INDEXAÇÃO AUTOMÁTICA E MANUAL: REVISÃO DE LITERATURA*

Simone Bastos Vieira
Subsecretária de Biblioteca
Senado Federal
70 160 Brasília, DF

RESUMO

Abordam-se as diversas pesquisas nacionais e estrangeiras que avaliam a qualidade da indexação manual e automática, em relação às técnicas e fontes empregadas para a extração dos termos significativos e a capacidade de recuperação da linguagem de indexação, nas bases de dados.

1 - INTRODUÇÃO

A indexação é uma técnica de análise de conteúdo que condensa a informação significativa de um documento, através da atribuição de termos, criando uma linguagem intermediária entre o usuário e o documento. E um dos processos básicos de recuperação da informação. Pode ser realizada pelo homem (indexação manual), ou por programas de computador (indexação automática).

Descrevem-se, nesta revisão, as várias pesquisas, estrangeiras e brasileiras, e seus resultados sobre análise comparativa entre indexação automática e manual, análise comparativa do uso eficiente do título, resumo, texto integral, citações e outras fontes para indexação e métodos estatísticos de avaliação de recuperação da informação, através dos vocabulários obtidos por indexação automática e manual.

Não se abordaram estudos teóricos, matemáticos, lingüísticos e históricos da indexação.

A revisão de literatura estrangeira, ao contrário da brasileira, não pretendeu ser exaustiva, devido ao grande volume de documentos. Abrangeu-se o período de 1970 a maio de 1 984.

2 - FUNDAMENTOS GERAIS DE INDEXAÇÃO

O Sistema Mundial de Informação Científica (UNISIST), em um de seus grupos de estudo. elaborou um documento com os princípios de indexação[1]. Esses princípios estavam voltados, especificamente, para a indexação manual, e pode-se dizer que são os mesmos adotados por vários autores brasileiros[2-3-4] e estrangeiros[5, 6, 7, 8, 9, 10]. Esse documento foi a primeira tentativa internacional de se normalizar o processo de indexação.

De acordo com o UNISIST[1], a indexação é a operação que descreve e identifica o conteúdo de um documento, através de termos. Os conceitos dos documentos podem ser representados por termos selecionados através da linguagem natural ou por símbolos.

A indexação está diretamente relacionada com a descrição física do documento, e ambos constituem um registro bibliográfico, proporcionando ao usuário informações físicas e de conteúdo do documento. Os dados são organizados da forma mais acessível para a recuperação da informação.

A indexação pode ser realizada em documentos, no seu todo ou em suas partes, e na estratégia de busca para recuperação em um sistema de informação.

Durante a indexação manual os conceitos são extraídos por um processo de análise intelectual, que compreende basicamente três fases:

1. compreensão do conteúdo do documento, através da leitura completa do texto ou do título, do resumo e de outras partes que compõem um documento. O UNISIST[1] recomenda o uso não apenas do título e/ou do resumo para indexar, pois nem sempre os mesmos contêm os termos que identificam, suficientemente, o conteúdo;

* Revisão de literatura extraída da dissertação *Análise comparativa entre indexação automática e manual da literatura brasileira de Ciência da Informação* aprovada pela Universidade de Brasília para obtenção do grau de Mestre em Biblioteconomia e Documentação, em dezembro de 1984.

2. identificação dos conceitos, estabelecendo o ambiente lógico e as diferenças entre os fenômenos, os processos, as propriedades, as operações, os equipamentos, etc.
3. seleção dos conceitos, observando alguns fatores como: exaustividade, especificidade e consistência.

O processo de indexação da estratégia de busca requer do profissional de informação a tradução da pergunta do usuário para a linguagem de indexação adotada. Normalmente se utilizam como instrumentos auxiliares tesauros, listas de cabeçalhos de assunto, listas de termos controlados, etc.

Nesse processo é essencial a interação usuário-indexador-sistema.

Alguns sistemas utilizam, ainda, pesos e elos na recuperação e indexação.

Essas fases podem ser melhor visulizadas na Figura 1.

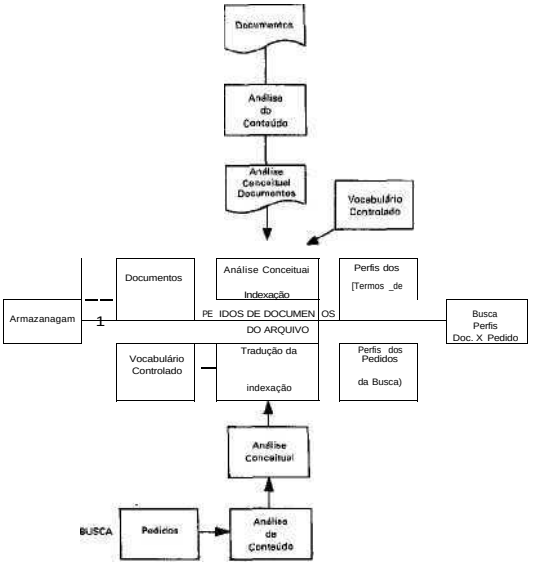

Figura 1 - Processo de indexação manual, segundo Lancaster[2] apud Cesarino & Pinto[4] (p. 33)

Pode-se dizer que a consistência é o ponto de estrangulamento na indexação manual. Segundo Pinheiro[11], ela reflete similaridades e diferenças de termos atribuídos pela subjetividade dos indexadores na fase de análise de conteúdo. Também está diretamente relacionada, quanto aos indexadores, no que se refere ao seu desempenho, à regularidade e imparcialidade de seleção dos conceitos, à experiência prévia, à técnica de indexação, ao conhecimento do assunto dos documentos, à submissão às diretrizes da indexação adotadas pelo sistema. Quanto aos instrumentos de indexação utilizados pelos sistemas de informação, estão relacionados com a sua qualidade[4, 8, 12].

O início da indexação automática deu-se no final da década de 50, quando Luhn[13, 14] desenvolveu a idéia de que o vocabulário existente em um documento deveria se constituir na base para a análise do seu conteúdo, sendo esta a melhor maneira de recuperá-lo.

A primeira aplicação realizada por Luhn[13,14] foi para a produção do índice *Key Word in Context* (KWIC), elaborado a partir da rotação automática das palavras significativas dos títulos.

O processo de indexação automática baseia-se, segundo Robredo[15], "na comparação de cada palavra do texto com uma relação de palavras vazias de significado, previamente estabelecidas, que conduz, por eliminação, a considerar as palavras restantes do texto como palavras significativas".

De acordo com o mesmo autor, esse processo pode identificar termos, pares de termos ou até frases significativas que expressem o conteúdo do documento, e pode-se dizer que é semelhante ao processo de leitura-memorização.

O processo de indexação automática encontra-se melhor visualizado e detalhado na Figura 2.

A indexação automática é um processo que pode utilizar diferentes métodos (ver item 3.3) desenvolvidos para programas de computador. Essa operação, ainda segundo Robredo[15], é objetiva, pois utiliza sempre os mesmos programas para extração de termos significativos dos documentos. O aspecto de inconsistência devido à subjetividade da indexação manual é eliminado, possibilitando uma melhor recuperação.

Os aspectos de especificidade e exaustividade estão diretamente relacionados com as características das necessidades de informação delineadas pelos usuários dos sistemas de informação e pela decisão administrativa dos referidos sistemas.

O processo de indexação automática da estratégia de busca é realizado pelos mesmos programas, assegurando assim a compatibilidade entre a linguagem utilizada na indexação e a utilizada na formulação da pergunta[16].

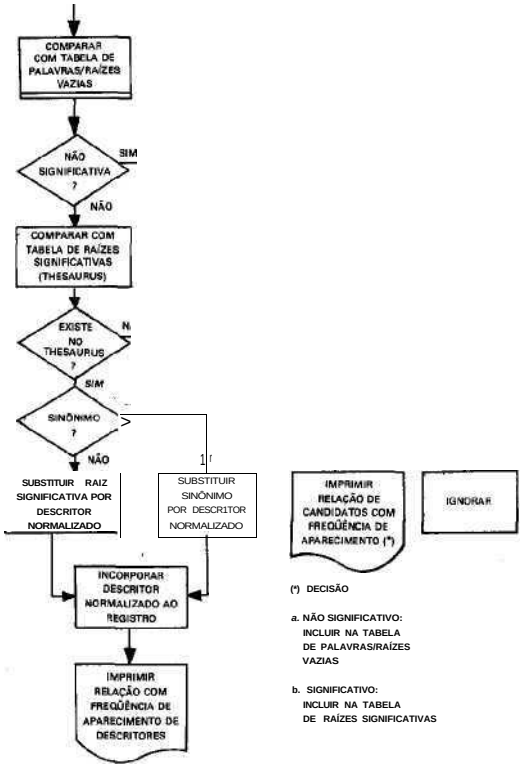

Figura 2 — Processo de indexação automática, segundo Robredo[15] (p. 247).

Foskett[6] define a exaustividade como sendo a extensão com que se analisa um documento, a fim de se estabelecer exatamente todos os assuntos que esse documento referencia, e a especificidade como a extensão em que um sistema de informação permite ser preciso ao se especificar o assunto de um documento.

Ambos os fatores, exaustividade e especificidade, estão relacionados, respectivamente, à renovação e relevância, que são as medidas de qualidade da recuperação da informação.

A relevância ou precisão é definida por Saracevic[17] como a medida de contato efetivo entre a fonte e o destinatário. Pode ser quantificada, tal como a revocação, através das seguintes fórmulas matemáticas, mencionandas por Robredo[16]:

Relevância = . ., sendo que:

a = número de documento relevantes selecionados;

b = número total de documentos selecionados.

Revocação =. -, sendo que:
a+ c

a = número de documentos pertinentes selecionados;
c = número de documentos não pertinentes selecionados.

Vickery[18] menciona as seguintes funções das linguagens de indexação:
a) recuperar documentos com conteúdo semelhante;
b) recuperar documentos relevantes sobre um assunto específico;
c) recuperar documentos por grandes áreas de assunto;
d) possibilitar a conversão dos termos de indexação entre diferentes linguagens; e
e) auxiliar na escolha do termo adequado para a estratégia de busca.

Cesariano & Pinto[4], posteriormente, abordaram ainda os seguintes aspectos:
a) representar o assunto de uma maneira consistente; e
b) permitir a compatibilidade entre a linguagem do indexador e a do pesquisador.

Segundo Robredo[16] a indexação pode ser realizada basicamente em três níveis, partindo do mais geral para o específico:
a) categorização - representa o assunto que predomina;
b) superficial - representa os conceitos principais de forma geral;
c) profunda - representa todos os conceitos fundamentais.

Os termos de indexação podem ser expressos através de linguagem:
a) natural ou livre, utilizando os mesmos termos do autor;
b) controlada, adotando termos aceitos e definidos previamente;
c) codificada, utilizando códigos previamente estabelecidos para expressar os conceitos significativos; e
d) coordenada as relações lógicas entre os termos, quando existentes, estabelecem-se através de:
- equivalência ou sinonímia entre os termos;
- subordinação ou hierarquia- parte do genérico para o específico e vice-versa; e
- coordenação ou associação - os conceitos estão relacionados à idéia de outro conceito.

Os termos podem, de acordo com VicKery[18], estar inter-relacionados por subgrupos de assunto, elos, pesos ou expressos em pequenas frases da linguagem natural.

3 - REVISÃO DA LITERATURA ESTRANGEIRA

3.1 -*ESTUDOS COMPARATIVOS ENTRE INDEXAÇÃO AUTOMÁTICA E MANUAL*

A comparação entre os dois tipos de indexação, automática e manual, é realizada para se verificar as diferenças e semelhanças entre os termos selecionados por programas de um computador e pelo homem. De acordo com os resultados obtidos, avalia-se a aplicabilidade de uma uma ou outra técnica.

Os testes de comparação podem ser divididos em testes de qualidade de indexação e de qualidade de recuperação.

A grande maioria de testes comparativos entre descritores atribuídos manual e automaticamente, segundo Salton[19, 20], chega a um resultado aproximado de 60% de compatibilidade entre uma linguagem e outra.

Salton[21], em um dos seus artigos, descreve uma fórmula matemática pela qual se obtém o coeficiente de avaliação entre dois vocabulários:

$$\frac{c}{a + m - c}$$

Onde:
q = valor comparativo entre dois vocabulários;
c = número de termos comuns;
a = número de termos atribuídos automaticamente;
m = número de termos atribuídos manualmente.

Carrol & Roeloffs[22] realizaram estudos comparativos entre indexação manual e automática na área de Ciência da Informação, aplicando a análise de correlação estatística. Verificaram que os termos obtidos pelos indexadores foram semelhantes aos da indexação automática, mas, levando-se em conta os custos de contratação, treinamento de mão-de-obra especializada e a inconsistência humana, a indexação automática é mais viável.

Vários testes comparativos de revocação e precisão foram realizados para verificar o desempenho da recuperação, através de termos atribuídos manual e automaticamente.

Salton[19] e Boyce & Lockard[23] realizaram suas experiências na área médica. O primeiro comparou os mesmos documentos indexados manualmente, utilizando vocabulário controlado, e indexados automaticamente, pelo programa *Automatic Document Analysis and retríeval System (SMART),* utilizando termos livres do resumo. Salton[19] verificou que numa indexação automática somente com truncagem de palavras, a indexação manual torna-se mais efetiva cerca de 15% a 20%. Quando se utiliza um controle através de tesauros e dicionários, a eficiência da indexação automática é semelhante à da manual. Aplicando-se no momento da recuperação a técnica de realimentação de relevância na pergunta, *a* indexação manual é menos eficiente.

Boyce & Lockard[23] aplicaram dois tipos de indexação manual e um tipo de indexação automática, em textos integrais. Seus resultados demonstram que a indexação automática foi mais consistente na revocacão com perguntas gerais e específicas. A indexação manual obteve melhores resultados de precisão com perguntas utilizando termos específicos, e *a* automática com termos gerais. A indexação automática é tão eficiente quanto a manual, concluíram ao final.

Van der Meulen & Janssen[24] avaliaram comparativamente a indexação automática do programa *Information ríetríeva/ System of Philips Research Laboratories (DIRECT),* que utiliza títulos e resumos, e a indexação manual desenvolvida pelo *Information Service for Physics, Electrotechnology and Control* (INSPEC). Criaram duas bases de dados, com os mesmos documentos, indexadas pelas duas técnicas, e realizaram duas perguntas-teste. Verificaram, ao contrário dos resultados mencionados anteriormente, que a indexação manual apresenta melhores índices de revocação e precisão, cerca de 20% em relação à automática. Os autores justificaram esse resultado devido ao pequeno número de perguntas-teste efetuadas.
Klíngbiel& Rinker[25] compararam a eficiência da indexação manual e da automática realizada em títulos e resumos pelo programa *Machine-Aided Indexing (MAI).* A indexação manual e a automática obtiveram os mesmos índices de revocação, mas a manual mostrou-se inferior na precisão.

Barnes, Costantini & Perschke[26] compararam a indexação manual e a automática em títulos e resumos do sistema SLC-II. O teste foi realizado em 5 000 documentos *do INIS Atomindex.* Na recuperação a indexação automática mostrou-se mais eficiente do que a manual, e os termos existentes em uma estavam compatíveis com os termos da outra. A revocação da indexação automática apresentou um Índice de 90%.

A maioria dos testes revela que a indexação automática produz resultados de recuperação, no mínimo, equivalentes aos obtidos pela manual.

Nos casos em que são aplicadas técnicas mais sofisticadas na recuperação, a indexação automática mostra-se, segundo Salton[19, 20], ainda mais eficiente.

Lancaster[27] aponta alguns dos problemas relacionados com indexação manual:

- falta de especificidade ou coordenação falsa entre termos no vocabulário;
- perguntas muito exaustivas ou muito específicas na formulação da pesquisa;
- insuficiência de exaustividade, ou exaustividade em excesso, ou ainda omissão de termos importantes na indexação dos documentos; e
- falta de interação do usuário com o sistema.

Wessel[28] acrescenta um outro problema. Na teoria, o indexador deveria produzir uma indexação superior à realizada por programas de computador, mas na prática a indexação manual apresenta muitas inconsistências, para produzir efetivos instrumentos de recuperação da informação.

3.2 -*ESTUDOS COMPARATIVOS ENTRE O TÍTULO E OUTRAS FONTES PARA INDEXAÇÃO AUTOMÁTICA*

Desde a criação do índice KWIC por Luhn[13, 14], a indexação automática em títulos tem sido questionada, tanto pela qualidade da indexação como pela qualidade da recuperação da informação.

Vários estudos foram realizados para testar a validade do título como fonte, para extração automática de palavras significativas, como, por exemplo, as pesquisas sobre o crescimento de palavras significativas em títulos, análises comparativas qualitativas entre a indexação e recuperação por títulos, resumos e texto integral, e a utilização de indexação por título para índices de serviços de alerta.

Um dos primeiros estudos foi realizado por Maizell[29], em 1960, testando o conteúdo dos títulos dos artigos *do Physics Abstracts.* Ele concluiu que 63% dos títulos continham informações suficientes para indexação.

Estudos semelhantes foram realizados por vários autores, comparando palavras de títulos, extraídas automaticamente, e descritores obtidos através de indexação manual, existentes nos índices de assunto de publicações secundárias.

Montgomery & Swanson[30], em análise realizada, encontraram um índice de 86% dos títulos incorporados no *Index Medicus* com conteúdo suficientemente significativo para serem utilizados em indexação automática.

Ruhl[31], em pesquisa na área de Química, encontrou 57% de títulos contendo todos os descritores existentes no índice de assunto do *Chemical Abstracts* e somente 12% dos títulos não continham três ou mais palavras significativas.

Kraft[32], em títulos de Direito, encontrou 64% de palavras significativas, e somente 10% não continham nenhuma palavra existente no índice do *Index to Legal Periodicals.*

Os títulos tornaram-se mais significativos, segundo Tocatlian[33], nos anos posteriores ao surgimento do KWIC. O crescimento de informações significativas em títulos estaria relacionado com a preocupação dos autores em torná-los mais relevantes, para serem utilizados em índices tipo KWIC. Bird & Knight[34] sugerem ainda outra justificativa: a necessidade de os autores tornarem seus títulos mais precisos, para poderem se sobressair em relação à explosão bibliográfica existente.

Ao examinar o crescimento de palavras em títulos, Ghosh[35] verificou um aumento significativo entre 1933 e 1972, quando 80% dos documentos poderiam ser recuperados por pesquisa somente no título.

Buxton& Meadows[36] compararam títulos, nas áreas de ciências exatas e ciências sociais, e constataram um significativo aumento de palavras substantivas entre 1947 a 1973. Para eles, o aumento de substantivos nos títulos pressupõe aumento de palavras relacionadas com o conteúdo dos documentos. Verificaram que os títulos de Química e Botânica possuíam maior valor para a recuperação do que os das áreas de Física, Medicina e História. Na área de ciências sociais é que se encontram títulos com menor valor para a recuperação.

Bloomfield[37] avaliou comparativamente a qualidade da recuperação da indexação manual, da indexação automática em títulos e do KWIC. Concluiu, entre outros resultados, que o uso do resumo para enriquecer a indexação por título gera um número elevado de descritores irrelevantes para a recuperação.

Svenonius[38] afirma que a indexação somente por título apresenta uma precisão maior do que por resumo ou texto integral. De acordo com a autora, o bom desempenho da recuperação não está relacionado com a quantidade de descritores atribuídos a um documento, mas sim à qualidade dos mesmos.

Salton[19, 39], ao contrário desses autores, não aconselha o uso somente de títulos para indexação, pois verificou que é menos eficiente para expressar o conteúdo, do que o uso, também, de resumos.

Barker, Veal & Wyatt[40], ao compararem a eficiência e o custo de busca bibliográfica em títulos, resumos e descritores na área de Química, verificaram que o resumo e os descritores, obtidos manualmente, aumentam, respectivamente, em 68% e 35% a revocacão, mas diminuem em 23% e 10% a precisão. Os títulos são relativamente mais precisos e menos exaustivos. Uma busca em linha utilizando títulos enriquecidos por resumo ou descritores aumenta cerca de 20% o tempo de uso do computador e, conseqüentemente, cresce o custo de impressão das referências recuperadas, que é inversamente proporcional ao índice de revocacão. O título é o mais indicado, pela sua maior precisão e menor custo final ao usuário.

Em 1975 Feinberg[41] escreveu um livro tratando de estudos comparativos entre o índice KWIC e o índice *Key Word out of Context* (KWOC), entre outros índices de palavras permutadas e índices elaborados por indexação automática em títulos. Afirma, como Svenonius[38], que o número de descritores atribuídos a um documento não está relacionado com a qualidade de indexação, e que um grande número de descritores pode, inclusive, prejudicar a recuperação. Segundo Feinberg[41], uma das vantagens da indexação a partir do título é a precisão.

Garfield[42] e Neufeld et alii[43] descreveram a aplicação de indexação automática em títulos para elaboração de índices de assunto em boletins de alerta. Segundo os autores, essa técnica foi a que ofereceu maior rapidez e precisão.

Kwok[44], partindo do pressuposto de que o título não possui palavras estatisticamente suficientes para indexação automática, propõe como fator de enriquecimento o uso de títulos citados nas referências. Após análises comparativas entre o uso de títulos citados com títulos e resumos e somente com títulos, verificou que o enriquecimento com títulos citados oferece uma representação de conteúdo mais compacta, uniforme e possibilita estabelecer, de forma adequada, relações associativas entre os descritores.

Garfield[45] propõe para a indexação automática o uso de títulos das citações existentes nos documentos. As citações, segundo o autor, são ilustrações ou complementações do que se deseja informar. São formas de estabelecerem-se relações entre trabalhos que possuam pontos em comum e, portanto, são ótimas fontes de indexação, melhores do que os títulos. É a técnica aplicada na elaboração do índice de assunto do *Citation Index.*

Um dos mais recentes trabalhos desenvolvidos para testar comparativamente o desempenho da recuperação da informação, através de palavras extraídas automaticamente em títulos, resumos, textos integrais e outras fontes, foi realizado por Cleveland, Cleveland & Wise[46].

Os autores desenvolveram esta pesquisa baseados no alto custo e na impossibilidade prática da indexação automática em texto integral. Verificaram, após testes de comparação entre oito combinações de indexação automática, que os índices de revocacão e precisão apresentados em documentos indexados em fontes como resumos e títulos são semelhantes aos obtidos por indexação em texto integral.

3.3 - *MÉTODOS DE INDEXAÇÃO AUTOMÁTICA*

A indexação automática é uma operação que identifica, através de programas de computador, palavras ou expressões significativas dos documentos, para descrever de forma condensada o seu conteúdo.

As palavras significativas são selecionadas automaticamente, através de metodologias específicas, adotadas de acordo com as políticas de indexação e recuperação da informação, desenvolvimento *de software* e capacidade de *hardware* dos sistemas de informação.

As políticas de indexação e recuperação variam, respectivamente, de acordo com a exaustividade e precisão da análise de conteúdo e com os índices de revocacão e precisão do resultado da pesquisa. Ambas as políticas dependem, diretamente, das necessidades de informação, caracterizadas pelos diversos tipos de usuários a que um sistema de informação atende.

Esta parte da revisão da literatura trata dos diversos métodos de indexação automática sem, no entanto, deter-se nos aspectos históricos de cada um. Esses aspectos encontram-se suficientemente analisados no artigo de revisão publicado por Batty[47].

3.3.1 — *Método de Freqüência ou Análise estatística*

O método de freqüência de palavras foi o primeiro a surgir. Foi proposto por Luhn[13, 14] em 1957 e 1958. O autor demonstrou

em seus trabalhos que a freqüência de uma palavra em documentos está diretamente relacionada com a capacidade dessa palavra de/para representar o conteúdo do documento, a nível de indexação e de recuperação da informação. As palavras mais adequadas para a indexação serão as que possuírem média freqüência.

O método de freqüência trata da contagem automática do aparecimento da palavra, que pode estar localizada, segundo Cleveland, Cleveland & Wise[46], no título, resumo, título das referências citadas, texto e em diversas combinações entre estas unidades, como, por exemplo, em título e resumo.

Garfield[45] acrescenta ainda a localização das palavras significativas através da freqüência no título das citações.

A contagem automática do termo é realizada através da ocorrência e / ou co-ocorrência da palavra. A freqüência pode ser estabelecida, de acordo com Soergel[48] e Sparck Jones[49], através da:

a. ocorrência total da palavra no documento - a palavra é contada todas as vezes que aparece, fazendo-se o somatório das vezes em que co-ocorre, posteriomente;
b. ocorrência única da palavra no documento — conta-se somente uma vez a palavra, independentemente do número de vezes que ela aparece;
c. ocorrência da palavra na coleção — a contagem é realizada somando-se seu aparecimento da coleção.

Soergel[48] diferencia, ainda, contagem de conceito. A contagem de conceito é o somatório das freqüências de ocorrência de todas as palavras que determinam aquele conceito. A freqüência de ocorrência de palavras será utilizada para desenvolver a estrutura terminológica, e a do conceito para desenvolver a estrutura classificatória.

A freqüência pode ser realizada, também, em palavras truncadas ou em raízes de palavras. Alguns sistemas utilizam esse tipo de freqüência para diminuir o ruído, evitando o aparecimento de mesmas palavras com diferentes desinências gramaticais[21, 50, 51, 52].

O método de freqüência possui outras aplicações, além da indexação automática.

Rosenberg[53] utilizou a freqüência de co-ocorrência de palavras como forma de aumentar o desempenho da indexação manual, fornecendo ao indexador uma lista de descritores candidatos extraídos automaticamente pela análise estatística combinada com a análise de associação entre palavras. Esses descritores, acompanhados de suas respectivas freqüências, serão utilizados para indexar novos documentos.

Henzler[54] aplicou a análise estatística em um estudo quantitativo comparando o vocabulário livre e o controlado, e concluiu que a linguagem controlada fornece uma maior perda de informação do autor para o usuário. O ideal seria combinar as duas linguagens de indexação.

A maior aplicação desse método, o de freqüência, é para realizar a seleção automática de descritores. A lei de distribuição de palavras em um texto, a lei de Zipf, surge como uma das técnicas que complementam a escolha do descritor. Vários autores aplicaram-na em seus experimentos.

Svenonius[38] aplicou a primeira lei de Zipf para verificar qual a freqüência que melhor se adapta à seleção automática de descritores. Os resultados encontrados demonstraram que palavras específicas, as de baixa freqüência, proporcionam maior precisão na recuperação; em contrapartida, as palavras de média freqüência proporcionam maior revocacão.

Schuegraf & Heaps[52], Pao[50] e Rowbottom & Willet[51] trabalharam com a lei de Zipf em raízes de palavras. Os dois primeiros autores propõem o uso de radicais de palavras para otimizar os custos da recuperação e, principalmente, de armazenamento da informação. Propõem, também, um algoritmo para fragmentar automaticamente palavras eqüidistantes.

Rowbottom & Willet[51] não aconselham a extração de palavras aplicando a lei de Zipf em pequenos textos, tais como títulos e resumos, pois a indexação não será suficientemente exaustiva e precisa.

3.3.2 - *Métodos de atribuição de peso*

O método de atribuição de peso aos descritores, segundo Salton[39], é uma forma de atribuir-lhes valores semânticos para torná-los mais precisos, sem no entanto diminuir sua capacidade de revocacão. É baseado na freqüência de cada descritor.

Luhn[13, 14] foi, novamente, o precursor deste método. Ele propôs um modelo relacionando diretamente a freqüência de uma palavra ou raiz de palavra ao valor dessa palavra para expressar o conteúdo dos documentos, ou seja, quanto maior a freqüência, maior peso a palavra receberá.

O peso pode ser atribuído, de acordo com Parker[55], Salton, Wu & Yu[56], Salton & Yang[57], Sparck Jones[58] e Soergel[48], por:

a) freqüência total ou freqüência única — a palavra recebe o mesmo valor do número de sua freqüência;
b) fonte — se a palavra se encontrar em um documento reconhecido como relevante, receberá um peso maior do que outra existente em um documento menos relevante;
c) por fonte e usuário — o usuário é quem julgará se o documento recuperado é relevante ou não. Se for, os descritores utilizados na estratégia de busca terão, posteriormente, seu valor aumentado; e
d) freqüência na coleção.

A indexação automática com pesos,proposta por Sparck Jones[58], denominada por Salton & Yang[57] de "freqüência inversa do documento", trabalha com a especificidade da palavra. As palavras de baixa freqüência são as mais específicas e recebem maior peso. As palavras de alta freqüência são os responsáveis pelo ruído da recuperação da informação. Segundo Sparck Jones[58], a extração de um número grande de palavras por documento aumenta a quantidade de freqüência, mas não a qualidade dos novos descritores.

Nesse mesmo trabalho Sparck Jones[58] realizou estudos comparativo-qualitativos sobre o desempenho da recuperação através da atribuição de pesos por freqüência de ocorrência e co-ocorrência de palavras em documentos e ocorrência de palavras na coleção. Verificou-se que o peso atribuído em relação à coleção é o mais problemático.

A análise discriminatória de documentos é uma das variações do método de atribuição de peso. SaltonS Yang[57] e Salton, Wu & Yu[56] aplicaram esta análise na freqüência de palavras na coleção para aperfeiçoar a revocação e precisão da recuperação. A melhor palavra será aquela que possuir capacidade de discriminação entre os vários documentos semelhantes de uma coleção.

Nesta técnica as palavras que possuem média freqüência são as mais indicadas para a indexação do documento ou da pergunta. As palavras de alta e baixa freqüência são, respectivamente, raras e gerais em termos de ocorrência e possuem um baixo poder de discriminação. O valor da palavra dependerá da maior ou menor distância que provocar entre os documentos da coleção. Esse valor é calculado através de uma fórmula matemática específica.

Dillon & Federhart[59] aplicaram em sua pesquisa um outro tipo de análise discriminatória para selecionar raízes de palavras relativamente freqüentes. Trabalharam, ao contrário dos outros autores, anteriormente citados, só com raízes que possuíam alta freqüência de ocorrência na coleção. As raízes foram analisadas de acordo com os vários significados semânticos e aplicou-se, posteriormente, uma função discriminatória para detectar, caracterizar e classificar os grupos semelhantes e diferentes.

Salton & Yang[57] e Salton, Wu& Yu[56] analisaram a teoria da relevância do usuário como método de atribuição de peso. Esta é uma técnica que aplica a freqüência de ocorrência de palavras no documento e na coleção. Requer uma realimentação constante, pois utiliza o julgamento da relevância do usuário para atribuição de pesos.

Robertson & Sparck Jones[60], Yu & Salton[61] e Harper & Van Rijsbergen[62] propuseram fórmulas matemáticas para atribuição de pesos baseados na teoria de relevância. Parker[55] desenvolveu um modelo matemático aplicado à atribuição de peso pelo usuário, no momento da pergunta, utilizando palavras extraídas por indexacão manual.

3.3.3 - *Método probabilístico*

O método probabilístico utilizado por Carrol & Roeloffs[22], Bookstein& Swanson[63] e Harter[64] aplica a freqüência de co-ocorrência em palavras truncadas automaticamente. As palavras truncadas são extraídas através de um critério estatístico de distribuição binominal, denominado distribuição de Poisson. Aquelas palavras cuja freqüência de distribuição for descrita pela função de Poisson serão não-significantes.

Carrol & Roeloffs[22] aplicaram esta técnica para comparar a indexação automática e a manual utilizando artigos da área de Ciência da Informação. Bookstein & Swanson[63] aplicaram, além desta técnica, a análise de "cluster", concluindo que as palavras significativas tendem a se aproximarem mais do que as não-significativas e concentram-se mais nas áreas de "cluster".

Harter[64] introduziu a noção de relevância, aperfeiçoando o modelo proposto por Bookstein & Swanson[63] e criou a distribuição de Poisson — 2 para analisar, com maior profundidade, palavras técnico-científicas. As palavras são tratadas em dois níveis - significantes e significantes especializadas.

Harter[64], baseado em modelo matemático por ele elaborado, definiu um algoritmo para medir a "indexabilidade" de uma palavra como reflexo do significado relativo das palavras.

3.3.4 - *Método matemático*

O método matemático é baseado na identificação da freqüência de co-ocorrência em pares de palavras em documentos, através de algoritmo.

Tanimoto[65], em 1958, foi o primeiro a propor este método, mas para a classificação automática. Esta idéia foi aplicada à indexação automática, segundo demonstra Batty[74], a partir dos anos 60.

Steinacker[66], em 1974, propôs um algoritmo para detectar frases ou grupos de palavras significativas. O algoritmo produz "cortes no texto" ao localizá-los e, posteriormente, ordena-os alfabeticamente construindo um índice rolado das várias combinações entre as palavras de um mesmo corte. Entre as aplicações dessa técnica, podem ser citadas: criação de dicionários, elaboração de tesauros e desenvolvimento, controle e manutenção de enciclopédias.

3.3.5 - *Análise de "cluster"*

A análise de "cluster" foi introduzida, segundo Batty[47], no início dos anos 60 pelo Cambridge Language Research Unit (CLRU), para a classificação de documentos e elaboração de esquemas de classificação.

A técnica se baseia, de acordo com Salton[21], no reconhecimento automático, em um grupo de documentos, dos subgrupos de assunto que mais se assemelham entre si e entre outros subgrupos.

Sparck Jones[67] realizou um projeto de indexação automática e recuperação baseado na análise de "cluster". E aplicada em *pattern of term,* termos simples, pares de termos isolados em correlação matricial, termo a termo.

3.3.6 - *Método de associação entre palavras*

O método de associação entre palavras, mencionado por Salton[21], utiliza a freqüência de ocorrência e co-ocorrência de palavras ou pares de palavras para identificar o conteúdo dos documentos. As palavras isoladas e as que se co-associam são identificadas em sentenças. Se as co-associações das mesmas palavras co-ocorrerem com determinada freqüência, então, serão consideradas "descritores associados".

O modelo associativo proposto por Jones, Giuliano & Curtice[68] parte do princípio de que todas as palavras significativas estão relacionadas linearmente. A primeira associação entre as palavras é denominada relação contínua, a segunda representa relações de sinonímia.

Lesk[69] comparou os resultados de recuperação em documentos indexados por análise de freqüência e associação, e verificou que o método associativo aumenta o desempenho da recuperação, além de poder ser utilizado em construção e normalização de terminologia para tesauros.

3.3.7 - *Experimentos em avaliação de recuperação da informação em linguagens de indexação*

Pesquisa sobre medidas de desempenho de recuperação em linguagens documentárias começaram a se desenvolver, de acordo com Regazzi[70], após a Segunda Guerra Mundial, devido a necessidade de selecionar a informação útil no caos documentário instalado pela explosão bibliográfica.

A comunidade de pesquisadores realizou vários estudos para descobrir a fórmula ideal de se medir a eficiência de sistemas de recuperação da informação e os instrumentos utilizados para a identificação do conteúdo dos documentos. Bourne[71] levantou, em sua revisão, as várias formas encontradas para quantificar o desempenho da recuperação em linguagens de indexação. Encontrou, como fatores mais citados, as medidas de revocacão e precisão.

Várias instituições avaliaram seus sistemas de recuperação e indexação utilizando as duas medidas. Bourne[71] apresenta um quadro histórico resumindo os projetos experimentais encontrados na literatura, a partir de 1954. Dentre estes destacam-se os projetos Cranfield e SMART, por serem os mais citados na literatura.

O projeto Cranfield, como menciona Bloomfield[72], subdivide-se em I e II; e ambos foram desenvolvidos sob a direção de C. W. Cleverdon, no College of Aeronautics, Inglaterra.

O Cranfield I, iniciado em 1957, tinha como objetivos testar e comparar a capacidade de recuperação de quatro sistemas de classificação: Classificação Decimal Universal, Alphabetic Subject Index, uma classificação facetada e o Uniterm System of Coordinate Indexing. Cleverdon[73] mediu a eficiência de cada linguagem através dos índices de revocacão e relevância.

Várias críticas foram dirigidas ao projeto[72, 74], tais como: algumas variáveis no julgamento de

relevância não foram suficientemente controladas, e o número mínimo de palavras, entre 20 e 60, para descrever de forma adequada o conteúdo dos documentos, tal como proposto por Cleverdon[73], não é adequado para armazenagem e indexação manual.

Em continuação às experiências do Cranfield I, conforme mencionam Bloomfield[72] e Simmons[75], Cleverdon criou, em 1967, o Cranfield II para testar três linguagens de indexação. O primeiro tipo de indexação utilizou termos livres, o outro, termos controlados e o terceiro, conceitos simples.

Os testes foram, novamente, baseados nas medidas de revocacão e precisão. Partiu da linguagem mais simples para a mais sofisticada, utilizando dicionários de sinonímia, associações entre conceitos e hierarquia de termos. Os resultados apresentados, segundo descreve Salton[19, 20], demonstraram que a indexação com termos simples e livres é mais eficiente do que as mais sofisticadas.

Cleverdon[73], através dessas experiências, propôs uma indexação, idealmente exaustiva, com 33 descritores por documento. Novamente, de acordo com Bloomfield[72], foi questionada a viabilidade de um número tão grande de descritores.

Segundo Regazzi[70], o maior mérito dos projetos Cranfield foi abrir as pesquisas na área de avaliação de recuperação da informação entre diversas linguagens documentárias, além de ter definido claramente como aumentar a revocacão e precisão dos sistemas de informação.

O projeto SMART foi desenvolvido, a partir de 1965, por Salton[21]. Foi elaborado para realizar avaliações de várias linguagens de indexação, em termos de revocacão e precisão. Contém, também, um conjunto de programas para realizar indexação automática em textos integrais.

Vários testes de avaliação de linguagem foram realizados no projeto SMART[21], e possibilitaram algumas conclusões:

a) o uso de termos com peso é, normalmente, mais efetivo do que termos sem peso;
b) o uso de dicionários de sinonímia é melhor do que o controle por palavras truncadas;
c) o uso de títulos é, normalmente, menos efetivo para a análise de conteúdo do que o resumo; e
d) a mais importante das conclusões, segundo Salton[21], e que foi, também, encontrada no Cranfield: as linguagens de indexação mais sofisticadas são menos eficientes do que as que utilizam termos livres e simples.

4 - REVISÃO DA LITERATURA NACIONAL

A literatura brasileira sobre indexação automática, comparada com a estrangeira é, significativamente, menor. O que, provavelmente, reflete o pouco desenvolvimento desta técnica no País.

A indexação automática de documentos, no Brasil, iniciou-se, praticamente, segundo Braga[76], com a utilização do programa KWIC para elaborar os índices das bibliografias especializadas que o Instituto Brasileiro de Bibliografia e Documentação (IBBD), atual Instituto Brasileiro de Informação em Ciência e Tecnologia (IBICT), publicava.

Os índices de assunto, conforme menciona Oliveira[77], eram elaborados por palavras-chave permutadas, retiradas, automaticamente, dos títulos das obras, como inclusão, se necessário, de termos preestabelecidos para enriquecer os títulos.

A primeira experiência da utilização do programa KWIC, de acordo com Zaher& Duarte[78] e Zaher et alii[79], foi para editar a *Bibliografia Brasileira de Física,* em 1968. Foi realizada por um grupo de especialistas do IBBD e do Centro Brasileiro de Pesquisas Físicas. Pode-se dizer que este fato marca o início da indexação automática no Brasil.

Após as experiências do antigo IBBD com o programa KWIC, surgiram trabalhos que o questionavam[80] ou o aplaudiam[78] e relatos de experiências de sua utilização, para a indexação e recuperação automática da informação em bibliotecas[81].

Esses trabalhos estão concentrados entre 1 960 e 1 970. Após 1 970 encontram-se algumas pesquisas que não tratam especificamente de indexação automática, mas que servem como apoio ao desenvolvimento dessa técnica.

Estudos sobre a utilização de títulos e/ou resumos para indexação automática foram realizados por Souza[82] e Braga[76]. A primeira autora analisou títulos de artigos de periódicos estrangeiros em Ciência da Informação e Biblioteconomia, simulando, manualmente, a técnica do KWIC. Constatou um crescente aumento do número de palavras significativas nos títulos dos artigos dos periódicos analisados, entre 1970 e 1980, anos posteriores à criação do KWIC, e concluiu que existe uma tendência de/para aumentar as palavras significativas nos títulos.

Braga[76] propõe, em sua dissertação, a utilização do resumo como fator de enriquecimento do título.

Os títulos, apesar de serem pontos de acesso ao conteúdo de documentos, são insuficientes para uma perfeita indexação e recuperação da informação. A autora concluiu que a proporção de palavras significativas do resumo é da ordem de doze para cada uma existente no título. Seu estudo foi realizado em títulos de periódicos científicos na área de Química. Ela aplicou a técnica do KWIC simulado: e para comparar as palavras dos títulos e dos resumos, elaborou tabelas de freqüência de palavras.

A análise de freqüência de palavras utilizando leis bibliométricas foi encontrada pela primeira vez na literatura brasileira em 1973. Maia[83] aplicou a primeira e segunda leis de Zipf, esta última na forma enunciada por Booth[34], assim como a fórmula de transição de Goffman[85], como fatores de análise da informação em língua portuguesa. As duas leis de Zipf estabelecem relações entre a ordem de série de uma palavra e a freqüência de seu aparecimento em texto suficientemente longo.

Booth[84] enuncia a primeira lei, através de uma fórmula matemática (r. f = c). Estabelece 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).

A fórmula de transição de Goffman[85] determina as ordens de série nas quais devem-se encontrar as palavras significativas de um texto em língua inglesa. Maia[83] concluiu que as leis são aplicáveis à língua portuguesa, apresentando para o português um valor diferente da constante "c", de língua inglesa.

Outros estudos bibliométricos foram realizados utilizando as leis de Zipf, sendo alguns *na* área de Lingüística documentária[86], e outros utilizando a lei de Bradford[11].

Robredo[87] utilizou a lei de Bradford, formulada por Brookes[88], como instrumento de controle terminológico estabelecendo "descritores de escopo" - termos de alta freqüência e baixa especificidade que caracterizam áreas do conhecimento, "descritores de facetas" - termos de média freqüência e especificidade, caracterizam subáreas de interesse, e os "descritores pontuais", os de baixa freqüência e alta especificidade, caracterizando um número limitado de documentos.

O método de freqüência de palavras em títulos e ou resumos, para determinação de descritores e construção de núcleos de termos, foi aplicado por Robredo em vários trabalhos[89, 90, 15, 91, 92].

O Centro de Informações Nucleares (CIN), segundo Barreiro[93], aplica a técnica de freqüência e uso de descritores para a seleção adequada do descritor em indexação manual e atualização do tesauro INIS.

A análise estatística e estudos de co-ocorrência de frases e palavras significativas foram utilizados como metodologia por Queiroz[94], para elaboração automática de resumos, e por Torres Filho95, na elaboração de índices automáticos de livros técnicos. Este autor propôs um algoritmo de freqüência.

A indexação automática, utilizando raízes vazias e raízes significativas foi proposta por Robredo[16, 90] para aumentar a rapidez de processamento e a a precisão da recuperação da informação.

Freund[96] descreve, em artigo, técnica semelhante. As raízes são extraídas automaticamente por análise estrutural, através de um algoritmo específico. A técnica é utilizada somente para a recuperação em linha, ou seja, para a indexação da pergunta e montagem da estratégia de busca. O autor fez, também, uma breve comparação entre a análise estrutural e o truncamento arbitrário de palavras.

Na literatura brasileira encontram-se alguns trabalhos que aplicam a lingüística computacional à indexação automática em textos integrais. Citamos Haller[97, 98] e Andreewsky & Ruas[99]. O primeiro autor desenvolveu seu programa na Universidade de Brasília. Esse programa consta de análises morfológica e sintática das palavras do texto, para extrair os descritores. Possui vários dicionários, entre eles o de freqüência de palavras vazias e significativas e o de raízes.

Andreewsky & Ruas" utilizam métodos lingüísticos e estatísticos de atribuição de pesos para as palavras significativas. O programa é uma adaptação para a língua portuguesa do *Système Syntaxique et Probablliste d' Indexation et de Recherche d' Informations Textuelles* (SPIRIT), desenvolvido pelo Centre National de Ia Recherche Scientifique (CNRS) para a língua francesa. Possui algoritmos de análise sintática e análise semântica, além de diversos dicionários.

5 - CONCLUSÃO

As técnicas de indexação automática e manual prendem-se em maior ou menor grau às características dos programas e da filosofia de recuperação dos sistemas de informação.

A aplicabilidade de uma técnica ou de outra foi testada através de vários experimentos, em

diversas áreas do conhecimento e em várias línguas. Ambas as técnicas foram consideradas eficientes. Em alguns casos há maior aceitação da indexação automática, em outros, da manual. Depende das línguas, das áreas do conhecimento em que foram aplicadas e das fontes de informação utilizadas na extração do termo que expressará o assunto do documento.

De maneira geral as pesquisas demonstraram um aumento de palavras significativas nos títulos após o surgimento do índice KWIC, principalmente, na área de ciências exatas. O que torna o título uma fonte a ser considerada para indexação.

O resumo é, também, fonte importante, mesmo que cause em determinadas situações uma menor precisão de recuperação.

O texto integral, para indexação automática, é, praticamente, inviável, devido ao alto custo de digitação e armazenamento.

As linguagens que utilizam termos livres, sem pré-coordenacão, possibilitam uma maior flexibilidade na montagem da estratégia de busca, fornecendo uma recuperação mais precisa.

A análise de freqüência e a atribuição de valores ou pesos aos termos e pares de termos são fatores que aumentam a precisão da resposta em buscas em linha, além de serem instrumentos válidos para a elaboração de tesauros.

No Brasil, os estudos experimentais iniciaram-se no final da década de 60, com a elaboração de índices KWIC para bibliografias, mas não foi dada continuidade a esses estudos.

No final da década de 70, as pesquisas de indexação automática recomeçaram através de estudos individuais, realizados em cursos de pós-graduação, concentrando-se na análise de freqüência, análise semântica 3 sintética do termo. As pesquisas de avaliação de linguagem de indexação para a recuperação estão menos desenvolvidas.

A tendência mundial, segundo Lancaster[100]: "será a do aumento contínuo de bases de dados textuais, com a eliminação da técnica de indexação manual e o desenvolvimento de vocabulários controlados *a posterior!".*

Artigo recebido em 9 de novembro de 1987.

REFERÊNCIAS BIBLIOGRÁFICAS

1 The UNISIST draft on indexing principies: test and comments. *International Classification,* 4(1):29-34, May 1977.

[2] CAVALCANTI, Cordélia Robalinho. *Indexação* & Tesouro; *Metodologia e Técnicas.* Ed. prelim. Brasília, ABDF, 89p.

[3] CAVALCANTI, Cordélia Robalinho. *Metodologia de indexação.* Brasília, 1 976. 8f.

[4] CESARINO, M. A. da N. & PINTO, M. C. M. F. Análise de assunto. *Revista de Biblioteconomia de Brasília,* S (1):32-43, Jan./Jun. 1980.

° BORKO, H. Toward a theory of indexing. *Information Processing and Management,* 73 (6):355-66, 1977.

6 FOSKETT, A. C. *The subject approach to Information. 3.* ed. London, C. Bingley, 1977. 476p.

7 FUGMANN, R. On the practice of indexing and its theoretical foundations. *International Classification, 7* j 1): 13-20, Apr. 1 980.

[8] HUTCHINS, W. J. *Languages of indexing and Classification:* a linguistic study of structures and functions. Stevenage, Peter Perenigrus Ltd., 1975. 148p.

^ JONES, Kevin P. How do we index a report of some Aslib Informatics Group Activity. *Journal of Documentation, 39* (1):1-23, Mar. 1983.

[10] JONKER, F. *Indexing theory, indexing methods and research services.* New York, Scarecrow Press, 1964. 124p.

11 PINHEIRO. Lena Vânia Ribeiro. Medidas de consistência indexação: interconsistência. *Ciência da Informação, 7* (2):109-14, 1978.

12 LEONARD, L. E. *Inter-indexer consistency studies. 1954 — 1975:* a review of the literature and summary of study results. Illinois, University of Illinois, Graduate School of Library Science, 1977. 51 p. (Occasional papers).

13 LUHN, H. P. The automatic creation of literature abstracts. *IBM Journal of Research and Development. 2:* 1 59-1 65, 1958.

[14] LUHN, H. P. A statistical approach to mechanized encoding and searching of literary information. *IBM Journal of Research and Development, 1* (4|:309-17. Oct. 1957.

15 ROBREDO, Jaime A indexação automática de textos: o presente já entrou no futuro. In: Machado, U. O., ed. *Estudos Avançados em Biblioteconomia e Ciência da Informação.* Brasília, ABDF, 1982. v. 1, p. 236-74.

[16] ROBREDO. Jaime *Documentação de hoje e de amanhã.* Brasília ABDF, 1976. VIII, 172p.

[17] SARACEVIC, Tefko. Relevance; a review of and *z* framework for think the notion in information science. *Journal of American Society for Information Science, 26* (6): 321-43, Nov./Dec. 1975.

VICKERY, B, C. Structure and function in retrieval languages. *Journal of Documentation, 27* (2):69-82, June 1971.

SALTON, G. A comparison between manual and automatic indexing systems, *Computing Reviews, 10* ,(6):274, June. 1969.

20 SALTON, G. A new comparison between conventional indexing and automatic text processing. *Journal of the American Society for Information Science, 23* (2):75-84, Mar./Apr. 1972.

[21] SALTON, G. Automatic text analysis: automatic document indexing and classification methods are examined and their effectiveness assessed. *Science, 168* (3929):335-43, 17 Apr. 1970.

[22] CARROLL, John M. & ROELOFFS, Robert. Computer selection of keywords using word-frequency analysis. *American Documentation, 20* (3):227-33, July 1969.

[23] BOYCE, Bert. & LOCKARD, Marta. Automatic and manual indexing performance in a small file of medicai literatura. *Bu/íetin of Medicai Library Association, 63* (4):378-85, Oct. 1975.

[24] VAN DER MEULEN. W. A. & JANSSEN. P. J. F. C. Automatic versus manual indexing. *Information Processing and Management, 13* |1):13-21. 1977.

[25] KLINGBIEL, Paul H. & RINKER, Catherine C. Evaluation of Machine-Aided Indexing. *Information Processing and Management, 12* (6):351-66, 1976.

[26] BARNES, C. l; COSTANTINI, L.& PERSCHKE, S. Automatic indexing using the SLC II *Sustem. Information Processing and Management, 14* (2):107-119, 1978.

[27] LANCASTER, F. W. *Evaluation of the operating efficiency of Medlars:* final report. Bethesda, National Library of Medicine, 1968.

[2]8 WESSEL, A. E. Indexing and analysis of information-some preliminary computs. In:_____*Computer/aided information retrieval.* Los Angeles, Melville Publishing Company, 1975, p. 1-10

[29] MAIZELL, R. Value of titles for indexing purposes. *Revue de la Documentation,* 27:1 26-7, 1 960.

[30] MONTGOMERY, C. & SWANSON, D. R. Title indexing. *American Documentation,* 73:359-64, 1962.

31 RUHL, M. J. Chemical documents and their titles: human concept indexing vs KWIC - Machine indexing./lmemjan *Documentation, 15* |2):1 36-41, Apr. 1964.

[32] KRAFT, D. H. A comparison of keyword in context (KWIC) indexing of titles with a Subject Heading Classification System. *American Documentation,* /5(1):48-52, Jan. 1964.

33 TOCATLIAN J. J. Are titles of Chemical papers becoming rnore informative? *Journal of the American Society for Information Science,* 27:345-50, 1970.

[34] BIRD. P. R.& KNIGHT, M. A. Word count statistics of scientific papers. *Information Scientist,* P(2):67-9,1975.

35 GHOSH, Jata S. Content representation in document titles: a case study with prostaglantin lit*erature.Aslib Proceedings,* 26(2):83-6, Feb. 1974.

[36] BUXTON, A. B. & MEADOWS, A. J. The variation in the information content of titles of research papers with time and discipline. *Journal of Documentation, 33* (1):46-52, Mar. 1977.

[37] BLOOMFIELD, M. Evaluation of indexing 2. The Simulated Machine Indexing Experiments. *Special Libraries, 61* (9|:501-7. Nov. 1970.

[38] SVENONIUS, Elaine. An experiment ín index term frequency. *Journal of the American Society for Information Science, 23* (2):1 09-21, Mar./Apr. 1972.

[3^] SALTON, G. Automated language processing. *Annual Review of Information Science and Technology,* 3:169-99, 1968.

[40] BARKER, F. H.: VEAL, D. C. & WYATT, B. R. Comparative efficiency of searching titles, abstracts and index terms in a free-text data *base. Journal of Documentation, 28* (1):22-36, Mar. 1972.

[41] FEINBERG, H. *Title derivatíve indexing techniques: a* comparative study. Metuchen, Scarecrow Press. 1973. 297 p.

[42] GARFIELD, E. A weekly subject index for Current Contents/ Life Sciences. In: ANNUAL MEETING OF THE MEDICAL LIBRARY ASSOCIATION, 71., San Diego, June 11-15, 1972.

[43] NEUFELD, M. L. et alii. Automatic title word indexing for a weekly current awareness service. In: ANNUAL MEETING OF THE AMERICAN SOCIETY FOR INFORMATION SCIENCE, 36, Los Angeles, October 21-25, 1973. v. 10. *Innovative Development in Information Systems:* their benefits and costs. Ed. by Helen J. Waldron & F. Raymond Long. Washington, DC, ASIS and Westport, Conn., Greenwood Press, 1973. pp. 167-8.

[44] KWOK, K. L. Cited titles: a new source of keyword extraction for automatic document classification and retrieval. In: ASIS ANNUAL MEETING, 37. Atlanta, 13-17 Oct., 1 974. *Proceedings.* Washington, ASIS, 1974. v. 11, pp, 56-57.

[45] GARFIELD, E. A conceptual review of citation indexing. In:_____*Citation indexing*_____its theory and application in Science Technology, and Humanities. New York, John Wiley and Sons, 1978, p. 1-5.

[46] CLEVELAND, D. B.; CLEVELAND, A. D. & WISE, O. B. Less than fulltest indexing using a non-boolean searching model. *Journal of the American Socienty for Information Science, 35* (1): 19-28, 1984.

[47] BATTY, C. D. The automatic generation of index languages. *Journal of Documentation, 25* (2):142-51, June 1969.

[48] SOERGEL, Dagobert. Automatic and semi-automatic methods as an aid in the construction of indexing languages and thesauri. *International Classification, 1* (1):34-9, May 1974.

[49] SPARCK JONES, Karen. Indexing term weighting. *Information Storage and Retrieval, 9* (11):61 9-33, Nov. 1973.

AO, Miranda Lee. Automatic text analysis based on GoffmarVs transition phenomena of word ocurrences. *Journal of the American Society for Information Science, 29* !3):121-4, May. 1978.

51 ROWBOTTOM, Mary E. & WILLET, Peter. The effect of subject matter on the automatic indexing of full text. *Journal ofthe American Society for Information Science, 33* (3):1 39-41, May 1982.

[52] SCHUEGRAF, Ernest. & HEAPS, l. Indexing for associative processing. *Canadian Journal of Information Science.* 5:93-101, May 1980.

53 ROSENBERG, Victor. A study of statistical measures for predicting terms used to index documents. *Journal of American Society for Information Science, 22* (1):41-50, Jan./Fab. 1971.

54 HENZLER, R. G. Free orcontrolled vocabularies: some statistical user-oriented evaluations of biomedical information systems. *International Ctassification, 5* (1):21-6, Mar. 1978.

55 PARKER, Lorraine M. PurgailÍ8. Towards a theory of document learning. *Journal of the American Society for Information Science, 34* (1): 16-21, Jan. 1983.

56 SALTON, G.; WU, H.& YU, C. T. The measurement of term importance in automatic indexing. *Journal of the American Society for Information Science, 32* (3):1 75-86, May 1981.

57 SALTON, G. & YANG, C. S. On the specification of term values in automatic indexing. *Journal of Documentatior, 29* (4):351-72, Dec. 1973.

58 SPARCK JONES, Karen. A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation, 28* |1):11-21, Mar. 1972.

59 DILLON, Martin» FEDERHART, Peggy. The use of discrimínant analysis to select content-bearing words. *Journal of the American Society for Information Science, 33* (4):245-53, July, 1982.

60 ROBERTSON, S. E. & SPARCK JONES, J. Relevance weighting of research terms. *Journal of the American Society for Information Science,* 27:129-146, 1976.

61 YU, C. T. & SALTON, G. Precision weighting-an effective automatic indexing method. *Journal of Association for Computing Machinery, 23:76-88,* 1 976.

62 HARPER, D. J.& VAN RIJSBERGEN, C. J. An evaluation of feedback in document retrieval using co-ocurrence data. *Journal of Documentation,* 34(3):189-216, 1978.

63 BOOKSTEIN, A. & SWANSON, D. R. A decision theoretic foundation for indexing. *Journal of American Society for Information Science, 26* (1):45-50, 1975.

64 HARTER, Stephen P. A probabilistic approach to automatic Keyword indexing. *Journal of the American Society for Information Science, 26* (4):1 97-206, July-Aug. 1975; *26* (5):280-289, Sept./0ct. 1975.

65 TANIMOTO, T. T. *An elementary mathematical theory of classification and predication,* IBM, 1958. n. p.

66 STEINACKER, Ivo. Indexing and automatic significance analysis. *Journal of American Society for Information Science, 25* (4):237-41, July/Aug. 1974.

67 SPARCK JONES, Karen. The role of automatic indexing in operational online retrieval systems. In: FID CONGRESS, 39., Edinburg, 25-28 September 1978. *New trends in documentation.* London, Aslib, 1980. p, 33-8.

68 JONES, P. E.; GIULIANO, V. E. & CURTICE, R. M. *Papers on automatic language processing — linear models for associative retrieval,* Report ESD-TR-67-202. Ad Little, Inc., Cambridge, 1967. v. 2

69 LESK, M. E. Word-word associations in document retrieval systems. In: CORNELL UNIVERSITY. Department of Computer Science. *Report nº ISR-13 to the National Science Foundation.* Ithaca, N. Y. Jan. 1968. Section 9.

70 REGAZZI, J. J. Evaluating indexing systems: a revíew after Cranfield, *The Indexer, 12* (1):14-21, Apr. 1980.

71 BOURNE, C. P. Evaluation of indexing systems. *Annual Peview of Information Science and Techonology,* 7:171-90, 1966.

72 BLOOMFIELD, M. Evaluation of indexing 4. A Review of the Cranfield Experimenta. *Speciai Libraries, 62* (1): 24-9, Jan. 1971.

73 CLEVERDON. C. W. *Report on the testíng analysis of an investigation into comparative efficiency of indexing systems.* Cranfield, College of Aeronautics, ASLIB, Cranfield Research Project, 1962. n.p.

74 CUADRA, C. A. & KATTER, R. V. *Experimental studies of relevance judgements.* Final Report, v. 1 Project. Summary. Santa Monica, System Development Corp., 1967. (TM-3520 0001 700).

75 SIMMONS, R. F. Automated language processing. *Annual Revíew of Information Science and Technology,* 7:137-69, 1966.

76 BRAGA, L. M. *Palavras de títulos e resumos como acesso ao conteúdo do documento:* uma análise numérica. Rio de Janeiro, URFJ/IBICT, 1982. 181 p. (Dissertação).

77 OLIVEIRA, Elvia A. Automação da Bibliografia Brasileira de Ciências Sociais. In: CONGRESSO REGIONAL DE DOCUMENTAÇÃO, 3., REUNIÃO FID/CLA 11., Lima, 1972. *Anais.* Rio de Janeiro, IBBD, 1972. p. 59-61.

78 ZAHER, C. R.& DUARTE, Y. C. Sistema Kwic versus descritores. In: CONGRESSO SOBRE DOCUMENTAÇÃO, 2.; FID/CLA Reunião 9., Rio *de* Janeiro, 1 969. /Ina/s. Rio de Janeiro, IBBD, 1969. p. 195-206.

79 ZAHER, C. L. et alii. Automação da informação em Física no Brasil. In: SEMINÁRIO SOBRE INFORMÁTICA, Rio de Janeiro, 1968. *Anais.* Rio de Janeiro, IBBD, 1969. p. 39-52.

80 KNIGHT, G. N. *Treinamento em indexação:* um curso da Society of Indexers, Rio de Janeiro, FGV, 1974. 216 p.

81 MACHADO, Norma & HAMAR, Alfredo A. Sistema de arquivamento *e* indexação por computador, do acervo de programas de um Centro de Processamento de Dados. In: CONGRESSO REGIONAL DE DOCUMENTAÇÃO, 2., REUNIÃO FID/CLA, 9., Rio de Janeiro, 1 969. *Anais.* Rio de Janeiro, IBBD, 1970. p. 237-41.

82 SOUZA, Elíana Santos. Estudo dos títulos de artigos de periódicos da área de Biblioteconomia e Ciência da Informação na década pós KWIC: 1960 a 1970. *Ciência da Informação, 7* (2):115-7. 1978.

83 MAIA, E. L e S. Comportamento bibliométrico da língua portuguesa como veículo de representação da informação. *Ciência da Informação, 2* (2):99-1 38, ! 973.

[84] BOOTH, A. D. A "Law" of ocurrences for words of law frequency. *Information and Control, W* (4):386-93, 1967.

85 GOFFMAN, W. A general theory of communication. In: SARACEVIC, T. *Introduction to Information science.* New York, Bowker, 1970. pp. 726-47.

[86] RIBEIRO, L. A. Aplicação dos métodos estatísticos e da teoria da informação e da comunicação na análise lingüística: estudo da linguagem jornalística. *Ciência da Informação, 3* (2):151-4, 1974.

[87] ROBREDO. J. Otimização dos processos de indexação dos documentos e recuperação da informação mediante o uso de instrumentos de controle terminológico. *Ciência da Informação, 11* (1]:3-18, 1982.

88 BROOKES, B. C. Bradford's law and the bibliographv of science. *Nature,* 224:953-56, 1 969.

89 BINAGRI. *Guia brasileiro de pesquisa agrícola em andamento.* Brasília, 1978. 2v. (Projeto PNUD/FAO/BRA/72/020. DOC./TEC. 78/061 e DOC./TEC./78062).

90 ROBREDO, J. A indexação automática como mecanismo básico no processo de transferência da informação In: CONGRESSO LATINO-AMERICANO DE BIBLIOTECONOMIA E DOCUMENTAÇÃO, 1., Salvador, 21-26 set., 1 980. *Anais,* 19 p.

[91] ROBREDO, J. *et alii.* Construção de um núcleo de thesaurus em agricultura baseado no uso real dos descritores In: REUNIÃO BRASILEIRA DE CIÊNCIA DA INFORMAÇÃO, 1.. Rio de Janeiro, 1975. *Anais.* Rio de Janeiro, IBICT, 1978. v. 1. pp. 289-303.

92 ROBREDO, J. *et alii. Elaboración de un thesaurus agrícola baseado en criterios de eficiencia del lenguaje en et proceso de comunicación.* Brasília, SNIDA, 1975. 23 p.

93 BARREIRO, S. C. experiência em indexação do Centro de Informações Nucleares. In: REUNIÃO BRASILEIRA DE CIÊNCIA DA INFORMAÇÃO,!., Rio de Janeiro, 1 975. .*Anais.* Rio de Janeiro, IBICT. 1978, pp. 237-45.

94 QUEIROZ, Mucio G. S. *Um estudo comparativo de processos estatísticos para obtenção automática de resumos.* Rio de Janeiro, PUC, 1973. (Dissertação).

[95] TORRES FILHO, Paulo Roberto Pinheiro. *Um sistema semi-automático para o apoio à indexação de documentos técnicos.* Rio de Janeiro, PUC, 1983. 82 p. *(Dissertação).*

96 FREUND, George Eduardo. Análise estrutural para aumentar a eficiência de pesquisa online. *Ciência da Informação. 11* (1):19-26, 1982.

97 HALLER, Johann. Análise automática de textos em sistemas de informação. *Revista de Biblioteconomia de Brasília, 11* (1):105-113, jan./jun. 1983.

[98] HALLER, Johann. Processamento de textos em linguagem natural. In: CONGRESSO NACIONAL DE INFORMÁTICA, 15., Rio de Janeiro, out. 1982. *(Trabalhos apresentados).* Rio de Janeiro, 1982. 9 p.

[99] ANDREEWSKY. Alexandre S RUAS, Vitoriano. *Indexação automática baseada em métodos lingüísticos e estatísticos e sua aplicabilidade à língua portuguesa* Rio de Janeiro, PUC - Dl. 1 982. 31 p.

100 LANCASTER, F. W. Trends in subject indexing from 1957 to 2000. In: FID CONGRESS, 39., Edinburgh, 25-28 September 1978. *New trends in documentation and Information.* London, Aslib, 1980. p. 223-33.

AUTOMATIC AND MANUAL INDEXING: **A LITERATURE REVIEW.**

ABSTRACT

Presents several national and foreign research works which evaluate the quality of the manual and automatic indexing related to the techniques and sources employed for the extraction of significant terms and the retrieving capabilities of the indexing language in the data bases.