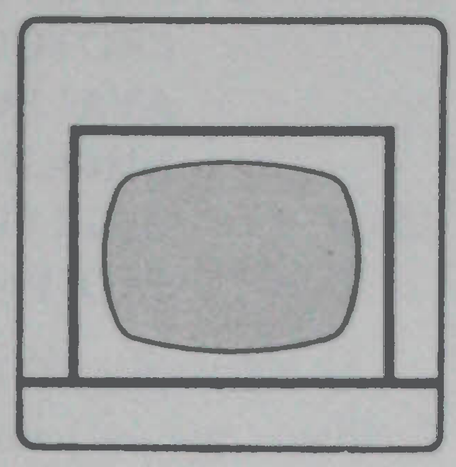

LITEREL, um sistema expedito e potente de indexação: As frequências literais relativas

J. Nunes da Costa (*)

resumo

O autor propõe um método de indexação baseado nas diferentes frequências literais relativas dos vocábulos, a que chama LITEREL, com recurso ao uso da informática.

abstract

The Author proposes a new method for indexation, based on the random different relative frequencies of the letters in the words, using the computer possibilites.

1. A teoria

As frequências com que as letras de um alfabeto surgem no vocabulário de uma dada língua não são iguais, como se sabe.

Tal facto é mesmo usado em questões de descriptagem, mediante sistemas mais ou menos complexos, mas não é este o aspecto que aqui queremos abordar. O que se irá tentar explanar é que o mesmo facto pode ser aproveitado para fins de indexação e pesquisa documentária, e é um método já bem antigo [1, 2, 3].

É certo que entretanto outros métodos foram desenvolvidos para a mesma finalidade: referimo-nos aos «Thesauri» e aos Dicionários. Porém, a elaboração dum «Thesaurus» é demorada, laboriosa, complexa, cara, e, não só não existe um «Thesaurus» Universal dos conhecimentos, como, para certas especialidades, os não há apropriados. Finalmente, o treino para as respectivas aplicações é laborioso, e, consequentemente, caro.

Por todas estas razões se entende que o método proposto, sendo simples, fiável, expedito, e de fácil aplicação, é, em muitos casos, suficiente e susceptível de larga utilidade.

Um exemplo (adiante aprofundado e pormenorizado) permitirá melhor elucidação:

um LIVRO
de José BARROS MOURA

(*) J. Nunes da Costa, Eng. Elect. (UP), Electricidade de Portugal, E.P.

c/ o título: REDES LOCAIS de COMPUTADORES;
PROTOCOLOS de alto nível e
AVALIAÇÃO do respectivo DESEMPENHO
Editor: McGRAW-HILL
SÃO PAULO, BRASIL

As ideias principais são assinaladas em maiúsculas. É óbvio que o leitor pode acrescentar outras que decorram da leitura da obra.

Ordenemos as ideias principais acima referenciadas em quatro colunas (que a prática mostrou serem as bastantes). Então obter-se-á o quadro 1.

QUADRO 1

1	2	3	4	
L	I	V	R	o
B	A	R	R	os
M	O	U	R	a
R	E	D	E	s
C	O	M	P	utadores
P	R	O	T	ocolos
A	V	A	L	iação
D	D	S	D	mpenho
M	C	G	R	aw-Hill
S	Ã	O		
P	A	U	L	o
B	R	A	S	il

Se, de uma maneira qualquer, se dispuser de quatro campos formados por um alfabeto por coluna, bastará então codificar, em cada campo alfabético, todas as letras da respectiva coluna, independentemente do número de vezes que a mesma letra apareça.

Eis, em termos simplistas, a «Teoria».

Vejamos agora como proceder para pesquisar uma informação e suponhamos que os dados fornecidos (mais incompletos que os codificados) eram os seguintes:

um Livro	B R A S ileiro
de um Tal Autor	M O U R a
com assuntos e ou	R E D E s
ideias	C O M P utadores
	A V A L iação
	F U N C ionamento

A última ideia, obviamente, não está codificada.

O exemplo foi propositadamente escolhido para mostrar que, ainda assim, o inconveniente não é relevante, dada a quantidade de informação já existente.

Para proceder à pesquisa seleccionar-se-iam (por meios manuais ou informáticos, consoante os casos) as letras nas respectivas colunas.

Um cálculo simples dar nos-á ideia da potência da selecção (Quadro 2).

QUADRO 2

	1	2	3	4
Letras diferentes codificadas	10	7	10	6
Letras diferentes seleccionadas	5	4	4	5
Proporção	50 %	57,1 %	40 %	83,3 %
Factor de ocorrência	9,5 %			

Isto significa que a probabilidade de aparecer o documento em causa, e só ele (mesmo com a quantidade restrita de dados que foi fornecida para a pesquisa), é de 90,5 %.

Uma consulta manual complementar, simples e rápida, permitirá seleccioná-lo.

Se, em contrapartida, quisermos pesquisar bibliografia sobre temas, o procedimento é análogo:

- alinham-se os temas em colunas;
- seleccionam-se por letras/colunas.

Não é possível apresentar aqui um cálculo, pois não há, obviamente, dados concretos. Porém a probabilidade é dada agora pelo somatório dos factores de ocorrência (tal como o calculado acima).

2. Elaboração

Quando o sistema surgiu, o seu suporte era constituído por Fichas Bibliográficas (75 × 125 mm) de extracção manual, como mostra a figura 1.

A comparação com o quadro anterior é tão clara e evidente que dispensa explicações supletivas.

É evidente que numa ficha manual (ainda por cima com as dimensões daquela) a quantidade de registos (=ideias principais) possível é muito restrita e nisso residia a limitação do sistema.

Com a respectiva informatização tal não sucederá.

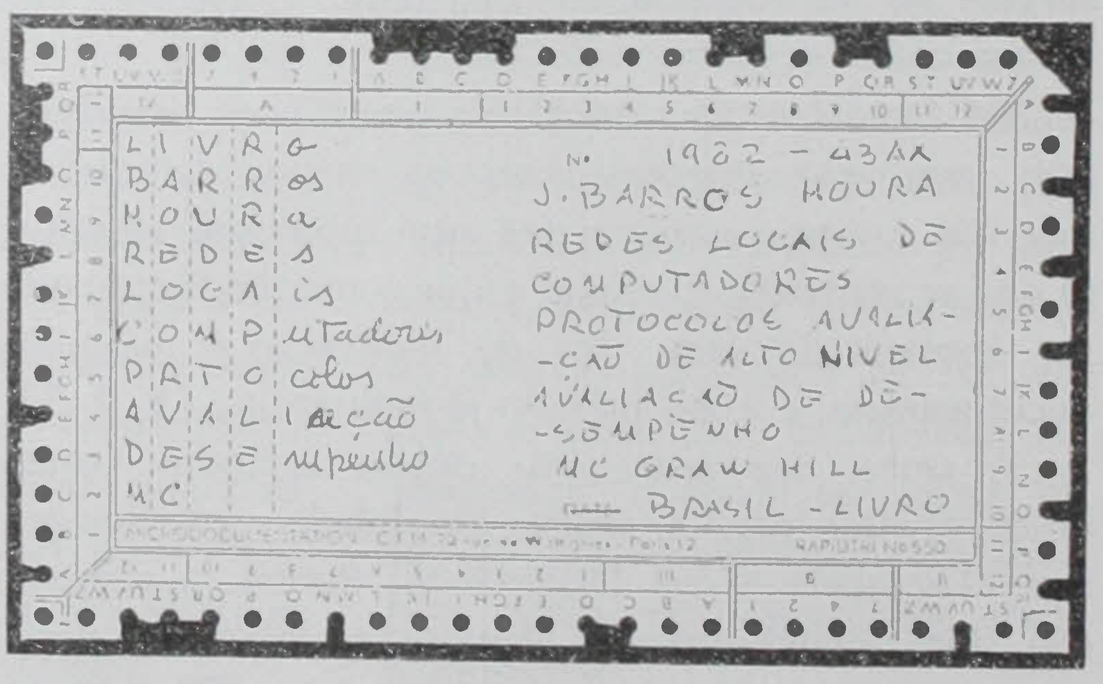

Fig. 1 — Exemplo de ficha bibliográfica de extracção manual

É agora chegado o ensejo de se falar da concepção do sistema, isto é, da base de dados que o irá gerir. Limitar-nos-emos apenas ao respectivo esquema lógico.

Um esquema lógico descreve a organização dos dados (deve ser tão inalterável quanto possível), não deve ser taxativo na maneira de ser utilizado e deve ser tão próximo quanto possível das exigências do utilizador.

Daqui se infere que ele depende da classe de documentos que vão constituir o banco de dados.

É porém possível ordenar as tarefas da sua elaboração em três grandes áreas, a saber:

A — Agrupamento dos Documentos, por afinidades entre si;
B — Ordenação, Arrumação e Localização respectiva;
C — Escolha dos Dados que interessará indexar.

A primeira área é prioritária. Como existem sempre franjas de agrupamento indeciso, este trabalho tem de ser tão exaustivo quanto possível e absolutamente exclusivo, sob pena de se cair, na prática, na indefinição dos casos de «fronteira».

A segunda área é clara. Bastará sublinhar (o que nunca é redundante) que indexar não é catalogar, arrumar, localizar, mas exige, isso sim, que estas tarefas tenham sido escrupulosamente cumpridas.

Quanto à terceira área, nela reside a maior das vantagens do sistema, pois dada a sua flexibilidade não é necessário que os dados sejam escolhidos sempre da mesma forma. Claro que há vantagem em seguir sempre uma mesma metodologia. Porém ela pode ser diferente de um grupo de documentos para outro.

Como exemplos, vamos abordar aqui 2 casos:

- um, corrente, é o da indexação de informações contidas em livros, revistas, e outros suportes similares;
- outro, menos vulgar, é o da indexação da informação contida na denominada Documentação Submersa (Directivas, notas de serviço, notas, telexes, etc.) a qual, normalmente, nunca é referenciada.

Tomemos o primeiro caso, isto é, o da construção do esquema lógico no caso de livros, revistas, e similares.

A. ÁREA AGRUPAMENTO

A.1 — Classe — Documentos Seriados

Tipos:
A.1.1 — Revistas
A.1.2 — Jornais
A.1.3 — Sumários
A.1.4 — Normas
A.1.5 — Catálogos Seriados
A.1.6 — Séries

A.2 — Classe — Documentos Não Seriados

Tipos:
A.2.1 — Livros
A.2.2 — Separatas
A.2.3 — Monografias, Teses
A.2.4 — Relatórios
A.2.5 — Catálogos não Seriados

B. ÁREA ORDENAÇÃO, etc.

B.1 — Número de Ordem
B.2 — Localização interna e/ou externa à entidade indexadora
B.3 — Arrumação

C. ÁREA ESCOLHA dos DADOS

C.1 — Designação (apenas se se trata dum seriado)
C.2 — Título (para ambos os casos)
C.3 — Autor(es)
C.4 — Língua
C.5 — Tipo do Documento
C.6 — Editor
C.7 — Ideias-Chave

O banco de dados será seguidamente construído, documento a documento, de acordo com estes critérios.

Abordemos agora o segundo exemplo, o da Documentação Submersa. Entende-se como tal aquela que:

a) normalmente não é objecto de arquivo e/ou classificação (por exemplo, notas manuscritas);
b) quando o é, é-o duma forma imperfeita (por exemplo, telexes);
c) fica englobada noutra de carácter mais importante (por exemplo, despachos exarados sobre documentos).

Todos temos tido, na prática, contactos com este tipo de informação: notas manuscritas, despachos exarados, telexes, folhetos, minutas iniciais, xerocópias, etc., etc.

Trata-se, das mais das vezes, duma fonte pessoal muito rica, directa e insubstituível, de acontecimentos que, ou se perdem, ou ficam na cabeça do interveniente, e que, «à la longue», se diluirão no esquecimento.

Nestes casos não há propriamente lugar para falar em Classes, mas apenas em Tipos. Eis uma sugestão aplicável a uma estrutura empresarial:

A. ÁREA AGRUPAMENTO

Tipos:
A. 1 — Apontamentos (do próprio)
A. 2 — Notas (de terceiros)
A. 3 — Comunicações (do próprio ou de terceiros)
A. 4 — Minutas
A. 5 — Exarados (noutros documentos)
A. 6 — Informações formais
A. 7 — Estudos
A. 8 — Telexes e suas Cópias
A. 9 — Fotos (referenciados pelas matérias nelas existentes)
A.10 — Xerocópias
A.11 — Folhetos

B. ÁREA ORDENAÇÃO, etc.

B.1 — Número de Ordem
B.2 — Localização interna e/ou externa à entidade indexadora
B.3 — Arrumação

C. ÁREA ESCOLHA dos DADOS

C.1 — Autor
C.2 — Destinatário
C.3 — Intervenientes
C.4 — Tipos de Documento
C.5 — Ideias-Chave

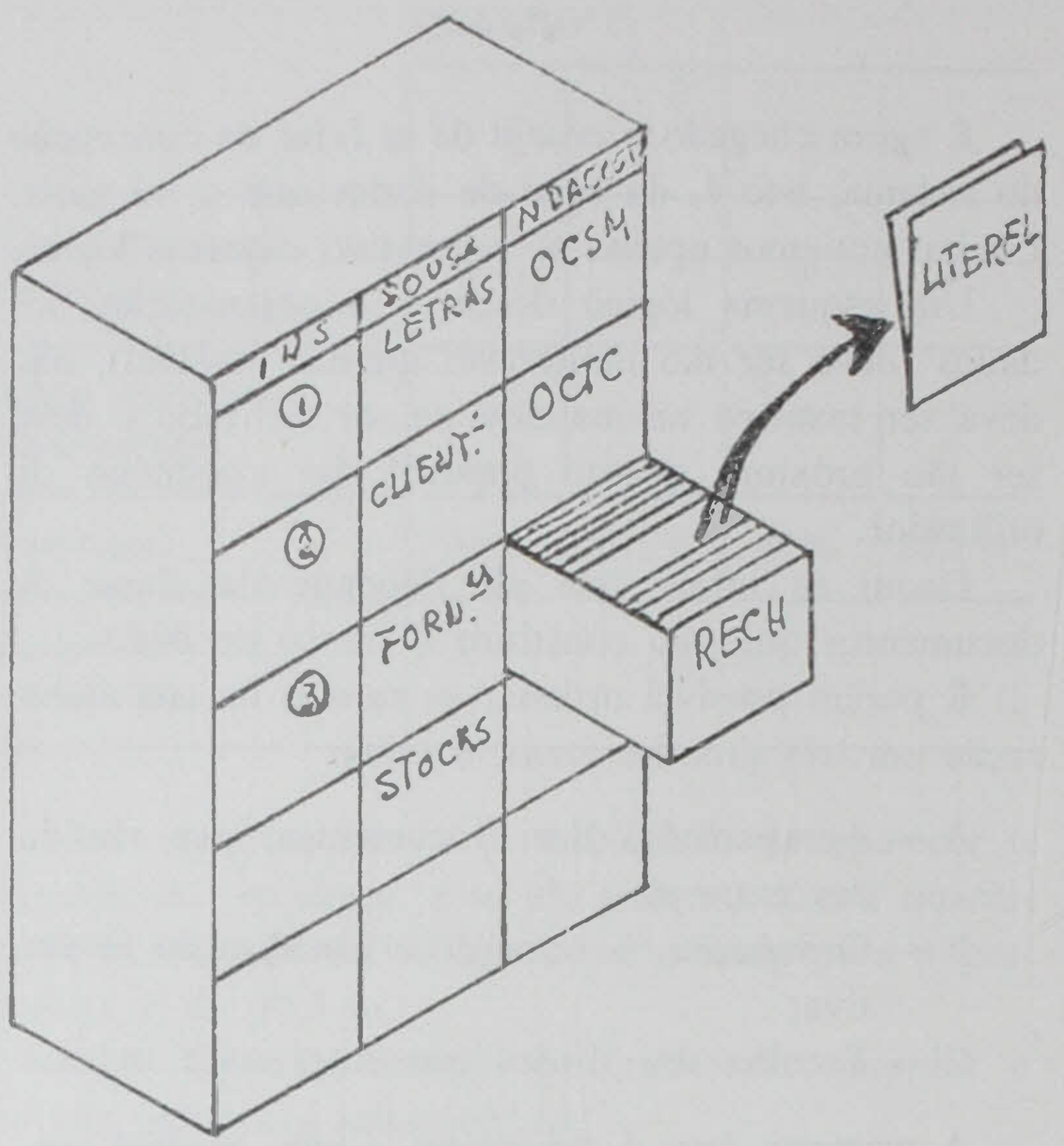

Fig. 2 — Esquema de ficheiro com «gavetões»

Os dois exemplos acabados de citar mostram, de facto, como a Base de Dados depende do tipo de classes dos documentos, mas também mostram como é possível, mesmo assim, estabelecer, usando uma mesma metodologia, construções lógicas flexíveis e diferenciadas.

Muitas outras utilizações poderão ser concebidas, coisa que deixaremos ao cuidado dos eventuais leitores interessados.

SUM.DE COLUNAS da TAB. LITEREL 16 Mar 88 17:04
Num. de colunas 110

Acao	Ã	Nome	Tipo	tam.	chave	Titulo
	1	GRTP	A	3		GRUPO E TIPO DO DOCUMENTO
	2	NORD	N	7		NUM. ORDEM,EXCLUSIVO DO DOC.
	3	LCLZ	A	30		LOCAL. DO DOCUMENTO
	4	ARRM	A	20		ARRUM. DO DOCUMENTO
	5	DSGS	A	20		DESIGN. SO P/ SERIADOS
	6	TTLO	A	30		TITULO DO DOCUM.
	7	AUTR	A	30		NOME(S) DO(S) AUTOR(ES)
	8	EDTR	A	30		NOME(S) DO(S) EDITOR(ES)
	9	DATA	N	4		ANO DA EDIçãO
	10	LING	A	10		LING. DO DOCUMENTO
	11	ICH1A	A	1		I.C.1 - coluna A
	12	ICH1B	A	1		I.C.1 - coluna B
	13	ICH1C	A	1		I.C.1 - coluna C
	14	ICH1D	A	1		I.C.1 - coluna D
	15	ICH1TEXT	A	16		I.C.1 - texto restante
	16	ICH2A	A	1		I.C.2 - coluna A
	17	ICH2B	A	1		I.C.2 - coluna B
	18	ICH2C	A	1		I.C.2 - coluna C
	19	ICH2D	A	1		I.C.2 - coluna D
	20	ICH2TEXT	A	16		I.C.2 - texto restante
	106		A	24		I.C.20 - coluna A
	107		A	24		I.C.20 - coluna B
	108		A	24		I.C.20 - coluna C
	109		A	24		I.C.20 - coluna D
	110		A	24		I.C.20 - texto restante

Fig. 3 — Sumário de campos do LITEREL

3. Informatização

Dois factores básicos (pelo menos) influem na informatização deste processo, a saber:

a) Documentação que se pretende indexar;

b) Equipamento e/ou programa a serem utilizados.

O exemplo que damos em seguida referir-se-á, quanto ao primeiro factor, a uma aplicação do Tipo bibliográfico, e quanto ao segundo, ao «package» VM/AS da IBM, correndo em grandes computadores.

É evidente que outras aplicações podem ser encontradas, e, também, que é possível conceber programas correndo noutros tipos de máquinas.

O que é e como funciona o VM/AS?

Trata-se de uma base de dados relacional, de concepção sofisticada, de uso muito inteligível e que, mediante um conjunto de ordens simples, permite criar e alterar ficheiros, introduzir, alterar ou apagar dados desses ficheiros e pesquisá-los combinatoriamente, de forma muito diversificada e clara.

O AS permite ainda executar outras funções variadas e importantes mas que, não relevando ao caso, não serão aqui referidas.

Podemos assimilá-lo a uma «sala de arquivo», com diversos «ficheiros» dos quais um, «denominado» «N DACOSTA» é «pertença» do A. Está fechado à chave e é o A. que detém a «pass-word».

O ficheiro dispõe de vários «gavetões» onde o A. mantém arrumada, segundo a sua própria óptica, a sua documentação.

Num deles, denominado RECH (de RECHercher), encontra-se uma «pasta» com o título LITEREL, que versa *exclusivamente* a documentação bibliográfica de que nos vamos ocupar (fig. 2).

Quizessemos porventura aplicar o método, a outras indexações, por exemplo documentação submersa, fotografias, mapas, etc., outras «pastas» novas teriam de ser abertas, com designações diferentes e próprias.

A figura 3 mostra o «Sumário de Campos» do LITEREL, em AS, definido de acordo com a metodologia indicada no parágrafo 2.

Prevê-se indexar um máximo de 20 ideias-chave, o que a experiência mostra ser amplamente suficiente.

Examinemos a figura 4. Trata-se de uma verdadeira ficha (IMAGE em linguagem AS), de clara simplicidade, o que possibilita, por teclagem directa, a introdução dos dados oriundos do técnico indexador.

Repare-se que, em cada ideia-chave, estão bem individualizadas as quatro primeiras posições, uma vez que é pela análise combinatória das frequências relativas, que se irá proceder à pesquisa selectiva.

Quando se trata de proceder a uma pesquisa bibliográfica, haverá, antes do mais, que alinhar em colunas as ideias-chave que ocorreram ao inquiridor.

Retomando o mesmo exemplo citado no parágrafo 1, teríamos agora a disposição do quadro 3, para um número (arbitrário) de 7 ideias-chave.

Numa segunda fase escolhe-se (por intuição ou não) as letras menos frequentes em cada coluna, mostrando a prática que bastará um conjunto total de 4.

Finalmente, há que dar à máquina uma ordem de selecção, o que obriga à construção de uma pequena

```
**************************************************************************
*          ENTRADA DE DADOS PARA PESQUISA BIBLIOGRAFICA                  *
*------------------------------------------------------------------------*
*   Caracterizacao do Documento                                          *
*                                                                        *
*   Numero de Ordem  : _____          Grupo e Tipo   : ___               *
*   Localizacao      : _____________________________                     *
*   Arrumacao        : _____________________                             *
*   Designacao (so para seriados) : _____________________                *
*   Titulo           : _____________________________                     *
*   Autores          : _____________________________                     *
*   Editor           : _____________________________  Data  : ____       *
*   Linguagem        : _____________                                     *
*------------------------------------------------------------------------*
*   Descricao do Documento por Ideias-Chave                              *
*                                                                        *
*   I.C.1 : _ _ _ _ _____________________  I.C.14: _ _ _ _ _____________ *
*   I.C.2 : _ _ _ _ _____________________  I.C.15: _ _ _ _ _____________ *
*   I.C.3 : _ _ _ _ _____________________  I.C.16: _ _ _ _ _____________ *
*   I.C.4 : _ _ _ _ _____________________  I.C.17: _ _ _ _ _____________ *
*   I.C.5 : _ _ _ _ _____________________  I.C.18: _ _ _ _ _____________ *
*   I.C.6 : _ _ _ _ _____________________  I.C.19: _ _ _ _ _____________ *
*   I.C.7 : _ _ _ _ _____________________  I.C.20: _ _ _ _ _____________ *
*   I.C.8 : _ _ _ _ _____________________                                 *
*   I.C.9 : _ _ _ _ _____________________                                 *
*   I.C.10: _ _ _ _ _____________________                                 *
*   I.C.11: _ _ _ _ _____________________                                 *
*   I.C.12: _ _ _ _ _____________________                                 *
*   I.C.13: _ _ _ _ _____________________  COMANDO LOCAL :                *
**************************************************************************
 COMANDO AS :
 ...PROGRAM  :
```

Fig. 4 — Ficha para indexação até 20 ideias-chave

QUADRO 3

Ideia-chave	A	B	C	D	Texto restante
1	B	R	A	S	ileiro
2	M	O	U	R	a
3	R	E	D	E	s
4	P	R	O	T	ocolo
5	C	O	M	P	utadores
6	A	V	A	L	iação
7	M	C	G	R	aw-hill

sub-rotina do tipo que se indica sob a forma de um fluxograma na figura 5.

Damos assim por concluída esta Nota prévia, com o voto de que ela possa vir a ser útil aos estudiosos desta matéria, tão rica, tão aliciante e até mesmo tão perigosa que é a indexação documentária.

BIBLIOGRAFIA

[1] J. NUNES DA COSTA, *Cartões perfurados de extracção manual,* Electricidade, n.º 18.

[2] Diversos autores, *Punched cards,* 2 nd edition, Reinhold Publishing Co, 1959, N. York.

[3] EDF, *Note sur l'elaboration et la redaction des fiches «Mecanalyse»,* Note bleue EDF/D3/C78/Dh, Jan. 60.

[4] F. VAN SLYPE, *Systemes de Communication de l'information,* 1979, Éditions d'Organizations.

[5] IBM, *A/S Manual,* Nov. 1985.

[6] C. IAULT, *Les bases de données relationelles,* Éditions d'Organization, 1986.

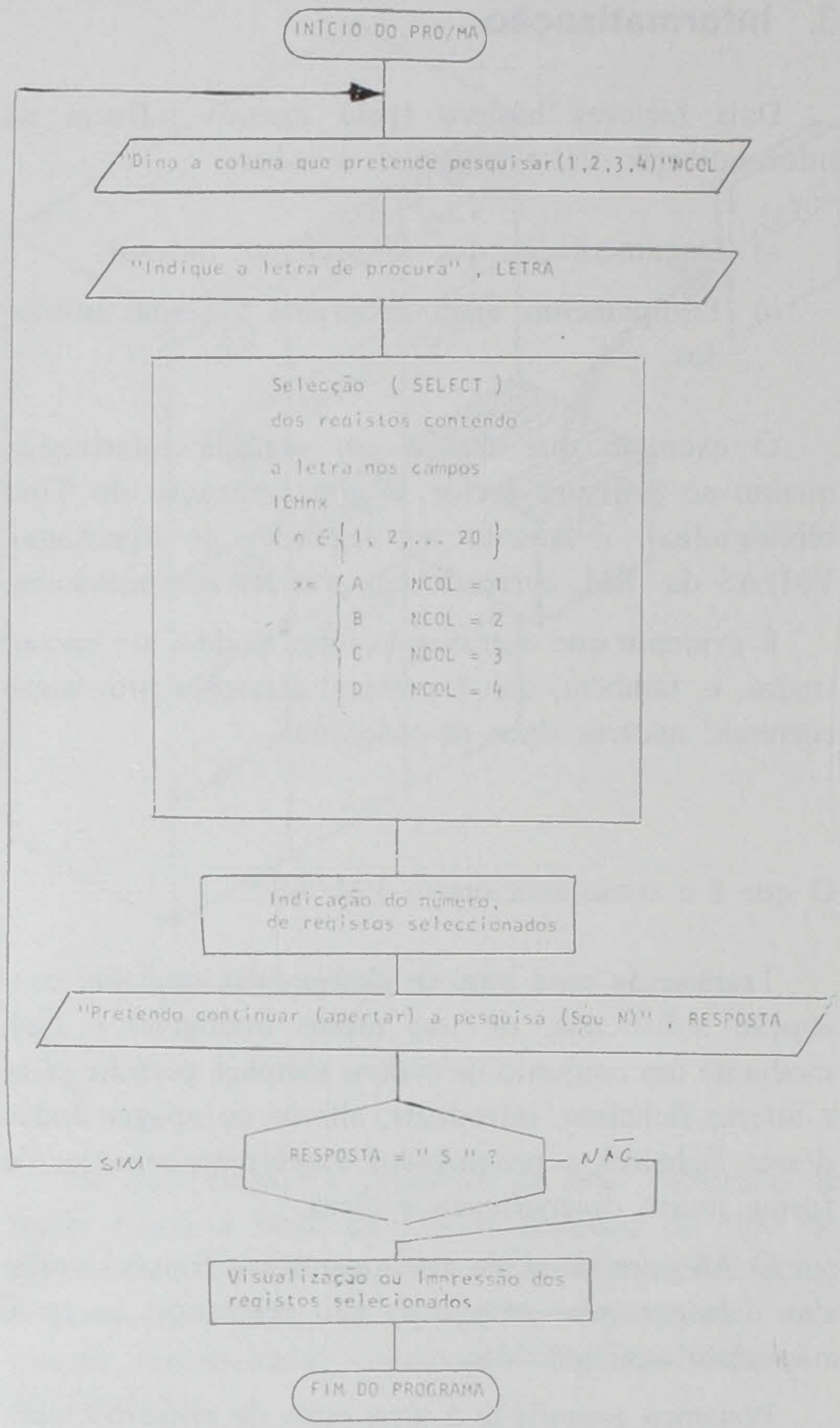

Fig. 5 — Fluxograma da sub-rotina de selecção