Representação de conteúdo via indexação automática em textos integrais em língua portuguesa

Flávia Pereira Braga Mamfrim

Resumo

Verifica-se a possibilidade da indexação automática derivativa de textos em língua portuguesa, a partir de seu texto integral. É aplicada a Fórmula de Transição de Goffman a 10 artigos na área de Bibliometria e formulado um algorltimo probabilístico de indexação. A Fórmula de Transição de Goffman ê perfeitamente aplicável à língua portuguesa, apontando para uma região de frequência de palavras onde estão concentradas as palavras indicativas do conteúdo dos artigos analisados.

Palavras-chave
Recuperação da informação; Indexação automática derivativa; Fórmula 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 setembro de 1990, sob a orientação da professora Gilda Maria Braga.

INTRODUÇÃO

A finalidade de um sistema de recuperação de informação é selecionar, dentre um considerável número de itens, aqueles que podem satisfazer a uma determinada necessidade expressa de informação.

Segundo Bookstein & Swanson[1], a fim de cumprir essa tarefa, o sistema precisa, dentre outras coisas, de um tipo de representação para o conteúdo de cada documento de sua coleção, o que é feito pela alocação de um conjunto de termos a cada documento.

Assim, indexação é um processo duplo, passando por uma primeira etapa de análise de conteúdo do documento para a extração de conceitos-chave do texto e uma segunda etapa de tradução desses conceitos para os termos de um vocabulário livre ou controlado.

Do ponto de vista do processo, a indexação pode ser atributiva ou derivativa. A indexação é derivativa quando os termos de indexação são derivados do próprio texto do documento analisado e é atributiva quando os termos de indexação são alocados independentemente dos termos do texto do documento.

Este é um processo bastante complexo, na medida em que está condicionado ao conhecimento do indexador, às características do sistema, do vocabulário utilizado, à complexidade do assunto, dentre outros fatores.

A indexação automática consiste na mecanização desse processo no todo ou em parte, visando a estabelecer rotinas que reduzam a interferência da subjetividade do indexador, tanto na análise do documento, quanto na seleção dos termos significativos.

Desde a década de 60, com a introdução do Computador na área de representação e recuperação da informação, uma série de avanços tecnológicos vem sendo observada, sem que, contudo, seja resolvido o problema central dos sistemas de recuperação de informação, qual seja, atingir e representar, de forma ideal, o conteúdo de um documento, possibilitando a combinação *(match)* entre ele e uma determinada necessidade expressa de informação.

Houve, na verdade, a automação do processo, sem que tenha havido uma mudança de filosofia da própria indexação. Apesar de automática, a indexação permanece com seu caráter manual, enfatizando a atribuição de termos via um vocabulário, agora dentro da máquina, e não a sua derivação completa a partir do próprio texto.

Neste sentido, vale lembrar a observação de Luhn[2] de que "o vocabulário existente em um documento deveria se constituir na base para a análise de seu conteúdo, sendo esta a melhor maneira de recuperá-lo".

A explosão da produção científica e suas diversas formas de divulgação, dentre outros fatores, vêm tornando a questão da indexação mais aparente e motivo de preocupação.

É fundamental que a área de indexação vá buscar instrumental em áreas afins como inteligência artificial, sistemas especialistas, sistemas de conhecimento etc.

O que parece ser necessário é uma mudança de filosofia, como, por exemplo, a extração automática dos termos, sem a interferência de vocabulários preestabelecidos, dentro da máquina.

Não se tenciona com isso omitir a questão do controle terminológico, o qual seria uma segunda fase do processo de indexação.

A partir dessa perspectiva, o processo de indexação seria composto por duas fases:

- *a* derivação automática de termos a partir do texto integral;

- o "controle" desses termos, objetivando *a* recuperação.

Toda essa problemática a respeito de formas alternativas de indexação é fundamental e encontra respaldo no fato de que, no Brasil, são poucos os serviços de informação organizados, com políticas e metodologias de indexação estabelecidas. Acresce ainda que quase não existem vocabulários especializados em língua portuguesa, o que dificulta e praticamente inviabiliza a própria indexação.

O mesmo não poderia ser proposto em países como os Estados Unidos, onde os serviços de informação são de tal forma complexos e organizados, que praticamente inexistem tentativas de implantação de sistemas alternativos de indexação.

Por outro lado, tais países desenvolveram de tal forma a linguagem de busca, que as deficiências da entrada são supridas pelas possibilidades de saída.

Desta forma, o Brasil, por sua carência no que diz respeito a serviços de informação organizados, constitui-se em um excelente laboratório que estimula o desenvolvimento de pesquisas que possam contribuir na organização de serviços e propor metodologias alternativas de indexação.

OBJETIVO

e Verificar a possibilidade da indexação automática derivativa a partir de texto integral, aplicando inicialmente a Fórmula de Transição de Goffman e, posteriormente, formulando outros algoritmos, se necessário.

s Analisar as diversas partes do texto, separadamente, objetivando verificar em quais delas se concentra o maior número de palavras de conteúdo semântico.

• Verificar, ainda, a aplicabilidade da fórmula à língua portuguesa, uma vez que a literatura nacional é bastante controversa quanto a este assunto.

FÓRMULA DE TRANSIÇÃO DE GOFFMAN

Vieira[3], em sua revisão da literatura de indexação automática e manual, identifica diversos métodos de indexação automática.

Dentre os métodos identificados pela autora, destaca-se aquele denominado Método de frequência ou Análise estatística. Este método, tendo Luhn como pioneiro, sugere que a frequência com a qual as palavras aparecem no texto fornece medida útil de sua importância.

Vários autores examinam a abordagem estatística aplicada à análise automática de textos (Van Rijsbergen[4], Swanson[5], Salton[6], Maron & Khuns[7], Stiles[8], Sparck Jones[9], Salton & Yang[10], Bookstein & Swanson[11], Harter[12], Das Gupta[13], Doyle[14], Borko[15] etc.).

Segundo Pão[16], os métodos estatísticos de indexação automática se baseiam na natureza estatística do uso das palavras.

Em 1948, George Zipf[17] formulou duas leis sobre distribuição de palavras em um texto - Primeira Lei de Zipf e Segunda Lei de Zipf.

A Primeira Lei de Zipf opera em relação às palavras de alta frequência. Segundo a lei, se palavras de um texto suficientemente longo forem colocadas em ordem decrescente de frequência, poder-se-á verificar que a ordem de série das palavras (R) multiplicada por sua frequência (F) produz uma constante (K).

$$R \times F = K$$

Para palavras de baixa frequência, Zipf propôs uma segunda lei, a qual foi aperfeiçoada por Boofh[18] e é conhecida como Lei de Zipf-Booth.

A lei é enunciada da seguinte forma:

$$In: \frac{2}{I1 \quad nX\,(n+1)}$$

onde:
- In é o número de palavras que ocorreram n vezes para **n**<5 ou **n**<6;
- 11 é o número de palavras que ocorreram uma única vez;
- 2 é uma constante atribuída à língua inglesa.

Embora constatadas empiricamente, tais leis não apresentam aplicabilidade no que refere a sistemas de informação.

Entretanto, a partir da observação de que ambas as leis que, na verdade, operam apenas em relação aos extremos da distribuição de palavras em um texto, Goffman, segundo Pão[19], sugeriu a existência dê um ponto onde haveria a transição das palavras de alta frequência para as palavras de baixa frequência, isto ô, onde o número de palavras tende para a unidade. Neste ponto estariam as palavras ditas representativas do conteúdo do documento em questão.

Neste momento, constata-se a possibilidade de se aplicar leis bibliométricas, que trabalham com frequência de palavras como instrumento de indexação em sistemas de informação. Este ponto comumente denominado Ponto T, é representado matematicamente como:

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

onde:
- I1 é o número de palavras que ocorreram uma única vez;
- 8 é uma constante derivada da língua inglesa;
- 2 é uma constante matemática da fórmula de Baskara, para resolução de equações de 2º grau.

Operacionalmente Goffman propôs que, uma vez identificado o ponto T, seria definida uma região dentro na qual estariam as palavras indicativas do conteúdo do documento. Esta região seria definida a partir de um ponto correspondente a uma frequência aproximada. Assim, a partir desta frequência são contadas as palavras entre o ponto T e a palavra de maior frequência. Este mesmo número de palavras é projetado para *abaixo* do ponto T, definindo uma região.

No Brasil, algumas experiências foram realizadas com a aplicação do ponto T à indexação.

Maia[20] analisou três textos na área de Bibliografia para verificar a aplicabilidade da Fórmula de Goffman à língua portuguesa.

Pinheiro[21] fez um estudo bibliométrico em um texto literário em língua portuguesa procurando verificar a aderência do texto às leis de Zipf e ao ponto T de Goffman.

Sepulveda e colaboradores[22], com o auxílio do microcomputador, aplicaram as leis de Zipf e a Fórmula de Transição de Goffman a um texto de proteção costeira, em idioma inglês, objetivando determinar os termos significativos do texto.

Mamfrim & Coelho[23] analisaram um mesmo texto, originalmente em inglês e sua respectiva tradução para o português, com o intuito de verificar a aderência à lei.

MATERIAL

O material utilizado na análise foi um conjunto de artigos em língua portuguesa, na área de Ciência da Informação, versando sobre Bibliometria,

Dentro de Ciência da Informação, a área Bibliometria foi escolhida em razão do "volume razoável" de artigos produzidos.

Os periódicos *Ciência da Informação, Revista de Biblioteconomia de Brasília* e *Revista da Escola de Biblioteconomia da UFMG* foram escolhidos por incluírem artigos sobre Bibliometria.

Foram levantados todos os artigos relacionados à área de Bibliometria redigidos em português no período de 1972 a 1988.

O critério utilizado para definir se um artigo se referia à Bibliometria foi a análise de seu título. Assim, para efeito desta pesquisa, são artigos de ou sobre Bibliometria aqueles que possuem no título:

a palavra bibliometria e suas variações (bibliométrico etc.);

- referência *a* leis especificas da Bibliometria ou processos bibliométricos (lei de Lotka, Bradford, produção de autores, dispersão da literatura, análise de citação etc.).

Ao todo foram encontrados 31 artigos de e sobre Bibliometria, dos quais foram analisados os seguintes:

1. VELHO, L M. A contemporaneidade da pesquisa agrícola brasileira como reflexo da distribuição da idade das citações. *Ciência da Informação,* v. 15, n. 1 p. 3-11, 1986.

2. ALVARADO, R. U. A Bibliometria no Brasil. *Ciência da Informação,* v. 13, n. 1, p. 91-107, 1984.

3. LIMA, R. C. M. Estudo bibliométrico: análise das citações no periódico Scientometrics, *Ciência da Informação,* v. 13, n. 2, p. 91-107, 1984.

4. GUSMÃO, H. R. Análise da literatura brasileira de Siderurgia. *Ciência da Informação,* v. 7, n. 1,p. 25-35, 1978.

5. MOTTA, D. F. Validade da análise de citação como indicador de qualidade da produção científica: uma revisão. *Ciência da Informação,* V. 12, n. 1, p. 53-60, 1983.

6. QUEIROZ, S. S. Bibliografia brasileira de Botânica, 1971-1972: Estudo bibliométrico, *Ciência da Informação,* v. 4, n. 1, p. 55-66, 1975.

7. CALDEIRA, P. T. Processo de crescimento epidemiológico aplicado à literatura brasileira de Doença de Chagas, *Ciência da Informação, v.* 4, n. 1, p. 4-16, 1975.

8. MULLER, S.P.M. Produtividade cientifica: uma análise parcial da literatura, *Revista da Escola de Biblioteconomia da UFMG,* v. 16, n. *2, p.* 218-240, 1987.

9. OLIVEIRA, S. M. A lei de Lotka sobre produtividade de autores: aplicabilidade do quadrado inverso. *Revista da Escola de Biblioteconomia da UFMG,* v. 13. n. 2, p. 207-233, 1984.

10. MOREL, R. L. & MOREL, C. M. Um estudo sobre a produção científica brasileira, segundo os dados do Institute for Scientific Information (ISI). *Ciência da informação,* v. 6, n. 2, p. 99-109, 1977.

MÉTODO

Os textos foram digitados em microcomputador com o auxílio de um processador de textos *(xy writer).*

A digitação foi feita em partes, respeitando a divisão proposta pelo autor. No caso de o artigo não apresentar qualquer tipo de divisão, este foi dividido de acordo com sua própria lógica interna.

Depois de contadas as partes do artigo, este foi recontado no todo. O artigo não foi redigitado; as diversas partes foram reunidas e recortadas.

A análise limitou-se ao corpo do artigo. Assim, foram digitados seu texto, título, subtítulos, legendas de tabelas e gráficos constantes do corpo do artigo. Foram desprezadas as notas, anexos, citações, referências, resumos e número em gráficos e tabelas.

A contagem dos textos deu-se com o auxílio de um programa especialmente elaborado para esse fim.

Para fins da contagem de palavras, convencionou-se que:

- palavra é uma sequência qualquer de caracteres entre espaços em branco ou entre pontuações (embora a pontuação tenha sido eliminada);

- palavras com hífen são consideradas como uma única palavra;

- diferentes flexões de uma mesma palavra são consideradas como palavras distintas;

- numerais, no corpo do artigo, foram considerados como palavras.

Foi eliminada qualquer pontuação do texto, como vírgula, ponto, parênteses etc., uma vez que o programa de contagem não estava preparado para compreender tais sinais.

Ao final de cada contagem, foi gerada uma lista decrescente de frequência das palavras. A cada palavra estava associada sua frequência. Este procedimento foi efetuado para o artigo no todo e para suas diversas partes.

A partir desta lista de frequências, foram elaborados quadros de distribuição das palavras no texto.

Os textos foram numerados e analisados individualmente, tanto numericamente quanto em relação a seu conteúdo.

A análise numérica consistiu no estudo da distribuição das palavras no texto. Foram identificados os totais de palavras distintas, totais de palavras, médias de ocorrências etc. Foram calculados os Pontos de Transição para cada um dos textos e suas partes, delimitadas as Regiões T, bem como identificadas as palavras contidas nestas regiões.

Além da Fórmula de Transição de Goffman, foi feita uma segunda tentativa aplicando-se aos artigos uma variação da Fórmula de Transição de Goffman. Nesta variação considerou-se que o Ponto de Transição (Ponto T2) seria aquele em que o número de palavras que ocorreram n vezes tenderia a 2, e não a 1, conforme o proposto por Goffman quando expôs sua Fórmula de Transição.

A partir do Ponto T foi estabelecida uma Região de Transição, que abrange da maior classe de frequência até o próprio Ponto T. O mesmo número de classes de frequência do Ponto T até a maior classe é projetado para abaixo do Ponto T, estabelecendo a Região de Transição.

A partir do Ponto T2 foram estabelecidas duas Regiões de Transição. A primeira, denominada Região T2 Restrita, é aquela que tem como limite superior o Ponto T. A segunda, denominada Região T2 Ampliada, tem como limite superior a maior classe de frequência.

Uma vez identificados os Pontos de Transição de todos os artigos no todo e de suas diversas partes, delimitadas as Regiões T, T2 Restrita, T2 Ampliada e identificadas as palavras contidas nestas regiões, foram elaborados quadros para cada artigo mostrando todas estas informações e que constam em anexo - Quadros Síntese dos Resultados.

Após a análise numérica de cada um dos textos selecionados, procurou-se fazer uma análise de conteúdo destes mesmos artigos, comparando-se as palavras obtidas na Região T com as palavras de seu título, resumo e descritores (quando existentes).

Procurou-se, ainda, identificar tais palavras no *Tesauro de Ciência da Informação,* versão preliminar, IBICT.

Nesta etapa da análise, o objetivo central foi tentar identificar as razões pelas quais uma determinada palavra consta da Região T, ou não, ou seja, tentar identificar o papel de determinadas palavras no conjun-

to do texto. Para isso foram identificados os termos que poderiam ser considerados **descritores qua descritores,** as associações entre os termos constantes da Região T que pudessem se constituir em termos de indexação.

Esta etapa da análise levou em consideração apenas as palavras obtidas na Região T do texto como um todo. Eventualmente foi feita uma comparação entre a Região T do texto como um todo com a Região T da parte referente aos resultados etc.

Após a análise dos 10 textos na área de Bibliometria, foi analisado um 11º artigo na área de Economia[24], com *o* intuito de testar, em outra área de assunto, os resultados obtidos.

DISCUSSÃO DOS RESULTADOS

Os resultados obtidos na análise individual dos textos e de suas partes encontram-se sintetizados nos Quadros SIntese, em anexo, onde podem ser verificados os totais de palavras distintas, totais de palavras, Ponto T, Ponto T2 e as palavras de conteúdo semântico contidas nas Regiões de Transição.

Quanto aos textos no todo, seu tamanho, isto é, o total de palavras no texto, é de em média 3 000 palavras com um desvio de 827, variando de 1 810 a 4 229 palavras. A diferença percentual é de 134.

O total de palavras distintas é de aproximadamente 900 palavras com um desvio de 189, variando entre 567 e 1 306. A diferença percentual é de 130.

O total de palavras ocorridas uma única vez ao longo do texto é de, aproximadamente, 590 palavras com um desvio de 123, variando de 347 a 831 palavras. A diferença percentual é de 139.

A média de ocorrências por palavra é de 3,5.

O total de classes de frequência está entre 29 e 42 classes de frequência por texto.

Apesar da grande variação no tamanho dos textos, pode-se observar que as palavras ocorridas uma única vez correspondem a aproximadamente 60% do total das palavras distintas. Este padrão faz lembrar a generalização de Price[25] em relação a Lei de Lotka: 60% dos autores contribuem com um único documento, em um determinado corpo de literatura. Embora isto possa ser uma curiosidade meramente numérica, a associação não poderia deixar de ser feita em relação às palavras que ocorrem uma vez e os autores que publicam um único documento.

O total de palavras distintas, por sua vez, corresponde *a* cerca de 30% do total de palavras.

O total de palavras ocorridas uma única vez corresponde a cerca de 19% do total de palavras, com desvio de 2,73.

Correlacionando-se *as* variáveis palavras distintas, total de palavras e palavras ocorridas uma única vez, obtêm-se as seguintes correlações:

- Correlação de Pearson (correlação entre números):
- palavras distintas em relação a palavras ocorridas uma única vez - 0,98;
- total de palavras em relação a palavras ocorridas uma única vez - 0,81;
- total de palavras em relação a palavras distintas - 0,87;
- Correlação por *ranking* (correlação por ordem de série):
- palavras distintas em relação a palavras ocorridas uma única vez - 0,93;
- total de palavras em *relação a palavras* ocorridas uma única vez — 0,85;

e total de palavras em relação a palavras distintas - 0,94.

Em todas as correlações feitas, observa-se que as variáveis total de palavras, palavras ocorridas uma única vez e palavras distintas mantêm entre si uma forte correlação positiva, demonstrando a grande dependência entre as variáveis e um padrão de construção dos textos.

A constatação dessas correlações entre as variáveis indica que um texto segue uma lógica interna de construção que independe de sua autoria. Com isso, observa-se a existência de um padrão que norteia a distribuição de palavras em texto.

Assim, para o conjunto de textos analisados, o texto padrão teria cerca de 3 000 palavras, 900 palavras distintas, 550 ocorrendo uma única vez ao longo do texto e oito palavras de conteúdo semântico.

Observando-se a estrutura de distribuição das palavras nos diferentes textos, nota-se que o número de palavras, associado à suas frequências, evolui muito lentamente nos primeiros 3/4 das distribuições. No último quarto, este número de palavras "explode", avançando rapidamente, chegando até mesmo a dobrar ou triplicar de uma classe de frequência para outra.

Fenómeno inverso ocorre com as frequências associadas ao número de palavras. No primeiro quarto da distribuição seu avanço em direção à unidade é bastante acelerado, chegando a ser reduzido à metade, quando passa-se de uma classe de frequência para a seguinte. Nos 3/4 finais, por outro lado, o avanço das frequências em direção à unidade é bastante lento.

Observa-se que a distribuição de palavras em um texto se conforma ao padrão das distribuições bibliométricas em geral: "poucos com muito e muitos com pouco". Assim, poucas palavras ocorrem muitas vezes, enquanto muitas palavras ocorrem poucas vezes ao longo do texto.

Nas porções superiores das distribuições, encontram-se predominantemente artigos, preposições e conjunções. De é a palavra de maior frequência obtida nas distribuições. O, a, e, em aparecem também com altas frequências de ocorrência em todos os textos. Tais palavras, de alta frequência, não apresentam conteúdo semântico, sendo incapazes de representar o conteúdo de um documento.

No extremo inferior das distribuições, o inverso ocorre, isto é, as palavras constantes das últimas classes de frequência, ou seja, as palavras que tendem a ocorrer uma vez ao longo de todo o texto são palavras que apresentam conteúdo semântico, são o que poderíamos chamar de "riqueza" da *linguagem.* Nestas últimas classes de frequência encontram-se substantivos, adjetivos, advérbios. Assim, nos extremos das distribuições de palavras tem-se o "cimento" e a "riqueza" da linguagem, indispensáveis à construção do texto, todavia dispensáveis para a sua representação, para fins de indexação.

Com isto constata-se não apenas um padrão quanto à estrutura e distribuição das palavras em texto, seja qual for o seu tamanho, como também comprova-se a própria proposição de Goffman quando se refere à porção intermediária da distribuição como a área de excelência para a localização das palavras de conteúdo semântico de um texto.

Este mesmo padrão de distribuição de palavras e estrutura de texto foi obseivado em relação a todas as partes dos diferentes textos, quando analisadas individualmente.

A fim de estabelecer esta área de excelência em que estariam as palavras de

conteúdo semântico representativas do conteúdo do texto, foram calculados os Pontos T e T2 e definidas as Regiões T, T2 Restrita e T2 Ampliada para todos os artigos, no todo, e em suas diversas partes.

No que se refere ao texto no todo, o Ponto T varia de 25,84 a 40,27. Sua variação percentual é de 56.

Apesar do tamanho bastante diversificado dos artigos, o Ponto T varia muito pouco em termos numéricos e percentuais. Esta variação de 56% não nos parece ser significativa, sobretudo quando comparada à variação de 134% em relação ao tamanho dos diferentes textos analisados.

O Ponto T mais baixo (T = 25,84) foi identificado no texto 6, cujo total de palavras é 1 810, e o Ponto T mais alto (T = 40,27) refere-se ao artigo 8, onde o total de palavras é 4 194. Enquanto o total de palavras praticamente triplica, o Ponto T não chega nem mesmo a dobrar. Isto ocorre porque, como foi dito anteriormente, o número de palavras ocorridas uma única vez permanece percentualmente constante em relação ao número de palavras distintas, aproximadamente 60%. Assim, sendo o Ponto T calculado com base no número de palavras ocorridas uma única vez, sua variação numérica é muito pequena.

O valor numérico do Ponto T também parece não ter qualquer influência no número de palavras de conteúdo semântico na Região T.

No texto 9, cujo Ponto T é 33,32, a Região de Transição apresenta 30 palavras de conteúdo semântico, enquanto no texto 5, cujo ponto T é 32,94, a Região T contém apenas oito palavras de conteúdo semântico em sua Região T.

Também não parece haver relação entre o valor numérico de T e o total de palavras na Região T (palavras sem e com conteúdo semântico). O total de palavras na Região T também não parece determinar o número de palavras de conteúdo semântico.

Quanto ao Ponto T2, este varia entre 18,13 e 28,33.

Observa-se que, de modo geral, as Regiões T2 Restrita e T2 Ampliada se mostram inadequadas para o estudo proposto. A Região T2 Restrita abrange um conjunto muito pequeno de palavras e que de modo geral está contido na Região T. A Região T2 Ampliada, por sua vez, abrange um total de palavras muito grande, que por vezes coincide com o total de palavras do texto. Assim, dadas as inadequações destas Regiões para o estudo, as análises basearam-se na Região T, conforme proposição original de Goffman, tanto para *a* análise dos textos na íntegra, quanto para as diferentes partes do texto.

A Região de Transição é composta basicamente por preposições, artigos, conjunções, substantivos e verbos.

Artigos, conjunções e preposições correspondem a, em média, 65% do total de palavras obtidas na Região T. Esta predominância deve-se ao fator da Região de Transição ter seu limite superior na classe de maior frequência, local por excelência de artigos, conjunções e preposições, ou seja, palavras sem conteúdo semântico.

Os substantivos correspondem a cerca de 30% do total de palavras da Região de Transição. Localizam-se em classes de frequência média, ou seja, na porção mediana da distribuição de palavras.

Os verbos são bastante infreqüentes na Região de Transição, representando em média 5% destas palavras. Os verbos mais frequentes são as conjugações do verso "ser".

Advérbios e adjetivos não são observados nas Regiões T. Sua localização de excelência é no extremo inferior da distribuição de palavras.

No que se refere a Região T do texto como um todo, foram identificadas palavras de conteúdo semântico em iodos os artigos analisados. Isto pode ser observado no exercício de análise de conteúdo realizado para cada texto.

Foram obtidas coincidências entre as palavras contidas na Região T e as palavras do titulo, resumo e descritores do artigo.

Considerando que título, resumo e descritores representam formas de indexação realizadas pelo autor e/ou comissão editorial do próprio periódico, estas coincidências proporcionam confiabilidade nas palavras obtidas.

Por outro lado, a comparação das palavras obtidas na Região T com o vocabulário constante do *Tesauro em Ciência da Informação* visou a verificar se tais palavras são consagradas como parte do vocabulário da área. Também nesse exercício foi obtido sucesso, ou seja, as palavras abrangidas pela Região de Transição, de forma isolada ou em conjunto com outras palavras da Região T, tendem a constar do tesauro, sendo, assim, reconhecidas como parte integrante da terminologia da área.

Eventuais deficiências na representação das palavras constantes na Região T podem ser supridas, se ao conjunto de palavras obtidas na Região T forem associadas as palavras de conteúdo semântico do título. Com este procedimento, baseando-se na própria indexação do autor (isto é, o título), é possível enriquecer o conjunto de palavras obtidas na Região T, aumentando seu poder de representação.

A Região T abrange em média oito palavras de conteúdo semântico.

Este padrão não foi observado em três artigos: artigo 2 - *A Bibliometria no Brasil;* artigo 9 - *A lei de Lotka sobre a produtividade de autores* e artigo 10 — *Estudo sobre a produtividade científica brasileira segundo os dados do ISI.* Os artigos 2 e 9 encontram-se bem acima deste padrão, abrangendo respectivamente 20 e 30 palavras de conteúdo semântico, e o artigo 10 encontra-se bem abaixo do padrão, abrangendo apenas duas palavras de conteúdo semântico na Região T.

Nota-se que os textos 2 e 9 apresentam uma característica que os diferencia dos demais artigos analisados. Tais artigos procuram retratar o estado-da-arte de uma determinada área de assunto, aproximando-se de um padrão de "artigo de revisão". O texto 2, embora não sendo um artigo de revisão, retrata o estado-da-arte na área de Bibliometria, identificando autores mais produtivos e áreas mais prestigiadas. O texto 9 é artigo de revisão que faz um levantamento do estado-da-arte no que se refere à *aplicação da* lei de Lotka.

Uma justificativa para o grande número de palavras de conteúdo semântico em artigos de revisão ou artigos com características de revisão pode ser a necessidade de se descrever com riqueza de detalhes o ambiente em que o estudo ou a revisão se desenrola.

No texto 9, isso pode ser observado pela presença de palavras como **biblioteconomia, computação, siderurgia, medicina brasileira, literatura médica espanhola** etc. - as quais se referem às áreas do conhecimento em que a lei de Lotka sobre produtividade de autores foi aplicada.

O mesmo pode ser observado no texto 2, onde palavras como **Braga, Goffman, Bradford e frente de pesquisa** foram encontradas. Tais palavras dizem respeito aos autores mais produtivos no período e às leis específicas e muito prestigiadas na área.

O que se pode observar por esta característica é que em artigos do tipo revisão, a literatura de e a literatura **sobre** estão re-

presentadas. Assim a Região de Transição agrupa tanto palavras relativas ao de quanto palavras relativas ao sobre. As palavras relativas ao sobre são exatamente aquelas que retraíam o ambiente em que o estudo ou a revisão se desenrola.

Em artigos científicos propriamente ditos, o que se encontra representado é a literatura de um determinado assunto. Assim a Região de Transição deste tipo de artigo tende a não agrupar palavras periféricas ao assunto principal.

Desta forma, pode-se dizer que o tipo e as características do artigo podem influenciar a categoria de palavras obtidas na Região T.

Vale notar que o texto 5 - *Validade da análise de citação como indicador de qualidade da produção científica: uma revisão,* apesar da palavra "revisão", não se conforma ao padrão obtido para artigos de revisão, por não sê-lo. Este artigo apresenta em sua Região T apenas nove palavras de conteúdo semântico, em contraposição aos demais artigos de revisão apresentam em média 25 palavras de conteúdo semântico.

Seu comportamento, todavia, adequa-se aos padrões obtidos para artigos científicos propriamente ditos, os quais abrangem cerca de oito palavras de conteúdo semântico em sua Região T.

A Região T foi delimitada para todas as partes dos diferentes artigos. Esta delimitação da Região T para cada uma das partes dos artigos analisados indica que a parte referente aos resultados retrata o comportamento do artigo como um todo, abrangendo de forma significativa palavras de conteúdo semântico candidatas a termos de indexação.

Nos 10 artigos analisados, as palavras constantes da Região T dos resultados estão contidas na Região T do texto no todo.

As demais partes, isto é, introdução, metodologia, conclusão etc., ou seus equivalentes, mostram-se inadequadas à indexação nos moldes propostos pela pesquisa, uma vez que não apresentam, de forma constante, um conjunto de palavras representativas do conteúdo dos artigos.

Dos 10 textos examinados, sete possuem uma parte especifica para "resultados". Em quatro destes sete artigos, a parte referente aos "resultados" é a maior em termos de tamanho, ou seja, em termos de número de palavras. Os demais três artigos apresentam outras partes maiores do que as de "resultados". Este fato, todavia, não impediu que a parte referente a "resultados" continuasse a repetir o padrão de distribuição do texto como um todo, contendo um grande número de palavras de conteúdo semântico.

Assim, pode-se dizer que o tamanho dos "resultados" não influi em seu poder de representação do conteúdo do artigo, o que tende a tornar a parte referente aos "resultados" uma parte densa em termos de informação; é seu conteúdo, ou seja, a literatura de, que encontra nos "resultados" seu lugar de excelência. Poder-se-ia, assim, inferir julgamentos de qualidade, ou seja, o que faz com que a parte "resultados" seja representativa é a sua qualidade, seu conteúdo, e não o seu tamanho em termos de total de palavras.

Esta característica observada em relação aos "resultados", isto é, sua tendência a reproduzir o comportamento do texto no todo, é de grande valia para a indexação, seja ela automática ou manual. Tal observação é cabível, uma vez que, para fins de indexação, indica o local do texto que tende a concentrar palavras de conteúdo semântico representativas. Havendo tabelas, nos resultados, seus títulos apresentam uma tendência a indicar o objeto central do artigo. Sua utilização para fins de indexação pode fornecer bons resultados. Nos artigos 5, 8 e 9, não foram identificadas partes especificas correspondentes a "resultados". Estes artigos apresentam de modo geral características de artigos de revisão, e não foi obtido padrão de distribuição das palavras de conteúdo semântico em suas partes, ou seja, as palavras de conteúdo semântico tendem a estar espalhadas peto corpo do artigo, e não concentradas em uma parte específica.

Assim, o tipo do artigo tende a influenciar não só o tipo de palavras obtidas na Região T do artigo como um todo, como também influencia na própria distribuição destas palavras por suas diferentes partes, ou seja, ao tango do próprio texto.

Após a análise dos 10 textos na área de Bibliometria, foi analisado um outro texto na área de Economia, com a intenção de comprovar os padrões e resultados obtidos na análise do conjunto dos 10 textos anteriores.

Este 11º artigo, ao qual chamaremos de artigo-teste, possui 5 718 palavras e 1 331 palavras distintas. O total de palavras distintas corresponde a aproximadamente 23% do total de palavras, o que se aproxima da média de 30% obtida para o conjunto de artigos na área de Bibliometria.

Este artigo-teste apresenta mesmo padrão de distribuição das palavras no texto. O número de palavras em relação a suas frequências associadas evolui exatamente da mesma forma observada no conjunto dos 10 artigos, ou seja, obedecendo a uma relação inversa.

As Regiões T2 Restrita e T2 Ampliada também se mostraram inadequadas por apresentarem respectivamente um número muito pequeno e um número muito grande de palavras.

O Ponto T localiza-se na frequência 39, 19.

A Região T abrange 25 palavras de conteúdo semântico, ou seja, um *número de* palavras de conteúdo semântico superior à média encontrada para Bibliometria. Todavia, como mencionado anteriormente, o artigo apresenta um componente teórico muito grande, o que pode ser responsável peto grande número de *palavras de conteúdo* semântico obtido. No que se relere à quantidade *de* palavras de conteúdo semântico em T, o artigo-teste aproxima-se dos artigos do tipo revisão; entretanto, no que se refere ao tipo de palavras obtidas, estas referem-se, na *totalidade, à literatura de, e* não a literatura sobre. Neste particular, difere dos artigos de revisão, os quais abrangem não só termos relativos à literatura de, como termos relativos à literatura sobre.

Dado o conjunto de palavras obtidas, na Região T, estas descrevem com grande confiabilidade o conteúdo do artigo, uma vez comparadas com os termos propostos pelo autor. É interessante notar que, se *associarmos ao conjunto de palavras de* conteúdo semântico obtidas na Região de Transição às palavras de conteúdo semântico do título, teremos ainda um indicador geográfico (Brasil), e um indicador temporal (anos 90), os quais posicionam o *estudo no tempo e no espaço. Neste* artigo, não há uma parte específica para "resultados". De suas sete partes, quatro apresentam palavras de conteúdo semântico em suas Regiões T. Assim, há uma tendência à dispersão das palavras de conteúdo semântico pelo corpo do artigo, e não sua concentração em uma parte específica. Nota-se que a estrutura do artigo-teste se conforma à estrutura obtida nos demais artigos analisados. O que se pode perceber com bastante clareza, a partir da desestruturação de um texto em palavras, é que, dado um determinado texto, aplicando-se a Fórmula de Transição de Goffman, tem-se um conjunto de palavras de conteúdo semântico, que de forma mais ou menos precisa, isoladamente, associadas ou em conjunto, sugerem pistas para indexação.

Uma vez discutida a estrutura de distribuição das palavras e observados alguns

padrões numéricos relacionados a esta estrutura de distribuição, é fundamental discutir o próprio método de indexação proposto.

Quanto a contagem de palavras, o programa contador considera como uma palavra um conjunto de caracteres entre espaços em branco, não reconhecendo, assim, a raiz das palavras e suas variações. Este programa poderia ser implementado de forma a reunir variações de uma mesma palavra. Isto agruparia, por exemplo, o plural e singular de uma mesma palavra, aumentando sua frequência. Esta observação é cabível na medida em que certas palavras de conteúdo semântico e representativas do conteúdo do artigo não se encontram na Região T, pois tiveram suas frequências divididas, como "bibliométrica" e "bibliométricas". O uso da truncagem para a reunião de variações de uma mesma palavra seria uma forma de aumentar o poder de representação da Região T.

Uma vez definida a Região de Transição, a eliminação daquelas palavras que não possuem conteúdo semântico poderia ser feita através do uso de uma lista ou vocabulário negativo em máquina. Na experiência realizada, esta exclusão foi feita manualmente, uma vez que este suporte não se encontrava disponível. O uso de um vocabular» negativo permitiria tão-somente a exclusão de conjunções, preposições e artigos.

Verbos, advérbios e adjetivos, em razão de sua variedade, não poderiam ser excluídos desta forma. Todavia, como notado anteriormente, quando nos referimos à composição da Região de Transição, sua quantidade é pouco significativa, uma vez que se localizam em peso no final da distribuição de palavras, e não em sua porção mediana. Assim a existência de um verbo ou advérbio na Região de Transição não traz danos à filosofia do método de indexação. Por outro lado, observa-se que a inclusão das palavras de conteúdo semântico do título enriquecem o conjunto, aumentando o poder de representação da Região de Transição. A inclusão das palavras de conteúdo semântico do título poderia ser suprimida no caso da recuperação automática, isto é, caso o programa de recuperação permitisse a associação dos termos de indexação com palavras do título. Assim a inclusão destas palavras se daria no momento da recuperação.

A partir destas observações é possível formular-se um algoritmo para indexação automática de textos integrais, que passaria pelas seguintes etapas:

— transferência do texto para o meio magnético;

- contagem automática de palavras;

- aplicação da Fórmula de Transição de Goffman;

— definição da Região de Transição;

- eliminação de conjunções, artigos e preposições via dicionário negativo;

- inclusão das palavras de conteúdo semântico do título ao conjunto de palavras obtidas.

O algoritmo resultante determina uma região de frequência onde existe a probabilidade de ocorrerem palavras de conteúdo semântico, representativas do texto. Ele reduz a incerteza quanto às palavras que efetivamente devem ser utilizadas como termos de indexação. O papel deste algoritmo como redutor de incerteza está muito claro quando a partir de textos de, em média, 3 000 palavras das quais 900 são distintas reduz esse universo de palavras a, em média, oito para artigos científicos e aproximadamente 25 quando artigos de revisão.

Além de reduzir incertezas, a indexação automática derivativa, com o algoritmo proposto, na medida em que retira do próprio texto palavras para a indexação, indexa com as próprias palavras do autor, diminuindo o hiato entre a linguagem do autor/usuário e a linguagem da indexação/sistema de informação.

CONCLUSÕES

A pesquisa objetivou verificar a viabilidade da indexação automática derivativa de textos integrais em língua portuguesa, através da aplicação da Fórmula de Transição de Goffman.

Sendo indexação um processo que culmina na representação do conteúdo de um documento através da alocação de um conjunto de termos, pode-se dizer que o objetivo foi atingido, constatando-se a viabilidade do método de indexação utilizado.

Em todos os artigos analisados, a Região de Transição funcionou filtrando, ou melhor, concentrando um conjunto de palavras que de forma isolada ou associada sugere sempre o conteúdo do artigo.

É importante ressaltar que o idioma dos textos não se constituiu em entrave à aplicabilidade da Fórmula de Goffman.

Esta questão deve ser considerada, uma vez que a Fórmula é derivada de outras duas, desenvolvidas com base na língua inglesa.

Assim, com base nas análises dos textos, pode-se dizer que a Fórmula de Transição de Goffman aplica-se sem maiores problemas à língua portuguesa.

O algoritmo proposto parece se adequar sem problemas aos artigos analisados.

É interessante notar que este algoritmo parte da quantidade de palavras com frequência igual a um, para chegar a uma região de frequência aproximada onde se encontram as palavras de conteúdo semântico do texto.

Duas conclusões são cabíveis:

- da quantidade de palavras emerge um conjunto de palavras de "qualidade", isto é, palavras de conteúdo semântico representativas do artigo. Mais surpreendente é que esta "qualidade" emerge justamente de um conjunto de palavras que não se mostra representativo do conteúdo do texto;

- o número de palavras com frequência igual a um aponta para uma frequência aproximada, isto é, demonstra uma relação entre a quantidade de palavras com frequência igual a um ("riqueza" da linguagem) e a frequência das palavras de conteúdo semântico representativas do texto ("riqueza" de representação ou meta-linguagem?").

Considerando que o texto padrão, para o conjunto de textos analisados, possui em média cerca de 3 000 palavras e 900 palavras distintas, das quais 550 ocorrem uma única vez, pode-se dizer que apenas 350 palavras se repetem. Partindo do princípio de que da quantidade emerge a qualidade, apenas 350 palavras são "candidatas" a termos de indexação. Desta forma, há uma tendência a se obter o máximo possível de expressão com o menor número de palavras possível, ou seja, utiliza-se uma pequena parte do vocabulário (cerca de 40% do total de palavras distintas) para se obter o maior nível de expressão de ideia.

Neste sentido, pode-se fazer um paralelo com o princípio do menor esforço, segundo o qual existe uma tendência humana em se obter o máximo com o menor esforço possível. Assim, a construção de um texto tende a seguir este princípio, ou seja, usa-se um pequeno número de palavras de "qualidade" para se obter um grande grau de expressão de uma ideia. Isso parece demonstrar que o vocabulário existente em um documento, ou seja, o vocabulário usado pelo próprio autor deve ser a base para a análise de seu conteúdo, sendo esta, talvez, a melhor forma de recuperá-lo. Nas ciências, isto é fundamental, na medida em que o autor é o próprio

usuário da informação, e vice-versa. Assim, subjacente ao processo derivativo de indexação, está a utilização do vocabulário consagrado pelo autor/usuário, o qual possibilitará certamente uma melhor recuperação, aumentando o grau de satisfação.

E importante notar que esse método de indexação talvez tenha sua *aplicação* limitada à áreas específicas e bem definidas. Sua utilização em sistemas mais amplos, multidisciplinares poderia não alcançar bons resultados, *não no* que diz respeito ao método de indexação em si, mas em relação à posterior recuperação. O uso deste método em sistemas multidisciplinares poderia ser responsável por falsas recuperações em excesso, uma vez que uma mesma palavra pode ter conotações diferentes, dependendo do seu contexto. Em sistemas especializados, ao contrário: o vocabulário sendo mais específico e definido, e estando as palavras ligadas a uma área específica, existe a tendência à obtenção de bons resultados.

Quando se propõe um método de indexação automática derivativa, via um algoritmo, onde não há a interferência do indexador, pensa-se imediatamente em sua recuperação. A forma mais apropriada seria uma recuperação também automática com o uso de soflwares que permitissem associações booleanas e não booleanas (conjuntos nebulosos, recuperação probabilística com indicadores *e* pesos etc.).

Apesar de aparentemente essa técnica de indexação representar um esforço de digitação muito grande, praticamente inviável, é importante notar que, com o desenvolvimento das *técnicas computacionais,* que vem possibilitando a transferência de caracteres gráficos diretamente do texto para o meio magnético, o esforço de digitação será plenamente dispensável tornando esta técnica não tão distante ou inviável.

Esse processo de indexação, além de possibilitar a indexação em si, permite que, através da frequência com *a* qual as palavras aparecem nos documentos, se possa vislumbrar o próprio vocabulário de uma área de assunto, o que pode ser de grande auxílio na construção de vocabulários especializados, tesauros, identificação *e* incorporação de novos termos, identificação de novas linhas de pesquisa dentro de uma área especifica e no acompanhamento de seu desenvolvimento.

O método tomou possível a observação de padrões de distribuição das palavras no texto que tendem a se repetir e a se relacionar não só à construção de um texto, como também a seu tipo e, talvez, qualidade.

Esses padrões observados implicam dizer que um texto não é apenas um conjunto de palavras com a finalidade de expressar uma ideia. E também uma estrutura que obedece a uma lógica interna de construção que parece independer de idioma, área de assunto e autoria.

Em relação à metodologia utilizada na indexação, pode-se concluir por sua validade. Todavia é importante fazer menção a suas limitações, impostas pela falta de recursos computacionais mais sofisticados, que possibilitem a incorporação de um vocabulário negativo em máquina, o uso de truncagem, para a incorporação de flexões de uma mesma palavra etc.

Quanto ao processo de indexação em si, foram obtidos conjuntos de palavras que concentram o conteúdo dos artigos. Constata-se, assim, conforme já foi dito, a proposição de Luhn[26], demonstrando que as palavras de um texto são a chave para sua indexação e futura recuperação.

Algumas sugestões podem ser feitas para a continuidade dos estudos, como, por exemplo, a incorporação de um certo controle terminológico no que se refere a um núcleo definido de palavras, a utilização das citações como fonte de termos de indexação etc.

Além de implicar um maior entendimento do processo de indexação em si, acredita-se que, com o desenvolvimento de estudos desta natureza, com a constatação de novos padrões de *distribuição e* frequência de palavras no texto, seja possível fazer derivações sobre conteúdo e qualidade da produção científica de áreas específicas.

REFERÊNCIAS *BIBLIOGRÁFICAS*

1. BOOKSTEIN, A., SWANSON, D. R. A decision theoretic foundation for indexing. *JASIS. v.* 26, n. 1, p. 45-50, 1975.

2. LUHN, H. P. The automatic creation of literature abstracts. *IBM Journal of Research and Development,* v. 2, p. 159-165, 1958.

3. VIEIRA, *S. B.* Indexação automática *s manual:* revisão da literatura. *Ciência da Informação,* v. 17, n. 1, p. 43-57, 1988.

4. VAN RIJSBERGEN, C. J. *Information retrieval, 2.* ed. London: Butterworths, 1979. p. 1-35.

5. SWANSON, D. R. Searching natural language by computer. *Science, v.* 132, n. 3 434, p. 1 099-1 104, 1960.

6. SALTON, G. Automatic text analysis. Science, *v. 168,* n. 3 929, p. 335-343, 1979.

7. MARON, M. E., KUHNS, J. L. On relevance probabilistic indexing and retrieval. *Journal of ACM, v. 7, n. 3,* p. 216-244, 1960.

8. STILES, H. F. The association factor in information retrieval. *Journal of ACM,* v. 8, p. 271-279, 1961.

9. SPARCK *JONES, K.* Automatic indexing. *Journal of Documentation, v. 30, n. 4,* p. 393-432, 1974.

10. SALTON G., YANG, C. *S.* On lhe specificiation of term values in automatic indexing. *Journal of Documentation,* v. 29, n. 4, p. 351-372, 1973.

11. *BOOKSTEIN,* A., SWANSON, D. R. Probabilistic models for automatic indexing. JASIS, v. 25, p. 312, 1974.

12. HARTER, S. P. A probabilistic approach to automatic keyword indexing. *JASIS,* v. 26, n. 4, p. 196-206, 1975; v. 26, n. 5, p. 280-289, 1975.

13. DAS-GUPTA, *P. An investigation into the Two-Poisson model of automatic indexing.* Syracuse, NY, Syracuse University, 1985 169p. (Ph. D. dissertation) Apud LANCASTER, F. W. Subject analysis. AR/ST, v. 24, p. 35-84, 1989.

14. DOYLE, L. B. Semantic road maps for literature searching. *Journal of ACM, v. 8,* p. 35-84, 1989.

15. BORKO, H. Toward a theory of indexing. *Information Processing and Management,* V. 13, p. 355-365, 1977.

16. PAO, M. L, *Concepts of information retrieval.* Englewood, CO: Libraries Unlimited, 1989, 285p.

17. LIPF, G. K. *Human behavior and the principie of least effort* Cambridge, MA, Addison-Wesley, 1949.

18. BOOTH, A. D. A "law" of occurrences of words of low frequency. Information and *Control, v.* 10, n. 4, p. 386-393, 1967.

19. PAO, M. L. Automatic text analysis based on transistion phenomena of word occurrencies. *JASIS,* v. 29, n. 3, p. 121-124,1979.

20. MAIA, E L S. Comportamento bibliométrico da língua portuguesa como veículo de representação de informação. *Ciência da Informação,* v. 2, n. 2, p. 99-139, 1973.

21. PINHEIRO, L. V. R. *Estudo bibliométrico em linguagem literária.* 1977. 17p. Trabalho não publicado apresentado à disciplina de Bibliometria da ECO/UFRJ.

22. SEPULVEDA, G, M. et al. Aplicação da lei de Zipf em um texto de proteção costeira através do uso do microcomputador. In: *Anais do 14- Congresso Brasileiro de Biblioteconomia e Documentação,* Recife, 1987. v. 2, p. 481-502.

23. MAMFRIM, F., COELHO, B. A. S. *Indexação automática derivativa: um estudo com o Ponto T.* Rio de Janeiro, 1988. 25p. Trabalho não publicado apresentado à disciplina de Bibliometria da ECO/UFRJ.

24. CARNEIRO, D. D., WERNECK, R. F. *Brasil: exercícios de crescimento para os anos 90.* Julho 1989. Mimeografado. (tradução Mariana de Moraes Bastos).

25. PRICE, D. J. de S. *O desenvolvimento da ciência.* Rio de Janeiro, Livros Técnicos e Científicos, 1976.

26. LUHN, H. P. A statistical approach to mechanized enconding and searching of literare information. *IBM Journal of Research and Development,* v. 1, n. 4, p. 309-317, 1957.

Artigo aceito para publicação em 5 de abril de 1991.

Flávia Pereira Braga Mamfrim

Mestre em Ciência da Informação pela Escola de Comunicação da Universidade Federal do Rio de Janeiro. Técnico em informação do Centro de Informação em Economia Internacional do Departamento de Economia da Pontifícia Universidade Católica, Rio de Janeiro.

Representation of contents **by** the **automatic** indexing **process of full** texts **in Portuguese language**

Abstract

Possibility of automatic derived indexing of full texts in Portuguese is verifyed. Ten papers in Bibliometrics were indexed and their different parts considered for quantitative and qualitative analysis. Structure and disíríbution patterns of words were studied. Goffman's transition formula proved to be adequate as a slarting point for the indexing algorithm, which yielded, in all papers, a concentration zone forsemantic loaded terms. The algorithm worked as an uncenainty reducer, feading to the semantically important words.

Key words
Information retrieval; Automatic derived indexing; Goffman's transition formula.

ANEXO

Quadro 1 – Síntese dos resultados do Texto 1

A contemporaneidade da pesquisa agrícola brasileira como reflexo da distribuição da idade das citações

	Total de palavras distintas	Total de palavras	Ponto T	Palavras de conteúdo semântico em T	Ponto T2	Palavras de conteúdo semântico em T2 restrita	Palavras de conteúdo semântico em T2 ampliada
Introdução	259	556	17,46	–	12,19	distribuição	idade, distribuição
Metodologia	292	603	20,05	–	13,68	–	artigos
Resultado e discussão	648	2 074	27,43	brasileiros, artigos, pesquisadores, literatura, trabalhos, países	19,28	pesquisadores, literatura, trabalhos, países, idade, referências, avançados, anos, citações	25 classes de freqüência e 52 palavras distintas
Conclusão	175	290	15,69	–	10,85	–	científica
Total	961	3 622	32,94	artigos, pesquisadores, brasileiros, países, idade, literatura, referências, trabalhos, citações, anos, avançados	23,14	países, literatura, idade, referências, trabalhos, citações, distribuição, cientistas, antigas, científica, pesquisa, avançados, anos	41 classes de freqüência e 402 palavras distintas

Quadro 2 – Síntese dos resultados do Texto 2

A Bibliometria no Brasil

	Total de palavras distintas	Total de palavras	Ponto T	Palavras de conteúdo semântico em T	Ponto T2	Palavras de conteúdo semântico em T2 restrita	Palavras de conteúdo semântico em T2 ampliada
Introdução	254	453	18,59	–	13,22	–	–
Material e Método	146	233	14,54	–	10,14	–	–
Resultados	587	1 770	25,54	lei, trabalhos, teses, segundo, bradford, autores, produção, tipo, literatura	17,91	teses	lei, trabalhos, teses, goffman, segundo, bradford, autores, produção, tipo, literatura, produtividade, brasileira, braga, aplicação, aplicações, tabela, zipf
Conclusão	181	316	15,63	–	10,91	–	
Total	832	2 771	31,34	lei, produção, bradford, teses, autores, literatura, informação, trabalhos, tipo, brasileira, aplicação, abordagem, bibliométrica, braga, segundo, frente, goffman, produtividade, pesquisa, ciência	22,53	produção	lei, produção, bradford, teses, autores, literatura, informação, trabalhos, tipo, brasileira, aplicação, abordagem, bibliométrica, braga, segundo, frente, ciência, goffman, produtividade, IBICT, pesquisa, zipf, bibliometria

Quadro 3 – Síntese dos resultados do Texto 3

Estudo bibliométrico: análise de citações no periódico Scientometrics

	Total de palavras distintas	Total de palavras	Ponto T	Palavras de conteúdo semântico em T	Ponto T2	Palavras de conteúdo semântico em T2 restrita	Palavras de conteúdo semântico em T2 ampliada
Parte 1	308	696	19,35	periódicos	13,94	periódicos	periódicos, informação, número, artigos, núcleo, lei, produtividade
Parte 2	533	1 543	25,77	periódicos citações citados	18,08	citados, número volumes títulos	periódicos, citações, citados, número, volume, títulos, total, volumes, tabela, bradford, produtividade, virtch's, artigos, fascículos, identificados, ordem, periódico, origem
Parte 3	344	790	20,38	periódicos	14,27	periódicos	periódicos, citados, tabela
Total	904	3 031	32,94	periódicos, citações, citados, número, títulos, tabela, artigos, produtividade	23,18	citados, número títulos, tabela	periódicos, citações, tabela, citados, número, títulos, artigos, produtividade, total, bradford, informação, volumes, volume, núcleo, periódico, dados, scientometrics, zonas, ordem, bibliometria, autores, longo, literatura, wirtch's, citado, estudos, estudo, fascículos, origem, relações, relação, tempo, ciência, estrutura, distribuição, etc., países, identificados, zona

Quadro 4 – Síntese dos resultados do Texto 4

Análise da literatura brasileira de Siderurgia

	Total de palavras distintas	Total de palavras	Ponto T	Palavras de conteúdo semântico em T	Ponto T2	Palavras de conteúdo semântico em T2 restrita	Palavras de conteúdo semântico em T2 ampliada
Introdução	342	819	21,32	–	14,93	número, periódicos, lei, bradford	número, periódicos, lei, bradford, artigos, autores, trabalhos
Objetivo e Material	105	164	12,15	–	8,45	–	siderurgia
Método	210	502	15,50	artigos, periódicos, lei	10,82	periódicos, lei	publicados, lei, número dados, tabela, autores periódicos, artigos
Resultados	184	327	14,60	artigos, periódicos, tabela, número, autores, citados	10,32	artigos, periódicos, tabela, número, autores, citados	–
Conclusão	257	455	19,25	–	13,47	–	–
Total	836	2 267	29,80	artigos, periódicos, número, autores, bradford, tabela, literatura, trabalhos	20,93	autores, bradford	artigos, periódicos, número, autores, bradford, tabela, literatura, trabalhos, lotka, produtividade, siderurgia, trabalho, produção, autor, distribuição, dados, núcleo, zonas, aplicação, citados, total, determinado, periódico

Quadro 5 – Síntese dos resultados do Texto 5

Validade da análise de citação como indicador de qualidade da produção científica - uma revisão

	Total de palavras distintas	Total de palavras	Ponto T	Palavras de conteúdo semântico em T	Ponto T2	Palavras de conteúdo semântico em T2 restrita	Palavras de conteúdo semântico em T2 ampliada
Introdução	129	218	13,14	–	9,15	–	–
Origem dos conveniências críticos	569	1 625	27,03	citações, trabalho, autor	18,97	trabalho, autor	citações, trabalho, autor, trabalhos, art análise, citação, autores
Citação, qualidade e produtividade científica	217	422	17,50	–	12,23	qualidade, citações	qualidade, citações, trabalho
Citar ou não citar	348	787	21,73	citar, trabalho	15,22	citar, trabalho	citar, trabalho, cons, autor, citações
Conclusões	155	258	14,34	–	10,50	–	–
Total	984	3 310	34,26	citações, trabalho, autor, citação, qualidade, análise, trabalhos, cole, autores	24,10	autor, citação, qualidade, análise	citações, trabalho, autor, citação, qualidade, análise, trabalhos, cole, autores, citar, art, artigos, cocos, uso, estudo, exemplo, literatura, razões, científica, citados, científicas, garfield, método, pesquisa, ciência, citado, lawani, importante, número periódicos

Quadro 6 – Síntese dos resultados do Texto 6

Bibliografia Brasileira de Botânica

	Total de palavras distintas	Total de palavras	Ponto T	Palavras de conteúdo semântico em T	Ponto T2	Palavras de conteúdo semântico em T2 restrita	Palavras de conteúdo semântico em T2 ampliada
Introdução e seleção da amostra	151	281	14,40	–	10,04	–	–
Metodologia	172	469	13,14	artigos, periódicos, total, autor, número	9,15	periódicos, total, autor, número, autores	artigos, periódicos, total, autor, número, autores, tabela, produtividade, relação, produção, distribuição, elaboração, artigo, divisão, decrescente, ordem, publicados, média, dados, incluindo, gráficos
Resultados	181	350	13,36	artigos, periódicos	9,31	–	artigos, periódicos, artigo, produzidos, total
Análise e conclusões	318	730	19,95	autores	13,96	autores	autores
Total	567	1 810	25,84	artigos, autores, total, periódicos, número, artigo, produtividade	18,13	periódicos, total, número, artigo, produtividade, autor, distribuição	132 palavras distintas

Quadro 7 – Síntese dos resultados do Texto 7

Processo de crescimento epidemiológica aplicado à literatura brasileira de Doença de Chagas

	Total de palavras distintas	Total de palavras	Ponto T	Palavras de conteúdo semântico em T	Ponto T2	Palavras de conteúdo semântico em T2 restrita	Palavras de conteúdo semântico em T2 ampliada
Introdução	127	204	13,43	–	9,32	–	–
Teoria do processo epidêmico	329	616	20,09	processo, população, tempo	14,08	processo	processo, população, tempo, trabalho, ex...ção
Doença de Chagas	219	390	17,94	chagas	12,54	–	chagas
Metodologia	183	354	15,06	chagas	10,51	chagas	chagas, bibliografia
Resultados e análise	279	762	17,39	autores, trabalho, ano, número	12,15	número	autores, trabalho, ano, variação, número, amostra, gráfico, chagas, trabalhos, diferença, período, processo, acompanhamento, grande, epidemiológico, infectados, valores, removidos
Conclusão	107	190	12,15	autores	8,45		autores, número
Total	649	2 708	31,84	autores, chagas, número, doença, trabalho, processo	22,37	número, doença, trabalho, processo, variação, chagas	autores, chagas, número, gráfico, doença, trabalhos, trabalho, processo, população, literatura, crescimento, tempo, assunto, ano, endêmica, período, mediana, brasileira, publicação, atoma, infectados, bibliografia, gramos, epidemiológico, removidos, variação, relação, gottman, produção, publicados

Quadro 8 – Síntese dos resultados do Texto 8

Produtividade científica: uma análise parcial da literatura

	Total de palavras distintas	Total de palavras	Ponto T	Palavras de conteúdo semântico em T	Ponto T2	Palavras de conteúdo semântico em T2 restrita	Palavras de conteúdo semântico em T2 ampliada
Introdução	178	301	16,53	–	11,56	–	–
Produt. Cient. e Fatores Comportam.	392	743	22,04	–	15,44	–	–
Produt. Cient. e Aspectos Ambientais	309	564	20,06	–	14,00	universidades	universidades, cientistas
Variáveis na Produtividade	258	493	18,05	produtividade	12,62	produtividade	produtividade, publicação
Prod. Cient. e Comunicação	356	747	22,46	–	15,75	informação	informação, científica, cientistas
Produtividade e Sistemas de Informação	395	842	23,25	–	16,30	–	informação
Conclusão	265	504	19,35	–	13,54	–	–
Total	1 306	4 194	40,27	científica, produtividade, pesquisa, informação, cientistas	28,33	pesquisa, informação, processo, cientistas	científica, produtividade, cientistas, pesquisa, informação, processo, comunicação, conhecimento, diferentes, fatores, ciência, informações, uso, bibliotecas, problemas, organização, desenvolvimento, pesquisadores, reconhecimento, estudos, produção, importância, científico, comunidade, características, principais, trabalho, universidades, cientistas

Quadro 9 – Síntese dos resultados do Texto 9

A lei de Lotka sobre a produtividade de autores: aplicabilidade do quadrado inverso

	Total de palavras distintas	Total de palavras	Ponto T	Palavras de conteúdo semântico em T	Ponto T2	Palavras de conteúdo semântico em T2 restrita	Palavras de conteúdo semântico em T2 ampliada
Introdução	236	420	18,26	–	12,77	–	–
Segunda parte	455	1 261	22,91	lei, lotka, literatura, ciências	18,08	ciências literatura ciência	lei, lotka, ciências, literatura, medicina, brasileira, assuntos, ciência, área, biblioteconomia, legal, médica, valências, voos, astronomia, estudos, geração vargas, resultados, siderurgia, história, produtividade, humanas, melhor, tecnologia
Terceira parte	615	1 877	23,33	autoria, ciências, price, lei lotka, ciência, grau, literatura, ideal, número, médica, assuntos	16,35	média, artigos, informação; biblioteconomia, especialista, medicina, tecnologia, história, price, lotka, ciência, assuntos, egéa, número, ciências, lei, literatura, grau, médica	30 classes de palavras e 231 palavras distintas
Quarta parte	120	250	12,0	história	8,3+	história, ciência	história, ciência, informação, biblioteconomia, egal, med, tecnologia
Quinta parte	204	421	15,87	lei	11,08	lei	lei, lotka
Total	984	4 229	33,32	lei, autores, ciência, ciências, assuntos, price, literatura, história, solla, número, medicina, grau, biblioteconomia, médica, tecnologia, informação, legal, brasileira, humanas, produtividade, trabalhos, relação, espanhola, exédia, artigos, vargas, siderurgia, computação, resultados	23,42	história, price, número, solla, medicina, grau, biblioteconomia, médica, tecnologia, informação, legal, brasileira, humanas, produtividade, trabalhos, média, relação, espanhola, artigos, vargas, siderurgia, computação, resultados	todas as palavras da distribuição

Quadro 10 – Síntese dos resultados do Texto 10

Um estudo sobre a produção científica brasileira segundo os dados do Institute for Scientific Information (ISI)

	Total de palavras distintas	Total de palavras	Ponto T	Palavras de conteúdo semântico em T	Ponto T2	Palavras de conteúdo semântico em T2 restrita	Palavras de conteúdo semântico em T2 ampliada
Introdução	213	391	15,58	–	11,78	ciências, ISI	ciências, ISI, dados, artigos
Metodologia	219	397	17,45	–	12,19	–	–
Resultados e conclusões	565	1.525	27,14	autores	19,05	autores, produção, instituições	autores, produção, instituições dados, política científica, tempo, ciências, número, país, 1974
Total	774	2 129	31,44	autores, produção	22,08	autores, produção, dados; ciências	autores, dados, produção científica instituições, países número, grafico, 1974

Quadro 11 – Síntese dos resultados do Texto 11

Brasil: exercícios de crescimento para os anos 90

	Total de palavras distintas	Total de palavras	Ponto T	Palavras de conteúdo semântico em T	Ponto T2	Palavras de conteúdo semântico em T2 restrita	Palavras de conteúdo semântico em T2 ampliada
Introdução	280	560	20,0		14,0	–	–
Experiência de crescimento e o investimento público	586	451	26,45	investimento, crescimento, anos	19,97	investimento crescimento	investimento, crescimento, anos, público, taxa
Um modelo de três hiatos	320	966	19,15	equação, investimento, nível, poupança, valor, público	13,40	nível, poupança	20 classes de frequência e 19 palavras
Valores possíveis	115	195	12,62	–		–	–
Cenários de crescimento	382	615	22,44	fiscal, crescimento	15,72	fiscal, crescimento	fiscal, crescimento, taxa, poupança, relação
Análise de sensibilidade	310	978	18,42	taxa, crescimento, cenário, investimento, valor, barra, relação, sensibilidade, capacidade, capacidade	12,88	crescimento, valor, cenário, relação, investimento	19 classes de frequência e 84 palavras
Conclusão	275	565	19,80		13,77	–	
Total	1 331	716	39,19	investimento, economia, fiscal, valor, setor, PIB, privado, aumento, décadas, utilização, crescimento, público, restrição, anos, externa, barra, taxas, taxa, nível, capacidade, equação, valores, economia, painel, setor	27,55	nível, aumento, privado, setor, valor fiscal, PIB, externa, anos, restrição, economia, valores, equação, capacidade	lista e distribuição de palavras