	0,55691	0,55727	0,55691	0,55691	0,55691	0,55691	0,55608
Physics	5	0,32472	0,336	0,32691	0,30789	0,23323	0,33443	0,30648
	10	0,42066	0,4375	0,42059	0,42236	0,34529	0,43412	0,42083
	20	0,51044	0,5288	0,49693	0,52465	0,46095	0,52819	0,51949
	30	0,54967	0,57261	0,52921	0,57138	0,51966	0,57269	0,56594
	40	0,57099	0,59877	0,59568	0,59822	0,56953	0,59896	0,59216
	50	0,58233	0,61611	0,61592	0,61598	0,58233	0,61647	0,60806
	60	0,6252	0,62803	0,62786	0,62786	0,6252	0,62825	0,62176
	70	0,63847	0,63847	0,63853	0,63853	0,63847	0,63861	0,63332
	80	0,64472	0,64443	0,64472	0,64472	0,64472	0,64472	0,64066
	90	0,65029	0,64983	0,65029	0,65029	0,65029	0,65029	0,64838

Tabela A.2: Valores de EMIM obtidos na variação de porcentagem de atributos selecionados para as bases Chemistry, CS e Physics

Apêndice

B

Tabelas com os Resultados da Avaliação Supervisionada

Este apêndice contém as tabelas com os valores de acurácia obtidos nas avaliações supervisionadas realizadas neste trabalho, bem como os respectivos desvios padrões. Os valores de acurácia preditiva contidos nas tabelas aqui apresentados foram utilizados como entrada para as análises estatísticas apresentadas na Seção 5.3.2 deste trabalho.

As Tabelas B.1, B.2, B.3, B.4, B.5 e B.6 contêm a compilação dos valores de acurácia e desvio padrão obtidos, pelos métodos que não fornecem ponto exato de corte, na variação de porcentagem de atributos selecionados. Para os métodos que sugerem pontos exatos de corte, os valores de acurácia são exibidos na Tabela B.7.

Por fim, nas Tabelas B.8, B.9 e B.10 são exibidos os resultados obtidos pelos métodos sem ponto exato de corte quando gerando subconjuntos com cardinalidades de atributos idênticas às geradas pelos métodos Luhn, LuhnDF e Salton, respectivamente. Nessas três tabelas, como o método ZSTF só pôde ser aplicado às bases *Chemistry*, *CS* e *Physics*, os valores para esse método são nulos para as bases *CIIS*, *IA* e *IFM*.

J48								
%	**RDF**	**RTF**	**TFIDF**	**TC**	**TV**	**TVQ**	**IG**	**GR**
5	85,50±5,74	86,38±3,86	86,08±4,01	86,52±4,15	86,67±4,01	86,23±4,46	88,15±2,86	86,37±4,26
10	84,60±3,49	84,16±3,60	84,01±3,23	84,01±3,23	84,60±2,93	84,60±2,93	88,15±2,86	88,15±2,86
20	83,86±4,25	83,86±4,25	84,15±4,35	83,86±4,25	84,31±3,99	84,01±4,66	88,15±2,86	88,15±2,86
30	83,71±4,17	83,71±4,17	83,71±4,17	83,71±4,17	83,71±4,17	83,71±3,72	88,15±2,86	88,15±2,86
40	83,56±4,14	83,56±4,14	83,56±4,14	83,56±4,14	83,71±4,17	83,56±4,14	88,15±2,86	88,15±2,86
50	83,56±4,14	83,56±4,14	83,56±4,14	83,56±4,14	83,56±4,14	83,56±4,14	88,15±2,86	88,15±2,86
60	83,56±4,14	83,56±4,14	83,56±4,14	83,56±4,14	83,56±4,14	83,56±4,14	87,11±2,81	87,11±2,81
70	83,56±4,14	83,56±4,14	83,56±4,14	83,56±4,14	83,56±4,14	83,56±4,14	84,60±3,85	84,60±3,85
80	83,56±4,14	83,56±4,14	83,56±4,14	83,56±4,14	83,56±4,14	83,56±4,14	83,71±4,17	83,71±4,17
90	83,56±4,14	83,56±4,14	83,56±4,14	83,56±4,14	83,56±4,14	83,56±4,14	83,71±4,17	83,71±4,17
100	83,56±4,14	83,56±4,14	83,56±4,14	83,56±4,14	83,56±4,14	83,56±4,14	83,56±4,14	83,56±4,14
SVM								
%	**RDF**	**RTF**	**TFIDF**	**TC**	**TV**	**TVQ**	**IG**	**GR**
5	86,22±3,42	86,67±3,99	88,60±4,22	88,45±4,09	88,45±2,93	88,15±2,29	90,37±3,58	88,59±5,26
10	86,37±3,16	86,08±4,66	86,81±2,92	86,67±3,74	88,44±4,61	89,33±4,17	90,22±3,30	90,07±3,43
20	87,26±3,37	86,08±4,52	86,82±4,21	86,97±4,59	86,51±4,00	88,44±3,70	89,93±3,94	89,93±3,94
30	86,96±3,87	87,55±3,65	86,66±2,98	87,10±3,65	86,82±2,73	86,96±3,70	89,33±5,37	89,33±5,37
40	86,95±3,44	87,25±3,47	87,10±3,72	87,40±2,76	88,14±3,37	88,59±3,51	90,52±4,76	90,52±4,76
50	87,25±3,17	87,39±3,69	87,25±3,40	87,25±3,17	87,99±3,82	87,39±3,69	90,22±3,78	90,22±3,78
60	86,51±3,34	86,81±3,56	86,96±2,89	86,51±3,34	86,66±3,53	86,81±3,56	88,58±5,01	88,58±5,01
70	87,25±4,29	87,25±4,29	87,25±4,29	87,25±4,29	86,51±4,26	87,25±4,29	88,44±4,42	88,44±4,42
80	87,10±4,34	87,10±4,34	87,10±4,34	87,10±4,34	86,95±4,40	87,10±4,34	87,69±4,17	87,69±4,17
90	87,25±4,63	87,25±4,63	87,25±4,63	87,25±4,63	87,10±4,35	87,25±4,63	87,99±4,07	87,99±4,07
100	87,69±4,83	87,69±4,83	87,69±4,83	87,69±4,83	87,69±4,83	87,69±4,83	87,69±4,83	87,69±4,83
KNN								
%	**RDF**	**RTF**	**TFIDF**	**TC**	**TV**	**TVQ**	**IG**	**GR**
5	57,33±6,83	62,81±7,67	65,19±7,86	64,45±7,56	69,50±6,17	74,22±5,55	76,44±4,44	77,20±6,93
10	38,82±7,74	41,49±6,73	45,05±7,76	42,82±6,70	44,61±6,51	55,13±8,81	70,37±4,09	71,41±3,17
20	43,73±9,16	43,58±8,76	43,87±8,76	47,88±9,95	36,30±6,58	34,68±8,43	61,33±4,19	61,33±4,19
30	38,52±9,83	41,06±9,65	41,21±9,67	41,21±9,41	41,95±9,35	42,26±8,93	63,56±3,97	63,56±3,97
40	31,40±5,03	30,81±4,22	30,81±4,22	31,26±5,20	36,90±6,64	30,82±2,95	60,29±3,98	60,29±3,98
50	32,59±7,46	29,92±4,46	29,92±4,46	32,59±7,46	29,63±3,67	29,92±4,46	57,33±5,98	57,33±5,98
60	33,03±8,37	33,03±8,37	33,03±8,37	33,03±8,37	33,03±8,37	33,03±8,37	51,84±6,28	51,84±6,28
70	29,04±3,33	29,04±3,33	29,04±3,33	29,04±3,33	33,03±8,39	29,04±3,33	32,89±10,19	32,89±10,19
80	27,12±1,32	27,12±1,32	27,12±1,32	27,12±1,32	41,36±13,60	27,12±1,32	29,77±7,02	29,77±7,02
90	26,52±0,45	26,52±0,45	26,52±0,45	26,52±0,45	40,17±9,42	26,52±0,45	29,62±6,93	29,62±6,93
100	27,99±3,64	27,99±3,64	27,99±3,64	27,99±3,64	27,99±3,64	27,99±3,64	27,99±3,64	27,99±3,64
Naïve Bayes								
%	**RDF**	**RTF**	**TFIDF**	**TC**	**TV**	**TVQ**	**IG**	**GR**
5	81,77±2,27	83,40±3,08	83,39±3,32	83,40±2,92	83,25±1,78	83,10±2,60	88,73±3,54	85,63±3,47
10	82,81±2,40	83,54±2,73	82,66±2,87	82,96±2,39	82,07±3,14	82,64±4,29	89,63±3,12	89,63±3,12
20	84,74±2,21	84,15±2,33	84,15±2,33	84,15±2,33	82,67±2,39	84,30±2,24	89,63±3,12	89,63±3,12
30	84,74±2,33	84,89±2,42	84,74±2,33	84,89±2,42	84,89±2,42	84,59±2,45	89,63±3,12	89,63±3,12
40	84,59±2,45	84,59±2,45	84,59±2,45	84,59±2,45	84,59±2,45	84,59±2,45	89,63±3,12	89,63±3,12
50	84,59±2,45	84,59±2,45	84,59±2,45	84,59±2,45	84,59±2,45	84,59±2,45	89,63±3,12	89,63±3,12
60	84,59±2,45	84,59±2,45	84,59±2,45	84,59±2,45	84,59±2,45	84,59±2,45	87,55±4,01	87,55±4,01
70	84,59±2,45	84,59±2,45	84,59±2,45	84,59±2,45	84,59±2,45	84,59±2,45	85,92±3,07	85,92±3,07
80	84,59±2,45	84,59±2,45	84,59±2,45	84,59±2,45	84,59±2,45	84,59±2,45	84,74±2,33	84,74±2,33
90	84,59±2,45	84,59±2,45	84,59±2,45	84,59±2,45	84,59±2,45	84,59±2,45	84,74±2,33	84,74±2,33
100	84,59±2,45	84,59±2,45	84,59±2,45	84,59±2,45	84,59±2,45	84,59±2,45	84,59±2,45	84,59±2,45

Tabela B.1: Acurácias preditivas e desvios padrões obtidos com variação de porcentagem de atributos na base de textos CIIS

J48								
%	**RDF**	**RTF**	**TFIDF**	**TC**	**TV**	**TVQ**	**IG**	**GR**
5	91.20±3.16	92.20±3.05	92.60±2.99	91.80±3.58	92.00±2.98	92.40±3.50	91.80±2.74	91.80±2.74
10	91.80±3.05	91.80±3.82	92.20±3.58	92.20±3.71	92.60±3.41	92.60±3.41	91.80±2.74	91.80±2.74
20	91.60±3.10	91.80±2.90	92.20±2.90	91.60±3.10	92.40±3.37	92.40±2.95	91.80±2.74	91.80±2.74
30	91.80±3.19	91.60±3.10	91.60±3.10	91.80±3.19	92.20±2.90	91.60±3.10	91.80±2.74	91.80±2.74
40	91.80±3.19	91.80±3.19	91.60±3.10	91.80±3.19	91.80±3.19	91.80±3.19	91.80±2.74	91.80±2.74
50	91.80±3.19	91.80±3.19	91.80±3.19	91.80±3.19	91.80±3.19	91.80±3.19	91.80±2.74	91.80±2.74
60	91.80±3.19	91.80±3.19	91.80±3.19	91.80±3.19	91.80±3.19	91.80±3.19	91.40±2.99	91.40±2.99
70	91.80±3.19	91.80±3.19	91.80±3.19	91.80±3.19	91.80±3.19	91.80±3.19	91.60±2.95	91.60±2.95
80	91.80±3.19	91.80±3.19	91.80±3.19	91.80±3.19	91.80±3.19	91.80±3.19	91.80±3.19	91.80±3.19
90	91.80±3.19	91.80±3.19	91.80±3.19	91.80±3.19	91.80±3.19	91.80±3.19	91.80±3.19	91.80±3.19
100	91.80±3.19	91.80±3.19	91.80±3.19	91.80±3.19	91.80±3.19	91.80±3.19	91.80±3.19	91.80±3.19
SVM								
%	**RDF**	**RTF**	**TFIDF**	**TC**	**TV**	**TVQ**	**IG**	**GR**
5	83.00±6.94	82.20±6.14	80.60±5.89	81.80±7.45	80.20±7.74	79.40±8.80	90.80±5.43	90.80±5.35
10	85.20±7.67	82.80±5.67	82.20±7.97	83.80±6.36	82.40±6.38	82.20±5.92	90.40±5.95	90.40±5.95
20	85.00±8.50	83.60±7.35	83.40±7.78	83.40±9.38	83.00±6.88	84.80±6.55	90.20±6.21	90.20±6.21
30	86.00±7.12	85.80±7.15	84.80±7.32	86.00±7.12	84.20±7.21	85.80±7.15	92.20±5.45	92.20±5.45
40	85.60±7.93	85.80±7.63	85.20±7.90	85.60±7.93	85.20±7.44	85.80±7.63	90.60±6.33	90.60±6.33
50	85.00±7.13	85.00±7.13	85.00±7.13	85.00±7.13	83.20±7.79	85.00±7.13	91.60±6.17	91.60±6.17
60	85.20±7.25	85.20±7.25	85.20±7.25	85.20±7.25	85.60±5.32	85.20±7.25	89.00±7.32	89.00±7.32
70	84.80±7.61	84.80±7.61	84.80±7.61	84.80±7.61	84.80±6.05	84.80±7.61	84.60±8.11	84.60±8.11
80	84.40±7.76	84.40±7.76	84.40±7.76	84.40±7.76	84.20±7.08	84.40±7.76	84.20±7.74	84.20±7.74
90	84.00±8.22	84.00±8.22	84.00±8.22	84.00±8.22	85.20±7.44	84.00±8.22	84.20±7.74	84.20±7.74
100	83.80±8.02	83.80±8.02	83.80±8.02	83.80±8.02	83.80±8.02	83.80±8.02	83.80±8.02	83.80±8.02
KNN								
%	**RDF**	**RTF**	**TFIDF**	**TC**	**TV**	**TVQ**	**IG**	**GR**
5	42.20±7.33	51.20±7.67	45.80±6.83	51.20±9.94	49.20±8.23	46.20±7.86	46.40±8.37	44.80±6.34
10	30.60±5.82	39.40±7.31	34.20±6.43	34.20±8.08	32.80±7.19	30.80±4.54	30.80±4.83	30.80±4.83
20	29.60±5.06	30.00±7.89	29.80±6.43	31.60±7.29	36.00±5.66	31.20±6.05	34.40±6.72	34.40±6.72
30	26.20±5.37	30.60±6.47	30.40±6.02	26.20±5.37	28.40±5.06	28.80±5.59	25.20±3.01	25.20±3.01
40	26.00±4.42	26.20±3.71	26.80±4.54	26.00±4.42	26.20±3.82	26.20±3.71	23.60±3.37	23.60±3.37
50	23.60±4.79	23.60±4.79	23.60±4.79	23.60±4.79	25.60±4.20	23.60±4.79	25.80±6.14	25.80±6.14
60	24.20±3.58	24.20±3.58	24.20±3.58	24.20±3.58	27.00±6.75	24.20±3.58	22.00±1.33	22.00±1.33
70	23.00±3.80	23.00±3.80	23.00±3.80	23.00±3.80	20.60±0.97	23.00±3.80	24.20±3.19	24.20±3.19
80	23.80±4.47	23.80±4.47	23.80±4.47	23.80±4.47	20.60±0.97	23.80±4.47	23.60±3.98	23.60±3.98
90	23.00±4.35	23.00±4.35	23.00±4.35	23.00±4.35	23.20±2.35	23.00±4.35	25.80±4.76	25.80±4.76
100	24.20±4.57	24.20±4.57	24.20±4.57	24.20±4.57	24.20±4.57	24.20±4.57	24.20±4.57	24.20±4.57
Naïve Bayes								
%	**RDF**	**RTF**	**TFIDF**	**TC**	**TV**	**TVQ**	**IG**	**GR**
5	83.20±5.59	84.60±7.18	83.20±7.79	85.40±8.00	84.40±8.10	84.20±7.63	90.40±5.40	90.40±5.40
10	83.80±6.14	84.40±8.26	85.20±7.50	84.20±7.27	84.60±8.38	86.40±8.47	90.40±5.40	90.40±5.40
20	83.20±6.27	83.80±5.77	83.60±6.17	82.80±5.98	84.20±7.08	83.20±6.20	90.40±5.40	90.40±5.40
30	83.20±5.75	83.40±6.47	83.40±6.74	83.20±5.75	83.00±6.13	83.40±6.47	90.40±5.40	90.40±5.40
40	82.60±6.04	83.20±5.51	83.60±6.31	82.60±6.04	83.40±6.19	83.20±5.51	90.40±5.40	90.40±5.40
50	82.80±5.83	82.80±5.83	82.80±5.83	82.80±5.83	82.80±5.83	82.80±5.83	90.40±5.40	90.40±5.40
60	83.60±5.80	83.60±5.80	83.60±5.80	83.60±5.80	83.60±5.80	83.60±5.80	89.80±5.20	89.80±5.20
70	83.60±6.24	83.60±6.24	83.60±6.24	83.60±6.24	83.60±6.24	83.60±6.24	86.00±6.32	86.00±6.32
80	83.20±6.27	83.20±6.27	83.20±6.27	83.20±6.27	83.20±6.27	83.20±6.27	82.60±6.60	82.60±6.60
90	83.20±6.34	83.20±6.34	83.20±6.34	83.20±6.34	83.20±6.34	83.20±6.34	83.20±5.59	83.20±5.59
100	83.00±6.34	83.00±6.34	83.00±6.34	83.00±6.34	83.00±6.34	83.00±6.34	83.00±6.34	83.00±6.34

Tabela B.2: Acurácias preditivas e desvios padrões obtidos com variação de porcentagem de atributos na base de textos IA

J48								
%	RDF	RTF	TFIDF	TC	TV	TVQ	IG	GR
5	79.88±8.52	81.44±3.52	83.50±4.83	83.67±4.95	83.15±5.86	79.19±6.55	85.04±5.24	86.42±4.99
10	81.43±5.60	83.67±4.43	82.63±5.63	83.49±3.88	82.12±5.24	82.47±4.89	84.53±4.45	85.56±3.75
20	86.25±3.83	86.42±3.59	84.87±4.23	86.25±3.65	84.87±4.38	84.87±4.38	84.36±4.63	84.53±4.45
30	86.59±3.54	86.42±3.69	85.90±3.90	86.42±3.69	85.91±4.14	86.08±3.51	84.53±4.74	84.53±4.74
40	86.42±3.41	86.08±3.60	86.42±3.41	86.42±3.41	86.25±3.36	86.25±3.65	84.53±4.74	84.53±4.74
50	86.42±3.41	86.42±3.41	86.42±3.41	86.42±3.41	86.42±3.41	86.42±3.41	84.53±4.74	84.53±4.74
60	86.25±3.65	86.25±3.65	86.42±3.41	86.25±3.65	86.42±3.41	86.25±3.65	85.22±3.58	85.22±3.58
70	86.25±3.65	86.25±3.65	86.25±3.65	86.25±3.65	86.25±3.65	86.25±3.65	85.90±3.82	85.90±3.82
80	86.25±3.65	86.25±3.65	86.25±3.65	86.25±3.65	86.25±3.65	86.25±3.65	86.25±4.00	86.25±4.00
90	86.25±3.65	86.25±3.65	86.25±3.65	86.25±3.65	86.25±3.65	86.25±3.65	86.42±3.78	86.42±3.78
100	86.25±3.65	86.25±3.65	86.25±3.65	86.25±3.65	86.25±3.65	86.25±3.65	86.25±3.65	86.25±3.65
SVM								
%	RDF	RTF	TFIDF	TC	TV	TVQ	IG	GR
5	76.45±5.24	75.94±5.81	75.42±5.72	76.79±5.62	75.58±5.48	75.93±5.62	80.05±3.73	78.33±4.76
10	77.99±6.34	77.83±5.33	77.65±5.72	77.83±4.67	76.45±5.44	77.31±5.59	81.25±5.03	80.73±4.99
20	78.17±3.18	78.69±4.34	79.38±3.83	79.03±4.18	79.72±3.57	79.54±4.16	79.88±4.37	80.05±4.54
30	78.17±3.94	78.34±3.97	78.34±4.14	78.00±3.65	79.71±3.93	78.17±4.72	79.88±3.14	80.22±4.24
40	78.85±4.20	79.03±3.76	78.68±3.98	78.85±4.35	78.52±2.53	79.37±4.80	79.19±3.85	79.19±3.85
50	79.54±4.82	79.37±4.20	79.54±4.24	79.37±4.65	79.02±4.71	79.37±4.20	78.51±3.55	78.51±3.55
60	79.54±4.54	78.85±4.73	79.02±4.34	79.54±4.54	78.85±4.29	78.85±4.73	78.34±3.44	78.34±3.44
70	79.02±4.50	79.02±4.50	78.85±4.45	79.02±4.50	78.51±4.30	79.02±4.50	78.16±4.29	78.16±4.29
80	78.68±4.45	78.68±4.45	78.51±4.39	78.68±4.45	76.96±3.80	78.68±4.45	77.65±4.28	77.65±4.28
90	78.34±4.22	78.34±4.22	78.16±4.13	78.34±4.22	77.48±3.98	78.34±4.22	78.17±3.95	78.17±3.95
100	78.34±4.37	78.34±4.37	78.34±4.37	78.34±4.37	78.34±4.37	78.34±4.37	78.34±4.37	78.34±4.37
KNN								
%	RDF	RTF	TFIDF	TC	TV	TVQ	IG	GR
5	69.38±7.80	68.35±7.10	69.39±5.74	69.57±5.70	73.68±6.70	71.80±6.29	73.17±6.93	57.21±2.80
10	65.29±5.00	63.76±3.73	64.45±5.11	63.07±5.13	62.88±7.54	66.49±7.66	64.95±5.87	58.57±6.24
20	63.38±3.97	63.91±4.21	63.57±3.45	63.05±3.48	64.94±3.56	64.25±4.12	62.55±6.23	62.03±6.18
30	62.69±3.82	59.62±3.94	59.44±4.67	60.98±3.84	59.11±4.96	63.39±3.21	59.62±6.75	59.11±7.38
40	60.97±5.01	58.75±3.50	59.09±3.74	57.71±8.04	59.61±4.21	52.20±8.32	57.73±6.28	57.73±6.28
50	53.92±9.54	53.41±9.05	53.07±8.82	53.92±9.54	52.89±8.41	53.41±9.05	58.59±6.82	58.59±6.82
60	54.09±9.84	53.58±9.46	53.75±9.54	54.09±9.84	53.75±9.16	53.58±9.46	55.15±8.97	55.15±8.97
70	48.95±11.66	48.95±11.66	48.95±11.66	48.95±11.66	53.75±9.12	48.95±11.66	57.74±8.78	57.74±8.78
80	47.57±12.32	47.57±12.32	47.57±12.32	47.57±12.32	53.75±9.37	47.57±12.32	53.94±8.10	53.94±8.10
90	46.88±12.92	46.88±12.92	46.88±12.92	46.88±12.92	47.23±12.78	46.88±12.92	47.93±10.77	47.93±10.77
100	46.88±13.00	46.88±13.00	46.88±13.00	46.88±13.00	46.88±13.00	46.88±13.00	46.88±13.00	46.88±13.00
Naïve Bayes								
%	RDF	RTF	TFIDF	TC	TV	TVQ	IG	GR
5	66.51±7.23	65.65±5.93	66.68±6.67	66.85±5.77	66.51±7.80	66.85±7.49	75.27±6.46	76.97±6.28
10	69.09±7.03	71.15±8.07	70.81±7.86	70.81±7.78	70.63±8.57	70.12±8.30	77.33±6.81	77.16±4.96
20	70.98±6.44	71.32±5.75	72.18±5.83	72.36±6.51	72.18±6.84	72.70±6.53	77.68±6.72	77.68±6.72
30	72.01±5.97	72.52±5.35	72.18±6.09	71.84±5.76	73.21±6.04	72.35±6.00	77.68±6.72	77.68±6.72
40	72.52±5.23	72.69±4.84	72.52±5.48	73.04±5.55	72.69±5.46	72.35±5.72	77.51±6.75	77.51±6.75
50	72.69±5.17	72.52±6.19	72.69±5.96	72.69±5.17	72.52±5.23	72.52±6.19	77.34±6.94	77.34±6.94
60	72.52±5.23	72.35±5.80	72.69±5.15	72.52±5.23	72.52±5.77	72.35±5.80	76.13±6.91	76.13±6.91
70	72.35±5.07	72.35±5.07	72.35±5.07	72.35±5.07	72.35±5.07	72.35±5.07	74.76±6.41	74.76±6.41
80	72.52±5.04	72.52±5.04	72.52±5.04	72.52±5.04	72.52±5.04	72.52±5.04	73.03±5.96	73.03±5.96
90	72.18±5.23	72.18±5.23	72.18±5.23	72.18±5.23	72.18±5.23	72.18±5.23	72.69±5.44	72.69±5.44
100	72.35±5.39	72.35±5.39	72.35±5.39	72.35±5.39	72.35±5.39	72.35±5.39	72.35±5.39	72.35±5.39

Tabela B.3: Acurácias preditivas e desvios padrões obtidos com variação de porcentagem de atributos na base de textos IFM

J48									
%	RDF	RTF	TFIDF	TC	TV	TVQ	IG	GR	ZSTF
5	99.74±0.83	99.74±0.83	99.74±0.83	99.74±0.83	99.74±0.83	99.74±0.83	99.74±0.83	99.74±0.83	99.74±0.83
10	99.74±0.83	99.74±0.83	99.74±0.83	99.74±0.83	99.74±0.83	99.74±0.83	99.74±0.83	99.74±0.83	99.74±0.83
20	99.74±0.83	99.74±0.83	99.74±0.83	99.74±0.83	99.74±0.83	99.74±0.83	99.74±0.83	99.74±0.83	99.74±0.83
30	99.74±0.83	99.74±0.83	99.74±0.83	99.74±0.83	99.74±0.83	99.74±0.83	99.74±0.83	99.74±0.83	99.74±0.83
40	99.74±0.83	99.74±0.83	99.74±0.83	99.74±0.83	99.74±0.83	99.74±0.83	99.74±0.83	99.74±0.83	99.74±0.83
50	99.74±0.83	99.74±0.83	99.74±0.83	99.74±0.83	99.74±0.83	99.74±0.83	99.74±0.83	99.74±0.83	99.74±0.83
60	99.74±0.83	99.74±0.83	99.74±0.83	99.74±0.83	99.74±0.83	99.74±0.83	99.74±0.83	99.74±0.83	99.74±0.83
70	99.74±0.83	99.74±0.83	99.74±0.83	99.74±0.83	99.74±0.83	99.74±0.83	99.74±0.83	99.74±0.83	99.74±0.83
80	99.74±0.83	99.74±0.83	99.74±0.83	99.74±0.83	99.74±0.83	99.74±0.83	99.74±0.83	99.74±0.83	99.74±0.83
90	99.74±0.83	99.74±0.83	99.74±0.83	99.74±0.83	99.74±0.83	99.74±0.83	99.74±0.83	99.74±0.83	99.74±0.83
100	99.74±0.83	99.74±0.83	99.74±0.83	99.74±0.83	99.74±0.83	99.74±0.83	99.74±0.83	99.74±0.83	99.74±0.83
SVM									
%	RDF	RTF	TFIDF	TC	TV	TVQ	IG	GR	ZSTF
5	97.84±2.48	97.86±2.11	96.79±2.99	97.85±1.70	96.24±3.59	94.63±5.03	97.58±2.69	97.85±2.13	97.30±2.21
10	96.79±3.72	97.32±2.19	96.24±2.27	97.31±2.21	96.51±2.56	95.97±3.16	98.12±1.82	98.39±1.89	96.51±2.52
20	95.44±2.53	95.70±2.61	95.97±2.92	95.70±2.61	95.97±3.16	96.51±2.86	98.66±1.42	98.93±1.39	95.70±2.60
30	96.24±2.59	96.78±3.06	96.78±3.06	96.24±2.90	96.78±3.06	96.78±3.06	98.93±1.39	98.93±1.39	95.97±2.61
40	96.24±2.90	96.51±2.86	96.78±3.06	96.24±2.90	96.78±3.06	96.51±2.86	98.66±1.42	98.66±1.42	95.97±2.92
50	96.24±2.90	96.24±2.90	96.24±2.90	96.24±2.90	96.51±2.86	96.24±2.90	98.66±1.42	98.66±1.42	95.70±3.15
60	96.24±2.90	96.24±2.90	96.24±2.90	96.24±2.90	96.24±2.90	96.24±2.90	98.39±1.38	98.39±1.38	95.97±2.92
70	96.24±2.90	96.24±2.90	96.24±2.90	96.24±2.90	96.78±2.47	96.24±2.90	97.31±3.82	97.31±3.82	95.97±2.92
80	96.24±2.90	96.24±2.90	96.24±2.90	96.24±2.90	96.78±2.47	96.24±2.90	96.24±2.90	96.24±2.90	96.24±2.90
90	96.24±2.90	96.24±2.90	96.24±2.90	96.24±2.90	96.51±2.56	96.24±2.90	96.24±2.90	96.24±2.90	96.24±2.90
100	96.51±2.56	96.51±2.56	96.51±2.56	96.51±2.56	96.51±2.56	96.51±2.56	96.51±2.56	96.51±2.56	96.51±2.56
KNN									
%	RDF	RTF	TFIDF	TC	TV	TVQ	IG	GR	ZSTF
5	65.33±5.63	74.75±6.39	73.93±6.50	75.28±7.05	74.47±7.11	77.14±6.15	70.46±5.41	70.72±4.50	73.41±8.49
10	57.55±3.69	61.04±4.61	59.42±3.88	62.12±4.58	61.30±5.15	62.65±5.43	53.78±2.77	54.59±3.39	61.56±3.91
20	55.38±1.57	54.84±1.46	56.74±4.29	55.92±2.33	57.81±3.60	56.19±2.39	46.78±5.43	47.32±5.75	55.11±1.54
30	53.77±2.62	53.77±1.41	54.04±2.12	53.77±1.41	55.11±2.69	53.23±2.79	45.17±5.82	43.82±5.08	53.51±2.23
40	51.61±1.88	54.04±1.70	54.04±1.70	51.61±1.88	52.15±2.13	54.04±1.70	39.54±5.96	39.54±5.96	52.97±2.37
50	48.66±3.75	52.42±1.78	52.96±1.25	48.66±3.75	48.12±4.04	52.42±1.78	32.53±4.30	32.53±4.30	52.70±2.53
60	48.66±4.06	47.30±2.70	44.35±4.50	48.66±4.06	49.46±3.32	47.30±2.70	28.77±2.28	28.77±2.28	48.40±3.95
70	42.50±5.13	42.50±5.13	42.50±5.13	42.50±5.13	36.31±5.55	42.50±5.13	32.80±7.30	32.80±7.30	40.35±6.12
80	45.17±4.95	45.17±4.95	45.17±4.95	45.17±4.95	44.36±7.33	45.17±4.95	30.40±5.95	30.40±5.95	43.55±6.66
90	45.95±6.28	45.95±6.28	45.95±6.28	45.95±6.28	42.20±8.00	45.95±6.28	26.88±1.31	26.88±1.31	43.29±6.83
100	33.89±8.15	33.89±8.15	33.89±8.15	33.89±8.15	33.89±8.15	33.89±8.15	33.89±8.15	33.89±8.15	33.89±8.15
Naïve Bayes									
%	RDF	RTF	TFIDF	TC	TV	TVQ	IG	GR	ZSTF
5	94.89±3.22	95.44±2.53	96.50±2.23	95.16±3.33	96.78±2.78	96.50±2.87	97.04±2.69	98.65±1.42	96.23±2.27
10	92.45±3.34	94.07±4.39	94.61±3.62	94.07±4.74	94.35±3.92	94.08±4.00	98.39±1.38	98.39±1.38	94.61±4.60
20	91.93±2.23	92.46±3.58	92.72±5.12	92.46±4.92	93.28±4.62	93.27±3.88	98.39±1.38	98.39±1.38	91.93±4.60
30	92.20±2.01	90.85±2.28	91.39±3.06	91.66±2.02	93.00±3.88	91.93±2.53	98.66±1.41	98.66±1.41	92.48±3.94
40	91.93±2.53	91.93±2.23	91.93±2.23	91.93±2.53	90.32±2.58	91.66±2.71	98.66±1.41	98.66±1.41	92.47±2.14
50	91.66±2.71	92.20±2.38	91.93±2.86	91.66±2.71	91.39±2.79	92.20±2.38	98.66±1.41	98.66±1.41	91.40±3.28
60	91.66±2.71	91.66±2.71	91.66±2.97	91.66±2.71	91.39±2.51	91.66±2.71	97.05±1.98	97.05±1.98	91.67±3.19
70	91.93±2.23	91.93±2.23	91.93±2.23	91.93±2.23	91.93±2.23	91.93±2.23	96.24±2.26	96.24±2.26	92.20±3.71
80	91.93±2.23	91.93±2.23	91.93±2.23	91.93±2.23	91.93±2.23	91.93±2.23	91.67±3.43	91.67±3.43	91.93±2.57
90	91.39±2.79	91.39±2.79	91.39±2.79	91.39±2.79	91.39±2.79	91.39±2.79	91.93±2.84	91.93±2.84	91.93±2.23
100	91.66±2.71	91.66±2.71	91.66±2.71	91.66±2.71	91.66±2.71	91.66±2.71	91.66±2.71	91.66±2.71	91.66±2.71

Tabela B.4: Acurácias preditivas e desvios padrões obtidos com variação de porcentagem de atributos na base de textos Chemsitry

J48									
%	RDF	RTF	TFIDF	TC	TV	TVQ	IG	GR	ZSTF
5	84.69±5.68	77.90±3.97	77.66±3.88	80.90±4.28	77.15±4.26	76.42±7.10	86.42±5.73	88.17±3.82	79.92±5.72
10	85.92±4.49	82.67±7.02	79.15±5.79	79.65±7.96	77.87±3.43	78.63±4.22	86.41±4.98	86.41±4.98	80.91±8.66
20	85.40±5.37	86.16±4.38	85.90±4.71	85.40±5.37	84.67±4.64	84.40±4.14	86.41±4.98	86.41±4.98	86.42±6.50
30	84.90±6.30	85.40±5.37	85.40±5.37	84.90±6.30	85.90±4.71	85.39±5.66	86.41±4.98	86.41±4.98	87.19±6.49
40	85.15±6.36	85.15±6.36	85.15±6.36	85.15±6.36	86.16±4.69	85.15±6.36	86.41±4.98	86.41±4.98	86.69±6.23
50	85.15±6.36	85.15±6.36	85.15±6.36	85.15±6.36	85.90±5.00	85.15±6.36	86.41±4.98	86.41±4.98	86.69±6.23
60	85.15±6.36	85.15±6.36	85.15±6.36	85.15±6.36	85.15±6.36	85.15±6.36	86.18±6.16	86.18±6.16	86.69±6.23
70	85.15±6.36	85.15±6.36	85.15±6.36	85.15±6.36	85.15±6.36	85.15±6.36	85.17±5.97	85.17±5.97	86.69±6.23
80	85.15±6.36	85.15±6.36	85.15±6.36	85.15±6.36	85.15±6.36	85.15±6.36	84.65±6.59	84.65±6.59	86.69±6.23
90	85.15±6.36	85.15±6.36	85.15±6.36	85.15±6.36	85.15±6.36	85.15±6.36	84.65±6.59	84.65±6.59	86.69±6.23
100	85.15±6.36	85.15±6.36	85.15±6.36	85.15±6.36	85.15±6.36	85.15±6.36	85.15±6.36	85.15±6.36	86.69±6.23
SVM									
%	RDF	RTF	TFIDF	TC	TV	TVQ	IG	GR	ZSTF
5	92.97±3.06	92.22±4.15	89.71±4.00	91.72±4.24	87.19±3.99	86.94±4.68	50.00±4.45	64.81±5.39	91.21±5.19
10	94.99±4.08	91.97±3.48	89.97±3.50	92.47±3.71	89.21±3.88	89.21±3.88	49.51±4.04	48.51±4.11	89.45±5.27
20	94.22±2.65	92.47±3.32	92.48±4.86	93.47±3.57	90.72±3.52	91.48±4.59	27.90±5.41	27.90±5.41	92.70±4.96
30	93.47±3.16	93.48±4.29	93.48±3.95	93.73±3.59	92.98±4.54	92.73±4.16	34.21±7.34	34.21±7.34	92.71±4.33
40	93.22±3.54	93.48±3.77	93.48±3.77	93.22±3.54	92.98±4.05	93.48±3.77	25.90±3.52	25.90±3.52	93.21±3.76
50	93.73±3.59	93.23±3.92	92.98±4.22	93.73±3.59	92.98±4.54	93.23±3.92	23.63±1.99	23.63±1.99	93.21±3.57
60	93.48±3.58	93.48±4.29	93.48±4.29	93.48±3.58	93.22±3.91	93.48±4.29	23.13±1.32	23.13±1.32	92.45±3.96
70	93.22±3.91	93.22±3.91	93.22±3.91	93.22±3.91	92.97±3.29	93.22±3.91	23.90±3.46	23.90±3.46	92.20±4.03
80	92.97±3.68	92.97±3.68	92.97±3.68	92.97±3.68	93.97±3.77	92.97±3.68	24.67±5.83	24.67±5.83	92.20±4.03
90	92.72±3.42	92.72±3.42	92.72±3.42	92.72±3.42	93.73±3.57	92.72±3.42	24.41±5.03	24.41±5.03	91.95±3.73
100	92.97±3.68	92.97±3.68	92.97±3.68	92.97±3.68	92.97±3.68	92.97±3.68	92.97±3.68	92.97±3.68	91.95±3.73
KNN									
%	RDF	RTF	TFIDF	TC	TV	TVQ	IG	GR	ZSTF
5	35.18±2.91	31.67±3.32	32.15±4.16	33.92±3.61	40.23±6.48	38.21±5.97	93.73±3.59	95.49±3.29	48.75±6.65
10	31.92±4.94	28.17±3.92	29.91±3.10	30.65±4.38	27.65±2.82	28.90±2.81	95.98±3.40	95.98±3.40	51.51±7.51
20	25.39±2.65	24.13±1.96	24.88±2.59	24.38±2.22	26.39±3.04	24.38±2.22	95.98±3.40	95.98±3.40	39.72±3.74
30	23.88±1.99	26.15±3.10	26.89±3.19	29.17±4.51	24.38±1.88	23.63±1.47	95.73±3.36	95.73±3.36	37.47±5.65
40	35.47±7.80	30.71±7.68	29.19±5.96	35.47±7.80	28.19±5.46	30.71±7.98	95.73±3.36	95.73±3.36	31.72±7.58
50	32.20±5.91	28.68±5.29	25.92±4.88	32.20±5.91	26.92±4.58	28.68±5.29	95.73±3.36	95.73±3.36	29.94±5.59
60	27.42±5.05	28.93±5.30	25.42±5.13	27.42±5.05	23.13±1.32	28.93±5.30	95.22±3.23	95.22±3.23	28.94±6.14
70	24.15±4.24	24.15±4.24	24.15±4.24	24.15±4.24	23.13±1.32	24.15±4.24	91.97±2.82	91.97±2.82	26.44±6.55
80	24.41±5.03	24.41±5.03	24.41±5.03	24.41±5.03	24.67±5.83	24.41±5.03	88.97±3.90	88.97±3.90	26.19±6.63
90	24.41±5.03	24.41±5.03	24.41±5.03	24.41±5.03	24.92±6.63	24.41±5.03	89.47±3.84	89.47±3.84	26.46±9.11
100	24.92±6.63	24.92±6.63	24.92±6.63	24.92±6.63	24.92±6.63	24.92±6.63	24.92±6.63	24.92±6.63	25.44±7.50
Naïve Bayes									
%	RDF	RTF	TFIDF	TC	TV	TVQ	IG	GR	ZSTF
5	86.94±4.83	91.47±4.47	91.72±4.71	90.72±4.99	90.71±3.91	91.46±3.94	96.24±2.71	95.49±2.83	88.46±5.16
10	90.72±4.07	88.96±4.59	88.71±3.35	89.21±4.89	89.71±3.18	89.72±3.59	96.49±2.68	96.49±2.68	89.22±4.74
20	88.71±3.92	88.97±3.33	88.72±4.25	89.47±3.46	90.72±4.24	90.22±4.13	96.49±2.41	96.49±2.41	89.71±3.82
30	88.22±4.83	88.72±4.08	88.97±4.86	88.47±4.24	89.46±3.03	88.47±4.40	96.99±3.08	96.99±3.08	89.47±4.07
40	89.22±4.05	88.47±4.56	88.47±4.56	89.22±4.05	88.97±3.90	88.47±4.56	96.74±2.90	96.74±2.90	89.72±4.01
50	89.47±3.84	89.22±4.05	88.72±4.41	89.47±3.84	88.97±4.08	89.22±4.05	96.49±2.68	96.49±2.68	89.72±4.01
60	89.22±4.38	89.47±3.84	89.47±3.84	89.22±4.38	89.22±4.38	89.47±3.84	95.24±2.99	95.24±2.99	89.72±4.01
70	88.97±4.41	88.97±4.41	88.97±4.41	88.97±4.41	88.97±4.41	88.97±4.41	94.24±3.54	94.24±3.54	89.72±4.01
80	88.97±4.41	88.97±4.41	88.97±4.41	88.97±4.41	88.97±4.41	88.97±4.41	92.97±4.37	92.97±4.37	89.72±4.01
90	88.72±4.57	88.72±4.57	88.72±4.57	88.72±4.57	88.72±4.57	88.72±4.57	93.47±3.76	93.47±3.76	89.72±4.01
100	88.47±4.85	88.47±4.85	88.47±4.85	88.47±4.85	88.47±4.85	88.47±4.85	88.47±4.85	88.47±4.85	89.72±4.01

Tabela B.5: Acurácias preditivas e desvios padrões obtidos com variação de porcentagem de atributos na base de textos CS

J48									
%	RDF	RTF	TFIDF	TC	TV	TVQ	IG	GR	ZSTF
5	96.09±3.31	95.82±2.82	93.49±2.78	93.49±2.78	95.58±3.23	95.57±2.15	97.13±2.86	97.13±3.58	95.05±3.37
10	96.87±3.44	97.66±3.11	97.91±2.70	97.66±3.11	95.57±2.46	96.09±2.22	97.13±2.86	97.40±3.00	94.79±3.46
20	97.13±3.12	97.13±2.86	97.40±3.00	97.13±2.86	97.40±3.00	97.13±2.86	97.40±3.00	97.40±3.00	96.62±2.74
30	97.13±3.12	97.66±2.57	97.40±3.00	97.40±3.00	97.40±3.00	97.40±3.00	97.40±3.00	97.40±3.00	97.13±3.12
40	97.40±3.00	97.13±2.86	97.40±3.00	97.40±3.00	97.40±3.00	97.40±3.00	97.40±3.00	97.40±3.00	97.13±3.12
50	97.40±3.00	97.40±3.00	97.13±3.12	97.40±3.00	97.13±2.86	97.40±3.00	97.40±3.00	97.40±3.00	97.40±3.00
60	97.40±3.00	97.40±3.00	97.40±3.00	97.40±3.00	97.40±3.00	97.40±3.00	97.40±3.00	97.40±3.00	97.40±3.00
70	97.40±3.00	97.40±3.00	97.40±3.00	97.40±3.00	97.40±3.00	97.40±3.00	97.40±3.00	97.40±3.00	97.40±3.00
80	97.40±3.00	97.40±3.00	97.40±3.00	97.40±3.00	97.40±3.00	97.40±3.00	97.40±3.00	97.40±3.00	97.40±3.00
90	97.40±3.00	97.40±3.00	97.40±3.00	97.40±3.00	97.40±3.00	97.40±3.00	97.40±3.00	97.40±3.00	97.40±3.00
100	97.40±3.00	97.40±3.00	97.40±3.00	97.40±3.00	97.40±3.00	97.40±3.00	97.40±3.00	97.40±3.00	97.40±3.00
SVM									
%	RDF	RTF	TFIDF	TC	TV	TVQ	IG	GR	ZSTF
5	96.85±3.47	95.28±5.22	93.45±5.96	95.03±4.71	92.17±5.39	92.69±4.59	98.42±2.22	98.69±1.38	95.28±4.25
10	97.90±2.72	97.37±2.77	96.07±3.57	97.90±2.72	93.45±5.60	94.51±4.88	98.68±2.24	98.95±1.84	97.89±2.72
20	97.64±2.62	97.90±2.42	98.16±2.16	98.16±2.16	96.33±3.97	97.38±3.28	99.21±1.27	99.21±1.27	97.12±3.15
30	98.43±2.22	97.64±3.39	97.64±3.39	97.65±3.59	97.38±3.50	96.86±3.23	99.21±1.27	99.21±1.27	97.90±2.99
40	98.16±2.79	97.90±3.46	98.16±2.79	98.16±2.79	97.90±2.72	98.16±2.79	98.95±1.35	98.95±1.35	97.65±3.14
50	98.16±2.79	98.43±2.22	98.16±2.16	98.16±2.79	97.90±2.72	98.43±2.22	98.95±1.35	98.95±1.35	97.90±2.99
60	98.16±2.79	98.43±2.22	98.43±2.22	98.16±2.79	98.43±2.22	98.43±2.22	99.21±1.78	99.21±1.78	97.90±2.99
70	98.43±2.22	98.43±2.22	98.43±2.22	98.43±2.22	98.43±2.22	98.43±2.22	98.95±2.22	98.95±2.22	97.90±2.99
80	98.16±2.16	98.16±2.16	98.16±2.16	98.16±2.16	97.90±2.99	98.16±2.16	98.43±2.22	98.43±2.22	97.64±2.90
90	98.16±2.16	98.16±2.16	98.16±2.16	98.16±2.16	97.90±2.99	98.16±2.16	98.68±2.24	98.68±2.24	97.90±2.42
100	98.16±2.16	98.16±2.16	98.16±2.16	98.16±2.16	98.16±2.16	98.16±2.16	98.16±2.16	98.16±2.16	98.16±2.16
KNN									
%	RDF	RTF	TFIDF	TC	TV	TVQ	IG	GR	ZSTF
5	84.87±8.77	86.43±3.18	79.12±7.97	78.62±8.78	76.25±9.37	77.02±7.10	77.53±6.49	56.16±3.25	85.92±4.60
10	74.39±7.06	69.70±12.80	79.66±8.47	72.85±10.21	80.43±5.88	74.96±8.33	65.51±5.24	64.51±5.64	80.20±9.48
20	58.99±8.08	63.16±9.14	68.93±13.58	63.68±9.74	66.30±13.45	58.50±10.32	64.77±5.77	64.77±5.77	66.34±10.87
30	70.78±4.34	52.48±10.10	48.56±10.86	53.76±8.79	44.12±12.81	46.23±6.58	61.87±4.19	61.87±4.19	60.58±9.66
40	47.56±9.13	61.37±5.23	63.20±6.50	49.91±7.83	56.15±6.28	57.42±8.10	63.99±4.57	63.99±4.57	53.00±13.05
50	49.14±7.54	59.02±3.67	60.84±5.59	49.14±7.54	64.25±5.36	58.76±4.23	42.07±10.69	42.07±10.69	59.01±3.79
60	56.66±4.76	56.67±5.50	57.97±5.61	56.66±4.76	62.92±6.93	56.67±5.50	61.64±6.20	61.64±6.20	56.65±6.08
70	53.25±7.92	53.51±7.27	54.30±5.69	53.25±7.92	53.78±8.04	53.51±7.27	63.78±8.85	63.78±8.85	53.79±7.16
80	51.72±6.70	51.72±6.70	51.72±6.70	51.72±6.70	50.70±6.49	51.72±6.70	50.41±6.92	50.41±6.92	47.27±4.24
90	47.76±9.83	47.76±9.83	47.76±9.83	47.76±9.83	59.24±17.12	47.76±9.83	60.32±8.74	60.32±8.74	46.17±8.93
100	44.07±15.51	44.07±15.51	44.07±15.51	44.07±15.51	44.07±15.51	44.07±15.51	44.07±15.51	44.07±15.51	44.07±15.51
Naïve Bayes									
%	RDF	RTF	TFIDF	TC	TV	TVQ	IG	GR	ZSTF
5	94.51±2.29	96.61±2.76	96.61±3.01	96.88±2.36	96.35±2.20	96.35±2.20	98.95±1.84	98.95±1.35	96.61±1.23
10	98.16±1.77	97.91±1.66	97.65±1.92	98.43±1.83	96.34±2.22	97.13±1.50	98.95±1.83	99.21±1.78	96.87±1.65
20	96.60±2.49	98.17±1.77	97.91±1.66	97.91±1.66	98.17±2.14	98.69±1.86	98.95±1.83	98.95±1.83	96.61±1.77
30	97.90±1.66	96.86±1.67	97.38±1.75	97.64±1.94	97.39±2.15	97.64±1.94	98.95±1.83	98.95±1.83	96.60±1.79
40	97.65±1.49	97.90±1.66	97.64±1.94	97.38±1.24	97.38±1.24	97.90±1.66	98.95±1.83	98.95±1.83	97.91±1.66
50	97.90±2.07	97.65±1.49	97.38±1.24	97.90±2.07	97.38±1.75	97.65±1.49	98.95±1.83	98.95±1.83	97.38±1.24
60	97.65±1.49	97.90±2.07	97.90±1.66	97.65±1.49	97.90±1.66	97.90±2.07	98.95±1.83	98.95±1.83	97.39±1.75
70	97.65±1.49	97.90±2.07	97.65±1.94	97.65±1.49	97.65±1.49	97.90±2.07	97.65±2.30	97.65±2.30	97.64±2.30
80	97.91±1.10	97.91±1.10	97.91±1.10	97.91±1.10	97.91±1.10	97.91±1.10	96.86±2.06	96.86±2.06	97.65±1.94
90	97.91±1.10	97.91±1.10	97.91±1.10	97.91±1.10	97.91±1.10	97.91±1.10	97.13±1.92	97.13±1.92	97.65±1.49
100	97.91±1.10	97.91±1.10	97.91±1.10	97.91±1.10	97.91±1.10	97.91±1.10	97.91±1.10	97.91±1.10	97.91±1.10

Tabela B.6: Acurácias preditivas e desvios padrões obtidos com variação de porcentagem de atributos na base de textos Physics

CIIS			
Algoritmo	**Luhn**	**LuhnDF**	**Salton**
J48	74,98±4,55	77,19±4,00	77,48±4,02
SVM	80,29±6,46	78,95±5,82	78,36±4,66
Naïve	76,01±4,5	75,27±5,45	74,82±4,84
KNN	36,75±9,19	31,69±5,05	33,77±5,53
IFM			
Algoritmo	**Luhn**	**LuhnDF**	**Salton**
J48	83,33±3,38	87,10±4,35	78,68±5,75
SVM	78,85±3,52	78,00±4,25	74,56±5,23
Naïve	73,72±4,99	72,52±5,71	72,70±6,26
KNN	59,10±3,30	63,90±5,29	46,38±8,56
IA			
Algoritmo	**Luhn**	**LuhnDF**	**Salton**
J48	71,20±5,27	71,20±4,54	72,20±5,03
SVM	81,80±7,97	80,80±7,32	79,80±7,39
Naïve	82,80±5,27	82,60±6,74	81,60±7,35
KNN	24,60±6,67	28,60±7,00	25,20±5,51
Chemistry			
Algoritmo	**Luhn**	**LuhnDF**	**Salton**
J48	95,45±3,28	99,74±0,83	77,13±8,22
SVM	95,97±2,62	95,73±3,79	91,41±5,34
Naïve	91,39±2,16	92,74±3,14	89,24±5,09
KNN	52,70±2,38	51,61±3,41	42,21±6,66
CS			
Algoritmo	**Luhn**	**LuhnDF**	**Salton**
J48	83,67±4,74	81,39±5,90	80,65±5,79
SVM	91,48±4,74	92,99±3,87	88,72±3,34
Naïve	88,24±6,11	87,70±4,16	87,21±5,67
KNN	25,13±3,19	24,13±2,28	25,40±3,27
Physics			
Algoritmo	**Luhn**	**LuhnDF**	**Salton**
J48	96,36±2,79	93,74±3,52	82,28±5,48
SVM	97,64±3,61	97,90±2,42	93,19±4,35
Naïve	97,64±1,94	97,64±1,94	97,15±3,07
KNN	54,58±8,37	47,02±9,94	49,34±7,29

Tabela B.7: Valores de acurácia obtidos com os diferentes classificadores para os subconjuntos gerados pelos métodos Luhn, LuhnDF e Salton

CIIS									
Alg.	**RDF**	**RTF**	**TFIDF**	**TC**	**TV**	**TVQ**	**ZSTF**	**GR**	**IG**
J48	83,71±4,17	83,71±4,17	83,71±4,17	83,71±4,17	83,71±4,17	83,71±3,72	-	88,15±2,86	88,15±2,86
SVM	87,26±3,72	86,82±3,45	87,10±3,44	87,11±3,13	86,81±2,93	86,81±3,93	-	89,48±5,11	89,48±5,11
Naïve	84,74±2,33	84,74±2,43	84,74±2,43	84,89±2,42	84,89±2,42	84,59±2,45	-	89,63±3,12	89,63±3,12
KNN	45,80±10,31	42,25±10,64	41,36±9,46	41,06±9,65	42,10±9,17	42,40±8,92	-	56,88±3,65	56,88±3,65
IFM									
Alg.	**RDF**	**RTF**	**TFIDF**	**TC**	**TV**	**TVQ**	**ZSTF**	**GR**	**IG**
J48	86,42±3,41	86,42±3,41	85,91±3,63	86,42±3,41	85,91±3,54	85,91±3,72	-	84,53±4,74	84,53±4,74
SVM	78,34±3,60	79,20±3,65	78,51±3,53	78,85±3,79	78,16±4,37	79,37±4,58	-	79,53±4,02	79,53±4,02
Naïve	72,86±5,42	72,35±5,24	72,18±6,02	72,69±5,02	73,38±5,98	72,18±6,06	-	77,51±6,75	77,51±6,75
KNN	60,63±4,62	59,62±3,33	59,78±3,58	58,40±3,68	58,76±4,61	54,10±7,83	-	58,94±6,28	58,94±6,28
IA									
Alg.	**RDF**	**RTF**	**TFIDF**	**TC**	**TV**	**TVQ**	**ZSTF**	**GR**	**IG**
J48	91,60±3,10	91,80±2,90	92,20±2,90	91,60±3,10	92,60±3,41	92,40±2,95	-	91,80±2,74	91,80±2,74
SVM	85,00±8,50	83,20±8,55	82,80±7,73	83,60±9,18	83,80±6,96	84,60±6,74	-	89,80±6,29	89,80±6,29
Naïve	83,20±6,27	83,80±5,77	83,40±6,11	82,80±5,98	83,60±6,79	83,20±6,20	-	90,40±5,40	90,40±5,40
KNN	29,80±5,29	27,60±6,85	29,00±5,75	31,60±7,29	35,40±9,62	31,00±6,27	-	34,40±6,72	34,40±6,72
Chemistry									
Alg.	**RDF**	**RTF**	**TFIDF**	**TC**	**TV**	**TVQ**	**ZSTF**	**GR**	**IG**
J48	99,74±0,83	99,74±0,83	99,74±0,83	99,74±0,83	99,74±0,83	99,74±0,83	99,74±0,83	99,74±0,83	99,74±0,83
SVM	95,97±2,92	96,51±3,13	97,05±3,22	96,51±2,86	96,78±3,06	96,78±3,06	96,77±2,49	98,93±1,39	98,93±1,39
Naïve	91,93±2,23	91,12±1,87	90,58±2,64	91,93±2,23	92,46±4,20	91,93±2,23	92,48±4,65	98,66±1,41	98,66±1,41
KNN	51,61±3,30	51,62±2,28	51,88±3,38	50,26±3,91	51,34±3,43	43,30±6,16	28,22±2,20	45,44±4,91	45,44±4,91
CS									
Alg.	**RDF**	**RTF**	**TFIDF**	**TC**	**TV**	**TVQ**	**ZSTF**	**GR**	**IG**
J48	84,90±6,30	85,40±5,37	85,40±5,37	84,90±6,30	85,90±4,71	85,39±5,66	82,45±6,28	86,41±4,98	86,41±4,98
SVM	93,72±2,95	93,48±4,29	93,23±4,10	93,73±3,59	92,98±4,54	92,73±4,16	91,73±4,41	96,99±3,08	96,99±3,08
Naïve	88,22±4,83	88,47±4,24	88,97±4,86	88,47±4,24	89,46±3,03	88,47±4,40	86,46±4,39	95,73±3,36	95,73±3,36
KNN	24,40±3,59	25,90±3,58	27,42±4,88	27,17±4,47	25,40±4,78	27,42±4,08	23,13±1,32	34,21±7,34	34,21±7,34
Physics									
Alg.	**RDF**	**RTF**	**TFIDF**	**TC**	**TV**	**TVQ**	**ZSTF**	**GR**	**IG**
J48	97,40±3,00	97,13±2,86	97,40±3,00	97,13±3,12	97,40±3,00	97,40±3,00	94,25±2,71	97,40±3,00	97,40±3,00
SVM	98,16±2,79	97,90±3,46	98,16±2,79	98,16±2,79	97,90±2,72	97,91±2,97	98,16±2,16	98,95±1,35	98,95±1,35
Naïve	97,65±1,49	97,90±1,66	97,64±1,94	97,12±1,50	97,38±1,24	97,90±1,66	97,38±2,48	98,95±1,83	98,95±1,83
KNN	51,21±7,94	54,30±4,24	54,56±5,33	54,05±10,14	57,42±5,54	55,09±2,87	53,56±11,31	65,03±4,44	65,03±4,44

Tabela B.8: Acurácia preditiva obtida pelos diferentes métodos de seleção de atributos com subconjuntos de cardinalidade igual ao método de Luhn

CIIS									
Alg.	**RDF**	**RTF**	**TFIDF**	**TC**	**TV**	**TVQ**	**ZSTF**	**GR**	**IG**
J48	83,86±4,25	83,86±4,25	84,15±4,35	83,86±4,25	83,86±4,06	84,01±4,66	-	88,15±2,86	88,15±2,86
SVM	87,26±3,37	86,08±4,52	86,96±4,61	86,97±4,64	86,07±3,79	88,29±3,60	-	90,08±3,82	90,08±3,82
Naïve	84,59±2,01	84,15±2,33	84,15±2,33	84,15±2,33	82,97±2,51	84,30±2,24	-	89,63±3,12	89,63±3,12
KNN	43,73±9,16	48,02±10,02	43,72±8,58	47,88±9,95	37,04±7,31	34,53±8,71	-	61,33±4,19	61,33±4,19
IFM									
Alg.	**RDF**	**RTF**	**TFIDF**	**TC**	**TV**	**TVQ**	**ZSTF**	**GR**	**IG**
J48	86,59±3,90	85,56±3,58	86,08±4,11	85,90±3,90	84,36±4,04	84,87±5,45	-	84,53±4,74	84,36±4,63
SVM	78,17±4,09	79,03±4,10	79,19±4,32	78,00±4,34	78,69±2,64	78,16±4,45	-	80,39±4,46	79,37±3,02
Naïve	72,00±5,12	72,86±6,09	72,86±6,05	72,35±5,91	73,38±6,45	72,70±6,93	-	77,68±6,72	77,68±6,72
KNN	63,90±3,56	63,39±3,60	63,90±3,60	63,04±3,35	59,12±5,36	62,02±3,04	-	59,80±7,34	60,31±6,39
IA									
Alg.	**RDF**	**RTF**	**TFIDF**	**TC**	**TV**	**TVQ**	**ZSTF**	**GR**	**IG**
J48	92,20±3,33	92,20±3,58	92,20±3,58	92,20±3,71	92,20±3,46	92,60±3,41	-	91,80±2,74	91,80±2,74
SVM	84,80±7,38	83,20±6,34	81,80±8,66	83,80±5,85	82,00±6,93	82,00±6,67	-	90,20±6,00	90,20±6,00
Naïve	83,60±6,72	84,80±8,07	85,20±7,84	84,00±8,22	85,00±8,29	85,40±9,09	-	90,40±5,40	90,40±5,40
KNN	32,20±5,85	37,00±6,41	35,40±6,60	38,60±9,14	33,80±7,27	30,60±4,33	-	30,20±4,66	30,20±4,66
Chemistry									
Alg.	**RDF**	**RTF**	**TFIDF**	**TC**	**TV**	**TVQ**	**ZSTF**	**GR**	**IG**
J48	99,74±0,83	99,74±0,83	99,74±0,83	99,74±0,83	99,74±0,83	99,74±0,83	92,99±3,19	99,74±0,83	99,74±0,83
SVM	96,00±3,80	96,24±2,90	96,77±2,79	96,78±2,78	95,97±2,89	96,24±2,90	89,50±5,34	98,39±1,89	98,39±1,89
Naïve	92,74±3,81	92,48±4,52	93,29±3,85	92,72±4,62	94,34±4,32	93,81±4,23	88,71±3,02	98,39±1,38	98,39±1,38
KNN	53,51±2,23	53,78±1,67	54,32±2,10	54,04±1,12	53,24±2,64	55,39±3,37	27,68±1,64	51,09±3,96	50,54±3,32
CS									
Alg.	**RDF**	**RTF**	**TFIDF**	**TC**	**TV**	**TVQ**	**ZSTF**	**GR**	**IG**
J48	85,15±5,45	86,16±4,38	85,41±4,45	85,90±4,71	83,42±5,18	84,67±4,79	76,61±5,54	86,41±4,98	86,41±4,98
SVM	94,73±3,21	91,72±3,90	90,72±3,52	92,73±3,62	90,21±4,14	91,23±5,03	86,96±3,65	96,24±2,12	96,24±2,12
Naïve	88,22±4,97	88,72±5,28	90,22±4,76	89,22±4,84	90,21±4,00	90,22±4,61	84,44±3,81	95,98±3,40	95,98±3,40
KNN	29,42±4,42	26,90±3,05	27,40±3,42	27,66±3,57	27,90±3,78	28,41±3,91	23,38±1,96	26,39±3,03	26,39±3,25
Physics									
Alg.	**RDF**	**RTF**	**TFIDF**	**TC**	**TV**	**TVQ**	**ZSTF**	**GR**	**IG**
J48	97,13±3,12	97,13±2,86	97,13±2,86	96,87±2,68	97,40±3,00	97,13±2,86	92,16±3,93	97,40±3,00	97,40±3,00
SVM	98,42±2,22	97,65±2,60	97,65±2,60	98,68±1,86	97,11±4,01	97,64±3,39	95,05±2,27	98,95±1,36	98,95±1,36
Naïve	97,12±1,50	97,39±1,24	97,91±1,66	97,12±2,31	98,43±2,19	98,17±1,77	98,18±1,74	98,95±1,83	98,95±1,83
KNN	52,75±6,52	58,97±9,35	56,65±10,32	62,87±8,11	54,80±9,62	60,84±6,16	77,00±6,34	63,47±4,54	63,47±4,54

Tabela B.9: Acurácia preditiva obtida pelos diferentes métodos de seleção de atributos com subconjuntos de cardinalidade igual ao método de LuhnDF

CIIS									
Alg.	**RDF**	**RTF**	**TFIDF**	**TC**	**TV**	**TVQ**	**ZSTF**	**GR**	**IG**
J48	83,86±4,25	83,86±4,25	84,15±4,35	83,86±4,25	84,31±3,99	84,16±4,61	-	88,15±2,86	88,15±2,86
SVM	87,26±3,51	86,37±4,26	87,26±3,87	86,53±4,40	86,36±3,83	87,55±4,11	-	89,78±3,86	89,78±3,86
Naïve	83,85±1,81	84,15±2,33	84,29±2,14	84,15±2,33	82,67±2,39	84,15±2,44	-	89,63±3,12	89,63±3,12
KNN	43,88±9,62	43,43±8,81	44,91±9,53	47,88±9,95	35,71±6,31	34,68±8,14	-	61,19±4,48	61,19±4,48
IFM									
Alg.	**RDF**	**RTF**	**TFIDF**	**TC**	**TV**	**TVQ**	**ZSTF**	**GR**	**IG**
J48	86,42±3,78	86,42±3,86	85,05±4,08	86,60±3,97	85,05±4,15	84,87±4,38	-	84,53±4,45	84,36±4,63
SVM	78,52±3,95	79,54±4,81	79,73±3,78	79,03±3,73	79,72±5,09	79,03±4,47	-	80,22±4,39	79,88±4,12
Naïve	71,15±6,32	72,01±6,03	72,70±5,74	71,84±6,55	73,04±7,36	72,18±7,34	-	77,68±6,72	77,68±6,72
KNN	63,22±3,10	63,57±3,54	63,57±3,98	63,74±3,20	67,00±5,08	66,82±6,29	-	62,37±5,86	63,40±5,53
IA									
Alg.	**RDF**	**RTF**	**TFIDF**	**TC**	**TV**	**TVQ**	**ZSTF**	**GR**	**IG**
J48	91,80±3,05	92,00±3,89	92,40±3,37	93,00±4,14	92,60±3,41	92,60±3,41	-	91,80±2,74	91,80±2,74
SVM	84,80±7,32	83,60±6,31	82,40±8,42	84,60±6,93	82,80±6,68	82,00±6,60	-	90,80±5,83	90,80±5,83
Naïve	83,80±6,43	83,80±8,13	85,20±8,01	84,20±7,69	85,20±8,12	85,60±8,42	-	90,40±5,40	90,40±5,40
KNN	30,60±6,47	36,00±7,66	33,60±5,87	33,00±8,60	31,40±7,60	30,00±4,22	-	35,20±6,88	35,20±6,88
Chemistry									
Alg.	**RDF**	**RTF**	**TFIDF**	**TC**	**TV**	**TVQ**	**ZSTF**	**GR**	**IG**
J48	99,74±0,83	99,74±0,83	99,74±0,83	99,74±0,83	99,74±0,83	99,74±0,83	96,76±3,78	99,74±0,83	99,74±0,83
SVM	95,71±2,57	95,70±2,88	96,24±2,90	95,97±2,62	95,97±3,16	96,24±3,14	94,89±3,24	98,93±1,39	98,66±1,42
Naïve	92,20±2,70	92,72±4,62	92,72±4,62	91,65±5,19	93,82±4,75	92,20±4,11	89,51±3,91	98,39±1,38	98,39±1,38
KNN	52,16±1,71	52,16±3,44	52,70±2,38	52,96±1,78	53,78±1,67	53,51±1,83	27,69±1,27	47,32±5,89	47,06±5,93
CS									
Alg.	**RDF**	**RTF**	**TFIDF**	**TC**	**TV**	**TVQ**	**ZSTF**	**GR**	**IG**
J48	85,15±5,45	85,91±4,52	85,41±4,45	85,90±4,71	83,42±5,18	84,42±5,11	76,85±7,34	86,41±4,98	86,41±4,98
SVM	94,23±2,64	91,97±3,68	90,21±3,40	92,23±3,81	90,47±4,51	90,48±5,74	86,71±3,70	96,24±2,42	96,24±2,42
Naïve	87,97±4,49	88,72±5,28	90,22±4,76	88,97±4,41	90,71±4,43	90,22±4,61	84,95±4,35	95,98±3,40	95,98±3,40
KNN	29,17±4,13	26,40±3,34	27,15±3,50	27,66±3,57	27,41±3,67	27,90±4,06	23,13±1,32	26,39±3,03	26,39±3,25
Physics									
Alg.	**RDF**	**RTF**	**TFIDF**	**TC**	**TV**	**TVQ**	**ZSTF**	**GR**	**IG**
J48	97,13±3,12	97,13±2,86	97,40±3,00	96,87±2,68	97,40±3,00	97,13±2,86	90,07±4,63	97,40±3,00	97,40±3,00
SVM	98,42±2,22	97,65±2,60	97,65±2,60	98,43±1,84	97,11±4,01	97,64±3,39	94,26±2,41	99,21±1,27	99,21±1,27
Naïve	96,86±2,06	97,65±1,94	98,17±1,77	97,39±2,15	98,69±1,86	98,69±1,86	97,91±1,63	98,95±1,83	98,95±1,83
KNN	57,15±6,92	61,84±9,01	57,19±8,13	58,18±9,44	63,43±9,32	58,70±10,68	73,64±5,51	65,28±4,16	65,28±4,16

Tabela B.10: Acurácia preditiva obtida pelos diferentes métodos de seleção de atributos com subconjuntos de cardinalidade igual ao método de Salton