CONCEITUAÇÃO DE UM PROGRAMA PARA INDEXAÇÃO AUTOMÁTICA DE TEXTOS

Jaime ROBREDO, Ex-Diretor do Projeto PNUD/FAO/BRA/72/020 (Sistema Nacional de Informação e Documentação Agrícola — SNIDA), Professor Titular do Departamento de Biblioteconomia, Universidade de Brasília, Brasília, DF.

José Adalberto de Paula FERREIRA, Analista - Programador, Biblioteca Nacional d Agricultura BINAGRI, Brasília-DF.

Apresenta a conceituação de um programa para a indexação automática de textos, tendo como principal característica a apresentação dos descritores de forma normalizada.

1. INTRODUÇÃO

As bases conceituais de qualquer programa de indexação automática são amplamente conhecidas, remontando, provavelmente, a primeira aplicação prática a geração de índices do tipo KWIC *(Key-word-in-context index)* de acordo com as idéias de LUHN (1). (2). Dentre a volumosa e as vezes contraditória literatura publicada sobre o assunto, alguns artigos e certos trabalhos de revisão contribuíram particularmente a firmar o interesse pelos métodos de indexação automática (3), (12), (14), (15), (18).

O princípio geral de um programa de indexação automática encontra-se esquematizado na Figura 1, extraída de uma publicação recente (19) de um dos autores desta comunicação, que tem participado anteriormente no desenvolvimento de um sistema que inclui rotinas de indexação e tradução automáticas (13), (16), (17).

Apresenta-se, de maneira sucinta, nesta comunicação a conceituação geral de um programa para indexação automática de textos, sendo sua principal característica a apresentação dos descritores em forma normalizada, o que diminui grandemente a dispersão das formas dos descritores representativos de um mesmo conceito e, consequentemente, aumenta a confiabilidade do processo de recuperação da informação.

2. CONCEITUAÇÃO DO PROGRAMA

O programo aqui apresentado denominou-se AUTOMINDEX/II, para deixar patente que se trata de um peerfeiçuqmedto do programa AUTOMINDEX, anteriormente definido, de acordo com os princípios básicos esquematizados na Figura 1, e correntemente utilizado pela Biblioteca Nacional de Agricultura (BINAGRI) (*).

Na versão inicial do programa, utiliza-se, de acordo com princípio bem conhecido, uma tabela de palavras vazias, que assegura a exclusão das palavras desprovidas de significado. Consideram-se como separadores de palavras diversos signos e caracteres previa-

(*) O programa AUTOMINDEX, além de permitir a indexação dos títulos incluídos na referencias de diversas bibliografias recentemente publicadas por este órgão, serviu para realizar a indexação por assuntos de mais de nove mil títulos de projetos de pesquisa agrícola em andamento, atualmente incorporados à base de dados explorada pelo Sistema BRACARIS (Brasilian Current Agricultural Research Information System) (20), (21), implantado e gerenciado pela BINAGRI.

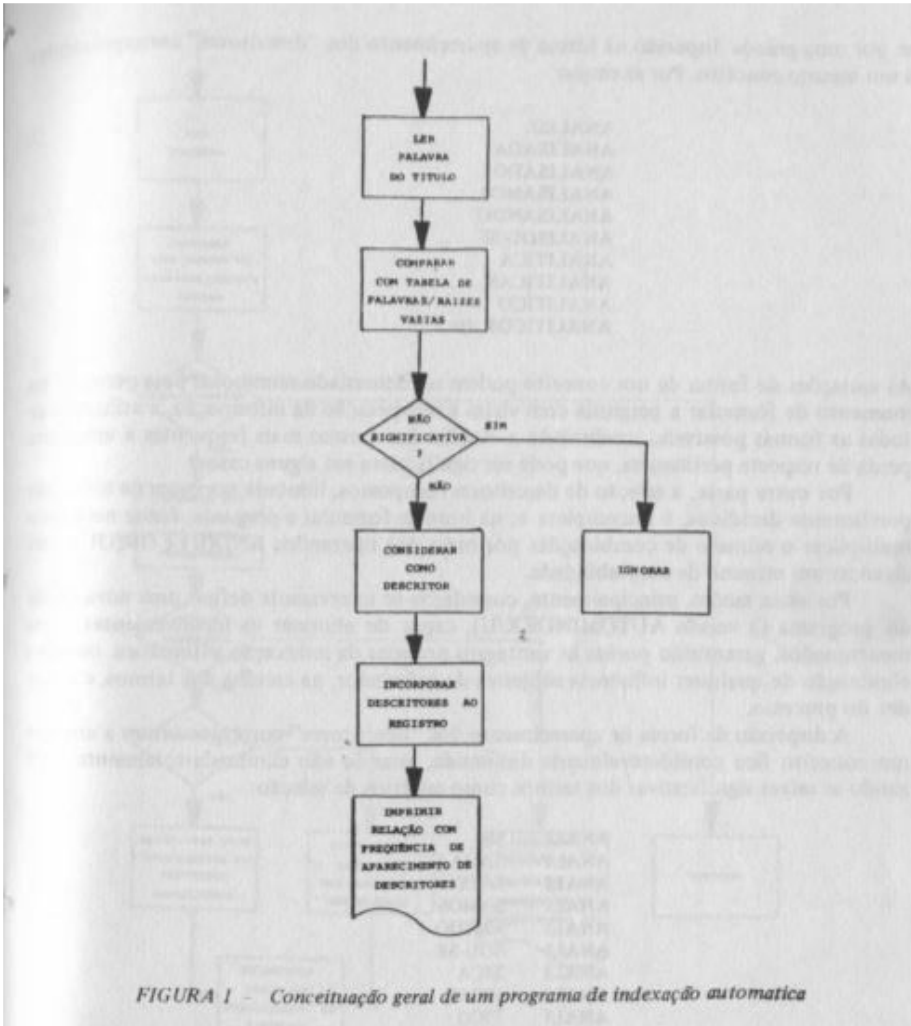

FIGURA 1 – Conceituação geral de um programa de indexação automatica

mente identificados (espaços, sinais ortográficos e de pontuação, etc.). Um mínimo de editoração dos títulos permite melhorar o nível da indexação. Assim, hifenando palavras que constituem um bloco de significado indivisível (por exemplo: nomes geográficos (São Paulo, Rio-Grande-do-Sul, Rio-São-Francisco), nomes científicos (Allium-Sativum. Ananas-Comosus, Hevea-Brasiliensis, etc.), obtém-se as expressões hífenadas como descritores. O programa gera uma tabela de freqüências de aparecimento dos descritores encontrados, em cada arquivo, o que representa uma notável ajuda no momento da escolha dos descritores na sua forma mais provável (ou mais frequentemente encontrada), quando se formula posteriormente a pergunta com vistas a recuperação da informação. De qualquer maneira, os resultados obtidos na indexação, com programas deste tipo, caracterizam-

se por uma grande dispersão na forma de aparecimento dos "descritores" correspondentes a um mesmo conceito. Por exemplo:

ANALISE
ANALISADA
ANALISADO
ANALISAMOS
ANALISANDO
ANALISOU-SE
ANALÍTICA
ANALÍTICAS
ANALÍTICO
ANALÍTICOS, etc.

As variações de forma de um conceito podem ser demasiado numerosas para permitir, no momento de formular a pergunta com vistas a recuperação da informação, a utilização de todas as formas possíveis, conduzindo a escolha das formas mais frequentes a uma certa perda de resposta pertinentes, que pode ser significativa em alguns casos.

Por outra parte, a seleção de descritores compostos, limitada aos casos de hifenação previamente decididos, é incompleta e, na hora de formular a pergunta, faz-se necessário multiplicar o número de combinações por meio dos operandos AND(E) e OR(OU), para alcançar um mínimo de confiabilidade.

Por estas razões, principalmente, considerou-se interessante definir uma nova versão do programa (a versão AUTOMINDEX/II), capaz de eliminar os inconvenientes acima mencionados, garantindo porém as vantagens próprias da indexação automática, ou seja a eliminação de qualquer influência subjetiva do indexador, na escolha dos termos, e a rapidez do processo.

A dispersão da forma de aparecimento dos "descritores" correspondentes a um mesmo conceito fica consideravelmente diminuída, quando não eliminada totalmente, utilizando as raízes significativas dos termos como critérios de seleção:

ANALI	**SE**
ANALI	**SADA**
ANALI	**SADO**
ANALI	**SAMOS**
ANALI	**SANDO**
ANALI	**SOU-SE**
ANALI	**TICA**
ANALI	**TICAS**
ANALI	**TICO**
ANALI	**TICOS, etc.**

Esta opção implica na preparação de uma tabelas que permite identificar as raízes consideradas significativas e, a partir destas, imprimir o descritor em forma normalizada (por exemplo: ANÁLISE), qualquer que seja a forma em que aparece no texto.

Na nova versão do programa, os termos que não figuram nem na tabela de palavras vazias, nem na tabela de raízes significativas, são gravados no registro indexado "candidatos a descritores", gerando-se, para cada arquivo indexado, uma listagem de descritores e de candidatos com suas respectivas freqüências de aparecimento, o que facilita consideravelmente a decisão de considerar o candidato como termo não significativo (incluindo-o na tabela de palavras vazias) ou como novo descritor (após inclusão da nova raíz

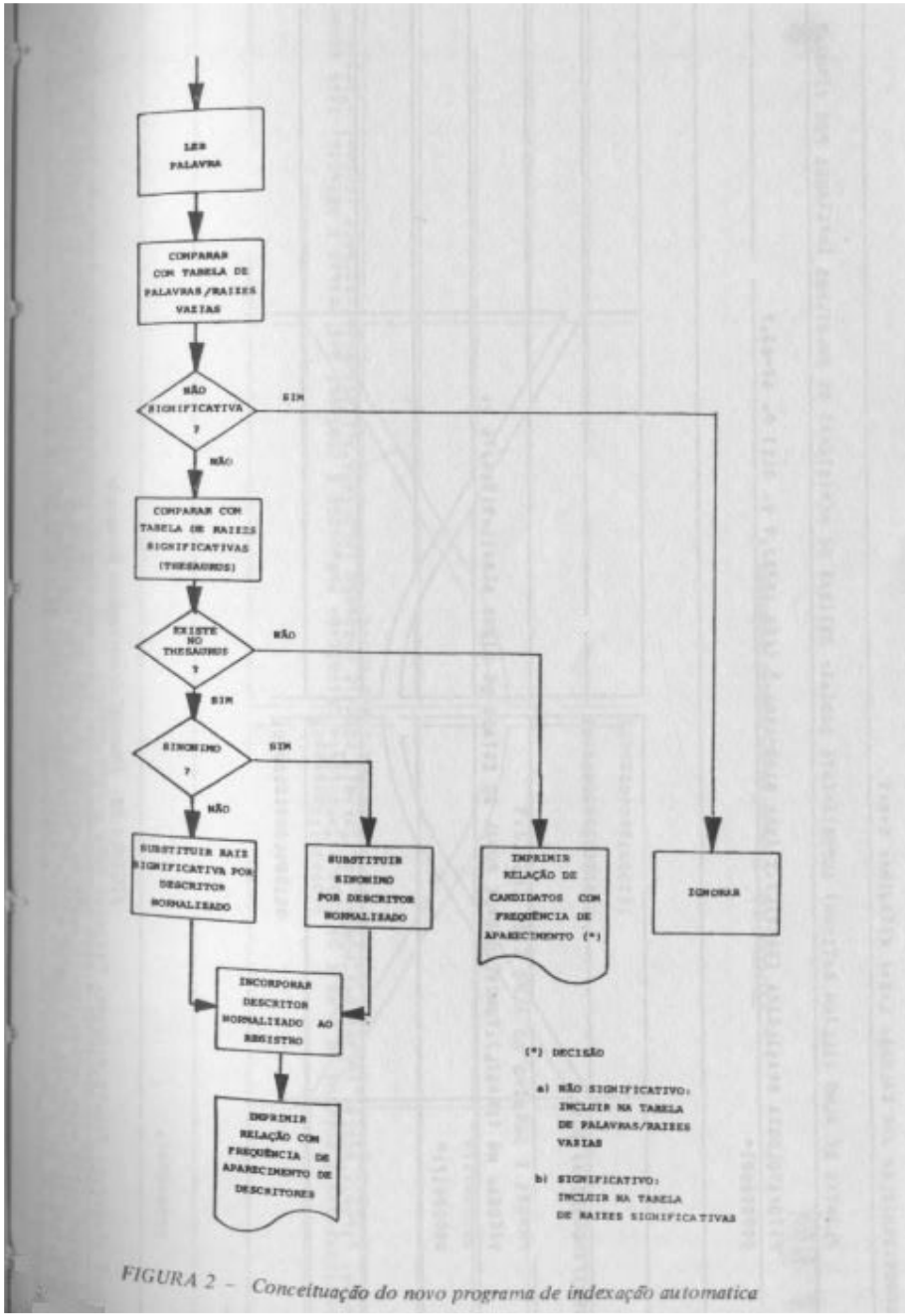

FIGURA 2 – Conceituação do novo programa de indexação automatica

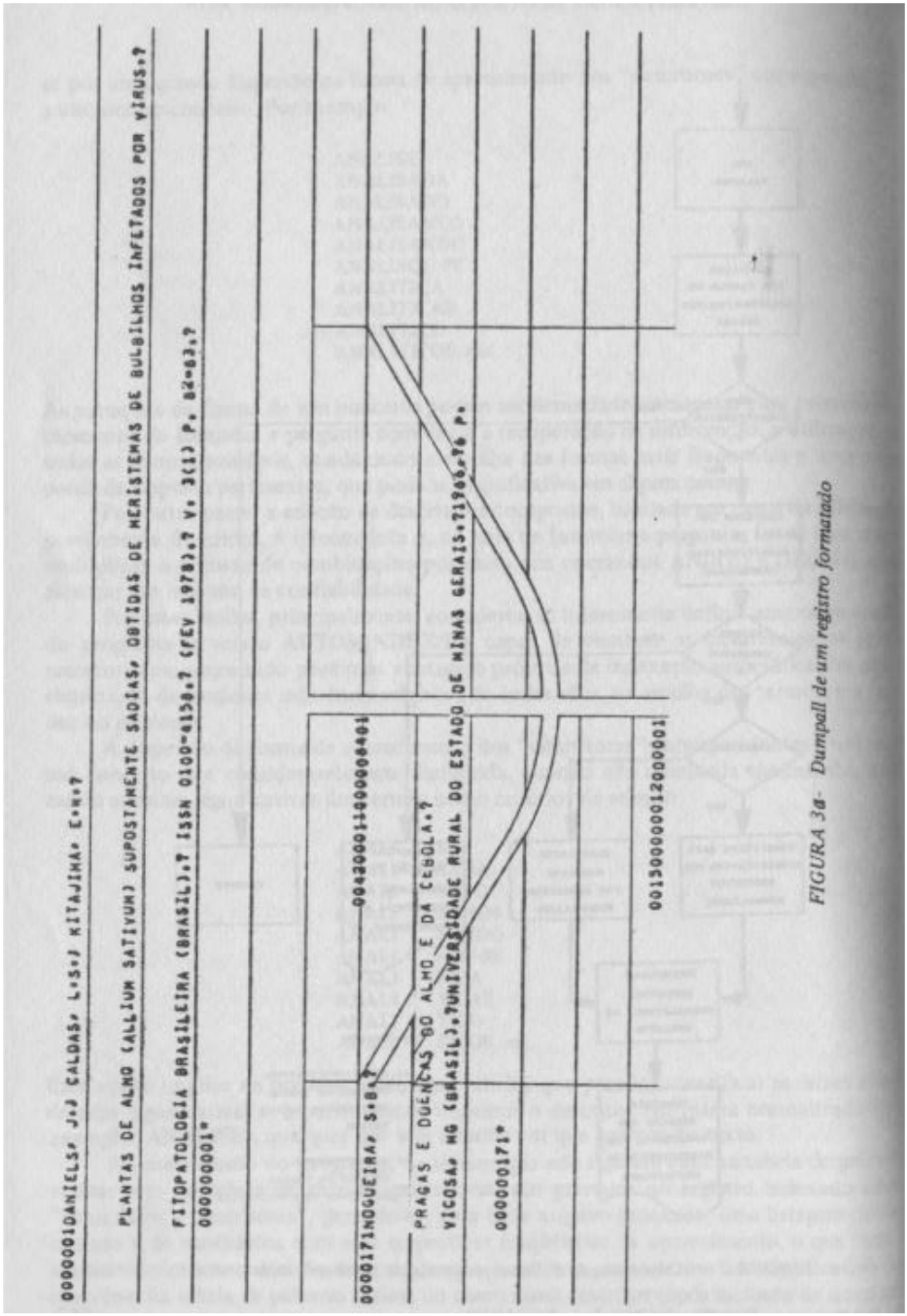

FIGURA 3a - Dumpall de um registro formatado

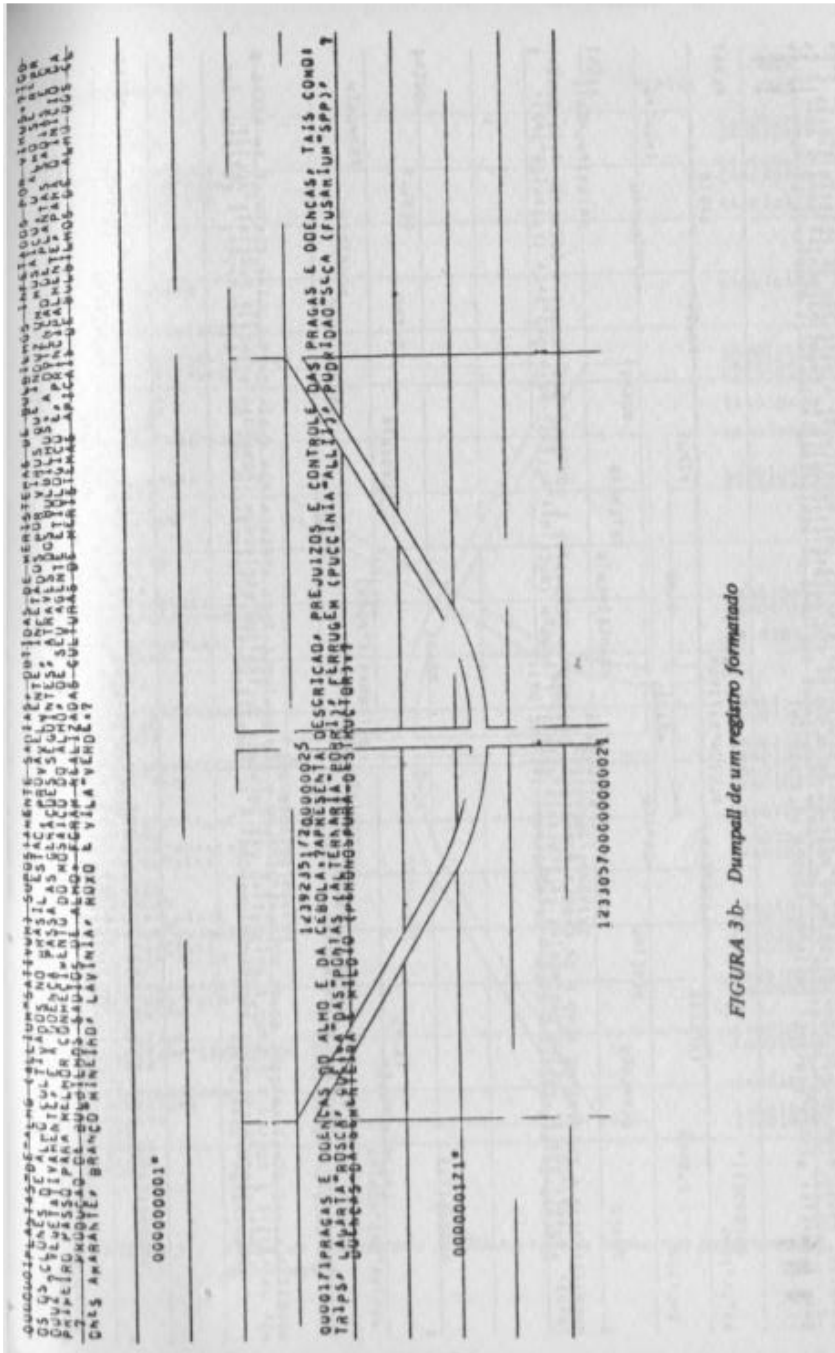

FIGURA 3b- Dumpall de um registro formatado

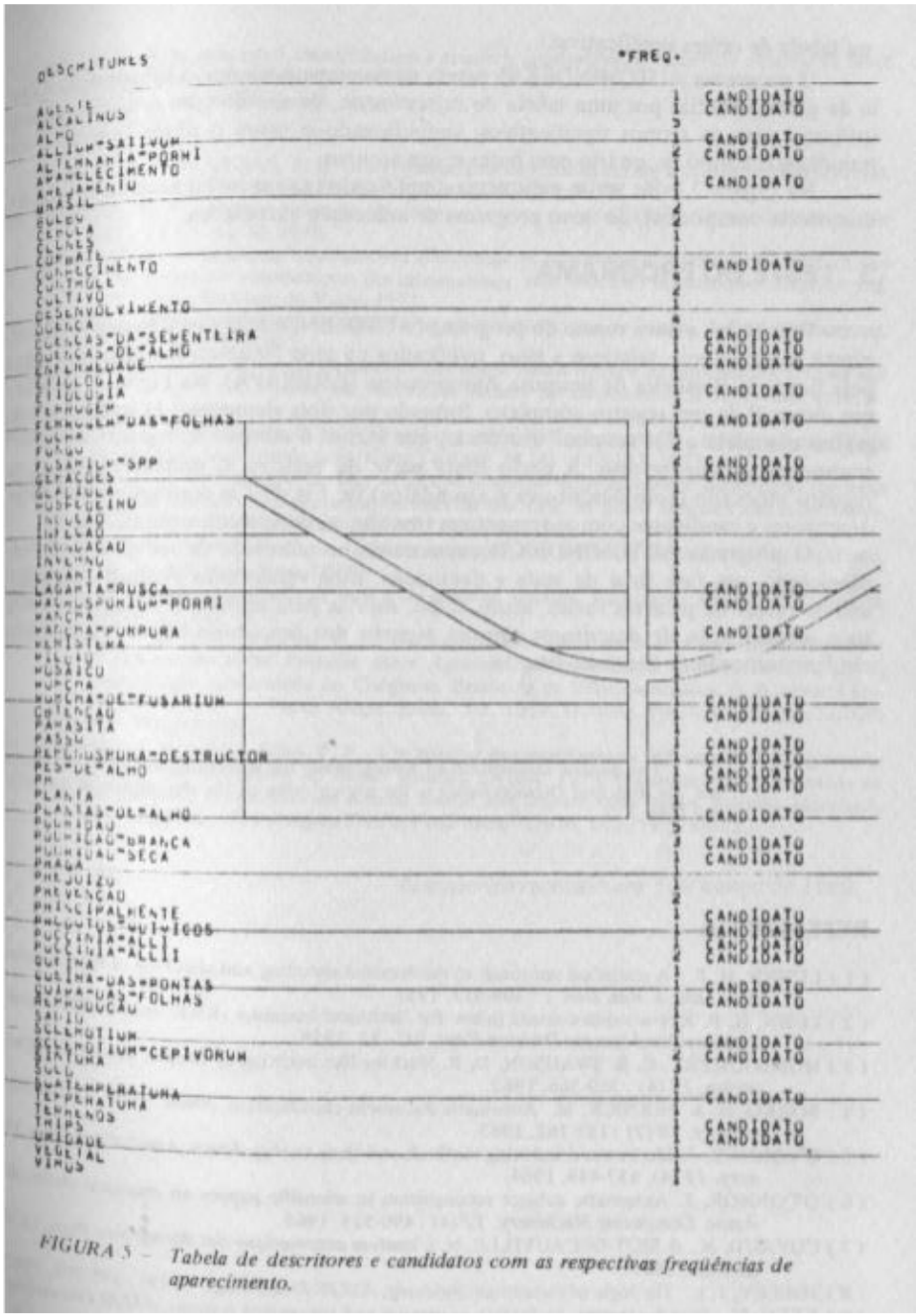

FIGURA 5 – *Tabela de descritores e candidatos com as respectivas frequências de aparecimento.*

na tabela de raízes significativas).

O programa AUTOMINDEX/II prevê, numa etapa posterior, a substituição da tabela de palavras vazias por uma tabela de raízes vazias, de acordo com o mesmo princípio utilizado para os termos significativos, simplificando-se assim o processamento e diminuindo-se o tempo necessário para indexar um arquivo.

Na Figura 2, pode ver-se o esquema simplificado (não se inclui a rotina de seleção de descritores compostos), do novo programa de indexação automática.

3. TESTE DO PROGRAMA

Para testar a nova versão do programa (AUTOMINDEX/II), escolheram-se alguns resumos informativos, relativos a alho, publicados na série *Resumos Informativos,* editada pela Empresa Brasileira de pesquisa Agropecuária (EMBRAPA). Na Figura 3 pode-se ver um *dumpall* de um registro completo, formado por dois elementos: 1) referência bibliográfica completa e 2) "resumo" a indexar, que inclui: o número de registro, o título, e o resumo propriamente dito. A partir desta parte do registro, o programa gera um novo registro indexado (com descritores e candidatos) (v. Fig 4) e as correspondentes listas de descritores e candidatos com as respectivas treqüências de aparecimento (v. Fig. 5).

O programa AUTOMINDEX/II encontra-se, no momento de redigir a presente comunicação, em fase final de teste e depuração. Uma versão mais avançada, incluindo o uso de raízes de palavras vazias, assim como. rotinas para seleção de descritores composto e identificação de descritores simples a partir dos descritores hifenados, encontra-se em fase avançada de desenvolvimento.

The general conception of a programme for automatic text indexing is presented. Its principal characteristics is the presentation of the descriptors in a standardized form.

REFERÊNCIAS

(1) LUNHN, H. P. A statistical approach to mechanized encoding and searching of literary information. *IBM. J. Res. Dev.* 1 : 309-317, 1957.

(2) LUHN, H. P. Key-word-in-context index for technical literature (KWIC index), *IBM Advanced Systems Development Division Rept.* RC - 12, 1959.

(3) MONTGOMERY, C. & SWANSON, D. R. Machine-like indexing by people. *American Documentacion, 13* (4) : 359-366, 1962.

(4) BORKO, H. & BERNICK, M. Automatic document classification. *Journ. Assoc. Computing Machinery. 10*(2) : 151-162, 1963.

(5) O'CONNOR, J. Mechanized indexing methods and their testing. *Journ. Assoc. Computing Machinery. 11* (4), 437-449, 1964.

(6) O'CONNOR, J. Automatic subject recognition in scientific papers: an empirical study. *Journ. Assoc. Computing Machinery. 12*(4) : 490-515, 1965.

(7) COYAUD, M. & SIOT-DECAUVILLE. N. *L'analyse automatique des documents.* Paris, La Haye, Mouton. 1967.

(8) JOLLEY, J. L. The logic of coordinate indexing. *ASL1B Proceedings. 19* (9) : 295-308, 1967.

(9) KEEN, M. Search strategy evaluation in manual and automated systems. *ASLIB Proceedings. 20* (1) : 65-87. 1968.

(10) DOYLE. L. B. Is automatic classiffication a research application of statistical analysis of text? *Journ. Assoc. Computing Machinery. 16* (4): 264-284, 1959.
(11) MARON, M. E. *& KUHNS, J. L. On relev*
(11) MARON, M. E. &
(11) MARON, M. E. & KUHNS, J. L. On relevance, probabilistic indexing and information retrieval. *Journ. Assoc. Computing Machinery. 1*7 (3): 216-244, 1970.
(12) SIMMONS. R. F. Natural language question-answering systems: 1969. *Comunications of the ACM. 13* (1) *:* 15-30. 1970.
(13) ROBREDO, J. Experiences comparatives d'indexage et possibilité d'indexage mécanisé en vue d'une recherche automatique des informations, sans barrières lingüistiques. *Rapport Inf.* 71/10, Paris, Institute du Verre, 1971.
(14) STIBIC, V. T. Het automatisch indexeren van documenten: pro en contra. *Informatie Jaargang.* 14 (11) : 516-522, 1972.
(15) WILLIAMS, M. E. Use of machine-readable data bases. *Anual Review of Information Science and Technology.* C. A. Cuadra, ed. Americam Society for Information Science, New York, *9 :* 221-284, 1974.
(16) ROBREDO, J. & BRISNER, O. An international computerised system for information retrieval in the glass and ceramic field. *Glass Technol. 14* (4) : 112-117, 1973.
(17) BRISNER, O. & BRUDAL, P. J. The ALLIANCE system with automatic indexing as the basis for international documentation services in the field of glass, ceramics and refractories. *Norsk Senter for Informatikk Rpt.* N° 7504, Oslo, 1975.
(18) BORKO, H. & BERNIER, C. L. *Indexing concepts and methods.* New York, São Francisco, London. Academic Press. 1978.
(19) ROBREDO, J. *Documentação de hoje e de amanhã.* Brasília, Associação dos Bibliotecários do Distrito Federal, 1978.
(20) ROBREDO, J. & CURVO Filho, P. F. *O Projeto BRACARIS como base do Sistema Brasileiro de Informação sobre Pesquisa sobre Agrícola em Andamento.* Brasília, SNIDA, 1977. Comunicação apresentada ao Congresso Brasileiro de Biblioteconomia, 9. & Jornada Sul- Rio-GRandense, 5., Porto Alegre, julho, 3-8, 1977. (Projeto PNUD/FAO/BRA/72/020. DOC/TEC/77/036.)
(21) ROBREDO, J. & CURVO Filho, P. F. *Um sistema automatizado de informação sobre pesquisa agrícola em andamento no Brasil.* Brasília, SNIDA, 1978). Comunicação apresentada ao Internacional Symposium on Animal Health and Disease Data Banks, Belsville, Maryland, dezembro, 4-6, 1978 (Projeto PNUD/FAO/BRA/72/020. Doc./Tec/78/033.)

Manuscrito recebido em 3 de março de 1980.