Dois novos métodos para seleção não-supervisionada de atributos em Mineração de Textos

Bruno Magalhães Nogueira[1], Solange Oliveira Rezende[1]

Laboratório de Inteligência Computacional (LABIC)
Universidade de São Paulo, Instituto de Ciências Matemáticas e de Computação
Caixa Postal 668, São Carlos, SP, Brasil - 13560-970
brunomn@icmc.usp.br, solange@icmc.usp.br

Abstract Feature selection is an essential task for the viability and efficiency of the Text Mining process. In unsupervised contexts, the feature selection is more difficult, since there aren't any general predefined feature quality measures for unsupervised methods. This work presents two new unsupervised feature selection methods proposed by the authors. These methods are compared with other eight methods found in the literature. The comparison is done in two ways: unsupervised, through Expected Mutual Information Measure, and supervised, through the accuracy of four classifiers (C4.5, SVM, KNN and Naïve Bayes). The statistical test of Kruskal-Wallis is applied to the results of these comparisons in order to determine the statistical significance of the performance of each method. The results point that there is no difference on the performance of the compared methods, allowing us to conclude that the proposed methods are as efficient as the other feature reduction methods.

1 Introdução

Ao lidar com contextos esparsos e de alta dimensionalidade de atributos, inerentes às tarefas de Mineração de Textos, o desempenho de algoritmos de aprendizado cai drasticamente. O excesso de atributos causa lentidão no processo de treinamento e decremento na qualidade do conhecimento extraído [18]. A fim de tratar o problema da alta dimensionalidade e viabilizar o uso de algoritmos de aprendizado de máquina, faz-se necessário um processo de redução de atributos, considerando apenas os termos mais representativos da coleção de textos.

Em contextos não-supervisionados, a redução de dimensionalidade de atributos é uma tarefa especialmente difícil, sendo um dos principais problemas defrontados em Mineração de Textos [3]. Em tarefas dessa natureza, a relevância de um atributo varia de acordo com o objetivo da tarefa a ser realizada. Assim sendo, são poucos os métodos de propósito geral que lidam com tarefas dessa natureza. Destes, uma classe muito interessante é a dos métodos de seleção de atributos, os quais selecionam um subconjunto dos atributos originais, contendo os mais importantes ao domínio da aplicação.

Dentre os métodos para seleção de atributos não-supervisionada encontrados, os mais simples utilizam medidas como a freqüência do termo na coleção para a determinação de sua importância. Outros métodos, mais custosos computacionalmente, analisam a variância dos termos na coleção. Este trabalho apresenta dois novos métodos, com baixo custo computacional: LuhnDF [9] e *Zone-Scored Term Frequency*. Estes métodos são comparados com outros encontrados na literatura quanto à eficiência na seleção de atributos em tarefas de Mineração de Textos: Método de Luhn [7], Método de Salton [13], *Ranking* por *Term Frequency*, *Ranking* por *Document Frequency* [17], *Term Frequency-Inverse Document Frequency* [11], *Term Contribution* [6], *Term Variance* [5] e *Term Variance Quality* [1]. É importante ressaltar que, embora sejam métodos relativamente antigos, são bastante utilizados até hoje, com bons resultados, como em [2] e [16].

Esses métodos são avaliados por medidas objetivas que mensuram a perda de informação acarretada pela eliminação de atributos sugerida por esses métodos, analisando quais métodos são mais propensos a eliminar termos muito importantes ao domínio. Com essa finalidade, foram aplicados dois métodos de avaliação: (1) não-supervisionado, por meio da medida estatística de *Expected Mutual Information Measure* (EMIM); e (2) supervisionado, realizado usando medidas de acurácia preditiva de quatro classificadores (C4.5, SVM, KNN e *Naïve* Bayes). Sob esses resultados, foi utilizado o teste estatístico de Kruskal-Wallis [4] para determinar significância estatística na diferença do desempenho dos métodos.

Este artigo está organizado da seguinte forma: na Seção 2 são apresentados os dois métodos de seleção de atributos propostos nesse trabalho, bem como os demais utilizados para a realização do estudo comparativo. Na Seção 3, a avaliação experimental para comparação dos métodos de seleção de atributos e os respectivos resultados são apresentados. Por fim, as conclusões deste trabalho são apresentadas na Seção 4.

2 Métodos não-supervisionados de seleção de atributos para Mineração de Textos

Os métodos não-supervisionado de seleção de atributos foram aqui 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 (incluindo os dois métodos aqui propostos), são apresentados nas seções a seguir. Para efeitos de padronização de notação, assume-se 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.

2.1 Baseados em Freqüência de Termos ou Freqüência de Documentos

Freqüência de termo (TF - do inglês *Term Frequency*), e freqüência de documentos (DF - do inglês *Document Frequency*), são duas medidas de relação entre

atributos e documentos. A TF é uma medida que contabiliza a freqüência absoluta de um termo ao longo da coleção de documentos. Já a DF mede o número de documentos que um termo ocorre. Os métodos que utilizam essa medida são descritos a seguir (à exceção daqueles que computam *rank* simples dessas medidas - RDF e RTF), incluindo o primeiro método aqui apresentado (**LuhnDF**).

O método de Luhn [7] baseou-se na Lei de Zipf [19], também conhecida como Princípio do Menor Esforço, para selecionar atributos. Nesse método, contabiliza-se a freqüência dos termos e ordena-se o histograma resultante em ordem decrescente, formando a chamada *Curva de Zipf*. Esse método propõe pontos de corte superior e inferior da *Curva de Zipf*, escolhidos subjetivamente (neste trabalho, em regiões próximas aos pontos de inflexão da *Curva de Zipf*), de maneira que termos com alta e baixa TF são descartados. Termos de alta TF são não relevantes por geralmente aparecerem na maioria dos textos. Já os termos de baixa TF são considerados muito raros e não possuem caráter discriminatório.

O método de Salton [13] baseia-se no poder de discriminação de um termo, ou seja, quão bem um termo é capaz de diferenciar 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, de maneira que os melhores termos são aqueles que atingem maior grau de separação. Para isso, esse método sugere considerar termos que não apresentem DF muito alta ou muito baixa, selecionando termos que apresentem DF entre 1% e 10% do total de documentos da coleção.

A TFIDF [11] é uma medida que visa ponderar a TF dos termos em função de suas distribuições na coleção, dando menor peso àqueles termos que aparecem em muitos documentos. Assim, termos muito comuns na coleção são considerados não discriminativos. Para isso, é introduzido o valor do inverso da freqüência de documentos (IDF - do inglês *Inverse Document Frequency*) para um atributo. A TFIDF resulta da multiplicação da TF de um termo pela sua IDF.

Com o objetivo de aproveitar a idéia de seleção de termos baseada na DF dos mesmos, apresentada no método de Salton, e a adoção de cortes subjetivos sugerida no método de Luhn, foi proposto pelos autores um novo método denominado **LuhnDF** [9]. Assim como no método de Salton, esse método seleciona termos cuja DF não seja tão grande, nem tão pequena. Desse modo, esse método pode ser visto como uma flexibilização do método de Salton, na medida em que os pontos de corte a serem adotados não são pré-fixados, ficando a cargo de uma análise subjetiva do analista. Para selecionar os pontos de corte, faz-se uma adaptação do método de Luhn para a freqüência de documentos, gerando-se os histogramas das DF dos termos de forma descendente, aplicando cortes de Luhn sobre esse histograma, escolhendo um ponto de corte superior e outro inferior, ambos próximos aos pontos de inflexão da curva de tendência do histograma.

2.2 Baseados em Variância de Termos

A variância mede a dispersão de uma variável em relação ao seu valor esperado, utilizada para avaliar a distribuição das freqüências de um termo ao longo da coleção. Três métodos foram encontrados na literatura e são descritos a seguir.

A Contribuição do Termo (TC) [6] é baseada no conceito de valor discriminativo de um termo [12], segundo o qual a importância de um termo pode ser vista como a contribuição desse para a similaridade de documentos. Embora não se valha explicitamente da medida estatística de variância, a análise que o método faz da distribuição da TF de um termo ao longo da coleção pode ser encarada como uma aproximação da variância do termo. Esse método provê maior *score* àqueles termos que aparecem em poucos documentos, ignorando atributos muito raros ou muito freqüentes. A contribuição do termo pode ser obtida utilizando a Equação 1, considerando x e y dois documentos da coleção:

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

A Variância do Termo (TV) [5] considera que os termos importantes são aqueles que não apresentam baixa freqüência de documentos e mantêm uma distribuição não-uniforme ao longo da coleção. Para a aplicação desse método, calcula-se a variância de todos os atributos do domínio, conforme a Equação 2:

$$TV_j = \sum_{i=1}^{N} \left[TF_{ij} - \bar{TF_j}\right]^2 \tag{2}$$

na qual $\bar{TF_j}$ é a média das freqüências do j-ésimo termo na coleção.

O método da Qualidade da Variância do Termo (TVQ) [1], 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:

$$TVQ_j = \sum_{i=1}^{N} TF_{ij}^2 - \frac{1}{N}\left[\sum_{i=1}^{N} TF_{ij}\right]^2 \tag{3}$$

2.3 Baseados em contexto

A idéia de contexto de um termo é oriunda da indexação por zonas [8], utilizada em recuperação de informação. Zonas são partes bem delimitadas de documentos constituídas por textos escritos. Por exemplo, em artigos científicos, o título, o resumo e a conclusão são zonas do documento. É fácil perceber que algumas dessas zonas trazem mais informação relevante que outras. Assim, parece natural ponderar a freqüência de um termo de acordo com a zona em que aparece.

Seguindo essa idéia, no presente trabalho é apresentado um método baseado em zonas do documento para seleção de termos denominado ***Zone-Scored Term Frequency* (ZSTF)**. 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 4

$$ZSTF_j = \sum_{i=1}^{N} \sum_{l=1}^{L} TF_{ijl} * P_l \tag{4}$$

na qual TF_{ijl} é a freqüência do j-ésimo termo na l-ésima zona do i-ésimo documento e P_l é o peso aferido à l-ésima zona dos documentos da coleção. A soma dos pesos de todas as seções deve ser igual a 1, ou seja, o peso de cada seção deve estar no intervalo entre 0 e 1.

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 ser mais representativos na coleção.

3 Avaliação experimental

Para avaliar a eficiência dos métodos propostos e os encontrados na literatura para seleção de atributos em contextos não-supervisionados, seis bases de textos foram utilizadas. Essas bases de textos foram montadas a partir de documentos científicos, como artigos e dissertações. Cada base é relativa a um domínio, sendo os documentos divididos em subdomínios, de acordo com o tema do mesmo. Esses subdomínios correspondem às classes utilizadas no processo supervisionado de avaliação, sendo desconsiderados no processo de seleção de atributos. Na Tabela 1 é apresentada uma descrição das bases de textos utilizadas. A escolha por documentos científicos foi tomada a fim de evitar problemas com a qualidade do vocabulário, muito comum em textos de outras naturezas, apesar dessa escolha trazer um limitante do número de bases disponíveis para a realização das comparações. Uma qualidade não controlada de vocabulário tem um impacto negativo nos resultados, como em [9].

Base	**Domínio**	**# Classes**	**# Docs**	**# Docs Classe Majoritária**
CIIS	Inteligência Artificial	4	675	276
IA	Inteligência Artificial	5	500	100
IFM	Inst. Fábrica do Milênio	4	591	291
CS	Ciência da Computação	4	408	127
Chemistry	Química	4	397	100
Physics	Física	4	391	117

Tabela 1. Descrição das bases de textos utilizadas

O pré-processamento dessas bases foi feito como sugerido em [9], gerando-se termos simples (*one-grams*) pela redução das palavras ao seu *stem* por meio do algoritmo de Porter [10]. Os métodos de Luhn, Salton e LuhnDF fornecem pontos exatos de corte de atributos, enquanto os demais fornecem *rankings* de atributos. Para os métodos que não fornecem pontos exatos de corte, variou-se a porcentagem de atributos selecionados (5%, 10%, 20%, ... , 90%), além de serem montados subconjuntos de atributos com as mesmas cardinalidades dos subconjuntos montados pelos métodos de Luhn, LuhnDF e Salton, para fins de comparação. Na Tabela 2 é possível observar o número de atributos gerados pelos métodos que fornecem pontos exatos de corte, bem como o número total de atributos para cada base.

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
100%	4101	72974	34747	23295	28194	22195

Tabela 2. Número de atributos nos subconjuntos de cada base de textos

O método ZSTF foi aplicado às bases de textos *CS*, *Chemistry* e *Physics*, por elas possuírem zonas bem delimitadas e serem as únicas bases disponíveis escritas, cada uma, sob um único modelo. Nessas bases, quatro zonas foram utilizadas e a elas foram atribuídos pesos: título, com peso 0,4; resumo, com peso 0,3; corpo, com peso 0,1; e conclusão, com peso 0,2.

A avaliação não-supervisionada da eficiência dos métodos de seleção de atributos se deu por meio da aplicação da medida estatística de *Expected Mutual Information Measure* (EMIM) [14], a qual tem por objetivo mensurar quão bem os termos selecionados em um determinado subconjunto de atributos consegue predizer o restante do vocabulário da coleção de documentos. 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 5:

$$EMIM_S = \sum_{j=1}^{M} \sum_{s=1}^{S} P(j,s) * log \frac{P(j,s)}{P(j) * P(s)} \tag{5}$$

Consegue-se, com essa medida, detectar a eliminação de termos 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. Assim os melhores métodos de seleção de atributos são aqueles que apresentam maior valor de EMIM para subconjuntos de mesma cardinalidade.

A avaliação supervisionada, por sua vez, foi efetuada por meio da avaliação da acurácia preditiva de classificadores. Adotou-se a utilização de quatro classificadores muito utilizados em tarefas de classificação textual: Árvores de Decisão C4.5, K-Vizinhos Mais Próximos (KNN, do inglês *K-Nearest Neighbor*, Máquinas de Vetor Suporte (SVM, do inglês *Support Vector Machines*) e Classificadores Bayesianos Simples. Foram utilizadas as implementações disponíveis no ambiente Weka [15], com processo de treinamento 10-*fold cross validation*.

Para efeitos de comparação de métodos de seleção de atributos, tem-se por hipótese que utilizando apenas atributos relevantes ao domínio do problema os classificadores conseguem obter maior acurácia. Embora a avaliação supervisionada possa ser considerada 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. Além disso, pode indicar a aplicabilidade desses métodos em contextos supervisionados.

Em ambos os tipos de avaliação dos subconjuntos de atributos gerados pelos diferentes métodos de seleção de atributos, foram efetuadas comparações es-

tatísticas a fim de verificar o impacto da aplicação dos dois métodos propostos, bem como os demais encontrados na literatura. Três grupos de comparações com diferentes hipóteses foram estabelecidos:

1. Variação da porcentagem de atributos selecionados, comparando métodos sem ponto exato de corte: visa verificar o impacto desse tipo de método, mais complexos, buscando por diferenças significativas no desempenho desses;
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;
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, mais simples, com métodos mais complexos, a fim de verificar se a aplicação do segundo grupo de métodos é mais eficaz, justificando o esforço 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 de Kruskal-Wallis [4], o mais indicado para amostras não emparelhadas e não paramétricas, como é a configuração exigida pelo nosso problema, aplicando o pós-teste de múltiplas comparações de Dunn.

Por limitações de espaço, apenas um exemplo dos resultados das avaliações é mostrado na Figura 1, a qual contém o resultado da avaliação não-supervisionada e da avaliação supervisionada obtida por meio da acurácia do classificador *Naïve Bayes*, ambos para a base de textos *Chemistry*[1]. Nessa figura, na qual se encontra os gráficos dos valores das avaliações para os diferentes métodos x número de atributos de cada subconjunto considerado, é possível perceber a grande semelhança entre os métodos nas avaliações, incluindo os métodos propostos.

Na Tabela 3 é possível observar os p-valores que foram obtidos nos diversos grupos de comparações. Pelo fato do número mínimo de amostras requerido pelo teste estatístico de Kruskal-Wallis ser igual a 8, não foi possível efetuar todos os testes estatísticos para todos os grupos de comparação. Para os casos em que não se pôde fazer testes estatísticos, a análise se deu subjetivamente pela diferença de desempenho dos métodos.

Todos os testes estatísticos realizados, para todos os grupos de comparações em ambas as avaliações apontaram que não há diferença estatística no desempenho dos métodos. A análise subjetiva das diferenças de desempenho entre os métodos também mostra que há grande similaridade entre estes. 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. Apesar disso, existem situações para as quais a aplicação

[1] As tabelas completas dos resultados aqui discutidos estão disponíveis em http://www.icmc.usp.br/∼brunomn/downloads/resultados_comparacao.pdf

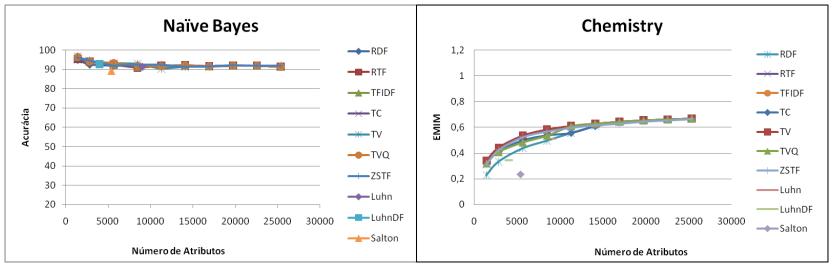

Figura 1. Acurácia preditiva do classificador Naïve Bayes e EMIM na base de textos *Chemistry*, para os diversos subconjuntos gerados

de um tipo de método é mais adequado. Em coleções com número reduzido de documentos, por exemplo, o uso de métodos baseados em DF não é indicado. Métodos mais maleáveis quanto aos pontos de corte, como Luhn e LuhnDF são indicados para processos onde o usuário tem conhecimento de domínio e possui uma expectativa prévia do número de termos, bem como quais termos seriam necessários para representar o domínio. Já métodos baseados em contexto mostram-se boas escolhas para coleções que possuam documentos com zonas de diferentes importâncias bem delimitadas.

Grupo1			**Grupo 2 - Supervisionado**	**Grupo 3 - Supervisionado**	
Base	**Não Sup.**	**Sup.**		**Cardinalidade Igual Luhn**	
CIIS	0,9966	0,9997	**Todas as bases**	Todas (sem ZSTF)	0,9523
IA	0,9989	>0,9999		Bases para ZSTF	0,9989
IFM	0,9989	>0,9999		**Cardinalidade Igual LuhnDF**	
Chemistry	0,9964	>0,9999		Todas (sem ZSTF)	0,9791
CS	0,998	0,9873	0,5748	Bases para ZSTF	0,924
Physics	0,9962	0,6515		**Cardinalidade Igual Salton**	
Todas (sem ZSTF)	0,9679	>0,9999		Todas (sem ZSTF)	0,4386
Bases para ZSTF	0,9685	0,9997		Bases para ZSTF	0,6166

Tabela 3. p-valores obtidos para os grupos 1, 2 e 3 de comparações

4 Conclusões

Neste trabalho, dez métodos não supervisionados foram comparados, sendo os métodos LuhnDF e ZSTF propostos pelos autores. A avaliação dividiu-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 os métodos que não fornecem ponto exato de corte (RTF, RDF, TFIDF, TC, TV, TVQ e ZSTF); confrontando 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 detectou diferença estatística no desempenhos dos métodos aqui comparados. 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, em ambas as avaliações. Com isso, pode-se afirmar que esses métodos possuem desempenho muito semelhante na preservação da informação em uma base textual e 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 **ZSTF** é o primeiro método de uma categoria de métodos que permitem ponderar zonas, indo ao encontro da análise subjetiva que um ser humano faz dos documentos. Como uma proposta inicial, seu resultado é satisfatório e promissor, embora o método falhe por não considerar o número de termos presente em uma seção para ponderação, o que pode minimizar o efeito da atribuição de pesos às zonas. A ponderação pelo número de termos de uma zona será futuramente investigada.

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 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**, proposto pelos autores 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.

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 preservar na base de textos a informação que a base de textos pode fornecer, e a estrutura das classes ou grupos que se delineiam na base.

Em suma, a alta eficiência demonstrada por todos os métodos aqui comparados comprova a hipótese de que se pode reduzir eficientemente o número de atributos de maneira não-supervisionada para tarefas de Mineração de Textos. De acordo com as avaliações estatísticas efetuadas, os testes aqui realizados apontam que os métodos mais simples, como o método **LuhnDF**, 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, aplicando os métodos Luhn, LuhnDF e Salton, já se sabe exatamente quantos e quais atributos devem ser mantidos. Quanto ao método **ZSTF**, este se mostrou uma eficiente opção para tarefas em que se deseja valorizar determinadas partes do documento no processo de seleção de atributos, apresentando desempenho sempre compatível com os melhores métodos, bem como um baixo custo computacional.

Referências

1. I. Dhillon, J. Kogan, and C. Nicholas. Feature selection and document clustering. In M. W. Berry, editor, *Survey of Text Mining*, pages 73–100. Springer, 2003.
2. L. Gonzaga, M. Grivet, and A. T. Vasconcelos. A simple and fast term selection procedure for text clustering. In *Proceedings of the VII International Conference on Intelligent Systems Design and Applications*, pages 777–781, Washington, DC, EUA, 2007. IEEE Computer Society.
3. I. Guyon and A. Elisseeff. An introduction to variable and feature selection. *Journal of Machine Learning Research*, 3:1157–1182, 2003.
4. W. H. Kruskal and W. A. Wallis. Use of ranks in one-criterion variance analysis. *Journal of the American Statistical Association*, 47(260):583–621, Dezembro 1952.
5. L. Liu, J. Kang, J. Yu, and Z. Wang. A comparative study on unsupervised feature selection methods for text clustering. In *Proceedings of IEEE International Conference on Natural Language Processing and Knowledge Engineering*, pages 597–601, 2005.
6. T. Liu, S. Liu, and Z. Chen. An evaluation on feature selection for text clustering. In *Proceedings of the XX International Conference on Machine Learning*, pages 488–495, San Francisco, CA, 2003. Morgan Kaufmann.
7. H. P. Luhn. The automatic creation of literature abstracts. *IBM Journal os Research and Development*, 2(2):159–165, 1958.
8. C. D. Manning, P. Raghavan, and H. Schtze. *Introduction to Information Retrieval*. Cambridge University Press, New York, NY, EUA, 2008.
9. B. M. Nogueira, M. F. Moura, M. S. Conrado, R. G. Rossi, R. M. Marcacini, and S. O. Rezende. Winning some of the document preprocessing challenges in a text mining process. In *Anais do IV Workshop em Algoritmos e Aplicações de Mineração de Dados - XXIII Simpósio Brasileiro de Banco de Dados*, pages 10–18. SBC, 2008.
10. M. F. Porter. An algorithm for suffix stripping. *Readings in Information Retrieval*, pages 313–316, 1997.
11. G. Salton and C. Buckley. Term weighting approaches in automatic text retrieval. Technical report, Ithaca, NY, EUA, 1987.
12. G. Salton and M. J. McGill. *Introduction to Modern Information Retrieval*. McGraw-Hill, Inc., New York, NY, EUA, 1986.
13. G. Salton, C. S. Yang, and C. T. Yu. A theory of term importance in automatic text analysis. *Journal of the American Association Science*, 1(26):33–44, 1975.
14. C. J. van Rijsbergen. *Information Retrieval*. Butterworth, 1979.
15. I. H. Witten and E. Frank. *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann Series in Data Management Systems. Morgan Kaufmann, second edition, 2005.
16. Z. Xu, R. Akella, M. Ching, and R. Tang. Semi-supervised clustering using bayesian regularization. In *ICDMW '07: Proceedings of the VII IEEE International Conference on Data Mining Workshops*, pages 361–366, Washington, DC, EUA, 2007. IEEE Computer Society.
17. Y. Yang and J. O. Pedersen. A comparative study on feature selection in text categorization. In *Proceedings of XIV International Conference on Machine Learning*, pages 412–420, Nashville, US, 1997. Morgan Kaufmann, San Francisco, US.
18. L. Yu and H. Liu. Efficient feature selection via analysis of relevance and redundancy. *Journal of Machine Learning Research*, 5:1205–1224, 2004.
19. G. K. Zipf. *Human Behavior and the Principle of Least Effort*. Addison-Wesley, 1949.