ESTUDOS

OTIMIZAÇÃO DOS PROCESSOS DE INDEXAÇÃO DOS DOCUMENTOS E DE RECUPERAÇÃO DA INFORMAÇÃO MEDIANTE O USO DE INSTRUMENTOS DE CONTROLE TERMINOLÓGICO.

Jaime Robredo
Chefe do Departamento de Biblioteconomia da Universidade de Brasília

RESUMO

Em qualquer área do conhecimento, os termos portadores de significado podem ser utilizados como descritores para representar o conteúdo dos documentos escritos, nos processos de indexação e organização da informação, assim como para formular as perguntas no processo de recuperação da informação. Quando ordenados em função de sua freqüência de uso, os descritores apresentam uma distribuição que obedece à lei de Bradford-Zipf. A aplicação das facilidades do processamento eletrônico de dados ajuda grandemente a estabelecer, para áreas específicas do conhecimento, instrumentos de controle terminológico, que permitem otimizar os processos de indexação e recuperação dos documentos, utilizando os termos e as associações entre estes que se destacam por sua riqueza de significado, para representar conceitos determinados. Foram estudados dois universos de termos significativos correspondentes a duas áreas diferentes do conhecimento (agricultura e política científica e tecnológica), resultantes da análise do conteúdo de conjuntos suficientemente grandes de documentos, indexados segundo princípios não subjetivos: indexação automática no primeiro caso e indexação automática simulada no segundo. Desse estudo resultam, para cada caso, listas de descritores estabelecidas a partir de suas respectivas freqüências de aparecimento e da aplicação de determinados conceitos de sinonímia e quase-sinonímia, que devem contribuir para otimizar os processos de indexação e recuperação da informação, tanto em sistemas manuais como automatizados. As conclusões estabelecidas no que diz respeito à otimização de indexação parecem confirmadas a partir dos conceitos da teoria da informação.

Descritores: Indexação; Controle terminológico; Recuperação da informação.

1 - INTRODUÇÃO

Os estudos que podem servir de base à construção de instrumentos de controle terminológico são numerosos em outros países* e, nos últimos anos, também foram publicados alguns trabalhos referentes à língua portuguesa, dentre os quais merece destacar-se o estudo de Lima e Maia[1]. As autoras estudaram a aplicabilidade a textos em língua portuguesa dos resultados de Zipf[2,3] e Goffman[4], que determinaram, respectivamente, a relação entre a ordem de série de uma palavra (em ordem de freqüência) e a freqüência de aparecimento em um texto "suficientemente longo", e as ordens de série nas quais devem encontrar-se as palavras significativas de um texto em língua inglesa.

* Ver, por exemplo: SARACEVIC, T. **Introduction to Information Science.** New York, Bowker, 1970, 751 p.

As referidas autoras aplicaram uma metodologia baseada nos estudos de Booth[5], que enuncia que, quando as palavras de um texto qualquer são ordenadas numa tabela, em ordem decrescente de freqüência de aparecimento, o produto da ordem na série (r) da palavra por sua freqüência (f) é uma constante (C):

$$r \times f = C \qquad [1]$$

chegando à conclusão de que a lei se aplica também à língua portuguesa, sendo, porém, diferente o valor da constante C para as duas línguas.

Estudos de freqüência de aparecimento dos termos significativos, nos títulos dos trabalhos publicados durante um determinado período, foram realizados pelo autor deste trabalho para construir um núcleo de thesaurus para indexação e. recuperação da literatura agrícola brasileira [6, 7].

A aplicabilidade universal das leis de Zipf e de Bradford confirma-se em numerosos estudos bibliométricos* Em áreas relacionadas com a lingüística documentária encontram-se, além dos trabalhos já citados, outros referentes a estudos de freqüência de palavras em determinados textos que comprovam, de maneira mais ou menos generalizável, a validade da lei de Zipf*!

A observação de que, na relação [1], os valores do produto r x f só são aproximadamente constantes na parte central da tabela em que se apresentam os termos ordenados de acordo com sua freqüência de aparecimento, existindo importantes desvios para os valores correspondentes aos termos de baixa e elevada freqüência, levou a considerar a possibilidade de utilizar a lei de Bradford, na sua formulação dada por Brookes[8], para estudar as leis que regem a distribuição das ocorrências dos descritores — e, eventualmente, de suas associações —, com vistas ao estabelecimento de vocabulários simplificados que permitam a otimização dos processos de indexação e recuperação, em áreas específicas do conhecimento[9].

Neste trabalho apresentam-se os resultados do estudo realizado sobre os termos significativos utilizados na literatura agrícola brasileira, de uma parte, e nas publicações brasileiras referentes a política científica e tecnológica, de outra parte[10, 11], ampliando e completando os resultados apresentados numa comunicação anterior[12].

2- METODOLOGIA

Para o estudo dos termos significativos referentes à agricultura, utilizaram-se as listagens de controle de freqüência de aparecimento dos descritores identificados no processo de indexação automática de cerca de 10.000 projetos de pesquisa, com um programa ao qual se faz referência em outra publicação[13] e que serviria de base ao processamento dos dados que figuram no Guia Brasileiro de Pesquisa Agrícola em Andamento, publicado pela Biblioteca Nacional de Agricultura[14]. As listagens acima referidas foram revisadas, para eliminar alguns erros de transcrição e corrigir algumas formas defeituosas dos descritores, assim como eliminar termos não significativos que não foram identificados anteriormente e identificar alguns sinônimos. A partir dessas listagens revisadas, foram obtidas novas, nas quais os descritores apareciam ordenados em ordem alfabética e em ordem decrescente de freqüência*

Para o estudo do universo de termos referentes à política científica e tecnológica, utilizaram-se documentos gerados pelo Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), existentes no Centro de Informação sobre Política Científica e Tecnológica (CPO) desse organismo. Os referidos documentos foram indexados manualmente, seguindo normas que visavam a simular o mais fielmente possível o processo de indexação automática. Os dados correspondentes a cada documento foram processados de maneira a se obter diversos produtos de interesse imediato para o CPO (índices de autores, índices de entidades, índices de assunto, listagens de referências bibliográficas e fichas catalográficas), assim como listagens de freqüência de aparecimento dos descritores e de freqüência de

* Ver, por exemplo, entre outros, os trabalhos de Solla Price, Kessler, Garfield, etc., citados por T. SARACEVIC, em Introduction to Information Science (New York, Bowker, 1970, p. 726-47), assim como os trabalhos de J. ROBREDO, La Dispersion des Information dans la littérature Verrière (Verres et Réfract 27 (3): 117-40, 1973), de G.H. BRAGA, Relações bibliométricas entre a frente de pesquisa (research front) e revisões da literatura: estudo aplicado à Ciência da Informação (Ci. Inf., 2 (1): 9-26, 1973), e de L.M. de FIGUEIREDO, Distribuição da Literatura geológica brasileira: estudo bibliométrico (Ci. Inf., 2 (1): 27-40, 1973).

** Ver, por exemplo: RIBEIRO, L.A. Aplicação dos métodos estatísticos e da teoria da informação e da comunicação na análise lingüística: estudo da linguagem jornalística. Ci. Inf., 3 (2): 151-54, 1974; PARKER -- RHODES, A.F. & JOYCE, T.A. Study of Word Frequency Distribution. Naturé, 178 (4545): 1308, 1956.

* O processamento, para obtenção dessas listas, foi realizado utilizando-se de programas colocados à nossa disposição pelo Departamento de Estatística da Universidade de Brasília, cuja contribuição e apoio agradecemos.

associação entre estes* Após revisão e depuração das referidas listagens, como no caso anterior, foram preparadas novas listas de descritores ordenados em ordem alfabética e em ordem decrescente de frequências de aparecimento.

Em ambos os casos, os dados assim apresentados serviriam de ponto de partida para a preparação de tabelas que indicassem o total de descritores que aparecem com uma determinada frequência.

Finalmente, partindo da ordenação dos descritores em série decrescente de frequência de aparecimento e dos sucessivos totais acumulados, correspondentes a cada termo da série, foram traçadas as curvas que representam graficamente a lei de Bradford, como se detalha a seguir.

3- RESULTADOS

As Figuras 1 e 2 mostram, respectivamente, a estrutura das tabelas de descritores ordenados alfabeticamente, com suas respectivas frequências de aparecimento, no caso de documentos referentes à pesquisa agrícola e à política científica e tecnológica.

As Figuras 3 e 4 permitem ver a estrutura das tabelas de frequência de aparecimento dos descritores, nos dois universos estudados.

As Figuras 5 e 6 permitem ver, respectivamente, -a estrutura das tabelas correspondentes aos dois universos estudados, que mostram o número de descritores diferentes que ocorrem um determinado número de vezes, indicando, ao mesmo tempo, o total acumulado de ocorrências. Os dados da terceira coluna dessas tabelas, ou seja, o número acumulado de descritores dispostos em ordem decrescente de frequência, são transportados sobre o eixo das abscissas, em escala logarítmica, num diagrama ortogonal; sobre o eixo das ordenadas são transportados os correspondentes totais acumulados de ocorrências. As curvas resultantes, traçadas nas Figuras 7 e 8, são a representação gráfica, segundo a lei de Bradford, da variação do número de ocorrências dos descritores em função do número de descritores identificados, em cada um dos casos estudados. Com efeito, de acordo com a formulação de Brookes[8], a lei de Bradford pode escrever-se em duas partes:

$$R(n) = \alpha n^{\beta} \quad (1 < n < c) \quad [2a]$$

$$= k \log n/s \quad (c < n < N) \quad [2b]$$

* Os programas utilizados integram o sistema BI B/ BATCH, desenvolvido pelo autor deste trabalho, operado pelo Departamento de Biblioteconomia da Universidade de Brasília, para fins académicos e de pesquisa.

que se referem, respectivamente, à parte curva inicial e à parte reta central de gráfico. O significado dos símbolos é o seguinte:

R (n) = totais acumulados de ocorrências,
α = número de vezes que ocorre o termo mais utilizado,
β = um número, sempre inferior a 1, que geralmente varia de um caso para outro,
n = número de termos com uma determinada frequência,
N = número de ordem, na série de descritores, do último descritor, no fim da parte reta da curva,
k = declive da parte reta da curva.

Pode observar-se que as curvas obtidas nos dois casos estudados apresentam o mesmo aspecto, com uma parte curva inicial de declive crescente, uma parte reta central e uma parte curva final com declive decrescente.

A análise, nas tabelas representadas nas Figuras 1, 2, 5 e 6, dos termos que aparecem em cada uma dessas três zonas, para cada um dos universos estudados, permite estabelecer algumas conclusões interessantes.

Os termos que se situam na primeira parte curva são termos de frequência de aparecimento muito elevada e são, em grande parte, os termos que definiriam a abrangência ou o escopo do universo de documentos considerados. Assim, no caso dos documentos de política científica e tecnológica, encontram-se, nessa zona, termos como:

CNPQ, PESQUISA, PROGRAMA, CIÊNCIA E TECNOLOGIA, etc.,

que são, obviamente, termos redundantes com o escopo escolhido. No caso dos descritores identificados a partir dos títulos das pesquisas agrícolas, encontram-se também termos que caracterizam o(s) assunto(s) coberto(s) pelas pesquisas consideradas:

PRODUÇÃO, MELHORAMENTO, CULTURA, BOVINO, CULTIVAR, FRUTO, SOJA, SEMENTE, MILHO, etc.

Estes tipos de termos que se situam entre os de mais elevada frequência poderiam ser denominados **descritores de escopo** e, como veremos depois, não parece que alguns dentre eles sejam necessários para caracterizar o conteúdo dos documentos no processo de indexação dos mesmos ou para formular as perguntas, uma vez que a abrangência do assunto coberto pelos documentos processados foi bem

definida. Seriam, pelo contrário, termos do maior interesse para caracterizar áreas específicas do conhecimento em acervos ou arquivos com caráter mais ou menos interdisciplinar. Assim, os descritores mais freqüentes identificados na análise dos documentos referentes a política científica e tecnológica indicam claramente que uma grande parte dos documentos se refere a PROGRAMAS DE PESQUISA subsidiados pelo CNPq em áreas de CIÊNCIA E TECNOLOGIA. No caso das pesquisas agrícolas, é evidente que a agricultura e pecuária são áreas de grande abrangência e, assim, os produtos encontrados mais freqüentemente aparecem como descritores de escopo dentro da variedade do conjunto de pesquisas agrícolas.

Os termos que se situam na parte central da curva, de acordo com sua freqüência de aparecimento, são, de toda evidência, nos dois casos estudados, aqueles que caracterizam as subáreas de interesse dentro de cada área identificada pelos descritores mais freqüentes acima referidos. Poderiam designar-se como descritores de facetas e são eles os que, em número relativamente limitado, permitem caracterizar e pré-selecionar os documentos de interesse com eficiência e rapidez. No caso dos documentos referentes a política científica e tecnológica, encontram-se, nessa zona, termos como:

RECURSOS HUMANOS, AGRICULTURA, INDÚSTRIA, COOPERAÇÃO INTERNACIONAL, ENERGIA, PÕS-GRADUAÇÃO, TRANSPORTE, etc.

que permitem segmentar o arquivo em grupos de documentos que tratam de aspectos bastante diferenciados. No caso das pesquisas agrícolas, encontram-se termos do tipo:

FISIOLOGIA, ALGODÃO, DOENÇA, IRRIGAÇÃO, PRAGA, MELHORAMENTO GENÉTICO, FERTILIDADE, ENTOMOLOGIA, FITOMELHORAMENTO, TRATO CULTURAL, ROTAÇÃO, CLONE, etc.,

que apresentam as mesmas características.

Deve-se sublinhar aqui que a fronteira entre os descritores de escopo e os descritores de facetas é impossível de se estabelecer de maneira absolutamente clara, já que depende da abrangência do assunto (agricultura é um tema muito mais amplo do que política científica e tecnológica) e do número de documentos considerados (alguns milhares no caso de pesquisa agrícola e algumas centenas no caso de política científica e tecnológica). Assim, pode acontecer que termos que no primeiro caso aparecem como pertencendo a um grupo se identificam, no segundo,como tendo mais afinidade com o outro.

Na parte curva final do gráfico aparecem numerosos termos de freqüência de aparecimento muito baixa, entre os quais figuram muitos vazios de significado, que não foram identificados anteriormente e que devem ser eliminados do vocabulário e introduzidos na tabela de palavras vazias, que utiliza o programa de indexação automática. Também se encontram termos sinônimos ou quase-sinônimos de outros mais freqüentemente usados, que aparecem entre aqueles que se situam nas duas zonas da curva anteriormente estudadas, e que podem, após introdução das remissivas correspondentes, ser eliminados do vocabulário. Os termos que restam são de alta especificidade e caracterizam, num acervo ou arquivo determinado, um número muito restrito de documentos. Poderiam denominar-se descritores pontuais. No caso dos documentos sobre política científica e tecnológica, identificam-se entre estes:

ADUBO ORGÂNICO, ADUBO QUÍMICO, BIOSFERA, DECOMPOSIÇÃO CATALITICA, DOENÇA DE CHAGAS, MICROCLIMA, RENDA FAMILIAR, etc.,

que "apontam" para assuntos muito específicos dentro do escopo geral dos documentos considerados. No caso da pesquisa agrícola, encontram-se nessa zona termos como:

AGROTÓXICOS, ALGODÃO ARBÓREO, CHUCRUTE, CLORETO DE POTÁSSIO, DEFOLIAÇÃO, GÁS METANO, LARANJA DOCE, MICROPLASMOSE, NOVILHA NELORE, etc.,

e, evidentemente, os nomes científicos:

ANANAS COMOSUS, CANAVALIA ENSIFORMIS, EUCALYPTUS ALBA, HEVEA BRASILIENSIS, MACRODON ANCYLODON, MANIHOT ESCULENTA, PINUS CARIBEAE, etc.,

que identificam produtos específicos ou correspondem a termos mais específicos relacionados com outros descritores incluídos nas zonas anteriores.

As Figuras 9 e 10 são uma representação gráfica das três zonas em que se localizariam os três tipos de descritores, na curva de Bradford-Brookes.

4- CONSIDERAÇÕES FINAIS

Os resultados apresentados levam naturalmente a questionar a validade de certos esforços de aprimoramento na construção de thesauri, nos quais

o empenho de completar as relações estruturais entre os termos leva à introdução contínua de outros novos, chegando-se em pouco tempo a conjuntos de vários milhares, dos quais a maioria nunca virá a ser utilizada. Em outro trabalho, já citado anteriormente[6], foi evidenciada a complexidade, em termos de lógica booleana, das estratégias de busca, quando se deseja cobrir todas as relações possíveis entre os termos, "para não se perder nada".

Mais prudente (e mais econômico) parece ser tentar desenvolver, em cada área do conhecimento, instrumentos de controle terminológico mais simples, organizados em níveis de especificidade/freqüência, facilmente diferenciáveis a partir de estudos do tipo aqui apresentado, que permitam indexar facilmente os documentos com flexibilidade *e,* sobretudo, formular as perguntas, no processo de recuperação da informação, com grande economia de termos e com grande precisão, combinando um ou dois escolhidos entre os descritores de facetas (ou entre os descritores de escopo), com um ou dois mais específicos, escolhidos entre os descritores pontuais*

A inclusão dos termos em uma ou outra das três zonas identificadas a partir da curva resulta de um fato experimental decorrente da freqüência de uso dos descritores e não tem nada a ver com qualquer conceito de classificação ou hierarquia. Por isso um mesmo termo pode aparecer em zonas diferentes, quando usado em diversas áreas do conhecimento'. Também termos da mesma natureza podem encontrar-se em zonas diferentes**

As técnicas de indexação automática de textos, ou mesmo as técnicas de indexação automática simulada, na medida em que não são subjetivas, podem ajudar de maneira decisiva na preparação de instrumentos de controle terminológico que reflitam, com razoável fidelidade, o conteúdo dos documentos que integram os arquivos processados e assegurem a obtenção de bons resultados no processo de recuperação da informação[15, 16], ao permitirem a atualização do vocabulário controlado ao mesmo tempo que se atualizam as bases de dados.

Os estudos de freqüência de aparecimento dos descritores e, eventualmente, de suas associações, aparecem uma vez mais como da maior importância na elaboração de instrumentos de controle terminológico para indexação dos documentos e recuperação da informação armazenada.

APÊNDICE

Cálculo, com base na teoria da informação, da informação máxima, da informação relativa e da redundância correspondentes ao uso dos instrumentos de controle terminológico estudados.

Os exemplos fornecidos por Moles[17], com base nas idéias de Shannon[18] e Guílbaud[19], da medida da informação em textos ou mensagens concretos podem servir de ponto de partida para analisar, mutátis mutandis, a taxa de informação das "mensagens" resultantes da indexação dos documentos mediante instrumentos de controle terminológico do tipo dos discutidos neste trabalho. Paralelamente, os resultados obtidos utilizando os princípios da teoria da informação deveriam servir de ponto de referência para avaliar as conclusões estabelecidas. Convém esclarecer o significado de alguns conceitos:

Informação (taxa de informação, taxa de originalidade):

A quantidade de informação transmitida por uma mensagem é o logaritmo binário do número de mensagens possíveis e com a mesma estrutura aparente, entre as quais o transmissor teve que escolher. No caso de uma mensagem de N elementos tirados de um repertório de n símbolos tendo probabilidades de ocorrência PJ a informação H é, em unidades de informação:

$$H = - N \; Z \; pi, \; \log_{2pi} \qquad [3]$$

Informação máxima:

A informação apresenta seu rendimento máximo H_m para um dado número de símbolos, se a estrutura da "linguagem" definida por esse repertório de símbolos e utilizada na composição da mensagem for tal que

* Vale a pena lembrar-se da analogia conceituai da indexação com dois tipos de descritores, com o princípio de indexação no sistema AGRIS, que utiliza as chamadas categorias de assunto e os códigos de objeto (ver, por exemplo: AGRIS- Categorias de Assunto (Rev. 1). Brasília, SNIDA, 1977. 153 p. Projeto PNUD/FAO/BRA/72/020, DOC/TEC/75/001 (Rev. D). A analogia, porém, não parece ir muito mais longe, já que nesse sistema tanto a escolha limitativa dos conceitos de indexação como da "classificação" numa ou noutra classe não parece basear-se em qualquer tipo de análise estatística da freqüência de uso dos mesmos. Isso dá lugar, na prática, como acontece com a maioria dos sistemas de classificação e thesauri até agora desenvolvidos, a que muitos termos ou conceitos nunca sejam usados ao mesmo tempo que se detecta a falta de outros "não-autorizados".

** Assim, por exemplo, três produtos agrícolas (SOJA, GIRASSOL, AÇAÍ) encontram-se cada um numa zona diferente.

cada símbolo tenha uma probabilidade de ocorrência igual (símbolos **equiprováveis**)*

Redundância:

Chama-se redundância (R) a quantidade:

$$R = 1 - \frac{H_1}{H_m} \qquad [4]$$

exprimindo (em porcentagem) o que é dito em demasia na mensagem, o "desperdício" de símbolos resultante de uma **codificação** defeituosa.
Redundância e informação fornecida por um dado tipo de mensagem são, por **definição,** independentes do **extrato** particular da mensagem escolhida, mas dependem do conjunto de conhecimentos comuns ao receptor e ao transmissor, conduzindo à ideia de uma informação diferencial, ao menos no caso dos receptores humanos.

Nos casos estudados, considerando como símbolos os descritores utilizados na indexação, teríamos, para cada área do conhecimento considerada, os seguintes dados e **resultados:**

PESQUISA AGRÍCOLA

Número de termos: n =5800
Número total de ocorrências: N= **75000**

Distribuindo as probabilidades de distribuição p_i dos termos em três **grupos,** de acordo com os valores da curva da Figura 7, temos:

I 40 termos com p_I = 0,62%
II 750 termos com p_{II} = **0,08%**
III 5000 **termos** com Pm = 0,003%

$40p_I$ = 26% de termos muito frequentes
$750p_{II}$ = 60% de termos frequentes
$5000p_{III}$ = **14%** de termos ocasionais

Donde a informação:

$$H = -75000 \sum_{i=I}^{i=III} p_i \log_2 p_i$$

$$= -75000 \sum_{i=I}^{i=III} 5800 \times 3{,}32\ p_i \log p_i = 190921$$

unidades arbitrárias de **informação***

A informação máxima que poderiam representar as 75000 ocorrências dos 5800 termos, se todos tivessem a mesma probabilidade de ocorrência (Pi = 1/5800) é:

$H_m = -75000 \log_2 1/5800$

= - 75000 x 3,32 **log 1/5800** = 933075 unidades arbitrárias de informação.

A redundância é:

$$R = 1 - \frac{H}{H_m} \cong 1 - 0{,}2 \cong 0{,}8 \cong 80\%.$$

POLÍTICA CIENTÍFICA E TECNOLÓGICA

Número de termos: n = 1900
Número total de ocorrências: N = 9400

Distribuindo as probabilidades de distribuição p; dos termos em três grupos, de acordo com os valores da curva da figura 8, resulta:

I 10 **termos** com p_I = 25%
II 450 termos com PU = 0,13%
III 1400 **termos** com Pm = **0,01%**

$10p_I$ = 25% de termos **muito freqüentes**
$400p_{II}$ = 55% de termos frequentes
$1400p_{III}$ = 20% de termos ocasionais

Donde a informação:

$$H = -9400 \sum_{i=I}^{i=III} 1900 \times 3{,}32\ p_i \log p_i = 58938$$

unidades arbitrárias de informação

A informação máxima:

H_m = - 9400 x 3,32 log **1/1900** = 102366 unidades arbitrárias de informação

A redundância

$$R = 1 - \frac{H}{H_m} \cong 1 - 0{,}6 \cong 0{,}4 \cong 40\%$$

* A informação é uma quantidade, essencialmente diferente da significação e independente desta. Uma mensagem de informação máxima pode parecer desprovida de sentido, se o indivíduo não for **suscetível** de decodificá-la para reconduzi-la a uma forma inteligível. De maneira geral, a intelegibilidade varia no **sentido** inverso da informação. Complexidade e informação de uma estrutura, de uma forma ou de uma mensagem são sinónimas.

* As fórmulas originais, quando aplicadas a caracteres, expressam a informação em **kits.** Em nosso caso, ao usarem-se termos em lugar de caracteres, poder-se-ia definir uma unidade da informação **múltipla** do bit (por exemplo, equivalente ao número médio de caracteres dos termos registrados). Essa unidade, porém, careceria de qualquer significado físico. Por isso, parece **preferível** utilizar a expressão "unidade arbitrária de informação". Por outra parte, essa aparente **indefinição** carece de importância já que, ao dividir H por H_m para calcular a redundância (R), o resultado é uma quantidade sem dimensões.

Os valores das originalidade» relativas H/H_m (respectivamente 20 e 60%), nos dois casos considerados, parecem corresponder aos valores delimitados pelas ocorrências relativas (em porcentagem) correspondentes aos respectivos termos ocasionais ou pontuais (limite m mimo) e pelos valores dessas mesmas ocorrências acrescidos dos correspondentes às ocorrências dos termos do escopo (limite máximo).

No caso dos documentos sobre pesquisas agrícolas:

limite mínimo:

Total de ocorrências dos termos pontuais - 10000 = 15%;
Total de ocorrências de todos os termos 74900

limite máximo:

Total de ocorrências dos termos pontuais
+ total de ocorrências dos termos de facetas - 29000 = 40%
Total de ocorrências de todos os termos 74900

(valor médio = 20%).

No caso dos documentos sobre política científica e tecnológica:

limite mínimo:

Total de ocorrências dos termos pontuais - 2000 = 22%;
Total de ocorrências de todos os termos 9400

limite máximo:

Total de ocorrências de termos pontuais
+ total de ocorrências de termos de facetas = 7500 = 73%
Total de ocorrências de todos os termos 9400

(valor médio ~= 50%)

Isso confirma a conclusão anterior de que o melhor grau de informação se obteria usando uma combinação de descritores de facetas e de descritores pontuais.

Por outro lado, o fato de ser a redundância do vocabulário utilizado para indexar os documentos sobre política científica e tecnológica menor do que a redundância do vocabulário agrícola pode explicar-se perfeitamente considerando que, ao ser no primeiro caso o escopo muito mais abrangente (interdisciplinaridade), a quantidade de descritores pontuais (baixa freqüência de uso) é relativamente muito mais elevada e, conseqüentemente, maior o poder informativo (especificidade) da maior parte dos termos.

AGRADECIMENTOS

Agradecemos a Luis Antônio Gonçalves da Silva, Chefe do Centro de Informação sobre Política Científica e Tecnológica do CNPq por seu constante apoio no desenvolvimento deste trabalho, e a Haruka Nakayama pela sua dedicação e eficiência liderando o grupo de estagiários e bolsistas do Centro. Agradecemos também a Adelaide Ribeiro Jordão e Selma Salim Silveira, bolsistas do CNPq junto ao projeto "Estudo das possibilidades de otimização da recuperação da informação a partir da racionalização do processo de indexação dos documentos", desenvolvido no Departamento de Biblioteconomia da Universidade de Brasília, e a Tânia Camargo Barcellos, Janete Miranda Torres e Maria do Carmo Ponte Soares, pela colaboração no registro e indexação dos documentos do CNPq e na transcrição dos dados.

REFERÊNCIAS BIBLIOGRÁFICAS

1 LIMA, E. & MAIA, S. Comportamento bibliométrico da língua portuguesa, como veículo de representação da informação. Ci. Inf., 2 (2): 99-138, 1974.

2 ZIPF, G.K.. The psycho-biology of language; an introduction to dynamic phylology. Cambridge, Mass., MIT Press, 1965. 336 p.

3 ———— The psycho-biology of language. Boston, Houghton Mifflin, 1935.

4 GOFFMAN, W. A general theory of communication. In: SARACEVIC, T. Introduction to information Science. New York, Bowker, 1970. p. 726-47.

5 BOOTH, A. D. A "law" of occurrences for words of low frequency. Information and Contrai, 10 (4): 386-93, 1967.

6 ROBREDO, J. et alii. Elaboracíón de un thesaurus agrícola basado en critérios de eficiência dei lenguaje en ei proceso de comunicación. Brasília, SNIDA, 1975. 23p. Comunicação apresentada no 5. World Congress l. A. A. L. D., México, 14 a 18 de abril, 1975.

7 ———— Construção de um núcleo de thesaurus em agricultura baseado no uso real dos descritores. Brasília, SNIDA, 1975. 15p. Preprint. Comunicação apresentada na 1. Reunião Brasileira de Ciência da Informação. Rio de Janeiro, 15 a 20 de junho de 1975.

8 BROOKES, B.C. Bradford's law and the bibliography of Science. Nature, 224: 453-956, 1969.

9 ROBREDO J. Estudo das leis que governam as associações entre descritores, com vistas à utilização de vocabulários simplificados e à otimização dos processos de indexação e recuperação. Brasília, BIB/FUB, 1980. 6p., anexos (BIB/PROJ/PESQ/80/02). Exemplar datilografado (difusão limitada).

10 ——Estudo das possibilidades de otimização da recuperação da informação a partir da racionalização do processo de indexação dos documentos. Brasília, BIB/FUB, 1980. 5p., anexos (BIB/PROJ/PESQ/80/02 (02)). Exemplar datilografado (difusão limitada). Projeto desenvolvido com auxílio do Conselho Nacional *de* Desenvolvimento Científico e Tecnológico.

11 CONSELHO NACIONAL DE DESENVOLVIMENTO CIENTIFICO E TECNOLÓGICO. Projeto de Implementação do Programa de Informação sobre Política Científica e Tecnológica. Brasília, CNPq, 1980. (difusão limitada).

12 ROBREDO, J.;SILVA, L.A.G. da. Estabelecimento de instrumentos de controle terminológico em áreas específicas do conhecimento. Brasília, BIB/FUB, 1981. 24p. Preprint. Comunicação apresentada na 33. Reunião Anual da Sociedade Brasileira para o Progresso da Ciência. Salvador, Ba. 8-15 de julho de 1981. Publicada também em: CIÊNCIA E CULTURA (Suplemento), 3 (7): 188, 1981 (Resumos da 33. Reunião Anual da SBPC).

13 ROBREDO, J.; FERREIRA, J.A. de P. Conceituação de um programa para indexação automática de textos. R. Bibliotecon, Brasília, 8 (2): 254-63, 1980.

14 GUIA BRASILEIRO DE PESQUISA AGRÍCOLA EMANDAMENTO. Brasília, BINAGRI, 1978/79. v.1 — Cadastro de instituições e pesquisadores, 262 p; v.2 — índice Geral, Partes l-111, fndice de assuntos, 2712 p.; Parte IV, Índice de instituições, pesquisadores, projetos novos e terminados, p. 2713-3352; Parte V, Listagem de referência dos projetos, p. 3353-3798.

15 ROBREDO, J. A indexação automática como mecanismo básico no processo de transferência de informação. Brasília, BIB/FUB, 1980. 20p. Preprint. Comunicação apresentada no 1º Congresso Latino-Americano de Biblioteconomia e Documentação, Salvador, Ba, 21 - 26 de setembro de 1980.

16———A indexação automática de textos: o presente já entrou no futuro. In: MACHADO, U. D., ed. Estudos avançados em Biblioteconomia e Ciência da Informação. Brasília, Associação dos Bibliotecários do Distrito Federal, 1982. v.1, p. 236-74.

17 MOLES, A. Teoria da informação e percepção estética. 2. ed. Rio de Janeiro, Tempo Brasileiro, Brasíia, Editora Universidade de Brasília, 1978. 308 p. (Biblioteca Tempo Universitário, 14). Tradução de Théorie de 1'information et perception esthétique. Paris, Flammarion, 1968. Obra fundamental; a composição tipográfica descuidada altera a inteligibilidade da parte matemática.

18 SHANNON, C. E. &WEAWER, W. The mathematical theory of communication. Urbana, 111., University of Illinois Press, 1949. 117 p.

19 GUILBAUD, T. La cybernétique. Paris, Presses Universitaires de France, 1954. 136 p. (Que sais-je?, 638).

ABSTRACT

In any área of knowledge, the meaning-carrier terms may be used as descriptors to represent the content of the written documents in the indexing and Information organizing processes, as well as to formulate the queries in the information retrieval process. When ranged in function of their frequence of use, the descriptors show a distribution which follows the Bradford-Zipfs law. The data processing facilities may be of great support to prepare, in specif ic áreas of knowledge, Instruments for terminology contrai which enable ones the optimization of the document indexing and retrieval processes by using, to represent specific concepts, the terms and their associations which stand out for their richness in meaning. Two groups of significant terms were studied, corresponding to two different áreas of knowledge (agriculture and scientific and technologic policy), resulting from the contents analysis of sufficiently large sets of documents, indexed according to non-subjective *principies:* automatic indexing in the first case and simulated automatic indexing in the second. In each case, with basis on their respective occurrences and on the application of synonymy and quasi-synonymy concepts, lists of descriptors were established, which must contribute to optimize the indexing *and information retrieval* processes, both in manual and automatic systems. The established findings concerning indexing optimization seem confirmed by the information theory concepts.

ABACATE	022
ABACATEIRO	006
ABACAXI	124
ABACAXICULTURA	001
ABACAXIZEIRO	006
ABACAXIZEIRO-CAYENNE	003
ABASTECIMENTO	018
ABASTECIMENTO-DE-AGUA	001
ABATE	020
ABELHA	013
ABELHA-AFRICANA	001
ABELHA-URUCU	001
ABELMOCHUS-ESCULENTUS	002
ABELMOCHUS - ESCULENTUS- MOENCH	001
ABÓBORA	012
ABOBOREIRA-CURBITA-MOSCHATA	001
ABOBRINHA	002
LYCUPERSIUM-ESCULENTUM-MILL	001
MABEA	001
MAÇA	029
MACA-DAMIA-INTEGRAFOLIA	001
MACA-DO-ALGODOEIRO	001
MACA-HELIOTHIS-SPP	001
MACADEMIA	006
MACADEMIA-INTEGRIFOLIA	001
MACAPA-OIAPOQUE	001
MACARRÃO	001
MACASSAR	002
MACHADO	003
MACHO	013
MACHO-CASTRADO	001
XYLEBORUS	001
YORKSHIRE	002
ZANGAO	001
ZEBU	023
ZEBU-BRANCO	001
ZEBUINO	011
ZERANOL	001
ZINCO	031
ZIRCONIO	001
ZONA	104
ZONA-ALGODOEIRA	001
ZONA-CLIMATICA	001
ZONA-DA-MATA	001
ZONEAMENTO	039
ZONEAMENTO-AGRICOLA	001
ZONEAMENTO-AGROCLIMATICO	002
ZONEAMENTO-ECOLOGICO	002
ZOOGEOGRAFIA	005
ZOOLOGIA	055
ZOOPLANCTON	001
ZOOSANIDADE	001
ZOOTECNIA	155
ZULIA-ENTRERIANA	006
ZULIA-ENTRERIANA-BERG	001

Figura 1

Fragmentos da listagem de descritores identificados no processo de indexação automática dos títulos dos projetos de pesquisa agrícola, ordenados alfabeticamente, com as respectivas freqüências de aparecimento.

ABASTECIMENTO	002
ABNT	001
ACADEMIA-BRASILEIRA-DE-CIENCIAS	001
AÇAO-GOVERNAMENTAL	001
AÇAO-PROGRAMADA	005
ACERVO	001
AÇO	001
ACOMPANHAMENTO	056
ACOMPANHAMENTO-FINANCEtRO	001
LICOPODIACEA-DO-DEVONIANO	001
LIDERANÇA-PARTIDARIA	001
LINGÜÍSTICA	002
LINHA-DE-PESQUISA	033
LINHITO	002
LISTA-DE-PARTICIPANTES	002
LISTAGEM	003
LISTAGEM-DE-PROJETO	020
LITERATURA-CIENTIFICA	002
UNIDO	004
UNITAR	001
UNIVERSIDADE	033
URBANISMO	013
URBANIZAÇÃO	007
URUCUIA	001
URUGUAI	001
USINA-DE-ETANOL	001
ZONA-FRANCA	003
ZONA-SEMI-ARIDA	001
ZOOLOGIA	010
ZOONOSE	002
ZOOTECNIA	004

Figura 2

Fragmentos da listagem de descritores identificados no processo de indexação de documentos referentes a política científica e tecnológica, ordenados alfabeticamente, com as respectivas freqüências de aparecimento.

PRODUÇÃO	1009
SOLO	989
MELHORAMENTO	929
CULTURA	784
BOVINO	706
CULTIVAR	661
CONTROLE	649
AVALIAÇÃO	567
FRUTO	502
SOJA	461
SEMENTE	449
MILHO	446
ECONOMIA	445
ADUBAÇAO	444
PLANTA	443
MICROCLIMA	007
MICROBIOLOGIA–DO–SOLO	007
LATICÍNIO	007
INCUBAÇÃO	006
PROSPECÇÃO	006
PROTECAO–CONTRA–VENTO–FRIO	006
MELHORAMENTO–DE–PROTEINA	006
MECANIZACAO–DA–CAFEICULTURA	006
HYPOTHENEMUS-HANIPEI	006
INDICATRIS	006
LATOSSOLO–ROXO	006
MOLUSCO	006
PRODUTOS-NATURAIS	006
LAMBARI	001
HYMENEA-SP	001
MASSONIANA - KHASYA - CARIBEAE	001
LESAO–HEPATICA–AGUDA	001
ITALICA-PLENCK	001
MAGNIFICA-WALKER	001
GALACTIA STRIATA	001
MATA–DE–DOIS–IRMAOS	001
MATA-DE-PERNAMBUCO	001
MATA-SERTAO-CENTRAL	001
FUSARIOSE DO ABACAXI	001
LAEVIGAIA	001
HYMENEOPTERA	001
LIMAO–SILICIANO	001
GAFANHOTO	001
MATERIA–GORDA	001
MACA–HELIOTHIS-SPP	001
HUMUS–MU–80	001
MATERIA–SECA	001
JABUTICABEIRO	001
MATERIAL-GENETICO	001
FUSARIUM SOLANI	001
HEMILINS	001
HEMOGLUTINA	001

Figura 3

Fragmentos da listagem de descritores identificados no processo de indexação automática dos títulos dos projetos de pesquisa agrícola, arranjados em ordem sequencial decrescente de ocorrências.

CNPQ	394
CIENCIA–E–TECNOLOGIA	280
PESQUISA	269
RECURSOS–HUMANOS	191
PROGRAMA	159
DESENVOLVIMENTO-CIENTIFICO	158
II–PBDCT	152
DESENVOLVIMENTO-TECNOLOGICO	145
INSTITUICAO–DE–PESQUISA	135
PROJETO	131
POS-GRADUACAO	124
TECNOLOGIA	106
SOCIOLOGIA	011
TECNOLOGIA-MINERAL	011
ADMINISTRAÇÃO	010
ÁLCOOL	010
BOLSA–DE–AUXILIO	010
PROJETO-FLORA-SUDESTE	002
PROJETO–FLORA–SUL	002
PROJETO-RADAM	002
PROJETO-SERTANEJO	002
PRONAPESA	002
PROTEC-RH	002
PSICOLOGIA	002
QUIMICA–ANALITICA	002
QUIMICA-INORGANICA	002
QUIMICA-ORGANICA	002
RADIOLOGIA	002
XILOTECA	001
XINGU	001
ZINCO	001
ZONA–SEMI–ARIDA	001

Figura 4

Fragmentos da listagem de descritores identificados no processo de indexação de documentos referentes a política científica e tcnológica, arranjados em ordem sequencial decrescente de ocorrências.

FREQUÊNCIA DE APARECIMENTO (EM ORDEM DECRESCENTE)	NÚMERO DE DESCRITORES QUE APARECEM COM UMA DETERMINADA FREQUÊNCIA	TOTAL ACUMULADO DE DESCRITORES	TOTAL ACUMULADO DE OCORRÊNCIAS
1009	0001	0001	01009
0989	0001	0002	01989
0929	0001	0003	02927
0784	0001	0004	03711
0737	0001	0005	05148
0706	0001	0006	05854
0661	0001	0007	06515
0649	0001	0008	07164
0072	0007	0214	41034
0071	0002	0217	41176
0070	0001	0218	41246
0069	0003	0221	41453
0068	0005	0226	41793
0067	0004	0230	42061
0066	0005	0235	42391
0065	0002	0237	42527
0006	0159	1401	67207
0005	0234	1635	68377
0004	0295	1930	69557
0003	0409	2339	70784
0002	0730	3069	72244
0001	2706	5774	74950

Figura 5

Fragmentos da tabela que relaciona o número de descritores identificados no processo de indexação automática de títulos de projetos de pesquisa agrícola, que ocorrem com uma determinada frequência (em ordem decrescente), com o total acumulado de ocorrências

FREQUÊNCIA DE APARECIMENTO (EM ORDEM DECRESCENTE)	NÚMERO DE DESCRITORES QUE APARECEM COM UMA DETERMINADA FREQÜÊNCIA	TOTAL ACUMULADO DE DESCRITORES	TOTAL ACUMULADO DE OCORRÊNCIAS
394	001	001	394
280	001	002	674
269	001	003	943
191	001	004	1134
159	001	005	1293
158	001	006	1451
152	001	007	1603
145	001	008	1748
014	005	117	5460
013	008	125	5564
012	002	128	5710
011	014	142	5938
010	012	164	6158
009	017	181	6314
008	027	208	6540
007	041	249	6827
006	041	290	7073
005	055	345	7348
004	070	415	7628
003	115	530	7773
002	261	791	8295
001	1102	1893	9391

Figura 6

Fragmentos da tabela que relaciona o número de descritores identificados no processo de indexação de documentos referentes a política científica e tecnológica, que ocorrem com uma determinada frequência (em ordem decrescente), com o total acumulado de ocorrências.

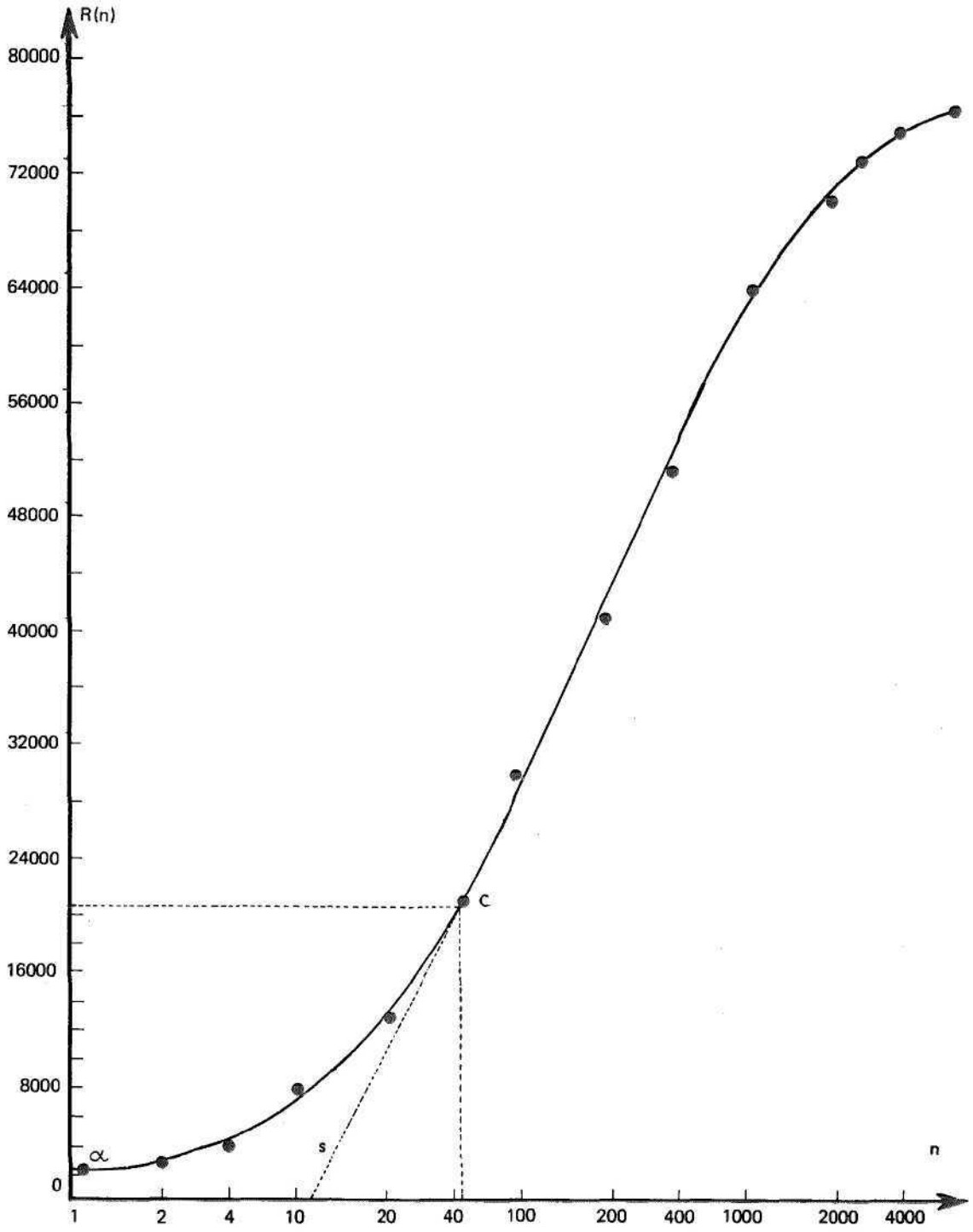

Figura 7

Representação da variação do número de ocorrências dos descritores em função do número de descritores identificados no processo de indexação automática dos títulos dos projetos de pesquisa agrícola.

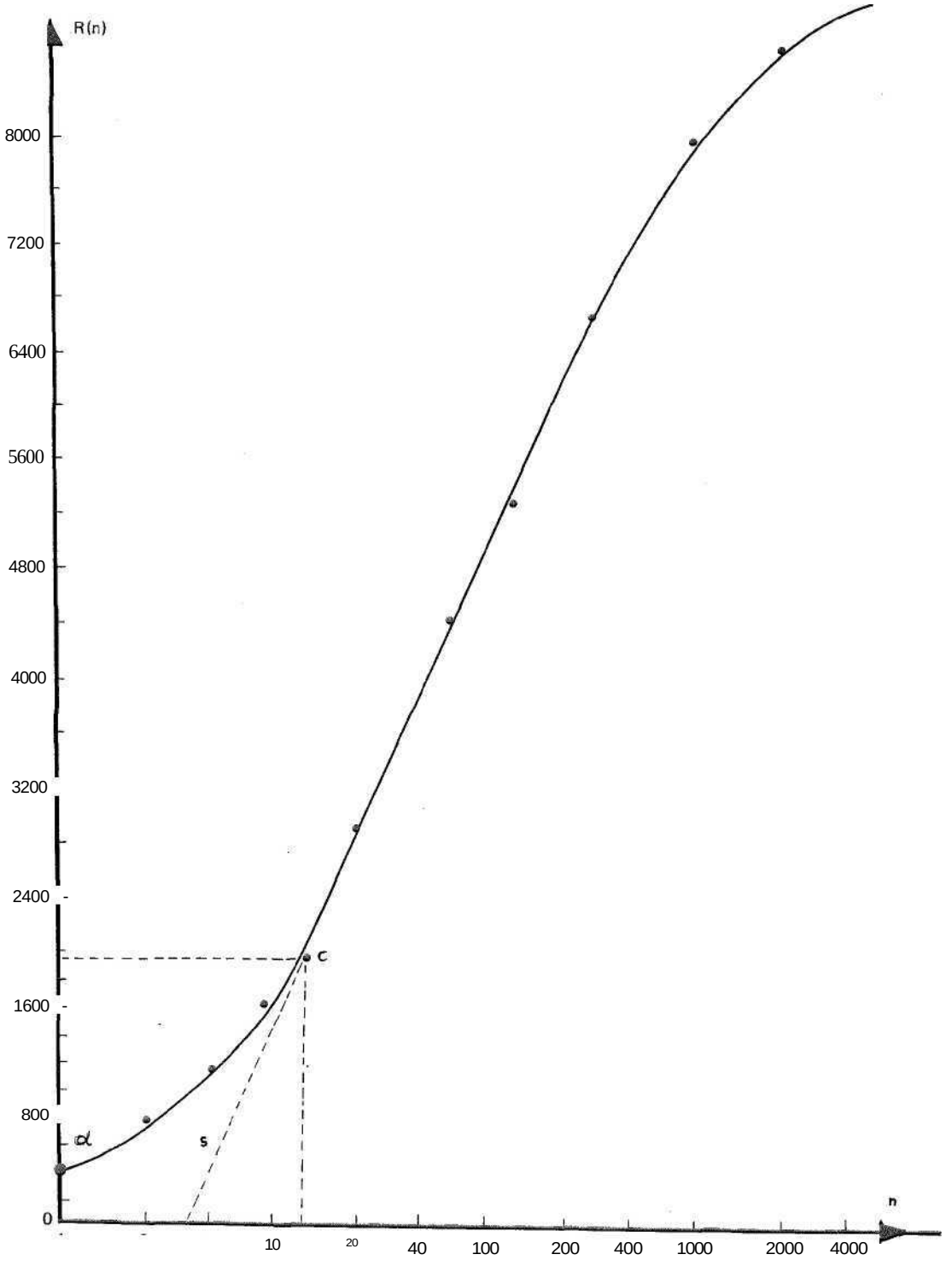

Figura 8

Representação *da variação do número de ocorrências* dos descritores em função do número de descritores identificados no processo de indexação de documentos referentes a política científica e tecnológica.

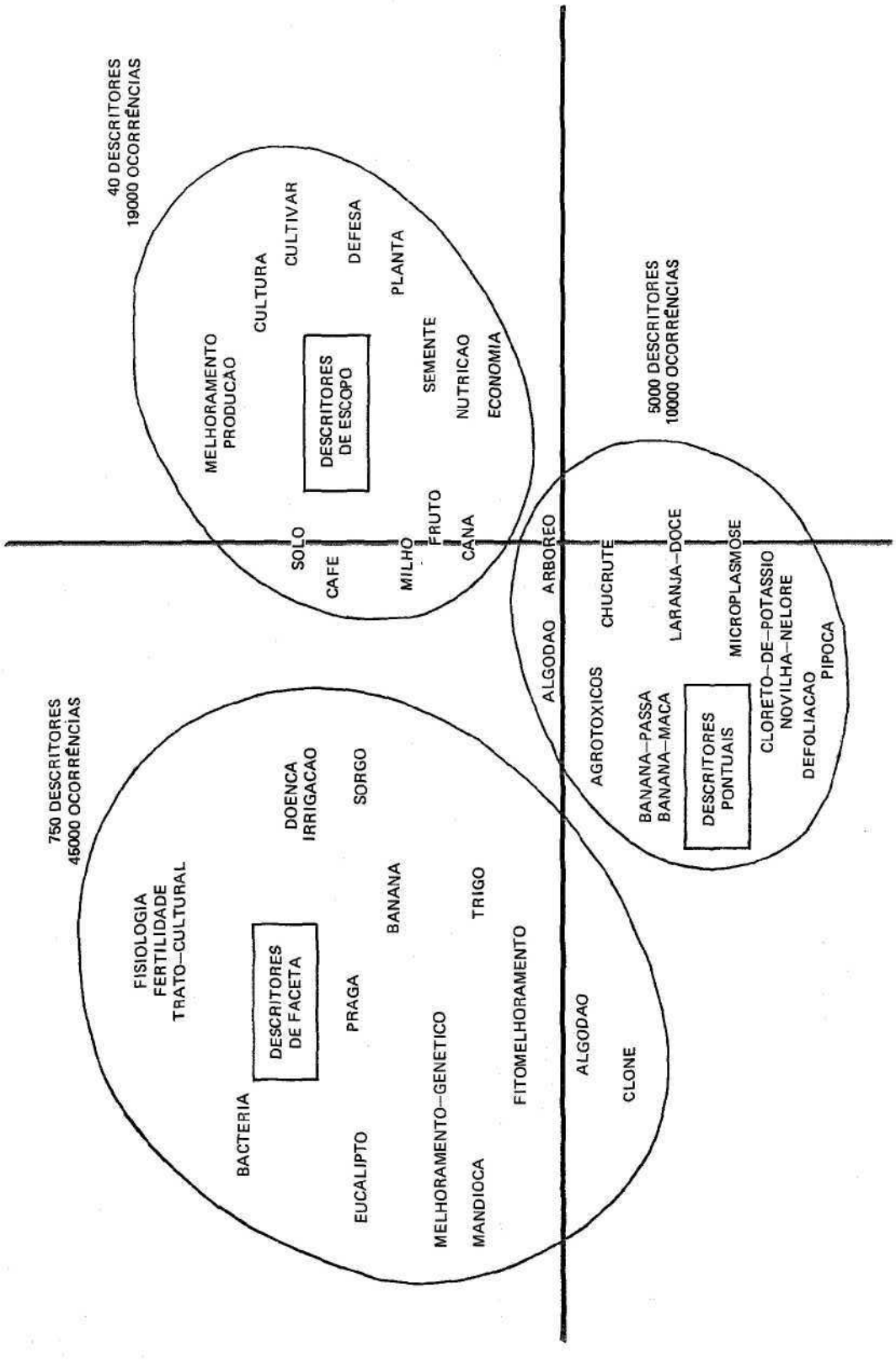

Figura 9

Representação das três zonas em que se distribuem os descritores identificados no processo de indexação *automática dos títulos dos projetos de pesquisa* agrícola.

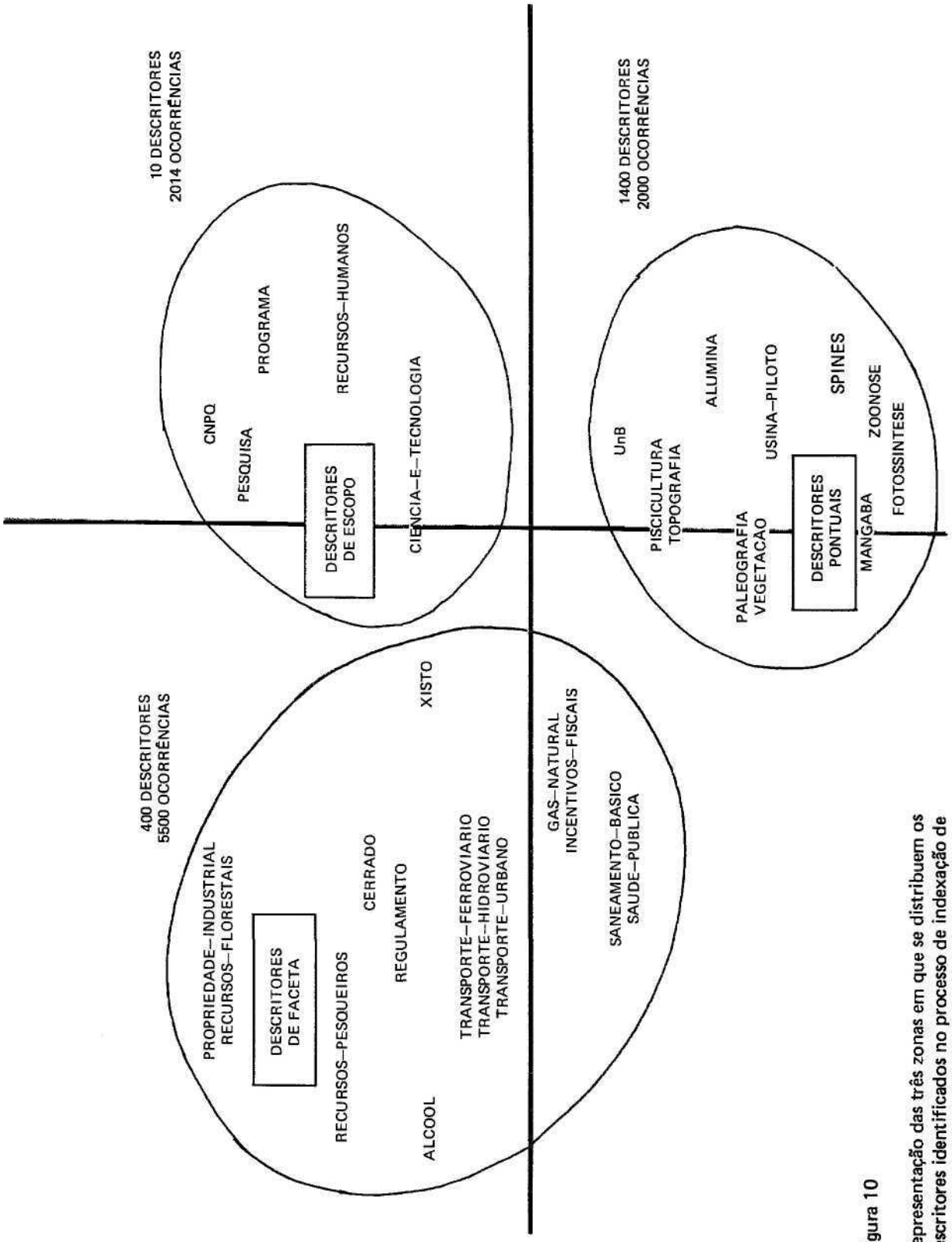

Figura 10

Representação das três zonas em que se distribuem os descritores identificados no processo de indexação de documentos referentes a política científica e tecnológica.