Estudo de um critério para indexação automática derivativa de textos científicos e tecnológicos

Vânia Lisbôa da Silveira Guedes

Resumo

Este trabalho é uma contribuição ao estudo da indexação automática fundamentado na íreqüência de ocorrência de palavras. Investiga a aplicação da 1ª e 2ª leis de Zipf e Ponto T de Goffman, em 11 notas técnicas de mecânica dos solos, engenharia civil. Entretanto, sugeriram a formulação de uma lei, aqui, denominada Lei da Amplitude de Variação de r.f. Além disso, através de um estudo estatístico da freqüência de ocorrência de palavras, mostra-se que é possível estimar a freqüência da palavra de maior ocorrência, e a distribuição de r.f. de um texto. Sugere-se, também, um critério matemático para delimitar a Região de Transição de Goffman, onde há concentração de palavras de alto conteúdo semântico. Finalmente, propõem-se futuras investigações para ratificar a lei de formação, citada anteriormente, e o critério de delimitação da Região de Transição, com o objetivo de contribuir para um mecanismo básico de indexação de textos científicos e tecnológicos.

Palavras-chave

Indexação automática, Recuperação da informação; Freqüência de palavras, Região de transição de Goffman.

Síntese da dissertação de mestrado em ciência da informação, aprovada pela Escola de Comunicação da Universidade Federal do Rio de Janeiro, em março de 1992, sob a orienta-ção da professora Rosali Fernandez de Souza, pesquisadora titular do IBICT/CNPq, e co-orientação do professor Ian Schumann Marques Martins, professor adjunto da Coppe/ UFRJ.

INTRODUÇÃO

A explosão bibliográfica e o tempo limitado de pesquisadores para buscar e assimilar informações são problemas fundamentais para a área de ciência da informação. Por outro lado, particularmente na área científica e tecnológica, a "vida média" de uma unidade da literatura é curta, em relação a outras áreas do conhecimento. Logo, para cientistas e tecnólogos, o rápido e eficiente acesso às informações recentes deve ser privilegiado.

Dentre as pesquisas desenvolvidas na área de ciência da informação, destacam-se aquelas dedicadas à representação da informação. Tal fato se deve à conscientização de que, em sistemas de recuperação da informação, a representação precisa do conteúdo temático de documentos é uma condição *sine-qua-non para a* recuperação de documentos relevantes. Nesse âmbito, como meio de representação da informação, a indexação tem sua função evidenciada.

Maron[1] infere que a capacidade íntima de reconhecer sobre o que trata o documento é a questão central do procedimento de indexação. Ele entende que, para fins de indexação, o(s) termo(s) selecionado(s) é a correlação comportamental sobre o que se pensa "sobre o que o documento trata", pois seria o termo usado para se procurar portal documento.

Entretanto, a "indexação-arte", como diria Borko[2], vem se revelando inadequada para minimizar a subjetividade inerente à indexação. O grau de consistência atingido depende, em grande parte, do grau de conhecimento que o indexador tem sobre o assunto determinado, assim como das políticas estabelecidas e recursos disponíveis. Além disso, o conhecimento é dinâmico, exigindo do indexador permanente atualização. Outro aspecto a considerar refere-se à inconsistência interindexadores (diferentes indexadores atribuindo diferentes termos-índice a um mesmo conceito/documento) e intra-indexador (o mesmo indexador atribuindo diferentes termos-índice a um mesmo conceito/documento, em diferentes momentos).

A indexação automática é a formalização e/ou mecanização do processo de indexação, em parte ou no todo, com o objetivo de reduzir a subjetividade do processo.

Dentre os estudos de indexação automática, são de interesse dessa pesquisa os estudos bibliométricos, fundamentados na freqüência de ocorrência das palavras, principalmente, nas leis de Zipf e Ponto T de Goffman.

Zipf observou que, em um texto suficientemente longo, o produto da ordem de série (r) de uma palavra (dada pela freqüência de ocorrência em ordem decrescente) pela sua freqüência de ocorrência (f) era aproximadamente constante. Enunciou, então, que

$$r.f = c,$$

expressão que ficou conhecida como Primeira Lei de Zipf (Booth[4]).

A Segunda Lei de Zipf enuncia que, em um texto, várias palavras de baixa freqüência de ocorrência (alta ordem de série) aparecem o mesmo número de vezes. Booth[4], ao modificá-la, a representa matematicamente por:

$$\frac{I1}{In} = \frac{n\,(n+1)}{2}$$

onde I1 é o número de palavras que têm freqüência 1, e In, o número de palavras que têm freqüência n.

Guedes e Valois[5] observam que a expressão $\frac{n\,(n+1)}{2}$ corresponde à soma dos n primeiros números naturais. Por exemplo, a soma de 1+2+3+4+5 =15 pode ser calculada levando-se em conta que n = 5, por $\frac{5\,(5+1)}{2}$ = 15. Assim, entende-se que a 2ª Lei de Zipf pode ser enunciada alternativamente como:

"a relação entre o número de palavras que aparecem uma só vez (I1) e o número de palavras que aparecem n vezes (In), em um texto, é igual à soma dos n primeiros números naturais, isto é, 1+2+3+...+n, representada matematicamente por $\frac{I1}{In}$ = 1+2+3 = ... +n".

Os comportamentos, inteiramente distintos, da 1ª e 2ª Leis de Zipf definem as duas extremidades da lista de distribuição de palavras de um texto. Assim, é razoável esperar uma região crítica, na qual há a transição do comportamento das palavras de baixa freqüência para as de alta freqüência.

Para se chegar a essa região de transição, a expressão da 2ª Lei de Zipf teria de fornecer o comportamento típico das palavras de alta freqüência, isto é, o número de palavras que têm freqüência n tenderia a 1. Substituindo-se, na expressão da 2ª Lei de Zipf, In por 1, obtém-se:

$$\frac{I1}{In} = \frac{n\,(n+1)}{2}$$

ou ainda rearranjando $n^2 = n\text{-}2\ I1 = 0$,

cujas raízes são $n = \frac{-1 \pm \sqrt{1+8\ I1}}{2}$

Da expressão anterior, interessa apenas determinar a raiz positiva

$$n = \frac{-1 + \sqrt{1+8\ I1}}{2}$$

Ao valor de n assim determinado dá-se o nome de Ponto de Transição de Goffman (T).

O Ponto de Transição de Goffman determina a vizinhança onde, de acordo com Goffman, devem estar incluídas as palavras de maior conteúdo semântico e, portanto, aquelas que seriam usadas para a indexação de um texto em questão.

Esta linha de raciocínio representa um passo importante na busca de um critério de indexação automática.

ESTUDOS BIBLIOMÉTRICOS DEDICADOS À INDEXAÇÃO DA INFORMAÇÃO

Os primeiros estudos bibliométricos fundamentados em freqüência de ocorrência de palavras, como medida de conteúdo temático, foram os de Luhn (1957) e Baxendale (1958).

Luhn[6,7] desenvolve uma abordagem estatística, com vistas à classificação e busca automáticas de documentos, e um método automático probabilístico, com vistas à criação de *abstracts.* Já Baxendale[8] analise comparativamente a eficiência de três métodos automáticos de indexação de artigos técnicos.

Seguem-se os estudos de Salton[9], dedicados a métodos automáticos de indexação e classificação, de Maia[10] acerca da adequação das Leis de Zipf e Ponto T de Goffman à indexação automática, os de Pinheiro[11], que verificam a aderência do método em questão a um texto literário, e os de Pao[12], que testam o método proposto por Goffman, em um artigo de Booth: *On geometry of libraries.*

Rowbotton e Willett[13] destacam o estudo de Boyce e Lockard, que testa a eficiência do método baseado nas Leis de Zipfe Pon-to T de Goffman para a recuperação de ter-mos-índice, na área médica, o de Kucera e Francis, que discorre acerca da quantidade de termos-índice produzidos, e o de Rowbotton, que avalia esse método para indexação automática.

Finalmente, vale destacar Sepulveda *et alii*[14] , que aplicam as leis de Zipf e Ponto T de Goffman a um texto de engenharia oceânica, em língua inglesa. Guedes e Valois[15], que aplicam o método visando à indexação automática de gêneses distintas de textos de mecânica dos solos; Manfrim e Coelho (*apud* Manfrim[16]), que desenvolvem uma análise comparativa da aplicabilidade desse método a um texto original em inglês e sua tradução para o português; e Manfrim[17], que verifica a possibilidade da indexação automática derivativa de textos de bibliometria, em inglês.

Outros trabalhos foram desenvolvidos, baseados nas leis de Zipf. Dentre esses, segundo Smith e Devine[18], citam-se os de Craig (1983), com vistas à seleção de frases, pares de palavras e expressões com múltiplas palavras.

O presente estudo se desenvolve a partir do potencial apresentado pelos estudos anteriormente citados. Nesse sentido, ele parte da premissa básica de que as freqüências de ocorrência de palavras de um determinado texto são medidas de significância dessas palavras em relação a esse mesmo texto. Em face de tal assertiva, desenvolve-se um estudo visando a um mecanismo matemático básico para o processo de indexação automática de textos científicos e tecnológicos.

MATERIAL

Procedeu-se à seleção de 11 notas técnicas, uma discussão e um artigo de mecânica dos solos. Os textos selecionados estão relacionados a seguir, sendo que a discussão e o artigo funcionam como textos-teste.

Procurou-se selecionar textos cujas palavras-chave poderiam ser facilmente atribuídas, o que foi feito por um especialista da área de mecânica dos solos. São eles:

1. O'REILLY, M. P. BROWN, S. F, OVERY, R. F. Viscous effects obseived in teste on an anisotropically normally consolidated silty day. *Géotechnique,* London, v. 39, n. 1, p. 153-158, 1989.
2. FELDKAMP,J. R Permeability measurement of clay pastes by a non-lineer analysis of transient seepage consolidation tests. *Géotechnique,* London,v.39,n. 1,p. 141-145,1989.
3. COLLINS, I. F. The nature of stress and velocity characteristics for critical stress states. *Géotechnique,* London, v. 40, n. 1,p. 125-129, 1990.
4. BOSSCHER, P. J, ORTIZ, G. C. Frictional properties between sand and various construction materials. *Journal of Geotechnical Engineering,* New York, v. 113, n. 9, p. 1 035-1 039, Sept 1987.
5. WANG, M. C., LIAO, W. P. Active length of laterally loaded piles. *Journal of Geotechnical Engineering,* New York, v. 113, n. 9, p. 1 044-1 047, Sept. 1987.
6. SULLY, J. P., CAMPANELLA, R. G., ROBERTSON, P. K. Over consolidation ratio of clays from penetration pore pressures. Journal of Geotechnical *Engineering,* New York, v. 114, n. 2, p. 209-215. Feb. 1988.
7. HANSMIRE, W. H., RAWNSLEY, R. P. Longterm tieback monitoring at Havard Square. *Journal of Geotechnical Engineering.* New York, v. 114, n. 3, p. 344-348, Mar. 1988.
8. DAKOULAS, P., GAZETAS, G. Vibration characteristics of dams in narrow canyons. *Journal of Geotechnical Engineering,* New York, v. 113, n. 8, p. 899-904, Ago. 1987.
9. PRATO, C.A., DELMASTRO, E. 1-D seismic analysis of embankment dams. *Journal of Geotechnical Engineering,* New York, v. 113, n. 8, p. 904-909, Ago. 1987.

10. DIYALJEE, V.A. Effects of stress histoty on ballast deformation. *Journal of Geotechnical Engineering,* New York, v. 113, n. 8, p. 909-914, Ago. 1987.

11. KAVAZANJIAN Jr., E., MITCHELL, J. K. Time dependence of lateral earth pressure. *Jornal of Geotechnical Engineering,* New York, v. 110, n. 4, p. 530-533, Apr. 1984.

12. LACERDA, W.A., MARTINS, I. S. M. Discussion by Willy A. Lacerda and Ian Schumann Marques Martins. *Journal of Geotechnical Engineering,* New York, v. 111, n. 10, p. 1242-1 244, Oct. 1985.

13. RAMALHO-ORTIGÃO, J. A., WERNECK, M. L. G., LACERDA, W. A. Embankment failure on clay near Rio de Janeiro. *Journal of Geotechnical Engineering,* New York, v. 109, n. 11, p. 1 460-1 479, Nov. 1983.

MÉTODO

Os textos foram digitados em *Wordstar,* obedecendo às seguintes normas:

- considerar textos, títulos, subtítulos de itens, legendas de figuras, legendas de gráficos e tabelas, citações no corpo dos textos, bem como pontos, quando usados para abreviatura de palavras (como, por exemplo, p.d.e. = *partial diferencial equation* (nota técnica 2).

- não considerar palavras-chave, notações, notas de rodapé, referências, agradecimentos, números, letras gregas, equações, expressões algébricas, tabelas, gráficos, ilustrações, autores de texto analisado, pontuações (exceto quando em abreviações e, inclusive, travessão.

Aos textos digitados, aplicou-se um programa de contagem de palavras, segundo sua freqüência de ocorrência, que considera:

- palavra como um conjunto de caracteres precedido e sucedido por espaço em branco, ou pontuação e espaço em branco;

- palavras hifenizadas como palavras únicas;

- diferentes formas flexionadas de uma mesma palavra como palavras distintas.

Para cada texto processado, foi produzida uma lista de freqüência de ocorrências decrescentes das palavras. Nessas listas, cada palavra foi associada à sua freqüência de ocorrência.

Às palavras distintas, com igual freqüência de ocorrência, atribui-se a ordem de série obtida pela média aritmética das ordens de série correspondentes.

Foram calculados a ordem de série (r) das palavras, a freqüência de ocorrência das palavras (f) e o produto r x f.

Foram construídos gráficos r x r.f de cada texto, para se verificar a adequação ou não da 1ª Lei de Zipf.

Para cada texto selecionado, aplicou-se a Fórmula do Ponto T de Goffman, tentando-se a identificação da região de transição.

Foi verificado se a região identificada inclui as freqüências que correspondem às palavras de maior conteúdo semântico, pré-selecionadas pelo especialista da área de mecânica dos solos.

DISCUSSÃO DOS RESULTADOS

Primeira Lei de Zipf

Observa-se que os desvios da 1ª Lei de Zipf podem ser quantificados pela diferença entre os valores máximo e mínimo do produto r.f. Define-se então a variável Δ (r.f) como:

$$\Delta \text{ (r.f)} = \text{(r.f) max.} = \text{(r.f) min.} \qquad (1)$$

Plotando-se os valores de Δ **(r.f) x** número de palavras dos textos estudados, obtém-se o gráfico 1 (ver Anexo).

No gráfico 1, observa-se que, quanto maior o número de palavras, maior o valor de Δ (r.f), isto é, quanto maior o texto, maior o desvio da 1ª Lei de Zipf. Por outro lado, uma outra característica observada, no gráfico 1, é a de que os pontos tendem a se alinhar ao longo de uma reta. Ajustando-se uma reta, pelo método dos mínimos quadrados, aos pontos do gráfico 1, obtém-se a seguinte equação:

$$\Delta \text{ (r.f)} = 53.5 + 1 \times \text{nº palavras} \qquad (2)$$

Para as notas técnicas de mecânica dos solos, parece que a equação do tipo $Y = A + Bx$ satisfaz a relação entre as variáveis Δ (r.f) e o número de palavras.

Observa-se, portanto, que a 1ª Lei de Zipf não se verificou para o conjunto de notas técnicas analisado.

Diante do comportamento verificado, a 1ª Lei de Zipf poderia ser substituída pela Lei de Amplitude de Variação de r.f. proposta pelo autor deste estudo:

"É contado o número de vezes que cada palavra ocorre em um texto. As palavras são ordenadas segundo sua freqüência de ocorrência decrescente. A ordem de qualquer palavra é chamada ordem de série (r), e o número de vezes que ela ocorre, freqüência (f). Calcula-se a diferença entre o r.f máximo e o r.f mínimo **[Δ (r.f)].** Verifica-se que o valor de Δ (r.f) é expresso por Δ (r.f) = A + B número de palavras, sendo A e B constantes."

Os textos 12 e 13, embora tenham sido digitados sem considerar as legendas das figuras, quadros e tabelas, enquadram-se no comportamento descrito anteriormente.

Análise da Distribuição de r.f *versus* log r

Nos gráficos 2 a 14 (em anexo), observa-se que a distribuição r.f *versus* log r pode ser aproximada por uma reta. Assim, pode-se escrever, de forma genérica, para os textos de mecânica dos solos analisados, que:

$$\text{r.f} = c + D \log r, \qquad (3)$$

onde C e D são constantes.

Ajustou-se, para cada texto analisado, utilizando-se o método do mínimos quadrados, uma reta do tipo da equação r.f = C + D log r, e obtiveram-se os valores do coeficiente linear (C) e do coeficiente angular (D).

Nesta equação, observa-se que, para r = 1, log de r = 0. Portanto, o coeficiente linear C representa o produto de r = 1 pela freqüência máxima de cada texto, ou seja, o coeficiente linear C é a própria freqüência máxima que ocorre em cada texto.

É razoável imaginar que o número de artigos e conectivos, em um texto, seja proporcional ao número total de palavras desse texto. Logo, o coeficiente linear C, dessa equação, que representa a freqüência máxima de um texto e que certamente está associado a um artigo ou conectivo, deve ser proporcional ao número total de palavras desse texto.

No gráfico 15 (em anexo), observa-se a proporcionalidade entre o valor de C (freqüência da palavra de r = 1) e o número total de palavras de cada texto analisado. Ajustando-se pelos pontos do gráfico 15 uma reta, utilizando-se o método dos mínimos quadrados, obtém-se a equação:

$$\text{F máx.} = C = 16{,}82 + 0{,}07 \text{ n° palavras} \qquad (4)$$

Entretanto, à medida que o número total de palavras decresce, tendendo a 0 (zero), o número de contectivos também tende a 0 (zero). Assim, a relação entre a freqüência máxima e o número de palavras de um texto pode ser inter-

pretada como uma reta passando pela origem. **Deste** modo, pode-se ajustar, **pelos pontos do gráfico 15, a reta da** equação:

F máx. = C = O + 0,078 x n° de palavras (5)

Para todos os textos analisados, observa-se que a freqüência máxima (r = 1) vem associada ao artigo "THE" e que a equação 5 se ajusta muito bem aos pontos do gráfico 15.

Os dois textos-teste, que são, dentre os textos analisados, os de menor e maior número de palavras, seguem o mesmo comportamento do conjunto de notas técnicas analisadas.

No gráfico 16 (em anexo), observa-se uma interdependência entre o coeficiente D e o número total de palavras, dos textos analisados. Os pontos do gráfico 16 podem ser aproximados por uma reta, significando que, quanto maior o número de palavras do texto, maior o coeficiente angular D da equação r.f = D log r. Assim, ajustando-se uma reta, pelo método dos mínimos quadrados, aos pontos do gráfico 16, obtém-se a equação:

D = 32,13 + 0,0368 n° de palavras (6)

Substituindo-se as expressões de C e D na equação r.f = C + D log r, obtém-se uma expressão da distribuição de r.f, apenas em função do número total de palavras, qual seja

r.f = (0,078 x n° de palavras) + [32,13 + (0,0368 x n° de palavras)] log r

Assim, a distribuição de freqüências e de *ranks,* de textos de mecânica dos solos, pode ser aproximadamente prevista, baseada apenas no número total de palavras de um texto em questão.

Palavras de alto conteúdo semântico em relação ao tema dos textos analisados

No que tange à Fórmula do Ponto T de Goffman, os resultados obtidos indicaram, na maior parte dos textos, uma proximidade do Ponto T com as freqüências associadas às palavras de alto conteúdo semântico. **Essa** assertiva é reforçada pelo fato de que essas palavras incluem-se entre as palavras-chave atribuídas aos textos analisados por um especialista em mecânica dos solos. As notas técnicas 3, 4, 8 e 10 apresentaram maior aderência ao Ponto T, ou seja, nessas notas técnicas, o Ponto T corresponde exatamente às freqüências associadas às palavras de alto conteúdo semântico. Entretanto, para efeito de indexação automática, o critério não se aplica, pois não pode ser generalizado.

A partir de então, decidiu-se analisar em que faixa de ordem de série se encontram as três palavras de maior conteúdo semântico, de maior freqüência de cada texto.

Em primeiro lugar, observa-se que nenhuma palavra de alto conteúdo semântico é encontrada com ordem de série igual ou inferior a 3.

Em segundo lugar, com o objetivo de identificar uma região de concentração de ocorrências de palavras-chave, analisam-se as relações entre as freqüências de ocorrência dessas palavras, suas ordens de série e o número total de palavras de cada texto. Para tal, apresenta-se o quadro 1, onde estão tabuladas as relações entre as ordens de série e freqüência de ocorrência das primeiras e terceiras palavras de alto conteúdo semântico, com o *rank* máximo e freqüência máxima de cada texto.

Os valores da relação f pc/f max. também não sugerem nenhuma tendência, o mesmo se aplicando para a relação f pc3/f max.

CONCLUSÕES E SUGESTÕES PARA FUTURAS PESQUISAS

Conclusões

Para os textos analisados, a 1ª Lei de Zipf não se verificou. Em contrapartida, verificou-se a Lei de Amplitude de Δ (r.f), proposta pelo autor deste estudo.

Os gráficos r.f x log r podem ser representados aproximadamente por uma reta, ou seja, pela equação r.f = C + C log r, em que C e D são funções lineares do número total de palavras de cada texto. Assim, as distribuições dos valores de *rank* e freqüência podem ser calculadas, para os textos analisados, a partir apenas do total de palavras de cada texto.

Quadro 1 - Relação entre os *ranks e* freqüências da primeira e terceira palavras de alto conteúdo semântico e o *rank e* freqüências máximos de cada texto

Texto	$\frac{r_{pc}1}{r_{max}}$	$\frac{r_{pc}3}{r_{max}}$	$\frac{f_{pc}1}{f_{max}}$	$\frac{f_{pc}3}{f_{max}}$	$\frac{r_{max}}{\text{Nº palavras}}$	$\frac{\text{nº}}{\text{palavras}}$
NT 1	2,17%	6,76%	20,5%	9,27%	19,7%	2 104
NT 2	3,51%	6,06%	8,21%	5,8%	21,7%	2 227
NT 3	2,99%	5,3 %	11,1%	6,8%	18,6%	2 516
NT 4	1,6 %	3,64%	20,5%	14,5%	22,3%	1 111
NT 5	1,46%	8,33%	28,46%	8,13%	19,59%	1 225
NT 6	0,91%	5,19%	28,2%	10,1%	20,23%	1 619
NT 7	2,41%	7,98%	26,1%	11,59%	26,09%	1 033
NT 8	2,49%	5,93%	25,3%	13,25%	24,12%	1 082
NT 9	6,15%	9,56%	7,08%	5,47%	22,23%	1 388
NT 10	1,67%	5,42%	39,25%	16,82%	17,95%	1 334
NT 11	1,76%	8,46%	40,58%	10,14%	25,34%	886
ART.	0,82%	2,06%	20,7%	8,6%	17,76%	4 771
DISC.	9,27%	12,75%	14,89%	10,64%	28,47%	606

r pc 1 = ordem de série da 1ª palavra de alto conteúdo semântico
r pc 3 = ordem de série da 3ª palavra de alto conteúdo semântico
r max = ordem de série máxima
f pc 1 = freqüência da primeira palavra de alto conteúdo semântico
f pc 3 = freqüência da terceira palavra de alto conteúdo semântico
f máx = freqüência máxima

Observa-se que os valores da relação r. pc1/r. max. não sugerem nenhuma tendência, talvez pelo baixo número de textos analisados. Por outro lado, os valores da relação r. pc3/r. max., com exceção da discussão, que possui apenas 606 palavras, não superam o valor de 10%.

Assim, pode-se definir a região de concentração de palavras de alto conteúdo semântico como aquela para a qual 5 > 3 e r. pc3/r max , 10%.

A aplicação da 2ª Lei de Zipf não se verificou satisfatoriamente.

Embora o Ponto T de Goffman tenha indicado, na maior parte dos textos, uma proximidade com as freqüências associadas às palavras de alto conteúdo semântico, o critério não se aplica à indexação automá-

tica dos textos analisados, já que não pode ser generalizado.

Para efeito de indexação, a região de concentração de palavras de alto conteúdo semântico situa-se na faixa de ordem de série tal que r > 3 e r.pc/r max < 10%.

Sugestões

Analisar um maior número de textos de mecânica dos solos, com vistas a ratificar a lei de formação, anteriormente citada, e reduzir a amplitude da região de ocorrência das três palavras-chave de maior freqüência. Nessa oportunidade, optar pela análise de conjuntos distintos, segundo tipo de documentos, tais como artigos, notas técnicas, discussões etc.

Aplicar esta análise para textos em áreas que não a de mecânica dos solos.

Study of a criteria for automatic indexing of scientific and technological texts

Abstract

This work is a contribution to the study of automatic indexing, based on word frequency of occurrences in texts. It investigates the application of Zipf's First and Second Laws and Goffman's Transition Point, in eleven technical notes of Soil Mechanics, Civil Engineering. The results showed a non-*conformity with the Zipfs Laws. However, it suggests the formulation of a new law here named The Amplitude of Variation of r.f. Beyond this, it is showed that the word of maximum occurrence and the distribution of r.f. can bem achieved through an statistical study of word frequency of occurrences. It also suggests a mathematical approach to define the transition region, proposed by Goffman, where the concentration of words of high semantic content probably occurs, i.e., those words most suitable as indexing terms. Finally, future research is suggested to ratify the obtained results and to improve the process in order to use it as a tool for automatic indexing of scientific and technological texts.*

Keywords:

Automatic indexing; Information retrievar, Word frequency. Goffman's transition point region: Scientific and technological texts.

REFERÊNCIAS BIBLIOGRÁFICAS

1. MARON, M. E. On indexing, retrieval and the meaning of about. *Journal of the American Society for Information Science,* New York, v. 28, n. 1, p. 38-43, Jan. 1977.

2. BORKO, Harold. Toward a theory of indexing. *Information Processing and Management.* Oxford, v. 13, p. 355-365, 1977.

3. BOOTH, Andrew D. A "law" of occurrences for words of low frequency. *Information and Control,* [s.1.],v. 10, n. 4, p. 366-393, Apr. 1967.

4. Idem.

5. GUEDES, Vânia Lisboa da Silveira, VALOIS, Eliana Candeira. *Adequação das Leis de Zipf (1ª e 2ª) e PontoT de Goffman à indexação de documentos científicos:* uma aplicação em mecânica dos solos (engenharia civil). Rio de Janeiro, 1988. Trabalho não publicado, apresentado à disciplina de bibliometria da ECO/UFRJ. 36 p.

6. LUHN, H. P. A statistical approach to mechanized encoding and searching of literary information. IBM Journal of *Research and Development,* New York, v. 1, n. 4, p. 309-317, Oct. 1957.

7. The automatic creation of literature abstracts *IBM Journal of Research and Development,* New York, v. 2, n. 2, p. 159-165, Apr. 1958.

8. BAXENDALE, P. B. Machine-made index for technical literature - an experiment. *IBM Journal of Research and Development,* New York, v. 2, p. 354-361, Oct. 1958.

9. SALTON, G. Automatic text analysis. *Science,* [s. 1.], v. 168, n. 3 929, p. 335, 1970.

10. MAIA, Elza Lima e Silva. *Comportamento bibliométrico da língua portuguesa, como veículo de representação da informação.* Rio de Janeiro: IBICT/UFRJ, 1973. 57 p. Dissertação (Mestrado em Ciência da Informação).

11. PINHEIRO, Lena Vania Ribeiro. *Estudo bibliométrico em língua literária.* Rio de Janeiro, 1977. 18 p. Trabalho não publicado, apresentado à disciplina Sistemas de Recuperação da Informação do IBICT/CNPq-ECO/UFRJ.

12. PAO, Miranda Lee. Automatic text analysis based on transition phenomena of occurences. *Journal of the American Society for Information Science,* New York, v. 29, n. 3, p. 121-124, May 1978.

13. ROWBOTTON, M. E., WILLETT, P. The effect of subject matter on the automatic indexing of full text. *Journal of the American Society for Information Science.* New York, v. 33, n. 3, p. 139-141, May 1982.

14. SEPÚLVEDA, G. M., COSTA, M. F. T. J. F., CUNHA, M. R. A. Aplicação da Lei de Zipf em um texto de proteção costeira através do uso de microcomputador. In: CONGRESSO BRASILEIRO DE BIBLIOTECONOMIA E DOCUMENTAÇÃO, 14., 1987, Recife. *Anais...* Recife: Comissão brasileira de Documentação em Processos Técnicos, 1987. v. 2, p. 481-496.

15. GUEDES, Vânia Lisboa da Silveira, VALOIS, Eliana Caldeira, op. cit., 1988.

16. MANFRIM, F. P. B. *Indexação automática derivativa em textos integrais em língua portuguesa.* Rio de Janeiro: IBICT/UFRJ, 1990. 250 p. Dissertação (Mestrado em Ciência da Informação).

17. Idem.

18. SMITH, F. J., DEVINE, K. Storing and retrieving word phrases. Information *Processing and Management,* Oxford, v. 21, n. 3, p. 215-224, 1985.

Artigo aceito para publicação em 25 de outubro de 1994.

Vânia Lisbôa da Silveira Guedes

Mestre em Ciência da Informação pela Escola de Comunicação (ECO), da Universidade Federal do Rio de Janeiro (UFRJ), Bibliotecária-documentalista da Coordenação dos Programas de Pós-Graduação de Engenharia (Coppe), da UFRJ.

ANEXOS

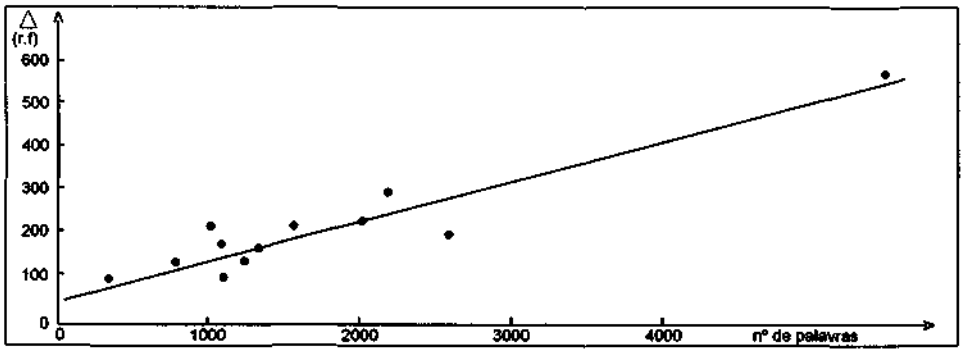

Gráfico 1 – Distribuição dos valores de △ (r.f) *versus* número de palavras dos textos analisados

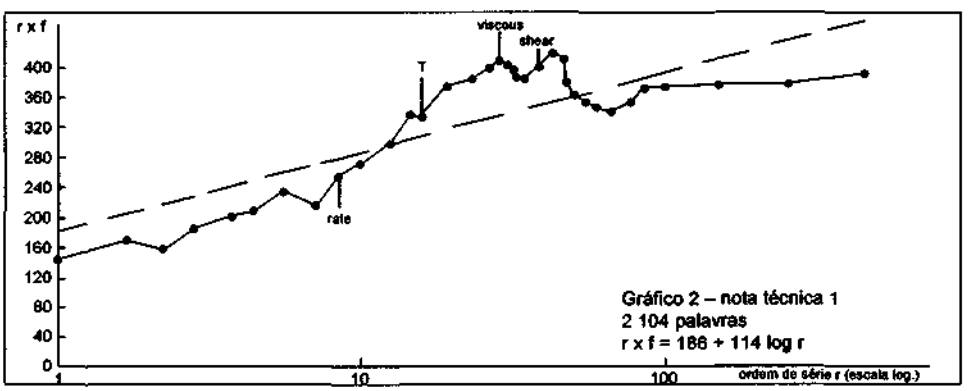

Gráfico 2 – Distribuição dos valores de r x f *versus* log. r da Nota Técnica 1

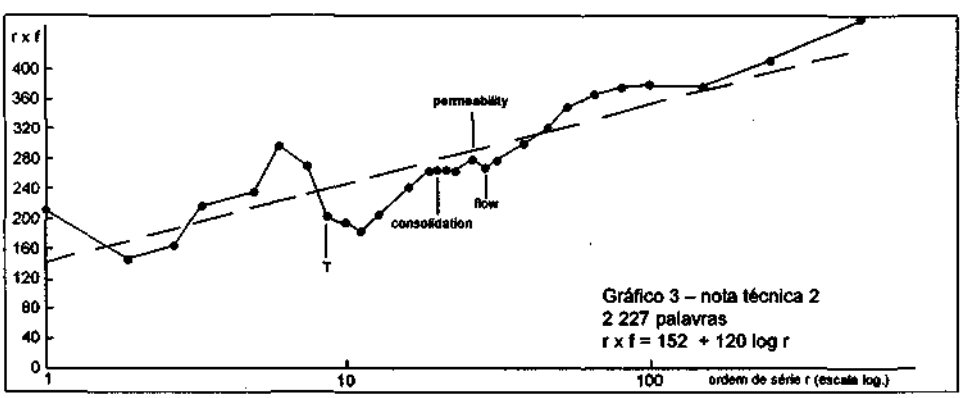

Gráfico 3 – Distribuição dos valores de r x f *versus* log r da Nota Técnica 2

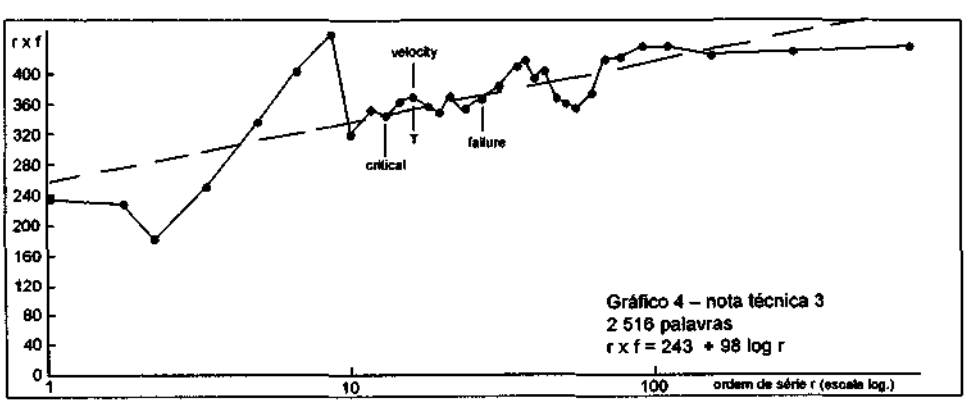

Gráfico 4 – Distribuição dos valores de r x f *versus* log r da Nota Técnica 3

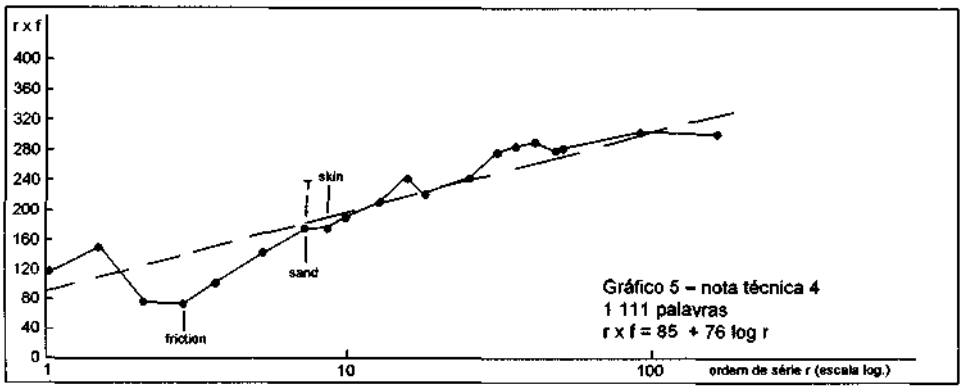

Gráfico 5 - Distribuição dos valores de r x f versus log r da Nota Técnica 4

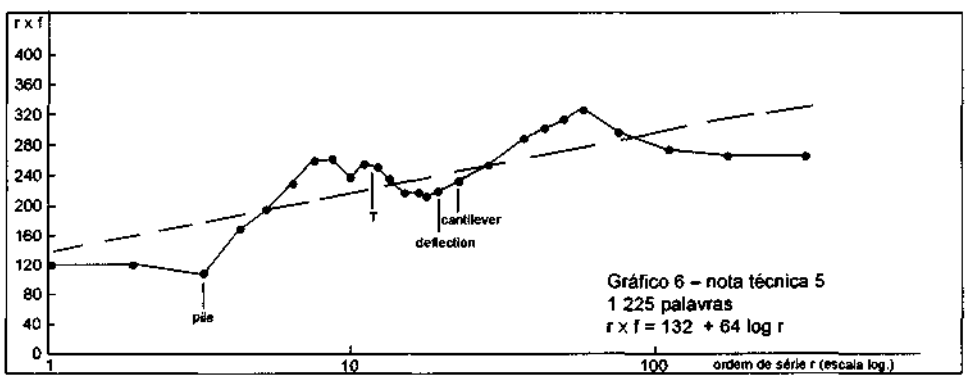

Gráfico 6 - Distribuição dos valores de r x f *versus* log r da Nota Técnica 5

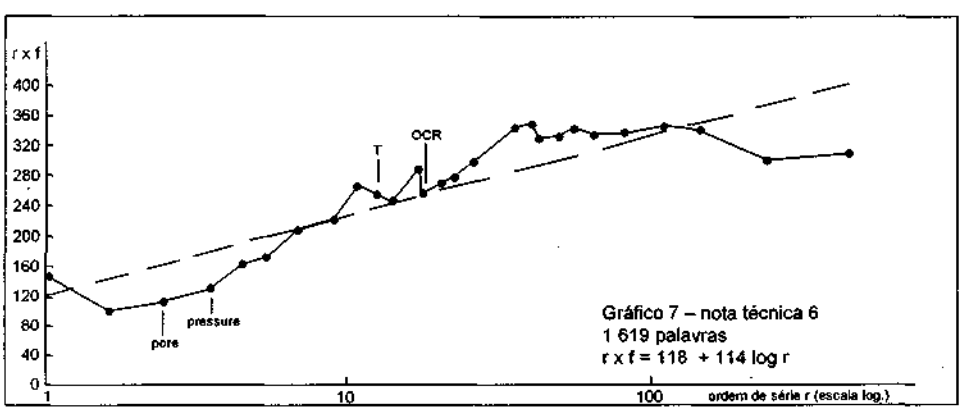

Gráfico 7 - Distribuição dos valores de r x f *versus* log r da Nota Técnica 6

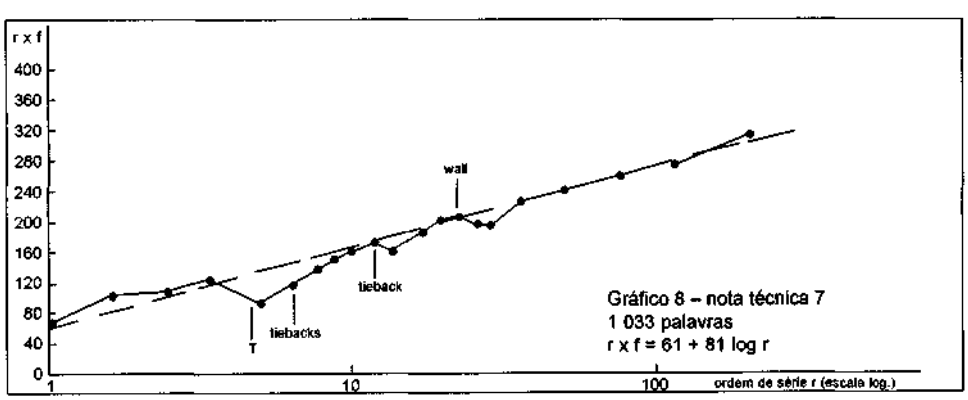

Gráfico 8 - Distribuição dos valores de r x f *versus* log r da Nota Técnica 7

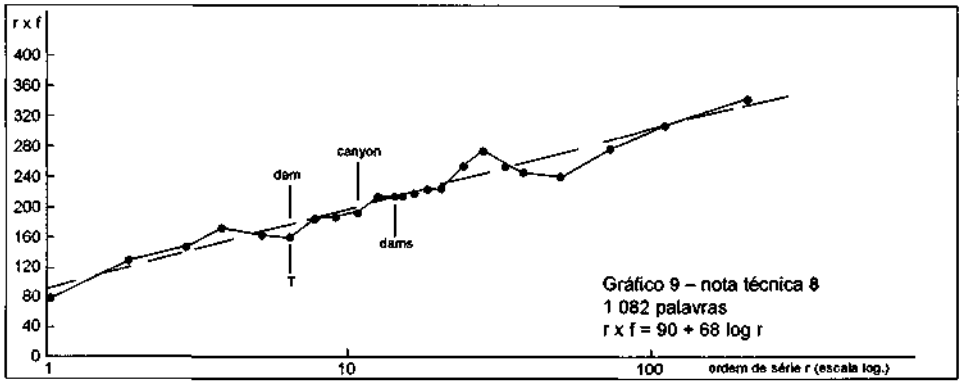

Gráfico 9 - Distribuição dos valores de r x f *versus* log r da Nota Técnica 8

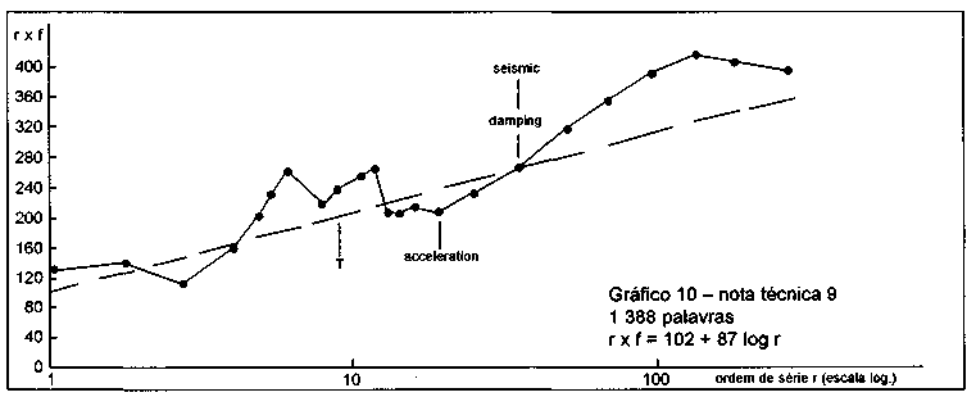

Gráfico 10 - Distribuição dos valores de r x f *versus* log r da Nota Técnica 9

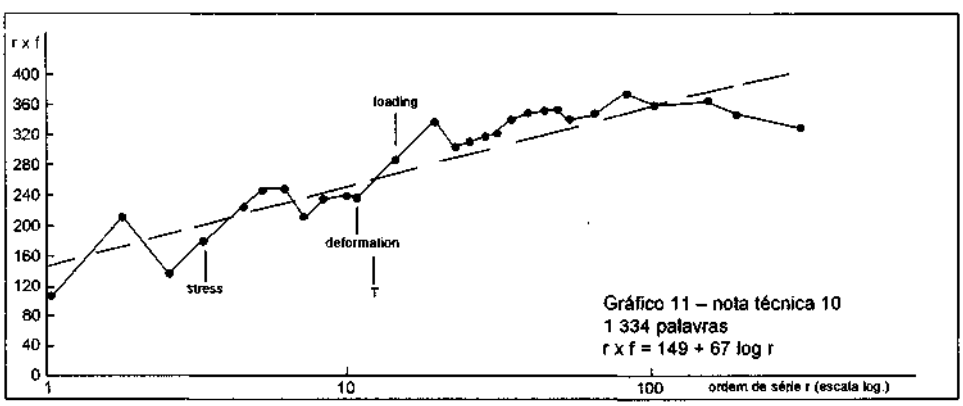

Gráfico 11 - Distribuição dos valores de r x f versus log r da Nota Técnica 10

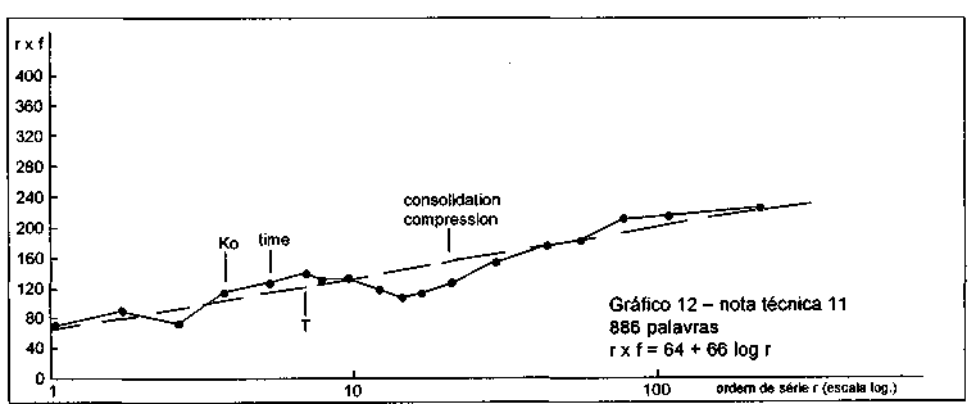

Gráfico 12 - Distribuição dos valores de r x f versus log r da Nota Técnica 11

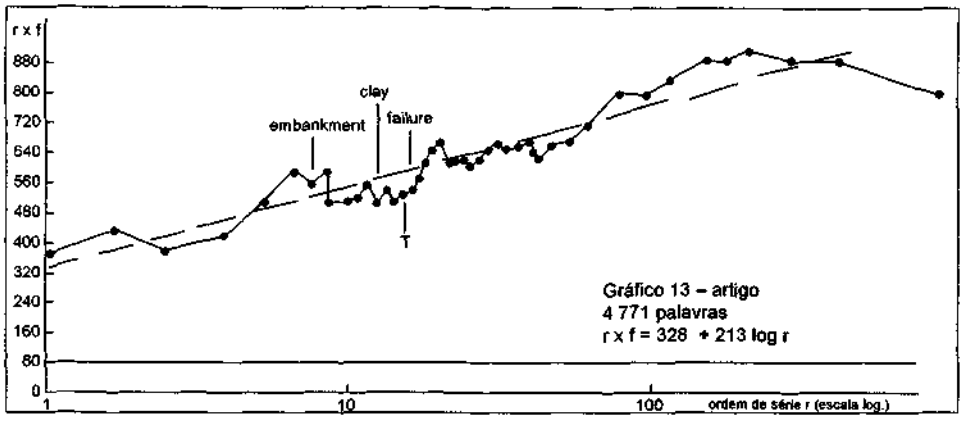

Gráfico 13 – Distribuição dos valores de r x f *versus* log r do Artigo

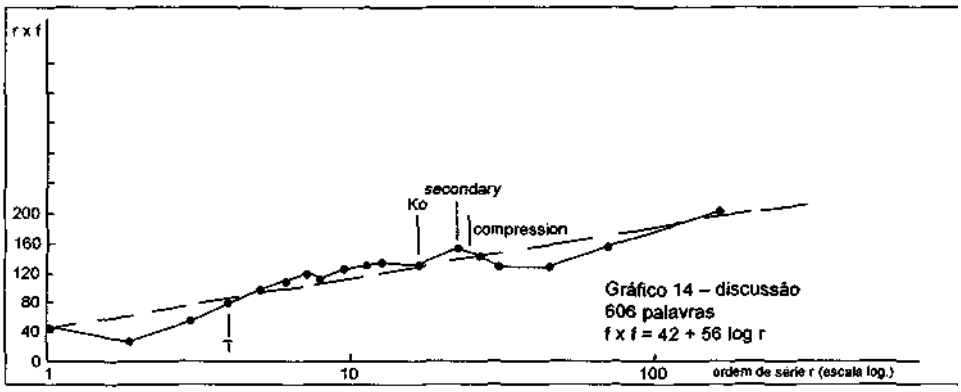

Gráfico 14 – Distribuição dos valores de r x f *versus* log r da Discussão

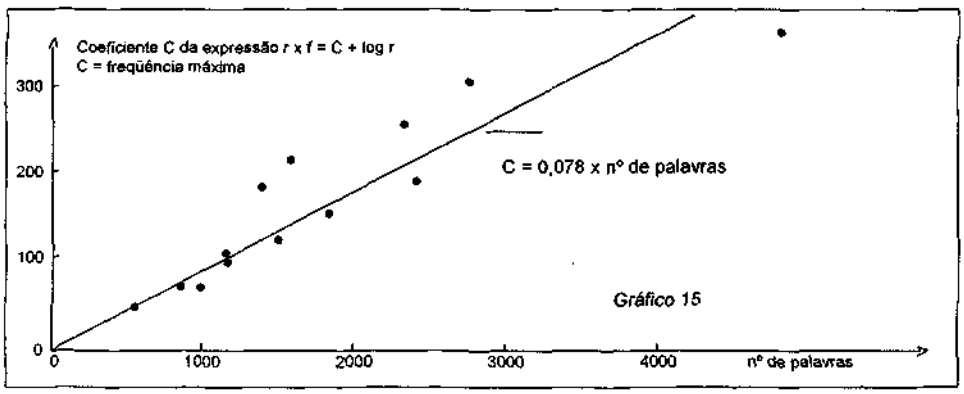

Gráfico 15 – Distribuição dos valores do coeficiente C *versus* número de palavras dos textos analisados

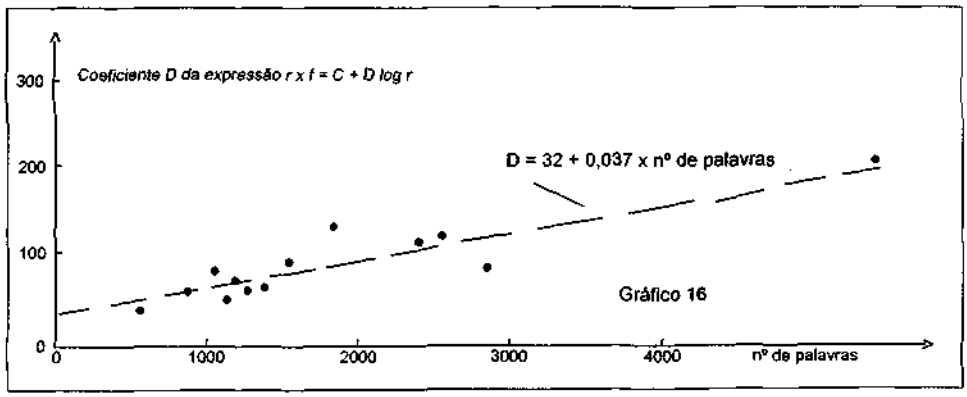

Gráfico 16 – Distribuição dos valores do coeficiente D *versus* número de palavras dos textos analisados