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

Bruno Magalhães Nogueira[1], Maria Fernanda Moura[2], Merley Silva Conrado[1], and Solange Oliveira Rezende[1]

[1] Universidade de São Paulo, Instituto de Ciências Matemáticas e de Computação. Caixa Postal 668, São Carlos, SP, Brasil - 13560-970
[brunomn,merleyc,solange]@icmc.usp.br
[2] Embrapa Informática Agropecuária. Caixa Postal 6041, Campinas, SP, Brasil - 13083-970
fernanda@cnptia.embrapa.br

Resumo Selecionar atributos em tarefas de Mineração de Textos é uma atividade essencial para viabilizar a obtenção de resultados válidos e compreensíveis. Em tarefas não-supervisionadas, esta é uma atividade mais difícil devido à falta de medidas objetivas para avaliar os subconjuntos gerados. Assim, no presente trabalho é avaliada a aplicação de seis diferentes métodos não-supervisionados para seleção de atributos em Mineração de Textos, sendo um destes proposto pelos autores. Estes métodos foram aplicados a coleções rotuladas, desconsiderando os valores de classe para a seleção dos subconjuntos. A representatividade dos termos foi medida por meio da acurácia de dois classificadores: C4.5 e SVM. Os resultados mostram que os métodos possuem um bom comportamento, especialmente aqueles baseados em similaridade de documentos e variância dos termos.
Palavras-chave: seleção de atributos, mineração de textos.

1 Introdução

Em um contexto em que cada vez mais dados textuais são armazenados pelas mais diferentes instituições, a Mineração de Textos (MT), por meio de técnicas computacionais de extração de conhecimento, atua como um agente transformador, obtendo dessa enorme quantidade de dados textuais conhecimento útil que pode ser usado como vantagem competitiva e suporte à tomada de decisão. O processo de Mineração de Textos pode ser visto como um caso particular de Mineração de Dados e é composto por cinco etapas: identificação do problema, pré-processamento, extração de padrões, pós-processamento e utilização do conhecimento. Cada um desses passos refere-se a etapas genéricas, as quais podem ser instanciadas de acordo com os objetivos do processo [10].

Dentre estas etapas, uma que demanda especial atenção é a etapa de pré-processamento. Embora seja uma etapa erroneamente considerada por muitos como de menor importância devido à falta de glamour técnico e excesso de trabalho manual, é no pré-processamento que se garante a qualidade dos dados

analisados, tratando os possíveis problemas encontrados e, conseqüentemente, assegurando maior fidelidade aos resultados obtidos pelos algoritmos de aprendizado. No pré-processamento, é também efetuada a estruturação dos documentos, transformando-os para um formato processável por algoritmos de aprendizado de máquina. A forma mais comum utilizada para essa estruturação é a abordagem *bag-of-words*, a qual é representada por uma matriz atributo-valor em que cada linha corresponde a um documento e cada coluna corresponde a um termo presente na coleção. Neste tipo de transformação, o número de palavras candidatas a atributos excede o número de documentos em mais de uma ordem de magnitude [3], gerando matrizes esparsas e de alta dimensionalidade.

Em geral, quando lidando com matrizes com estas características, o desempenho de algoritmos de aprendizado cai drasticamente. Além disso, a grande dimensionalidade exige um alto tempo de processamento. Desta maneira, 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 mais representativos da coleção de textos. A escolha de bons métodos de seleção de atributos é vital para uma correta delimitação do domínio do problema, bem como para a melhoria da eficiência dos algoritmos de aprendizado de máquina. Neste sentido, neste trabalho é apresentada uma comparação de métodos para seleção de atributos em Mineração de Textos.

Este trabalho está inserido em um contexto cujo objetivo é a extração de uma taxonomia de tópicos por meio da metodologia proposta por Moura [8]. Essa metodologia extrai uma taxonomia de tópicos a partir de uma coleção de documentos relativa a um domínio específico, possibilitanto a organização hierárquica desses documentos e da informação neles contida. Para tal, utiliza algoritmos de agrupamento hierárquico que provêem uma versão satisfatória da taxonomia, a qual pode ser editada pelo usuário final. Assim, um aspecto importante que deve ser observado no pré-processamento dessa metodologia é que os atributos mais representativos da coleção sejam considerados, buscando eliminar os atributos irrelevantes. Como essa metodologia lida com coleções de documentos não rotulados, as soluções para seleção de atributos apresentadas neste trabalho consistem de métodos não-supervisionados.

Este trabalho está dividido da seguinte forma: na próxima seção, a metodologia de pré-processamento e seu processo de validação são apresentados. Na seção subseqüente, são apresentados os experimentos realizados e seus resultados. Por fim, são discutidas as conclusões e os trabalhos futuros.

2 Metodologia para Pré-processamento

No pré-processamento de coleções de documentos não rotulados para a obtenção de uma organização hierárquica e descritores de tópicos identificados, além de efetuar a estruturação dos dados a fim de que se tornem processáveis por algoritmos de extração de conhecimento, busca-se garantir a representatividade dos termos considerados, por meio de seleção não-supervisionada de atributos.

Entretanto, medir a representatividade de subconjuntos de atributos em tarefas não-supervisionadas não é uma atividade trivial. A fim de possibilitar a avaliação dos subconjuntos de atributos obtidos por diferentes técnicas de seleção, neste trabalho, aplica-se o processo aqui adotado a coleções rotuladas. O objetivo é medir a representatividade dos subconjuntos de termos por meio da obtenção de acurácia em atividades de classificação de textos. É necessário ressaltar que os rótulos dos documentos são desconsiderados na geração e na seleção de atributos, sendo utilizados apenas para a validação dos subconjuntos obtidos. Os passos deste pré-processamento são apresentados nas subseções a seguir.

2.1 Padronização da Coleção

Nesse passo, analisa-se a coleção com a qual se irá trabalhar, principalmente no que remete a dois aspectos: verificação de representatividade da coleção, identificando se os documentos disponíveis são suficientes e representativos sobre o domínio; e se os documentos não possuem problemas, como caracteres corrompidos. Assim, aplica-se uma seqüência de passos de forma recorrente, a fim de que a coleção de textos trabalhada apresente os requisitos necessários.

Em um primeiro momento, efetua-se a **conversão dos documentos**, colocando-os na forma de texto plano, descartando aqueles que não puderem ser convertidos. Logo após, faz-se a **padronização dos caracteres**, removendo todos os caracteres desnecessários, como acentos, pontuação, cedilhas, números e *underlines*. Todos os caracteres restantes são convertidos para a sua forma minúscula. Feito isso, deve-se fazer uma **verificação de informação pré-existente**, como título, autor e idioma, inserindo-os no documento. Por fim, faz-se uma **avaliação** subjetiva da coleção disponível. Se a base de dados disponível for considerada insuficiente, esta deve ser, de alguma maneira, completada. Este processo de padronização da base de dados pode ser repetido várias vezes, até atingir um ponto considerado satisfatório.

2.2 Geração de Atributos

Uma vez que a coleção é considerada satisfatória, passa-se, então, à geração dos atributos que representam a coleção, os quais são termos de interesse do domínio. Considera-se, aqui, termos como palavras simples (*one-grams*) reduzidas ao seu *stem*, isto é, a forma de palavra despida de inflexões.

Este passo pode ser efetuado com o suporte da ferramenta PreText [7], a qual identifica os termos e aplica a eles o processo de *stemming*. Para este processo de *stemming*, esta ferramenta baseia-se no algoritmo de Porter [9], o qual foi adaptado para três idiomas: Português, Espanhol e Inglês. Todos os termos são gerados considerando cada texto como uma *bag-of-words*, sem considerar informações a respeito do contexto em que este se encontra.

2.3 Seleção Não-Supervisionada de Atributos

O número de atributos gerados em tarefas de Mineração de Textos é muito grande e a maioria destes termos está presente em poucos documentos, resul-

tando em uma representação esparsa das freqüências dos termos nos documentos. Assim, o uso de métodos eficazes para seleção de atributos torna-se essencial para a garantia da validade e da eficiência do processo de extração de conhecimento, na medida em que delimita o domínio a ser tratado pelos algoritmos.

O método de seleção de atributos para Mineração de Textos mais comumente usado é o corte de Luhn [6]. Nesse, são sugeridos dois pontos de corte para os atributos de acordo com a freqüência dos termos (*term frequency* - TF: número de ocorrências um termo em um determinado documento). Para encontrar estes pontos de corte, gera-se o histograma da freqüência dos termos de forma descendente, adotando como pontos de corte os dois pontos de inflexão da curva de tendência. Entretanto, estes pontos não são exatos, ficando a cargo da subjetividade do especialista aproximá-los.

Outro método bastante difundido é o de Salton [11], o qual usa a medida de DF (*document frequency*: número de documentos nos quais um determinado termo aparece) para a seleção dos termos. Nele é sugerido considerar termos que possuam DF entre 1% e 10% do número total de documentos, sendo considerado um corte bastante agressivo, reduzindo muito o número de termos.

Aproveitando as idéias dos cortes de Luhn e Salton, propõe-se neste artigo a utilização de um novo método, o qual denominou-se "Luhn-DF". Neste, gera-se os histogramas das DF dos termos de forma descendente, efetuando os cortes nos pontos de inflexão da curva de tendência, tal qual o primeiro método supracitado. Este método seleciona, assim como o de Salton, termos cuja DF não é tão grande, nem tão pequena, sendo, porém, menos agressivo.

Além destes três métodos, neste trabalho são comparados outros três métodos de seleção de atributos, os quais fornecem *rankings* de atributos, a saber:

1. *Term Contribution* - TC [5]: representa o quanto um termo contribui para a similaridade entre documentos na coleção. É medida de acordo com a Eq. 1:

$$TC(t_k) = \sum_{i,j \cap i \neq j} f(t_k, D_i) * f(t_k, D_j) \tag{1}$$

   na qual, $f(t_k, D_i)$ é a TF-IDF do *k-ésimo* termo no *i-ésimo* documento.
2. *Term Variance* - TV [4]: esta medida calcula a variância de todos os termos da coleção, atribuindo os maiores *scores* àqueles termos que não possuem baixa freqüência em documentos e possuem uma distribuição não-uniforme ao longo da coleção. Pode ser expressa como mostrado na Eq. 2:

$$v(t_i) = \sum_{j=1}^{n} \left[f_{ij} - \bar{f}_i \right]^2 \tag{2}$$

   na qual f_{ij} é a freqüência do *i-ésimo* termo no *j-ésimo* documento e $\bar{f}_i$ é a média das freqüências do *i-ésimo* termo ao longo da coleção.
3. *Term Variance Quality* - TVQ [2]: é bastante similar à TV, usando a variância no cálculo da qualidade dos termos, como mostrado na Eq. 3

$$q(t_i) = \sum_{j=1}^{n} f_{ij}^2 - \frac{1}{n} \left[\sum_{j=1}^{n} f_{ij} \right]^2 \tag{3}$$

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

2.4 Validação dos Subconjuntos de Atributos

Com os atributos selecionados, constrói-se a matriz atributo-valor resultante. A cada célula interna c_{ij} é atribuído o valor da freqüência absoluta do *j-ésimo* termo no *i-ésimo* documento. Adiciona-se, então, uma última coluna que contém os valores dos rótulos de cada documento.

A fim de estabelecer uma avaliação não-subjetiva dos resultados do pré-processamento e obter uma validação dos subconjuntos de atributos obtidos, optou-se por utilizar coleções rotuladas, aplicando sobre cada subconjunto de atributos dois classificadores largamente utilizados: árvores de decisão C4.5 e *Support Vector Machines*, utilizando *10-fold cross validation* para medir a acurácia dos mesmos. É importante ressaltar que somente neste momento do processo se utiliza o valor dos rótulos, de maneira que todas as etapas do pré-processamento aqui apresentadas são aplicados às coleções desconsiderando os valores dos rótulos. Embora seja uma análise simplificada se comparada às técnicas de análise de agrupamentos, a análise de acurácia de classificadores é utilizada em outros trabalhos na literatura, como [1] e [12], podendo-se considerar os resultados limitados, porém válidos.

3 Experimentos e Resultados

Nesta seção, os experimentos efetuados na avaliação dos seis métodos de seleção de atributos apresentados na Seção 2.3 são exibidos.

3.1 Bases de Dados Utilizadas

Três bases de dados textuais compostas por artigos científicos de diferentes domínios foram utilizadas. A primeira possui artigos em Português do Instituto Fábrica do Milênio (IFM)[3], uma instituição brasileira cujo foco é a busca por soluções manufatureiras para as necessidades das indústrias. A segunda é composta por documentos em Inglês sobre quatro subdomínios de Inteligência Artificial: *Case Based Reasoning, Inductive Logic Programming, Information Retrieval* e *Sonification* (CIIS)[4]. A última base de dados contém artigos em Inglês sobre outros cinco subdomínios de Inteligência Artificial: *Agents & Multiagents*, *Fuzzy Logic*, *Machine Learning*, *Planning & Scheduling* e *Robotics* (IA)[5].

A todas essas bases de textos foi aplicado o pré-processamento apresentado na Seção 2. Após isso, aplicou-se a essas bases as técnicas de seleção de atributos previamente discutidas na Seção 2.3. Um sumário do resultado do pré-processamento dessas bases, bem como o número de atributos selecionados pelas técnicas que fornecem um ponto de corte, podem ser observados na Tabela 1.

[3] http://www.ifm.org.br
[4] http://infoserver.lcad.icmc.usp.br/infovis2/PEx
[5] http://labic.icmc.usp.br/projects

Base	#Docs	#Classes	#Docs. Classe Maj.	#*Stems*	#Luhn	#Salton	#Luhn-DF
IFM	582	4	291	34747	12551	6553	8881
CIIS	681	4	276	20495	6789	2464	3561
IA	500	5	100	72936	18061	7527	7005

Tabela 1. Descrição das bases textuais.

3.2 Validação dos Subconjuntos de Atributos

Os resultados aqui expressos foram obtidos utilizando a ferramenta WEKA [13] para induzir os classificadores, adotando os parâmetros *default* e *10-fold cross validation*. Na Tabela 2 é possível observar os resultados de validação obtidos. É importante ressaltar que pretende-se apenas avaliar como diferentes algoritmos de seleção de atributos se comportam sob um mesmo processo de avaliação, não comparando o desempenho dos classificadores. Para as técnicas que não fornecem um ponto exato de corte, ordenou-se os atributos de acordo com o *ranking* por eles obtidos e variou-se a porcentagem do número de atributos selecionados de 10% a 90% do total de atributos, sempre com um incremento de 10%. Por limites de espaço, mostra-se na referida tabela somente os resultados com os subconjuntos de 10, 50 e 90%. Para efeito de comparação, assume-se como método base o de Luhn, por ser este o mais freqüentemente usado na área.

		Luhn	Salton	Luhn-DF	Perc.	TC	TV	TVQ
C4.5	IFM	83.33±3.38	78.68±5.75	87.10±4.35	10%	83.49±3.88	82.12±5.24	82.47±4.89
					50%	86.42±3.41	86.42±3.41	86.42±3.41
					90%	86.25±3.65	86.25±3.65	86.25±3.65
	CIIS	84.74±4.99	84.30±4.75	86.82±2.91	10%	92.89±3.04	92.89±3.04	93.04±2.90
					50%	92.75±3.00	92.75±3.00	92.75±3.00
					90%	92.75±3.00	92.75±3.00	92.75±3.00
	IA	71.80±6.76	71.80±4.66	71.20±4.54	10%	93.00±4.14	92.60±3.41	92.60±3.41
					50%	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
SVM	IFM	78.85±3.52	74.56±5.23	78.00±4.25	10%	77.83±4.67	76.45±5.44	77.31±5.59
					50%	79.37±4.65	79.02±4.71	79.37±4.20
					90%	78.34±4.22	77.48±3.98	78.34±4.22
	CIIS	94.37±1.69	93.03±2.23	93.03±2.12	10%	95.85±2.30	95.56±2.71	94.82±3.23
					50%	95.56±1.72	95.55±1.72	95.56±1.72
					90%	94.66±2.14	94.51±2.45	94.66±2.14
	IA	82.40±7.59	80.00±7.12	80.40±7.04	10%	83.40±5.97	82.60±6.26	82.00±6.11
					50%	85.00±7.13	83.20±7.79	85.00±7.13
					90%	84.00±8.22	85.00±7.62	84.00±8.22

Tabela 2. Acurácia em porcentagem para a validação dos subconjuntos de atributos.

Por meio dos resultados, é possível perceber que todos os métodos aplicados apresentam um comportamento satisfatório no processo de avaliação. Entre-

tanto, pode-se dizer que os métodos TC, TV e TVQ conseguem uma redução mais eficiente, uma vez que, com um número bem inferior de atributos selecionados em relação aos demais, estes métodos conseguem acurácia, se não melhor, ao menos equiparada. Em processos de Mineração de Textos esta é uma característica de suma importância, pois uma maior redução do espaço de atributos, mantendo a representatividade, possibilita processos mais rápidos e eficientes.

Quando esses três cortes são comparados ao método de Salton, nota-se que esses conseguem a mesma eficiência com aproximadamente o mesmo número de atributos. Porém, na análise subjetiva preliminar dos *stems* não selecionados, foi possível perceber que o corte de Salton elimina grande quantidade de *stems* considerados fortes candidatos a descritores, o que não é interessante, de acordo com o contexto deste trabalho. Quanto ao método Luhn-DF, embora esse selecione um número de atributos superior ao demandado pelos métodos TC, TV e TVQ para obter acurácia semelhante, esse torna-se um método interessante na medida em que possui um custo computacional muito inferior a esses três. Quando comparado aos métodos com um custo computacional semelhante (Luhn e Salton), Luhn-DF consegue, mantendo a representatividade, selecionar menos atributos que o primeiro e eliminar menos candidatos a descritores que o segundo.

4 Conclusões

Selecionar bons subconjuntos de atributos, especialmente em tarefas não-supervisionadas, não é um trabalho trivial. Até hoje, são encontrados poucos trabalhos na área de seleção não-supervisionada de atributos voltados para Mineração de Textos. Neste trabalho, foram apresentados seis eficientes métodos para tal tarefa, sendo um deles aqui proposto, o qual foi denominado "Luhn-DF".

Analisando os resultados, é possível perceber que métodos baseados em variância de termos e similaridade de documentos são boas escolhas. Os métodos expostos (TC, TV e TVQ) mostraram uma tendência a selecionar melhores subconjuntos, uma vez que com menores conjuntos de atributos conseguem obter resultados equiparados às outras técnicas, especialmente quando o classificador utilizado é o C4.5. Entretanto, todas as técnicas comparadas apresentaram bons resultados, mostrando-se opções válidas para seleção de atributos de acordo com a situação. Por exemplo, em casos em que o número de atributos for muito grande e a aplicação de técnicas computacionalmente mais custosas for inviável, técnicas como Luhn ou Luhn-DF podem ser úteis.

Quanto ao método proposto (Luhn-DF), é possível perceber que esse geralmente seleciona um número de atributos menor que o corte de Luhn tradicional, mantendo a representatividade. Além disso, por ser um corte menos agressivo que o de Salton, o corte proposto elimina menos *stems* considerados fortes candidatos a descritores dos conteúdos dos textos.

Em trabalhos futuros, pretende-se adicionar outras bases de dados à análise, bem como outras técnicas de seleção de termos. Neste sentido, será testada a variação dos percentuais de DF e TF para seleção de termos. Além disso, planeja-se realizar uma avaliação subjetiva da representatividade dos termos por meio de

especialistas dos domínios. Por fim, pretende-se adicionar a noção de contexto aos termos, aplicando pesos diferentes para termos de acordo com a parte do texto em que estes ocorrem.

Agradecimentos

Os autores agradecem ao CNPq, à Capes e ao Projeto IFM pelo apoio.

Referências

1. M. Dash, H. Liu, and J. Yao. Dimensionality reduction of unsupervised data. In *ICTAI '97: Proceedings of the 9th International Conference on Tools with Artificial Intelligence*, pages 53–5392, Washington, DC, USA, 1997. IEEE Computer Society.
2. 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.
3. G. Forman. An extensive empirical study of feature selection metrics for text classification. *The Journal of Machine Learning Research*, 3:1289–1305, 2003.
4. L. Liu, J. Kang, J. Yu, and Z. Wang. A comparative study on unsupervised feature selection methods for text clustering. *Natural Language Processing and Knowledge Engineering, 2005. IEEE NLP-KE '05. Proceedings of 2005 IEEE International Conference on*, pages 597–601, 30 Oct.-1 Nov. 2005.
5. T. Liu, S. Liu, Z. Chen, and W.-Y. Ma. An evaluation on feature selection for text clustering. In *Proceedings of the Twentieth International Conference on Machine Learning (ICML)*, pages 488–495. AAAI Press, 2003.
6. H. P. Luhn. The automatic creation of literature abstracts. *IBM Journal os Research and Development*, 2(2):159–165, 1958.
7. E. Matsubara, C. Martins, and M. Monard. Pretext: uma ferramenta para pré-processamento de textos utilizando a abordagem bag-of-words. Technical Report 209, Inst. de Ciências Matemáticas e de Computação – USP – São Carlos, 2003.
8. M. F. Moura. Uma abordagem para a construção e atualização de taxonomias de tópicosMonografia de Qualificação de Doutorado, Inst. de Ciências Matemáticas e de Computação – USP – São Carlos.
9. M. Porter. An algorithm for suffixing stripping. *Program*, 14(3):130–137, 1980.
10. S. O. Rezende, J. B. Pugliesi, E. A. Melanda, and M. F. Paula. Mineração de dados. In S. O. Rezende, editor, *Sistemas Inteligentes: Fundamentos e Aplicações*, chapter 12, pages 307–335. Manole, 1 edition, 2003.
11. 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.
12. N. Wiratunga, R. Lothian, and S. Massie. Unsupervised feature selection for text data. In *ECCBR 2006: Proceedings of 8th European Conference on Case-Based Reasoning*, pages 340–354, Heidelberg, 2006. Springer Berlin.
13. 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, June 2005.