ANÁLISE COMPARATIVA ENTRE INDEXAÇÃO AUTOMÁTICA E MANUAL DA LITERATURA BRASILEIRA DE CIÊNCIA DA INFORMAÇÃO.

SIMONE BASTOS VIEIRA
Senado Federal
Subsecretaria de Biblioteca
70160, Brasília, DF

Texto parcialmente extraído da dissertação **Análise comparativa entre indexação automática e manual da literatura brasileira de Ciência da Informação.** UnB – Curso de Mestrado em Biblioteconomia e Documentação, dezembro 1984. Orientação: Jaime Robredo. Foi realizado um estudo comparativo da qualidade de indexação manual e automática, na área da Ciência da Informação, num conjunto de documentos em língua portuguesa, constituído por artigos publicados no periódico **Ciência da Informação.** Na indexação automática utilizaram-se os títulos e os resumos, e na indexação manual os textos dos artigos. A qualidade dos vocabulários foi avaliada em função do grau de coincidência dos termos em ambos os vocabulários, e de suas respectivas freqüências de aparecimento. A qualidade de recuperação da informação em linha, nas bases de dados, formadas com os mesmos registros dos referidos artigos indexados automaticamente e manualmente, foi medida através dos respectivos índices de precisão da recuperação. A indexação automática apresenta menor redundância no vocabulário, e permite maior precisão na recuperação, especialmente quando se aplica a truncagem dos descritores.

1. INTRODUÇÃO

A informação é considerada como um produto tão valioso quanto os recursos minerais e energéticos em um país. Para se medir o desenvolvimento de uma nação, basta verificar a quantidade e a qualidade de informações que seus habitantes geram e consomem.

A tarefa de tornar acessíveis as informações relevantes requer uma série de atividades que compõem o que se denomina ciclo documentário. Essas atividades

compreendem, basicamente, a seleção, aquisição, registro, descrição física, análise de conteúdo, armazenamento, recuperação e disseminação de informação.

Dentre as diversas formas de análise de conteúdo, a indexação é a técnica que parece oferecer uma melhor condensação do assunto do documento, e o faz mediante a atribuição de descritores, possibilitando aumentar a capacidade de armazenamento e o desempenho de recuperação.

A indexação é uma das operações significativas que compõe o ciclo documentário. Pode-se dizer que é uma atividade-meio, que possibilita ao usuário o acesso adequado ao conteúdo dos documentos. Usualmente, é considerado um dos pontos de estrangulamento dos sistemas de informação.

A escassez de pessoas especializadas em indexar rapidamente e com eficiência os documentos é uma realidade brasileira. A formação de profissionais nesta área não é muito adequada. Só a partir do novo currículo mínimo, aprovado, em 1982, os cursos de graduação em Biblioteconomia passaram a incluir, de maneira mais acentuada, a indexação como uma das técnicas a serem ministradas. Mesmo assim, não se pode acreditar que, em áreas tecnológicas muito especializadas, os bibliotecários possam indexar com perfeição, pois a formação acadêmica que receberam está, tradicionalmente, voltada mais para a área de ciências sociais.

A indexação manual é uma tarefa que requer conhecimento do assunto do documento, consistência técnica e desenvolvimento de linguagens de indexação apropriadas a cada sistema de informação. É uma técnica que exige do profissional um tempo razoável de dedicação por documento. Todos esses aspectos apresentados tornam a tarefa de análise manual cada dia mais dispendiosa.

A indexação automática de títulos e resumos apresenta-se como uma opção rápida, eficiente e de baixo custo, a longo prazo, para a análise do conteúdo dos documentos. É uma técnica que prescinde, de certa forma, da presença do homem para a realização intelectual da atividade. O computador, programado especialmente e alimentado com títulos e resumos a serem indexados, pode efetuar a indexação em espaço de tempo bem inferior, além de permitir maior consistência.

Este estudo irá fornecer diretrizes que avaliarão a eficiência da indexação automática em títulos e resumos em relação à indexação manual, em língua portuguesa e na área de Ciência da Informação.

2. METODOLOGIA

Dentre as várias áreas do conhecimento existentes, foi selecionada a literatura brasileira referente à Ciência da Informação, por ser relativamente nova, carente de estudos sistemáticos quanto à sua terminologia, e por ser a área de atuação dos

profissionais de informação, bibliotecários, cientistas da informação, documentalistas e tantos outros.

A presente pesquisa foi realizada utilizando como universo de estudo os artigos publicados no periódico **Ciência da Informação**, no período de 1972 a 1983.

As etapas do trabalho experimental foram as seguintes:

1. elaboração da descrição bibliográfica de cada artigo do periódico **Ciência da Informação**, no formato utilizado pelo sistema BIB/DIÁLOGO (5), que é baseado no formato sugerido pela UNESCO(10);

2. indexação manual, em linguagem livre, devendo a escolha dos descritores recair na forma mais aproximada da utilizada pelo autor. Utilizaram-se o título, o resumo e o texto. A operação foi realizada por uma estudante do Departamento de Biblioteconomia da UnB, devidamente orientada;

3.digitação, validação e formatação de uma base de dados para interrogação em linha, denominada CINFORM, a partir dos registros indexados manualmente e dos resumos já existentes nos artigos;

4. geração de instrumentos de controle e de análise comparativa:

- índice tipo KWIC;
- listagem de referências bibliográficas;
- índices de autores pessoais e institucionais;
- listagem alfabética dos descritores manuais;
- listagem de freqüência de aparecimento dos descritores.

5. cópia do arquivo CINFORM; eliminam-se os descritores introduzidos manualmente para gerar um novo arquivo, idêntico, com registro a serem indexados automaticamente;

6. indexação automática dos títulos e resumos hifenados previamente, para evitar a perda de palavras significativas;

Exemplo: indexação automática.

Neste processo as palavras dos títulos e resumos são comparadas com duas tabelas: uma com palavra e a outra com raízes, ambas vazias de significado. Excluem-se as palavras coincidentes das tabelas e os descritores são as palavras não eliminadas;

7. formatação e geração de uma outra base de dados para interrogação em linha, a partir dos registros indexados automaticamente, denominados INDEXCI, e criação dos mesmos instrumentos de controle e de análise comparativa mencionados no item 4;

8. análise comparativa entre os descritores obtidos por indexação manual, automática e índice KWIC, através da verificação da coincidência de termos e da aplicação da lei de Bradford, nas freqüências dos descritores;

9. interrogação das bases de dados CINFORM e INDEXCI a partir de 33 buscas bibliográficas sugeridas por oito professores do Departamento de Biblioteconomia da UnB.

Foram montadas 72 estratégias de busca, utilizando-se os operadores booleanos **e, ou,** e truncagem à direita dos descritores. As perguntas às duas bases de dados foram realizadas de forma interativa, dirigindo-se sempre a mesma estratégia às duas bases. O resultado de cada estratégia foi avaliado de acordo com a precisão da resposta, aplicando-se a fórmula:

$$\text{Precisão} = \frac{\text{número de referências relevantes}}{\text{número de total referências recuperadas}}$$

Para a digitação, formatação, processamento, indexação automática, formatação das bases de dados e recuperação da informação foram utilizadas as facilidades oferecidas pelo sistema BIB/DIÁLOGO (5), implementado no Departamento de Biblioteconomia da UnB, para computadores Burroughs B6700, e terminais Burroughs, modelo TVA 800/10, com a linguagem de controle CANDE[3].

3. RESULTADOS

3.1 – Análise comparativa entre os vocabulários obtidos por indexação automática e manual.

A verificação da coincidência entre os vocabulários foi realizada agrupando-se em uma única lista alfabética os descritores obtidos por indexação manual (base de dados CINFORM), indexação automática (base de dados INDEXCI) e pelo índice KWIC, com suas respectivas freqüências, perfazendo um total de 837 termos. A partir dessa lista elaborou-se uma outra, de raízes, com truncagens para eliminar as variações morfológicas das palavras, com um total de 371 radicais e truncamentos.

A tabela 1 demonstra os totais dos descritores coincidentes entre as duas bases de dados e entre estas e o índice KWIC.

TABELA 1 – Descritores coincidentes

Base de dados	Descritores comuns no total de 875 termos	%
INDEXCI e CINFORM	203	23,6
INDEXCI e KWIC	264	30,8
CINFORM e KWIC	172	20,0

O índice de coincidência de termos entre as bases de dados INDEXCI e CINFORM é relativamente baixo (23,6%), principalmente se comparado ao índice de 60% mencionado por Salton (8) como o normalmente encontrado.

Na tabela 2 verifica-se que os termos do índice KWIC estão incluídos nos vocabulários de INDEXCI e CINFORM.

TABELA 2 – Descritores não-coincidentes

Base de dados	Descritores não-coincidentes	%
INDEXCI	252	29,4
CINFORM	237	27,6
KWIC	1	0,1

Essa total coincidência significa que os títulos dos artigos, em língua portuguesa, do periódico **Ciência da Informação**, apresentam-se com palavras suficientemente significativas para representar o conteúdo dos documentos e, por conseqüência, são fontes importantes para indexação automática.

Na tabela 3, o índice de coincidência aumenta cerca de 20% entre INDEXCI E CINFORM. Essa tabela representa a comparação entre descritores com freqüência (f) de ocorrência, nas duas bases de dados, maior ou igual a quatro.

TABELA 3 – Descritores com freqüência 4

Base de dados	Descritores f 4	%(*)
INDEXCI e CINFORM	65(**)	43,0
CINFORM	19	12,5
INDEXCI	39	25,0

Verifica-se, neste trabalho, que a freqüência está diretamente relacionada com a coincidência, ou seja, à medida em que a freqüência aumenta, cresce o número de descritores coincidentes. Estes são, também, os que contêm maior conteúdo significativo para representar o assunto dos artigos registrados em INDEXCI e CINFORM.

(*) Cálculo realizado sobre o valor 151, que representa o total de descritores em INDEXCI e CINFORM com freqüência maior ou igual a quatro.

(**) Este valor equivale ao total de termos comuns em INDEXCI e CINFORM, com freqüência maior ou igual a quatro em pelo menos um deles.

Um outro fator que aumenta a coincidência dos vocabulários é a truncagem dos termos. As tabelas 4 e 5 apresentam, respectivamente, os totais de coincidência e não-coincidência dos radicais e truncamentos dos descritores.

TABELA 4 – Raízes e truncamentos coincidentes

Base de dados	Raízes e truncamentos comuns	%
INDEXCI e CINFORM	161	43
INDEXCI e KWIC	172	46
CINFORM e KWIC	121	32

Os percentuais da tabela 4 foram calculados sobre o valor 371, total de raízes e descritores truncados.

TABELA 5 – Raízes e truncamentos não-coincidentes

Base de dados	Raízes e truncamentos não-coincidentes	%
INDEXCI	97	26,1
CINFORM	40	10,7
KWIC	1	0,26

Observa-se que o total de raízes e termos truncados é 371, e o total de descritores obtidos por indexação manual e automática é 857. Isto pode significar que existem 57% de termos a serem considerados redundantes, devido, principalmente, aos descritores compostos e pré-coordenados encontrados em INDEXCI (147) e em CINFORM (312), sendo que a indexação manual apresentou um índice de 52,9% de pré-coordenação a mais do que a indexação automática.

A redundância em CINFORM ocorreu porque a combinação dos termos foi normalmente realizada com a repetição de um descritor de mesma raiz com outros. Ex.: autor - autor brasileiro
- autores transientes
- autoria...

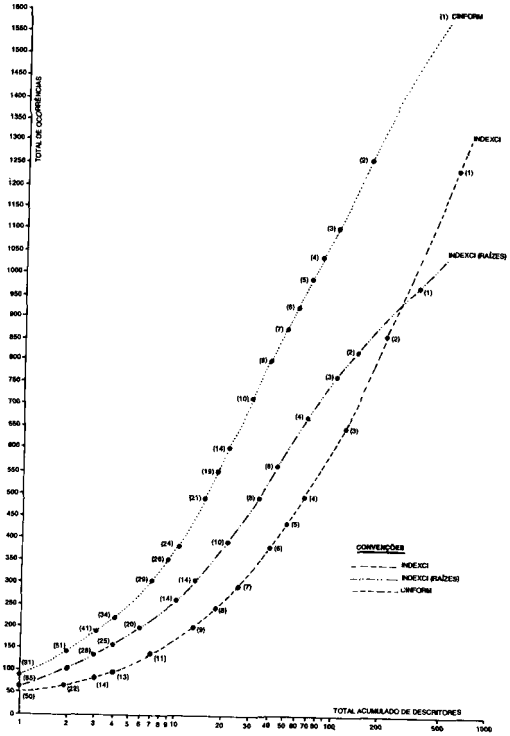
A figura 1 apresenta graficamente a aplicação da lei de Bradford aos vocabulários de INDEXCI, CINFORM e à lista de raízes e palavras truncadas de INDEXCI.

No eixo das abcissas encontra-se o total acumulado de ocorrências (coluna AXB), e no eixo das ordenadas figura o total acumulado de descritores (coluna B). Esse tipo de gráfico já foi delineado anteriormente por Robredo (6) de acordo com a adaptação à formulação de Brooks (2).

A dispersão em CINFORM caracteriza a redundância dos termos pré-coordenados e a existência de várias sinonímias. Em INDEXCI, a dispersão deve-se ao significativo número de termos muito específicos, tais como siglas, datas e nomes de instituições de países, que estão relacionados com a área da Ciência da Informação. Deve-se também às variações morfológicas dos termos (ex.: autor e autoria) e à ocorrência de algumas palavras pouco significativas que não foram eliminadas através do filtro de qualidade, mas que facilmente podem ser acrescentadas às listas de palavras e raízes vazias.

FIGURA 1 – Representação em escala semilogarítmica da variação do número de ocorrências dos descritores em função do número de descritores identificados, nas bases de dados CINFORM e INDEXCI

(OS NÚMEROS ENTRE PARÊNTESES INDICAM A FREQÜÊNCIA CORRESPONDENTE DOS DESCRITORES)



Na curva da tabela de raízes e palavras truncadas de INDEXCI verifica-se que praticamente não existe dispersão, pois o radical e o truncamento eliminam as redundâncias e as variações morfológicas das palavras na base de dados INDEXCI.

Verifica-se, novamente, a concentração e dispersão de termos nas freqüências elevadas e baixas.

Utilizando-se terminologia adotada por Robredo (6), os termos que se localizam na parte inicial da curva, com freqüência alta, são os descritores de escopo, que caracterizam subáreas da Ciência da Informação ou as categorias desta área, como, por exemplo, os descritores Análise da Informação, Automação, Informação Cientffica, Documentação, Biblioteconomia, Transferência de Informação e tantos outros.

Os termos que se situam na parte central da curva, entre as freqüências 10 e 4, são os descritores de facetas, que representam assuntos mais específicos, tais como Análise Bibliométrica, Bibliotecas Especializadas, Catalogação, Intercâmbio de Informações, Estudo de Usuários e outros. Estes são os descritores mais indicados para selecionar com rapidez e precisão documentos relevantes em uma busca bibliográfica, pois são eles que melhor caracterizam o conteúdo dos documentos.

Em alguns casos, os descritores de escopo podem se confundir com os descritores de facetas, dependendo da área de abrangência do assunto a ser pesquisado e do grau de revocação desejado.

Na última parte da curva encontram-se os descritores pontuais, que são numerosos e possuem freqüência abaixo de 3. Esses descritores são os maiores responsáveis pela dispersão e caracterizam sinonímias, quase-sinonímias, as variações morfológicas, assim como descritores com pouco valor significativo, que devem ser eliminados mediante um aprimoramento das listas utilizadas como filtros. Mas esses descritores representam, também, descritores de alta especificidade e relevância.

Com estes dados pode-se afirmar que a indexação automática, aplicada aos títulos e resumos dos 144 registros bibliográficos existentes na base de dados, constituída pelos artigos publicados no periódico **Ciência da Informação**, entre 1972 e 1983, identificou, de maneira equivalente à da indexação manual, os termos significativos que caracterizam essa base.

3.2. Análise comparativa da qualidade de recuperação em linha entre as bases de dados indexadas automaticamente e manualmente.

Para melhor comparação entre a precisão e o ruído obtidos através de buscas bibliográficas em linha, realizadas nas bases de dados INDEXCI e CINFORM, elaborou-se o teste estatístico Mann-Wilcoxon com o objetivo de calcular o índice mé-

dio de precisão das duas bases. Para a primeira encontrou-se o valor de 0,713 (ou 71,3%), e para a segunda 0,577 (ou 57,7%).

A precisão e o ruído são inversos, significando que o índice médio de ruído apresentado em INDEXCI é o 0,287 (ou 28,7%), e em CINFORM 0,423 (ou 42,3%).

Como se pode observar na figura 2, o índice médio de precisão apresentado em INDEXCI é significativamente superior ao encontrado em CINFORM. Isto se explica pelo próprio vocabulário existente em uma base e na outra.

Demonstrou-se, anteriormente, que INDEXCI apresentou menor redundância, o que acarretará, por conseqüência, uma recuperação mais precisa.

A redundância em CINFORM, em contrapartida, fornecerá um número maior de referências recuperadas, o que não significa, necessariamente, serem elas relevantes.

Proporcionalmente ao total de referências recuperadas, o número de referências relevantes em INDEXCI foi significativamente superior ao de CINFORM, que apresentou 50,4% de referências irrelevantes recuperadas.

Verificou-se, também, no item 3.1, que a compatibilidade de termos entre ambas é relativamente baixa, havendo maior coincidência em termos com freqüência de ocorrência acima de 3. Isto acarreta diferentes resultados na recuperação, pois nem sempre um termo escolhido para a formulação da estratégia de busca em uma base é o mais adequado para a outra.

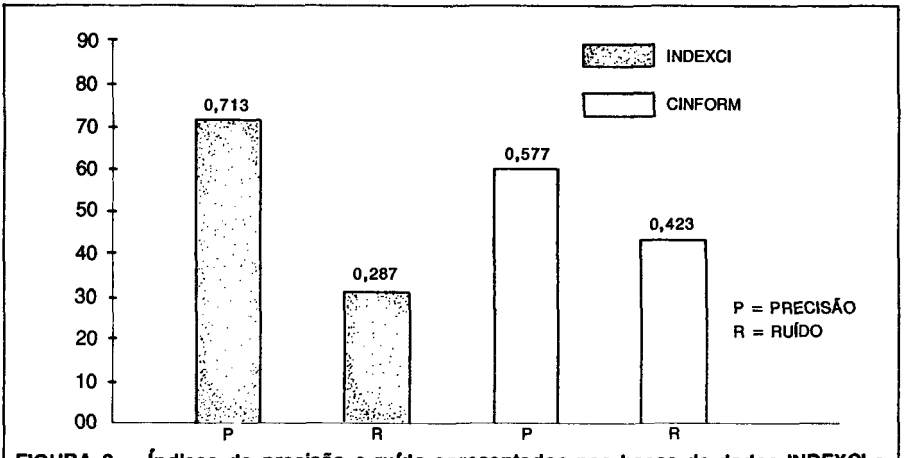


FIGURA 2 – Índices de precisão e ruído apresentados nas bases de dados INDEXCI e CINFORM.

O conhecimento das diferenças entre os vocabulários é importante para selecionar precisamente o descritor e formular-se a estratégia de busca.

A pré-coordenação existente no vocabulário da base de dados CINFORM é

um fator que possivelmente deveria tornar a recuperação mais precisa, pois torna o descritor mais específico. Nesta pesquisa a pré-coordenação foi um fator que dificultou a elaboração da estratégia de busca, apresentando ruídos no resultado final. É possível que, se a indexação manual tivesse sido realizada através de um tesauro para controlar a pré-coordenação, o resultado seria, provalvelmente, um pouco diferente.

O vocabulário formado pela indexação automática foi mais adequado para a busca de linha, pois os termos oferecem maior flexibilidade na formulação da estratégia, através de lógica booleana, realizando-se a coordenação adequada e desejada entre os descritores, no momento da pergunta e de acordo com o assunto.

Isto indica que, no caso deste trabalho, o uso da combinação de termos livres e simples foi mais eficiente para a recuperação em linha do que o uso de vocabulários pré-coordenados.

Este resultado foi o mesmo a que chegaram o projeto CRANFIELD II (1), as experiências de Salton (8) e Robredo (6).

Ocorreu falsa recuperação em somente uma busca de INDEXCI. Isto se deve ao uso de palavras ambíguas e homônimas no resumo ou no título. No caso da busca sobre estágio de estudantes de Biblioteconomia, todas as referências recuperadas foram irrelevantes, pois esta palavra foi utilizada como sinônimo de **situação atual**, que não expressa o conceito desejado.

Observa-se então que, para tornar a indexação automática mais eficiente, devem-se adotar algumas diretrizes simples no sentido de evitar palavras ambíguas e erros de digitação no título e no resumo.

O resultado negativo de busca bibliográfica em ambas as bases indica a inexistência do termo nos respectivos vocabulários. Quando negativo em INDEXCI, pode significar que, em algumas buscas, as palavras utilizadas como descritores estão erroneamente incluídas nas listas de palavras não-significativas, ou, então, hifenadas de forma incorreta.

Os resultados demonstram que o bom desempenho da precisão e da revocação estão intrinsicamente relacionados com a seleção correta do descritor para identificar o conteúdo dos documentos.

4. CONCLUSÃO

A indexação automática aplicada aos títulos e resumos dos artigos do periódico **Ciência da Informação**, entre 1972 e 1983, identificou, de maneira equivalente à da indexação manual, os descritores que caracterizam a base de dados formada com os referidos artigos.

Os dois tipos de indexação apresentaram, basicamente, as mesmas características em seus vocabulários, quando aplicada à análise de freqüência nos descritores.

Constatou-se, ao comparar os vocabulários das bases CINFORM (indexada manualmente) e INDEXCI (indexada automaticamente), que na primeira se encontram descritores com freqüência de aparecimento superior aos de INDEXCI, o que parece contribuir para aumentar a redundância no vocabulário e o ruído na recuperação. Encontrou-se também maior quantidade de sinonímia e quase-sinonímia na base CINFORM.

O índice de coincidência entre os vocabulários das bases INDEXCI e CINFORM foi baixo, ocorrendo maior coincidência com descritores de freqüência maior do que 3, indicando que o aumento da freqüência está relacionado com a coincidência.

Demonstrou-se, nesta pesquisa, que os descritores de facetas (média freqüência) são os mais significativos para a identificação do conteúdo dos documentos. Estes, combinados com os descritores pontuais (baixa freqüência), possibilitam uma recuperação mais precisa. Os descritores de escopo (alta freqüência) permitem uma indexação macrocategorizada. Quando se consideram as raízes dos termos e as palavras truncadas, diminui-se a dispersão entre os vocabulários, aumentando a coincidência e concentração dos termos significativos. Isto indica que o uso das raízes e palavras truncadas também é aconselhável nas buscas, quando se desejam melhores índices de precisão e revocação.

Verifica-se que a análise de freqüência das palavras, em língua portuguesa, contribui para os estudos semânticos de vocabulários formados por indexação automática ou manual. Através do número de ocorrências do termo identifica-se o seu nível de especificidade e o seu valor para a recuperação. O uso deste filtro de qualidade, a freqüência, pode contribuir para os estudos terminológicos de linguagem de indexação e elaboração de tesauros.

A qualidade de recuperação apresentada na base de dados INDEXCI, medida em termos de precisão, apresentou índice superior ao de CINFORM.

Os termos livres e simples do arquivo INDEXCI apresentaram melhor flexibilidade para a formulação da estratégia de busca. Isto significa que a indexação automática, utilizando linguagem livre e pós-coordenação dos descritores, no momento da recuperação, ofereceu melhor resultado de recuperação.

O emprego da indexação automática parece aconselhável, face ao grande número de documentos existentes, tornando cada dia mais difícil realizar a indexação manual com um mínimo de qualidade requerida para assegurar o acesso posterior à informação. Vários estudos estrangeiros (1,4,7,9) já revelaram a sua validade e, neste trabalho, confirma-se a eficiência da indexação automática, quando aplicada aos títulos e resumos em língua portuguesa.

As técnicas de indexação automática se prendem em maior ou menor grau às características dos programas. Para se obterem melhores resultados, faz-se mister seguir algumas regras, específicas de cada sistema, como, por exemplo, na prepa-

ração dos registros e na conceituação das tabelas de palavras e raízes não-significativas.

Abstract

Comparative analysis of automatic and manual indexing of the Brazilian literature on information science.

A comparative study of the quality of manual and automatic indexing was carried out, in the area of information science, on a population of documents written in Portuguese, integrated by the papers published in the Brazilian journal **Ciência da Informação.**

In the case of automatic indexing, the titles and the abstracts of the papers were considered, and in that of manual indexing the full text of the papers.

The evaluation of the quality of the vocabularies was made in function of the coincidence of the terms in both vocabularies, as well as in function of the respective frequencies of occurrence of the terms.

The quality of the on-line information retrieval, in the data bases integrated by the records of the same papers, indexed either manually or automatically, was established by using the precision measurements.

Automatic indexing leads to a lower redundancy in the vocabulary and a higher precision in the retrieval process, especially when roots or truncated descriptors are used.

REFERÊNCIAS

1. BLOOMFIELD, M. Evaluation of indexing. The simulated machine indexing experiments. **Special Libraries, 61**(9): 507-7, Nov. 1970.
2. BROOKS, B. C. Bradford's law and bibliography of Science. **Nature, 224:**5 3-6, 1969.
3. BURROUGHS CORPORATION. **B 6700/B 7700 Command and edit (CANDE) language:** information manual. 1972.
4. KWOK, K.L. Cited titles: a new source of keyword extraction for automatic classification and retrieval. In: ASIS ANNUAL MEETING, 37. Atlanta. **Proceedings.** Washington, ASIS, 1974. V. 11 p. 56-7.
5. ROBREDO, J. **BIB/BATCH, manual de registro bibliográfico.** Brasília. UnB. Departamento de Biblioteconomia, 1981.
6. 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.
7. SALTON, G. Automated language processing. **Annual Review of Information Science and Technology, 3:** 169-99, 1968.
8. SALTON, G. A new comparison between conventional indexing an automatic text precessing. **Journal of the American Society for Information Science, 23** (2): 75-84, Mar/Apr. 1972.
9. SVENONIUS, E. An experiment is index term frequency. **Journal of the American Society for Information Science, 23** (2): 109-21, Mar/Apr. 1972.
10. UNESCO. **UNISIST guide to standards for information handeling.** Paris, 1980, 304 p.