INFORMÁTICA: O SISTEMA DE PALAVRAS-CHAVE DO CONTEXTO (K.W.I.C.)

*Antônio Carlos Marques Mattos**

1. Introdução

Apresentamos, neste artigo, um sistema de arquivamento e de recuperação de informações *(information retrieval)* que, embora de aparecimento relativamente recente (1959), já se tem mostrado bastante superior dos demais (ver referências bibliográficas 10).

O sistema de palavras-chave do contexto *(Key Word in Context* — K.W.I.C.), como é chamado, é o que melhor se adapta a um computador. No entanto, pode ser implantado manualmente, desde que o tamanho do banco de dados não ultrapasse certos pontos, pois do contrário as limitações humanas eliminarão boa parte das vantagens que o K.W.I.C. oferece.

A grande vantagem do K.W.I.C. é que, em princípio, pode ser recuperada, em tempo hábil, *toda e qualquer informação armazenada no banco de dados.*

Como sempre ocorre, também existem desvantagens: o tamanho dos arquivos é maior que o dos sistemas convencionais, assim como é maior o tempo de busca e pesquisa. São essas desvantagens que limitam a utilização manual do K.W.I.C.

Assim, o *que se ganha em recuperação perde-se em tempo e espaço físico.* Mas, desde que o sistema seja operacionalizado por computadores, as desvantagens tornam-se desprezíveis. E nesse caso nenhum outro sistema consegue suplantá-lo.

Se o sistema for manual, o K.W.I.C. deixa de ser prático a partir de um certo número de documentos a serem arquivados, sendo mais conveniente outros sistemas. Mas, essa maior conveniência é ilusória. Na verdade, o que se faz é *reduzir a capacidade de recuperação para compensar as deficiências humanas.*

Um exemplo talvez esclareça melhor esse ponto. Suponhamos um banco de dados constituído por uma biblioteca. O sistema tradicional, o de classificação decimal universal (C.D.U.), possui baixa capacidade de recuperação, em relação ao K.W.I.C., por dois motivos, basicamente:

a) só se consegue recuperar com eficiência as informações armazenadas que dizem respeito ao nome do autor e ao título do documento (da obra). As demais informações que estão armazenadas no banco, como é o caso dos livros editados em 1960 (por exemplo), não podem ser encontradas (recuperadas) com eficiência;

b) depende de catálogos que, além de não estarem sempre atualizados, não permitem uma classificação correta em muitos casos. Tal é o caso do livro *Formal organizations*, por exemplo. Esta obra trata de vários assuntos: teoria da aprendizagem (psicologia), cadeias de Markov (teoria da probabi-

* Professor do Departamento de Métodos Quantitativos da Escola de Administração de Empresas de São Paulo, da Fundação Getulio Vargas, e Engenheiro Eletricista (E.P.U.S.P.).

lidade), Organogramas (administração geral), "confiabilidade dos sistemas" (engenharia de sistemas), e vários outros. Como não é possível, pelo C.D.U., ordenar essa obra em todos esses assuntos, classifica-se em apenas um ou dois (por exemplo, administração geral), reduzindo-se desta forma o livro a cerca de 30% do seu conteúdo original. *Isto é perda de informação na classificação.*

O sistema K.W.I.C., como veremos, não apresenta esses inconvenientes.

2. Palavras-chave

Todo o sistema K.W.I.C. está baseado no conceito de palavras-chave e no emprego de estruturas de lista, esse último a ser exposto no item 3 deste trabalho.

Pode-se definir palavra-chave como sendo todo vocábulo ou conjunto de vocábulos de um documento que possua *conteúdo informático*.

Quando constituída de um único vocábulo, a palavra-chave deverá ser um substantivo ou um verbo. Assim, as palavras "de", "também", "ela", "pequeno", etc., não possuem conteúdo informático, agindo apenas como elementos de ligação e de qualificação.

Em caso de ser constituída de um conjunto de vocábulos, a palavra-chave, para possuir conteúdo informático, deve expressar, *auto-suficientemente*, uma idéia completa. Como exemplos temos:
Palavra-chave
recuperação de informação
cadeias de Markov
classificação decimal universal
Pedro Álvares Cabral

Como podemos notar, o "índice analítico" que aparece nas últimas páginas de um livro é um conjunto de palavras-chave desse livro.

Um conjunto ordenado de palavras-chave denomina-se THAESAURUS, que nada mais é do que um "dicionário de idéias afins". No dicionário comum temos estruturado um sistema no qual, a partir de um vocábulo qualquer, obtemos o significado semântico desse vocábulo. Já no caso de um *thaesaurus*, a situação é exatamente a inversa. Possuímos uma determinada idéia ou significado e queremos encontrar o vocábulo que expresse essa idéia. O *thaesaurus* permite encontrar esse vocábulo. Sua utilidade é grande, pois permite uniformizar a terminologia empregada, ou seja, evitar o emprego de vários sinônimos para expressar a mesma idéia. Isso aumenta a eficiência de recuperação do sistema K.W.I.C.

Na fase de classificação dos documentos, que veremos a seguir, o *thaesaurus* será utilizado sempre que tivermos mais de um vocábulo expressando, igualmente, um dos conteúdos de um documento. Em outras palavras, devemos dar preferência, na classificação, aos termos que já constam do *thaesaurus*, nada impedindo, no entanto, que novas palavras-chave sejam introduzidas.

3. Montagem dos arquivos

O banco de dados será constituído de:

a) quatro arquivos característicos; com

b) uma sistemática de manipulação desses arquivos (estruturas de lista).

O conjunto *a* e *b* constituirá o sistema de recuperação de informação K.W.I.C.

3.1 SISTEMA MANUAL

Para fins didáticos e, em alguns casos, práticos, o sistema manual é o mais conveniente, pois é o de mais fácil visualização.

O sistema é constituído dos seguintes arquivos:

I) arquivo dos documentos originais;

II) arquivo dos resumos dos documentos;

III) arquivo de classificação dos documentos ou classificador;

IV) arquivo de palavras-chave ou indexador.

O processo de arquivamento inicia-se com a localização física dos documentos originais, cada um referido por meio de um número-código (arquivo I). Esse arquivo deve ser tal que, dado o código de um documento, seja possível determinar, fácil e inequivocamente, a sua localização física.

A segunda etapa do processo, que, juntamente com a terceira, são as mais importantes, consiste na elaboração do resumo do documento em fase de classificação. Este resumo será um dos componentes do arquivo II. Por ocasião da consulta ao sistema, esse *abstract* irá facilitar bastante a seleção dos documentos relevantes, pois economizará o tempo de busca e consulta aos originais.

Esta etapa, como também a terceira, deverá ser realizada por técnicos de alto nível e conhecedores do assunto tratado pelo documento. Somente assim a classificação poderá ser correta, pois está envolvido aí um processo de *síntese*, onde a elaboração mental é bastante complexa.

O resultado da segunda etapa traduz-se pela emissão da ficha de resumo, com o aspecto seguinte:

Figura 1 — Ficha 1

Código do documento	0013
Identificação do documento	Mattos, Antônio Carlos M., Revista X, O sistema de palavras-chave do contexto (K. W. I. C.), São Paulo, 1971, págs., 27-40
Resumo do documento	O artigo apresenta o sistema de recuperação de informação baseado em palavras-chave (K. W. I. C.). Descreve o sistema manual e mecanizado, detalhando as técnicas de montagem dos arquivos e a sistemática de utilização. Contém descrições de sistemas em funcionamento e várias aplicações.

A terceira etapa do processo de montagem dos arquivos consiste na determinação de um conjunto de palavras-chave que expressem, sinteticamente, todo o conteúdo do documento. Essa síntese é crucial, pois dela dependerá a boa eficiência de recuperação do sistema. As palavras-chave podem ser obtidas, com vantagem, do resumo já feito. Obtemos, assim, o documento classificado, expresso formalmente por meio de uma ficha da seguinte forma exemplificada:

Figura 2 — Ficha 2

Código do documento	0013
Palavras-chave que sintetizam o documento 0013.	Mattos, ACM. Editora X 1971 PALAVRAS-CHAVE K. W. I. C. Recuperação de Informações Informática

Uma vez concluídos os trabalhos de classificação dos documentos, disporemos de um conjunto de fichas do tipo 1 e um outro do tipo 2. Para facilitar a descrição a seguir, convencionaremos chamar de KW1, KW2, . . . as várias palavras-chave utilizadas, e de CD1, CD2, . . . os códigos dos documentos classificados.

Isto posto, admitamos dado o seguinte conjunto de fichas do tipo 2:

A figura 4 mostra os arquivos já montados, com a estrutura de lista indicada para a palavra-chave KW1.

O arquivo IV é constituído pelas palavras-chave KW1 até KW6, arranjadas de forma *ordenada* (por exemplo, em ordem alfabética). A cada palavra-chave correspondem dois indicadores ou ponteiros. Por exemplo, à KW1 correspondem os ponteiros CD4 e CD1. O ponteiro à direita (CD1) indica, no arquivo de classificação dos documentos (arquivo III), qual a primeira obra a ter KW1 referenciada, como indicado pela flecha.

O arquivo de classificação dos documentos (arquivo III) contém, pela ordem dos códigos dos documentos, as fichas obtidas na terceira etapa de amostragem já descrita, com as informações adicionais que se seguem. A ficha KW1 do arquivo IV contém o ponteiro CD1. No arquivo III, a ficha CD1 contém KW1. Nesse arquivo, a *próxima obra a conter KW1 é a CD2*. Por esse motivo, o ponteiro à direita de KW1 em CD1 é CD2.

Esse processo é repetido em todo o arquivo III, até chegar a última obra que contém KW1 como palavra-chave. Atingido esse "fim-de-linha" da lista, lançamos na ficha KW1 do arquivo IV, à esquerda, a obra correspondente ao fim da linha, CD4, no caso. Com isso, fechamos o ciclo da lista. Evidentemente, a cada obra que entre no arquivo III e que contenha KW1, corresponderá uma alteração no ponteiro à esquerda de KW1, sendo o novo valor igual ao código da obra entrante.

O arquivo III é constituído dos documentos classificados, onde cada ficha possui dois ponteiros para cada palavra-chave. O ponteiro à direita, como já explicamos, corresponde ao próximo documento que tem a mesma palavra-chave referenciada. O ponteiro à esquerda de cada palavra-chave indica o documento *anterior* que tem a mesma palavra-

Figura 3

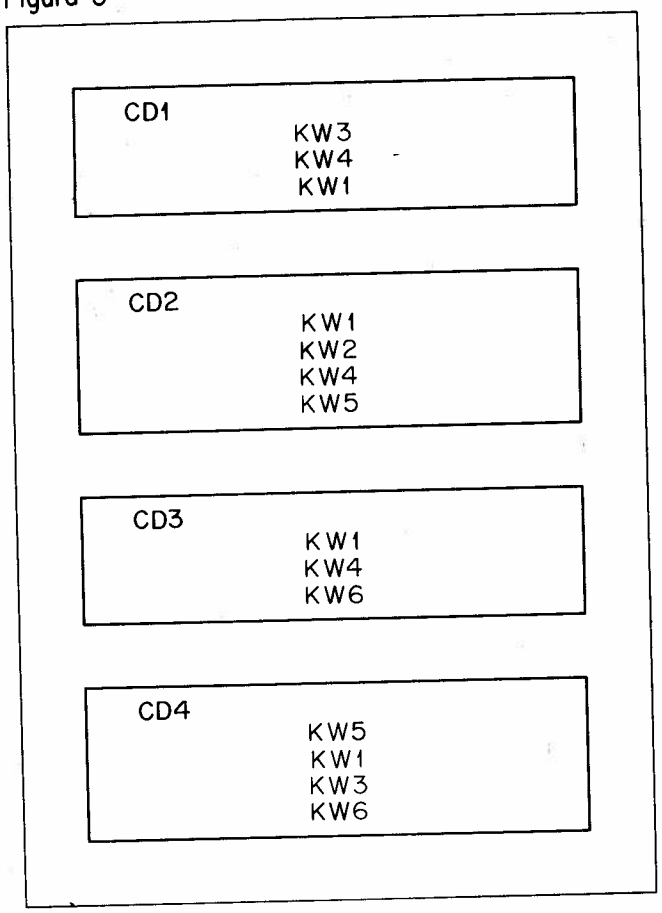
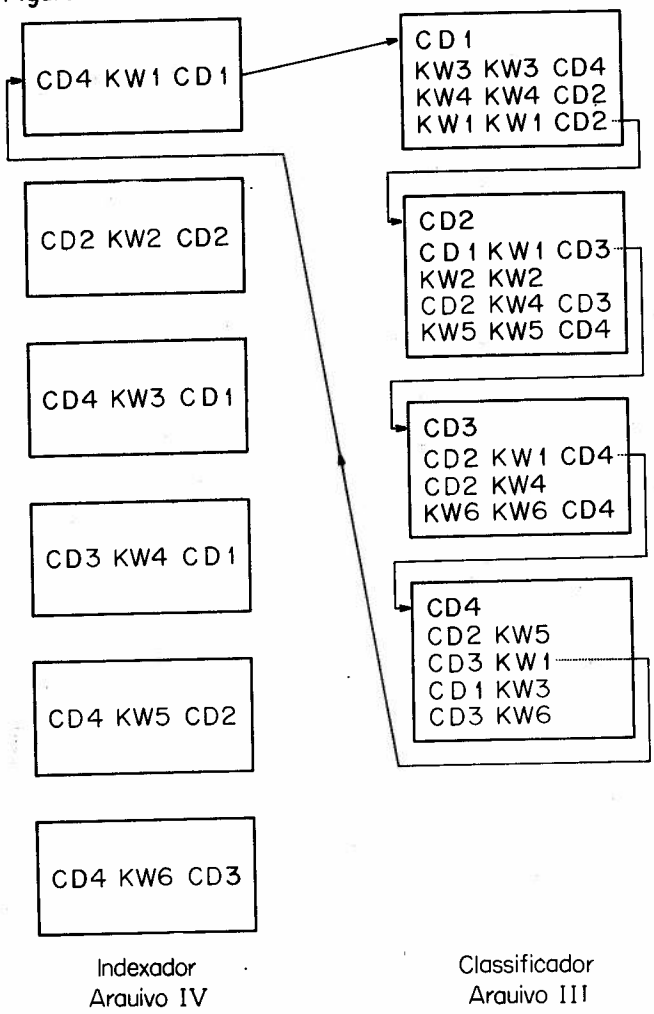


Figura 4



chave referenciada. No caso de KW1, esse fato está mostrado pelas flechas.

Cada novo documento que entra no arquivo III, impõe os lançamentos correspondentes: a) à esquerda das palavras-chave do arquivo IV por ele indexadas, e b) à direita das palavras-chave correspondentes do arquivo III, que estavam indicadas no arquivo IV.

Como exemplo, admitamos que um novo documento, CD5, seja incorporado ao sistema. Seja esse documento classificado como mostra a ficha a seguir:

O sistema sofre, agora, as alterações seguintes:

a) Arquivo I:
é introduzido o documento original CD5;

b) Arquivo II:
é introduzido o resumo de CD5;

c) Arquivo III:
é introduzida a ficha de classificação de CD5, e são lançados os ponteiros à direita de KW2 na ficha CD2, de KW4 em CD3, e de KW5, KW3 e KW6 em CD4;

d) Arquivo IV:
é introduzida a ficha de KW7 e os ponteiros da esquerda de KW2, KW3, KW5 e KW6 são alterados para CD5.

Os arquivos III e IV ficarão como segue, onde está mostrado, por meio de um gráfico, a estrutura de lista para KW2, *duplamente orientada*.

Figura 5

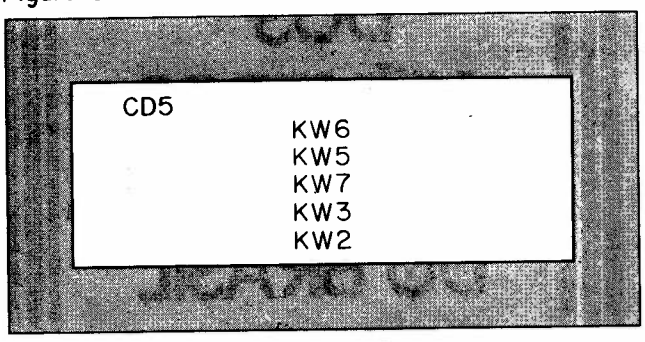
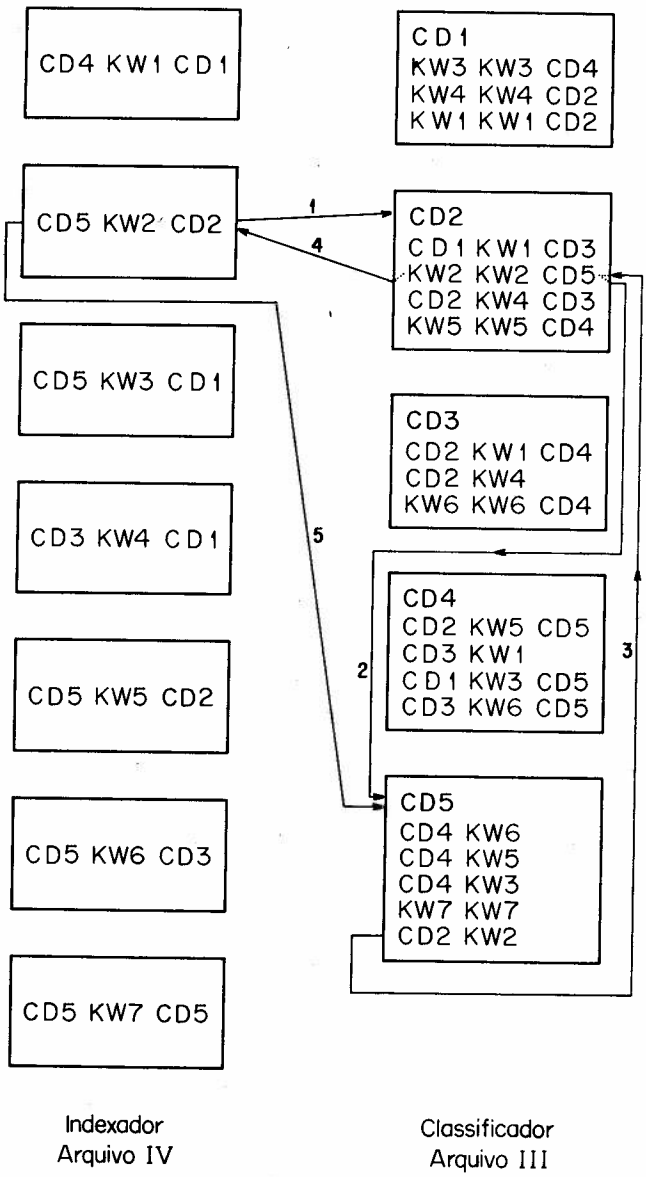


Figura 6



3.2 SISTEMA DE ARQUIVAMENTO MECANIZADO

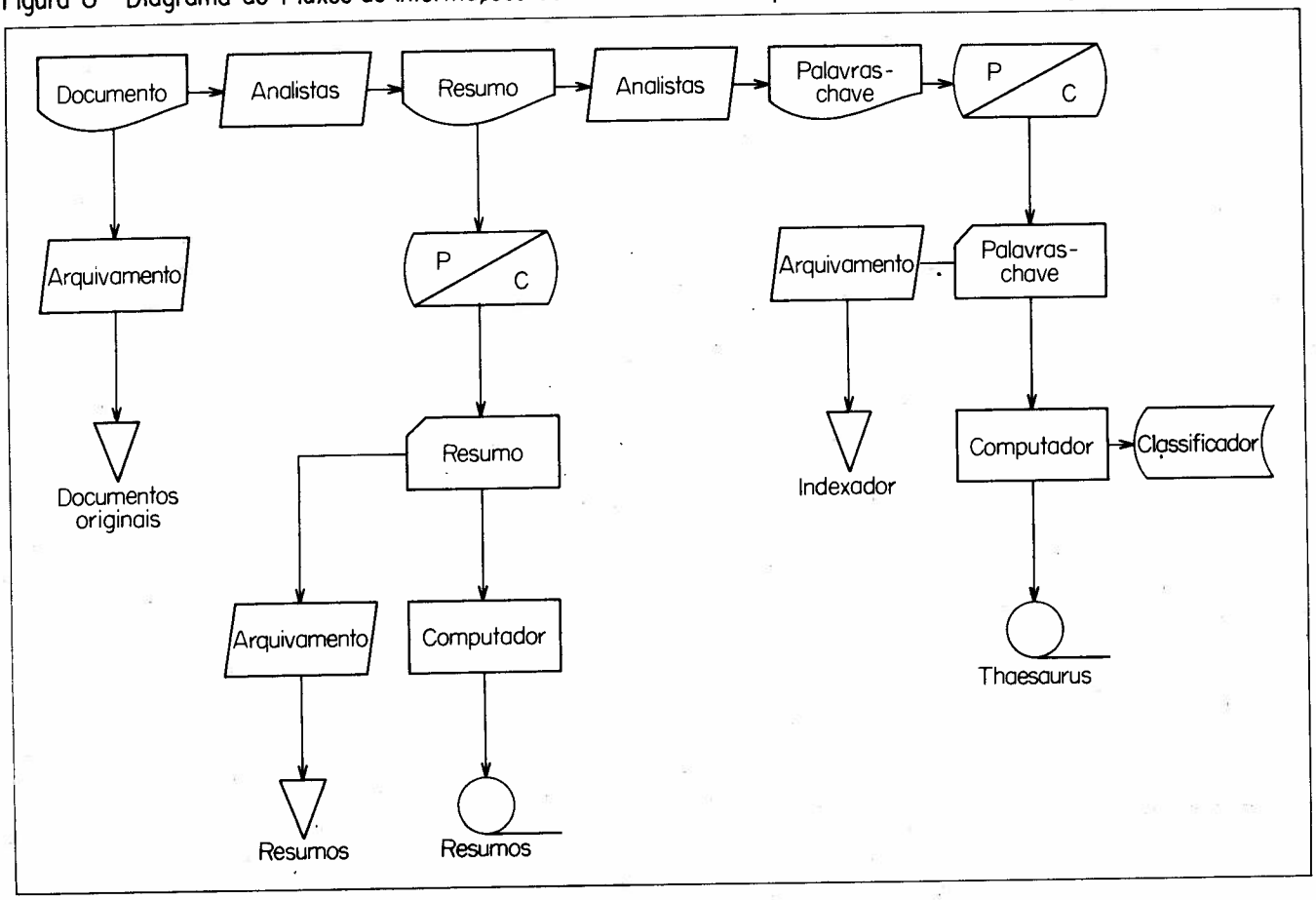
No caso de se ter um sistema mecanizado de arquivamento, esse deverá ser estruturado como mostrado no diagrama de fluxos da figura 8.

A convenção de símbolos utilizados neste artigo é descrita na figura 7 (ver referências bibliográficas 15).

Figura 7 — Símbolos para fluxogramas

NOME	SÍMBOLO	UTILIZAÇÃO
Processamento		Qualquer função principal de processamento
Cartão perfurado		Todas as variedades de cartão perfurado
Documento		Documentos e relatórios de todas as variedades
Fita magnética		Quando utilizada em arquivos em linha
Memória fora de linha		Arquivos fora de linha em fichas, cartões, fitas, discos, etc.
Teclado em linha		Informação recebida de, ou fornecida a um computador, via dispositivo c/teclado
Operação manual		Qualquer processamento manual, sem intervenção de dispositivos eletromecânicos
Entrada/Saída		Qualquer tipo de documento ou dados de entrada ou de saída
Acesso aleatório		Arquivos em linha e discos, tambores, etc.
Comunicação à distância		Transmissão automática de dados, por *time-sharing*, teleprocessamento, etc.
Fluxo de informações		Sentido do percurso das informações, dado pela flecha
Decisão		A direção a seguir depende do critério de decisão especificado no símbolo
Operação de teclado		Uma operação realizada em um dispositivo fora de linha, com teclado

Figura 8 – Diagrama de Fluxos de Informações de um sistema de arquivamento mecanizado dirigido para o K.W.I.C.



Como observamos, o computador, nesse caso, é programado de modo a montar os arquivos III e IV a partir da ficha de classificação, apenas. O resumo é gravado diretamente na fita, sem necessidade de processamentos intermediários.

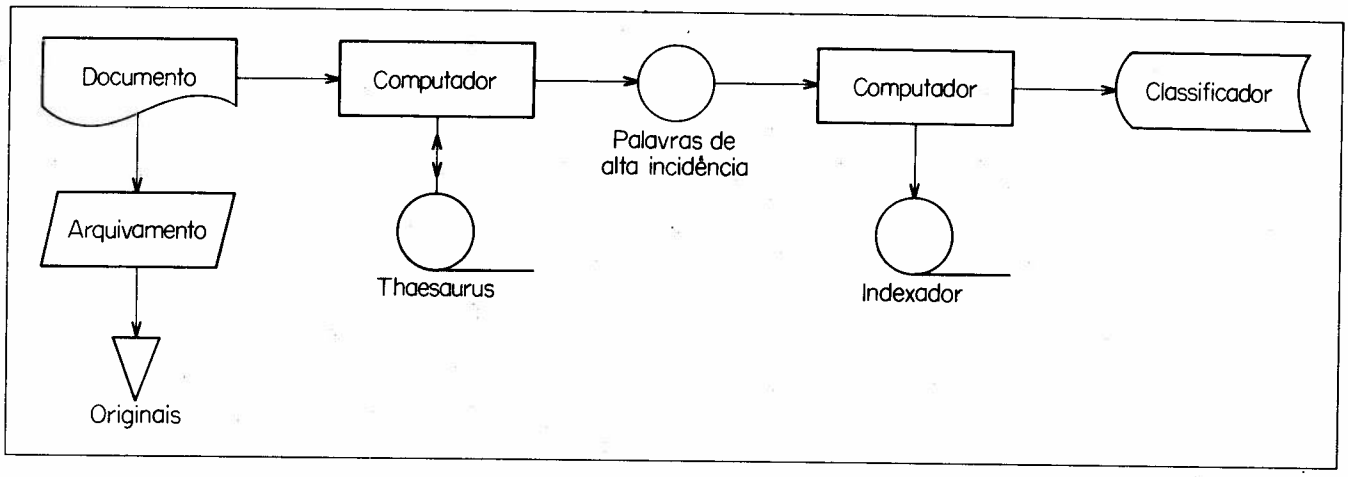
A fase manual do sistema consiste no arquivamento, em armários, dos documentos originais, e na elaboração dos resumos e das fichas de classificação. Essa última fase é, como já dissemos, a mais importante de todo o sistema, dela dependendo seu bom funcionamento.

Existe um outro sistema de arquivamento ainda mais automatizado que o já descrito. Embora nesse caso o processamento manual (e intelectual) seja praticamente eliminado, temos nossas dúvidas quanto à sua maior eficiência. A substituição do trabalho de síntese humano pela máquina é algo ainda muito sujeito a críticas, sendo poucos os casos bem sucedidos (traduções). Seja como for, passaremos a descrevê-lo.

O sistema esquematizado a seguir, na figura 9, não gera resumos, mas somente pesquisa; no documento original, há incidência das várias palavras com conteúdo informático. Findo o processo, são relacionadas as palavras que possuem uma freqüência relativa de incidência superior a um certo limite (por exemplo, 10%). Essa é a sistemática básica, embora possa sofrer certas sofisticações.

As palavras assim encontradas vão constituir a ficha de classificação do documento. O restante do processo é igual ao do sistema anteriormente examinado.

Figura 9 – Sistema de Arquivamento Automático (K.W.I.C.)



4. O sistema de recuperação de informações

Uma vez montados os arquivos, teremos no banco de dados um conjunto de informações pronto para ser utilizado pelos usuários do sistema. A probabilidade desse usuário obter as informações de que necessita aumentará à medida que a quantidade de informações depositadas no banco for tornando-se maior com o tempo.

Os arquivos foram formados segundo um critério que permitirá recuperar as informações, ali depositadas, com bastante eficiência. O sistema que passaremos a analisar possibilitará essa recuperação eficiente.

4.1 O SISTEMA DE RECUPERAÇÃO MANUAL

O problema básico de recuperação pode ser enunciado assim: "qual é o documento que trata, simultaneamente, dos assuntos A, B, C, etc., onde A, B, C, etc. são palavras-chave dadas?"

Descreveremos, a seguir, um sistema que utilizando-se da mesma estrutura de lista dos arquivos já montados, permitirá solucionar esse problema. Para ilustrar melhor seu funcionamento, vamos supor que desejamos encontrar, nos arquivos já montados, os documentos que tratem ao mesmo tempo dos assuntos KW1, KW4 e KW6.

Começamos por localizar, no indexador, a palavra-chave KW1 (ver figura 10).

Seguindo os ponteiros da direita, percorreremos, a partir de KW1 no arquivo IV, todo o arquivo III, como indicado. Nesse percurso, iremos anotando as incidências em cada uma das fichas do arquivo III. A indicação feita na figura corresponde à lista descrita na figura 11 a seguir.

Repetindo o processo para KW4 e KW6, obtemos as seqüências descritas na figura 10. Isto feito, voltamos as atenções para o número de incidências em cada ficha do arquivo III. Temos:

Figura 10

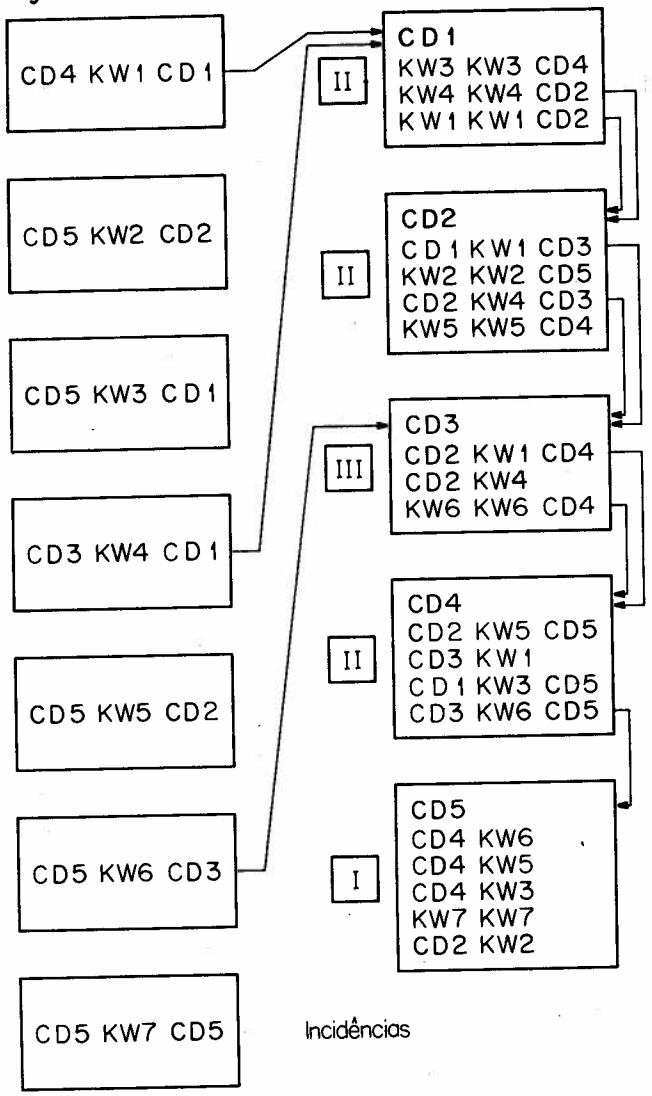


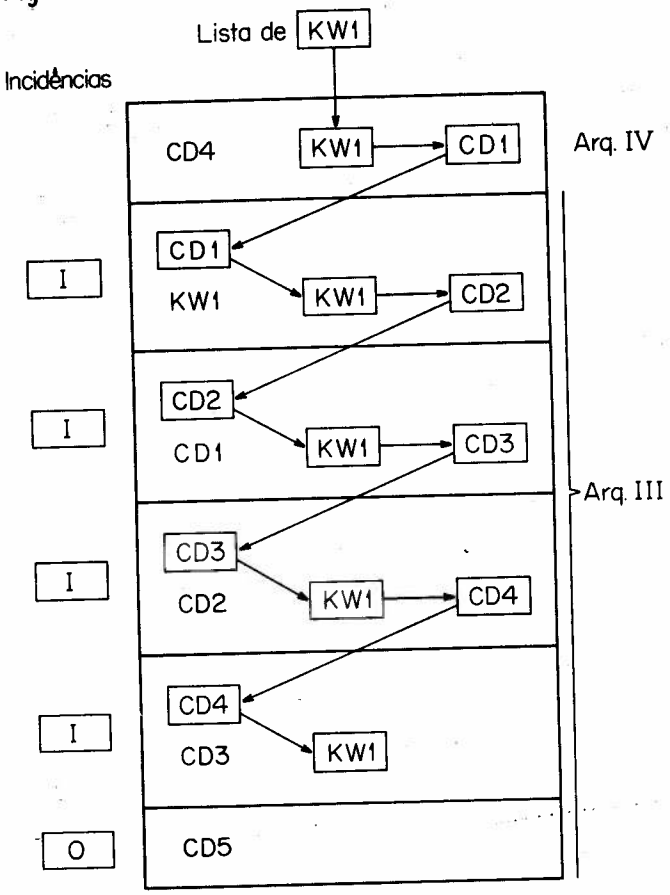
Tabela 1

Fichas	N.º de incidências
CD1	2
CD2	2
CD3	3
CD4	2
CD5	1

Os documentos procurados deverão ter três incidências, no total, pois fizemos a pesquisa nos arquivos a partir das três palavras-chave: KW1, KW4 e KW6.

Assim, observando a tabela 1, obtemos a obra desejada, que é a de código CD3, que resolve o problema.

Figura 11



4.2 CRITÉRIOS DE RELEVÂNCIA

No item anterior, ao recuperar os documentos do arquivo, usamos um certo *critério*, que nos permitiu encontrar CD3, e que consistiu em considerar *relevante* o documento que contivesse palavras-chave iguais às fornecidas pelo usuário do sistema, KW1, KW4 e KW6, no caso.

Tal critério, no entanto, nem sempre é satisfatório.

De fato, isso ocorreria, por exemplo, com o documento CD9, um livro cujo título é *O modelo matemático dos juros*, e que seria referenciado por meio das palavras-chave:

Modelo matemático
Juros contínuos
Juros discretos

Se um usuário estivesse buscando uma bibliografia que tratasse do assunto "teoria dos modelos matemáticos", ele forneceria ao sistema as palavras-chave, para consulta aos arquivos, como segue:

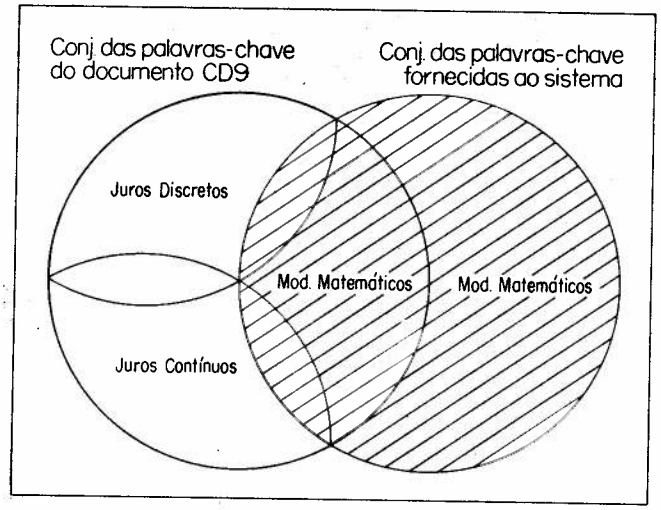
Modelo matemático.

Nessas circunstâncias, o sistema recuperaria, entre outros, o documento CD9, já que o número de

incidências é igual ao de palavras-chave fornecidas (uma). Esse livro, entretanto, não trata da teoria dos modelos matemáticos, mas sim das aplicações dessa teoria aos juros. Notamos, assim, que esse documento não é relevante, no caso, e o critério não é satisfatório.

Em termos da teoria dos conjuntos as palavras-chave fornecidas constituem um subconjunto de CD9, e pelo critério usado, CD9 foi considerado relevante, embora contenha outros assuntos nos quais o usuário não está interessado (por hipótese). A figura 12 mostra esses conjuntos.

Figura 12



Existem vários critérios de relevância, cada um com suas vantagens e desvantagens.[1] Apresentaremos, a seguir, um critério bastante usado na prática.

Antes de estabelecer esse critério, entretanto, são necessárias algumas definições preliminares.

Chamaremos de K_u o conjunto das palavras-chave fornecidas pelo usuário do sistema, e de $K_d(i)$ o conjunto das palavras-chave dos documentos de código CDi. Por $|A|$ entenderemos o número de elementos do conjunto A (finito).

Com essas definições, o que temos denominado de "incidências" pode ser expresso como segue:

$$\text{Incidência no documento} — i = |K_u \cdot K_d(i)| \quad (1)$$

ou seja, o número de incidências é igual ao número de elementos do conjunto intersecção de K_u com $K_d(i)$. O critério de relevância utilizado no item anterior também pode ser expresso por

$$|K_u| = |K_u \cdot K_d(i)| \quad (2)$$

Definamos, agora, o índice de aproximação semântica de um documento i, R(i), com relação ao conjunto de palavras-chave K_u, por

$$R(i) = \frac{|K_u \cdot K_d(i)|}{|K_u + K_d(i)|} \quad (3)$$

e a distância semântica d(i) por

$$\boxed{d(i) = 1 - R(i)} \quad (4)$$

O critério de relevância prática, ao qual nos referimos, pode, então, ser dado por: "Será considerado relevante todo o documento (i) cuja distância d(i) seja menor que uma constante C prefixada," isto é,

$$\text{Critério de relevância} \quad \boxed{d(i) \leqslant C} \quad (5)$$

A vantagem das medidas assim definidas é possuírem certas propriedades que nos são convenientes. Essas propriedades encontram-se demonstradas no apêndice.

A primeira propriedade da distância é ser limitada superiormente e inferiormente. Isso possibilita o estabelecimento do critério (5), estabelecendo que o documento relevante é todo aquele cuja distância semântica do assunto fornecido pelo usuário não exceda C.

A segunda propriedade (ver apêndice) garante-nos que um documento com distância nula coincide, conceitualmente, com o solicitado ao sistema.

A título de ilustração, vamos calcular as distâncias semânticas no exemplo do item anterior. Nesse caso, tínhamos:

K_u = (KW1, KW4, KW6)
$K_d(1)$ = (KW1, KW3, KW4)
$K_d(2)$ = (KW1, KW2, KW4, KW5)
$K_d(3)$ = (KW1, KW4, KW6)
$K_d(4)$ = (KW1, KW3, KW5, KW6)
$K_d(5)$ = (KW2, KW3, KW5, KW6, KW7)

Com esses dados, construímos a tabela 2 a seguir:

Tabela 2

i	$\lvert K_u \cdot K_d\ (i) \rvert$	$\lvert K_u + K_d\ (i) \rvert$	R(i)	d(i)
1	2	4	0,50	0,50
2	2	5	0,40	0,60
3	3	3	1,00	0,00
4	2	5	0,40	0,60
5	1	7	0,14	0,86

Se estabelecermos que os documentos relevantes deverão estar a uma distância não superior a 20% — valor de C em (5) — teremos como resultado do processo de recuperação, apenas o documento CD3.

Se adotarmos C = 50%, então serão relevantes, no caso em foco, os documentos CD3 e CD1.

Com uma outra ilustração, apliquemos a definição de distância ao exemplo da figura 12.

Neste caso:

K_u = (modelos matemáticos)

K_d = (modelos matemáticos, juros discretos, juros contínuos)

Número de elementos do conjunto intersecção =

$= |K_u \cdot K_d| = 1$

Número de elementos do conjunto reunião =

$= |K_u + K_d| = 3$

$R = 1/3 = 33{,}3\%$

Distância semântica = d = 66,6%

Assim, o valor de R informa-nos que o documento CD9 contém apenas 33% do assunto que nos interessa (modelos matemáticos); por conseguinte, somente será recuperado se fizermos

$C = 67\%$

no critério (5).

Por outro lado, se estivermos interessados no modelo matemático dos juros contínuos, teremos:

K_u = (modelos matemáticos, juros contínuos)

e $R = 2/3 = 67\%$

e $d = 33\%$

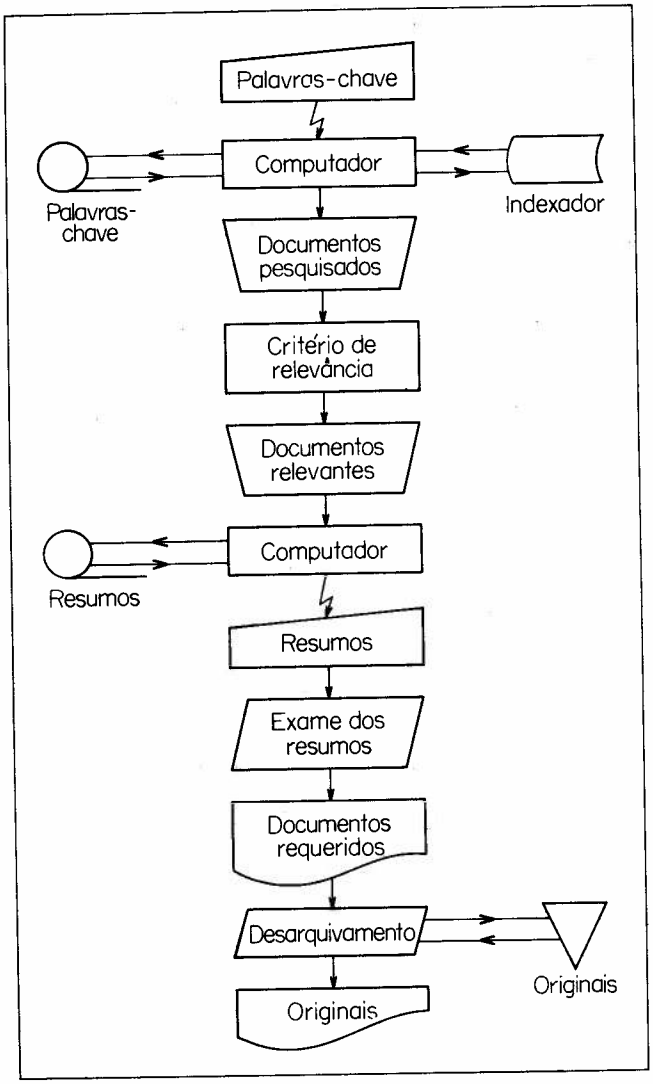
e o documento CD9 conterá 67% do que procuramos. Por conseguinte, para recuperá-lo, bastará fazer C = 33% em (5).

Desta forma, podemos observar como varia a quantidade de documentos recuperados pelo sistema, quando alteramos o valor de C.

4.3 O SISTEMA K.W.I.C. DE RECUPERAÇÃO AUTOMÁTICA DE INFORMAÇÕES

O sistema anterior, quando automático, está esquematizado na figura 13.

Figura 13 – Sistema de Recuperação de Informações (K.W.I.C.)

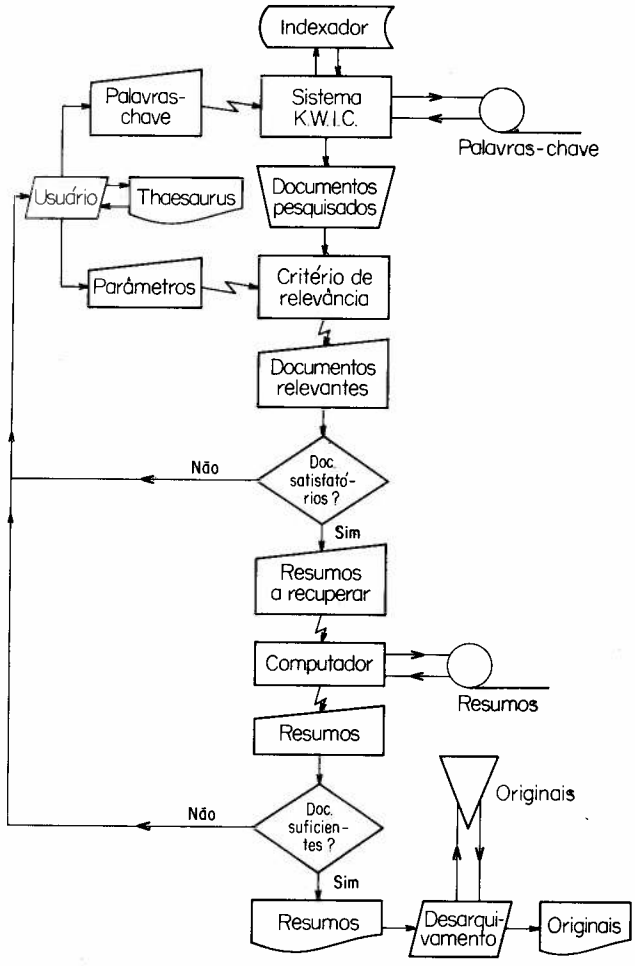


O funcionamento do sistema de recuperação tem início com a relação das palavras-chave fornecidas pelo usuário. Com esses dados é feita a pesquisa aos arquivos classificador e indexador, resultando o conjunto dos códigos dos documentos pesquisados, isto é, aqueles que possuam pelo menos uma das palavras-chave fornecidas. Esses códigos são submetidos a um critério de relevância, o qual desempenha o papel de um filtro informático, restando os códigos dos documentos agora considerados relevantes. Tais códigos são usados para a busca dos respectivos resumos no arquivo, que constituem

o relatório de saída do sistema, que vai ter ao usuário. Este, após relacionar os resumos que lhe interessam, terá os originais desses resumos recuperados manualmente do arquivo, o que finda o processo.

O sistema descrito é o *básico*, podendo evoluir para outros mais sofisticados. Por exemplo, podemos ter um sistema de recuperação *conversacional*, isto é, onde se realiza um diálogo homem-máquina, tornando mais eficiente o processo de busca. Tal sistema está esquematizado na figura 14.

Figura 14 - Sistema de recuperação de informações conversacional



O sistema da figura 14 é mais eficiente porque, permitindo amplo diálogo com o usuário, possibilita sua máxima utilização.

Como é fácil de ver, o processo pode ser reiniciado tantas vezes quanto for julgado necessário pelo usuário, ocasião em que novas palavras-chave e novos parâmetros — o valor da constante C na relação (5) — podem ser fornecidos. E, é claro, cada ciclo do processo corresponde à obtenção de novas informações.

5. Sistemas em funcionamento

Vários sistemas, semelhantes ao descrito em nosso artigo, já se encontram em funcionamento em países estrangeiros (e, esperamos, também em breve no Brasil).

Um desses sistemas, o do Centro de Informações Científicas e Tecnológicas do Japão, está bem descrito num artigo do jornal *O Estado de São Paulo* (ver referências bibliográficas 15), ao qual remetemos o leitor.

Um outro sistema, chamado Eclair, existe implantado na França, o qual tivemos oportunidade de ver funcionando. Encontra-se na biblioteca do Instituto de Pesquisas em Informática e Automática (I.R.I.A.), em Rocquencourt. Essa biblioteca, que contém cerca de 5.000 documentos (livros, revistas, etc.), é especializada em informática, e contém um *thaesaurus* bastante completo sobre esse assunto.

No processo de classificação das obras, são gerados os documentos mostrados na figura 15. Para cada obra, vemos o número do documento, o número do inventário, a categoria, o tipo de obra, o ano de edição, a língua utilizada, etc. Notamos também as palavras-chave *(mots-clé)* que refletem o conteúdo informático de cada obra.

A partir das palavras-chave, que constituem o arquivo classificador (ver figura 10), é montado, via computador, o arquivo indexador (cf. figura 10). Uma parte desse arquivo está mostrada na figura 16, e a entrada é feita pelas palavras-chave selecionadas pelo usuário, obtendo-se então os números dos documentos referenciados por tais palavras.

Entretanto, o sistema de recuperação lá existente não é automático, mas sim manual. Seu funcionamento realiza-se como segue. Suponhamos que se queiram obras sobre hidrodinâmica. Na figura 16 encontramos assinalados os documentos números 00049 e 00056. Na figura 15 vemos o documento 00049, contendo "Hidrodynamique" como palavra-chave associada, sendo este um dos documentos recuperados.

Figura 15

	LEBESQUE
	ORLICZ

DOCUMENT NUMERO 0000000049

NUM INVENTAI	00049
CATEGORIE	A
TYPE OUVRAGE	T
DATE PARUTIO	67
LANGUE	A
COTE RAYON	T007
AUTEUR INDEX	DESBARD
AUTEURS	LICHNEROWICZ
	ANDRE
TITRE	RELATIVISTIC HYDRODYNAMICS AND MAGNETOHYDRODYNAMICS
	LECTURES ON THE EXISTENCE OF SOLUTIONS
EDITEUR	BENJAMIN
	NEW YORK
MOTS CLE	HYDRODYNAMIQUE
	MAGNETOHYDRODYNAMIQUE
	CONDUCTIVITE
	RELATIVITE

DOCUMENT NUMERO 0000000050

NUM INVENTAI	00050
CATEGORIE	A
TYPE OUVRAGE	T
DATE PARUTIO	61
LANGUE	A
T-VOL-FASC	2
COTE RAYON	G011
DIRECTEUR PU	SNEDDON I. N.
AUTEUR INDEX	HERATCHIAN
AUTEURS	FUCHS
	B. A.
	LEVIN
	V. I.
TITRE	FUNCTIONS OF A COMPLEX VARIABLE
	FONCTIONS D'UNE VARIABLE COMPLEXE
EDITEUR	PERGAMON PRESS
	NEW YORK
MOTS CLE	FONCTION COMPARAISON
	DOMAINE STABILITE
	THEOREME WEIERSTRASS
	FONCTION CYLINDRIQUE
	FONCTION BESSEL
	FONCTION ALGEBRIQUE
	EQUATION DIFFERENTIELLE
	TRANSFORMATION LAPLACE
	INTEGRATION
	HURWITZ
CLASSIFICATI	C2
	E1
	G

DOCUMENT NUMERO 0000000052

Figura 16

```
**********HURWITZ
          *        00050*

**********HUYBENS-FRESNEL
          *        02641*

**********HYDRAULIQUE
          *        02175**        02175*

**********HYD ODYNAMIQUE
          *        00049**        00056*

**********HYDROLOGIE
          *        00445*

**********HYPERBOLE
          *        01393*

**********HYPERBOLIQUE
          *        00366**        00687*

**********HYPERFREQUENCES
          *        02690*

**********HYPERGEOMETRIE
          *        00162*

**********HYPERGEOMETRIQUE
          *        00057**        00157**        01270*

**********HYPOTHESE
          *        01810*

**********HIPOTHESE DU CONTINU
          *        00378*

**********HYPOTHESES DU CONTINU
          *        02136*

**********H1
          *        01142*

**********IBM
          *        02045*
```

6. Conclusões

O sistema descrito nesse artigo tem inúmeras aplicações nas mais variadas áreas da arquivística.

Nos departamentos de investigações dos serviços de segurança, tal sistema é extremamente útil. Montando um arquivo de palavras-chave, constituído das características de cada pessoa identificada — nome, filiação, dados somáticos, ficha datiloscópica, etc. — um indivíduo poderia ser mais facilmente localizado, a partir de quaisquer dados pessoais. O arquivo de resumos seria constituído das fichas de antecedentes (e outros dados) de cada cidadão. Se houvesse uma Central Nacional de Informações Policiais a eficiência do Banco de Dados seria ainda maior. Segundo estamos informados, existe, nos EUA, um sistema em que a própria fotografia da pessoa é arquivada segundo determinado código. E, entre os dados de entrada, no caso de consulta ao sistema, pode ser fornecida a "fotografia falada" (um desenho) da pessoa procurada. Um subsistema, chamado de "reconhecimento de amostras" *(Pattern recognition)*, seleciona as fotos que se assemelham com essa amostra.

Na área do direito, o K.W.I.C. pode também mostrar-se útil. Cada decreto ou lei seria reduzido a um conjunto de palavras-chave e arquivado. O sistema assim montado permitiria descobrir, de uma forma eficiente e rápida, que leis, decretos, etc. envolvem determinados assuntos solicitados pelo consulente do Banco de Dados. O problema atual da contradição entre as leis, isto é, a existência de leis e decretos que se negam mutuamente, poderia ser assim resolvido. A tarefa de verificação da inconstitucionalidade das leis seria também bastante facilitada.

O K.W.I.C. possui uma característica particularmente útil nos Serviços Nacionais de Informações. Como em qualquer organização burocrática (no sentido de racional-legal), existe uma hierarquia. E em função do nível hierárquico do indivíduo, este terá acesso a alguns tipos de informação, apenas (note-se a relação informação-poder!). Assim, estabelecendo-se certas palavras-código (chaves), somente os conhecedores desses códigos terão acesso a determinadas informações. Pode-se, desta forma, estabelecer um processo seletivo de obtenção de informações do Banco de Dados. E isto se aplica também às empresas, de um modo geral.

Na área da medicina, o chamado "diagnóstico por computador" torna-se bastante viável com o K.W.I.C. Assim, cada anomalia que possa existir no corpo humano pode ser expressa por uma série de palavras que traduzam as características desse estado anormal (sintomas). O arquivo de resumos conterá informações sobre cada anormalidade, incluindo tratamentos, medicamentos, etc.

Nas bibliotecas, o K.W.I.C. é revolucionário. Na medida em que os livros e artigos não mais são catalogados por títulos e autores, mas sim com base no *índice* e no próprio *conteúdo* da obra, esse sistema é muito mais potente que os tradicionais. Embora o título seja, na maioria dos casos, constituído de palavras-chave associadas ao conteúdo da obra, é bastante reduzida a capacidade desse título sintetizar, satisfatoriamente, os assuntos tratados; principalmente quando ele tem caráter predominantemente *comercial*.

Os exemplos de utilização do K.W.I.C. mencionados são uma amostra muito pequena da aplicabilidade prática do sistema, mas servem para ilustrar algumas áreas de aplicação.

Outros aspectos do sistema são mostrados a seguir.

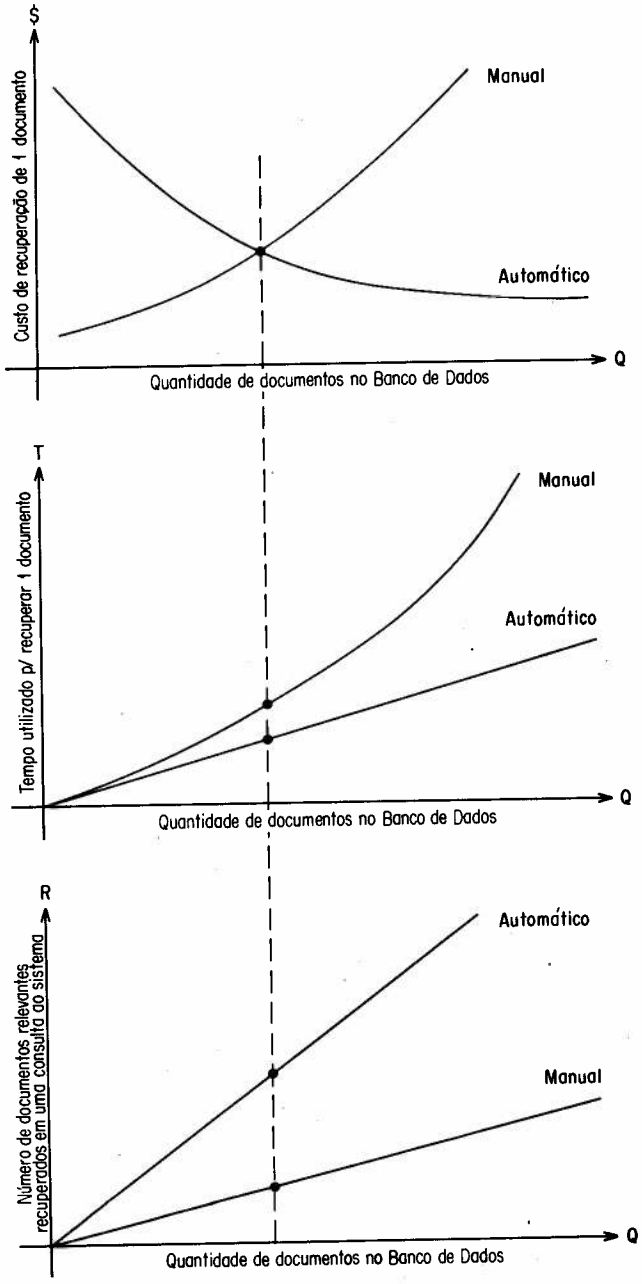
Nos comentários que se seguem, estaremos supondo que os sistemas manuais não empregam o K.W.I.C., pois este não é muito adequado a esse tipo de processo, e que os sistemas automáticos uti-

lizam-se do K.W.I.C. Embora pareça mais conveniente comparar sistemas manuais com e sem o K.W.I.C., fazendo o mesmo para os automáticos, tal conveniência é mais didática que prática.

As comparações a seguir são meramente qualitativas. Dados reais podem ser encontrados nas referências bibliográficas 2, embora nem sempre sejam aplicáveis no caso das condições brasileiras.

A figura 17 relaciona a quantidade de documentos do Banco de Dados (Q) com o custo de recuperação desses documentos ($), com o tempo despendido com essa recuperação (T), e com a quantidade de documentos relevantes recuperados numa consulta (R).

Figura 17



Os seguintes fatos são notórios:

1. Quando o arquivo atinge um dado tamanho, o sistema manual torna-se mais oneroso que o automático. Esse ponto de "*break-even*" é característica de todas as comparações entre sistemas mecanizados e automáticos (gráfico $XQ);
2. O tempo gasto na recuperação manual aumenta rapidamente com a quantidade de documentos no Banco. Este é um dos motivos da lentidão observada em muitas repartições públicas que lidam com arquivos (gráfico TXQ);
3. O número de documentos relevantes obtidos do Banco é muito maior no sistema automático. Isto significa que, muitas vezes, não se encontra um documento no arquivo, não porque ele não esteja nesse arquivo, mas porque a classificação é deficiente (gráfico RXQ).

De acordo com a finalidade do Banco de Dados, um desses fatores suplanta totalmente os outros. Assim, se o Banco tem em vista integrar um sistema de defesa antiaérea (como é o caso do sistema americano SAGE — *Semi-automatic ground environment*) o tempo de resposta do sistema (2.° gráfico) e o número de documentos relativos obtidos (3.° gráfico) são fundamentais, sendo o custo (1.° gráfico) secundário (exceto se atingir valores proibitivos).

Já no caso de se ter uma Biblioteca Nacional, a variável básica é a expressa por R, no 3.° gráfico, sendo as demais secundárias (na medida em que não tornem o sistema proibitivo).

APÊNDICE

Neste apêndice serão estudadas algumas propriedades do módulo de um conjunto, citado no item 4.2.

Definição 1

Sejam A e B dois conjuntos finitos, e Ø o conjunto vazio.

Denomina-se módulo (ou número de elementos) de A, ao número inteiro, representado por $|A|$, que satisfaz às propriedades

i) $|\varnothing| = 0$ (1)

ii) $B \supset A \longrightarrow |B| > |A|$ (2)

iii) $|A + B| = |A| + |B| - |AB|$ (3)

(Fim da definição 1)

Note-se a semelhança formal desta medida com a medida de probabilidade de um conjunto de pontos amostrais.

Definição 2

Chama-se "índice de congruência" de dois conjuntos A e B, o número real R(A,B) dado por:

$$R(A,B) = \frac{|AB|}{|A + B|} \quad (4)$$

(Fim da definição 2)

O número R(A,B) costuma receber diferentes designações, de acordo com as particulares interpretações que pode assumir nas aplicações. Tal foi o caso do "índice de relevância".

Entre as propriedades que (4) possui, algumas são demonstradas a seguir.

Propriedade 1

Se A e B são dois conjuntos quaisquer satisfazendo $A + B \neq \emptyset$ então $0 \leqslant R(A,B) \leqslant 1$

Demonstração

a) Como

$$\emptyset \subseteq A \;\forall\; A$$

tem-se que, usando (1) e (2):

$$|A| \geq 0 \;\forall\; A \qquad (5)$$

b) Pela definição de intersecção de dois conjuntos,

$$AB \subseteq A$$
$$AB \subseteq B$$

usando (2):

$$|AB| \leq |A|$$
$$|AB| \leq |B|$$

Somando essas desigualdades membro a membro, nem

$$|AB| \leq |A| + |B| - |AB|$$

ou, usando (3)

$$|AB| \leq |A + B|$$

c) Dessa desigualdade e de a), obtemos

$$0 \leq |AB| \leq |A + B|$$

Uma vez que, por hipótese

$$|A + B| \neq |\emptyset| = 0$$

obtemos, em definitivo:

$$\boxed{0 \leqslant \frac{|AB|}{|A + B|} \leqslant 1} \qquad (6)$$

(C.Q.D.)

Propriedade 2

A condição necessária e suficiente para que $A = B$, é que $R(A,B) = 1$. Formalmente:

$$R(A,B) = 1 \longleftrightarrow A = B$$

Demonstração $R(A,B) = 1 \longleftarrow A = B$

a) *Necessidade:*

É imediato, bastando fazer $A = B$ em (4).

b) *Suficiência:* $R(A,B) = 1 \longrightarrow A = B$

Se $R(A,B) = 1$, então, de (4):

$$|AB| = |A + B| \qquad (7)$$

Por outro lado, partindo da identidade

$$A + B = AB + A(U - B) + (U - A)B$$

onde U é o conjunto universo, e notando-se que os conjuntos AB, A(U − B), e (U − A)B são disjuntos, obtemos, aplicando (3):

$$|A + B| = |AB| + |A(U - B)| + |(U - A)B|$$

introduzindo agora (7), vem

$$|A(U - B)| + |(U - A)B| = 0$$

Usando a propriedade já demonstrada, de que o módulo é um número não negativo, a igualdade anterior implica necessariamente em:

$$\begin{cases} |A(U - B)| = 0 \\ |(U - A)B| = 0 \end{cases}$$

Como (1) e (2) garantem que

$|A| = 0 \longleftrightarrow A = \emptyset$, temos:

$$\begin{cases} A\,(U - B) = \emptyset \\ (U - A)B = \emptyset \end{cases}$$

ou, desenvolvendo

$$\begin{cases} A = AB \\ B = AB \end{cases}$$

Pela transitividade da igualdade, obtemos, finalmente,

$$\boxed{A = B}$$

(C.Q.D.)

Definição 3

Denomina-se *distância* entre dois conjuntos A e B, ao número d(A,B) dado por

$$d(A,B) = 1 - R(A,B)$$

Propriedade 3

$$0 \leq d(A,B) \leq 1$$

Propriedade 4

$$d(A,B) = 0 \longleftrightarrow A = B$$

As propriedades 3 e 4 decorrem, trivialmente, das 1 e 2 e da definição 3.

REFERÊNCIAS BIBLIOGRÁFICAS

Livros

1. O'Vrien, James Jl. *Management information systems*. New York, Van Nostrand Reinhold Company, 1970.
2. Saracevic, Tefko, ed. *Introduction to information science*. New York, R.R. Bowker Company, 1970.
3. Vickery, B. C. *On retrieval system theory*. London, Butterworths, 1961.
4. Library of Congress. *Subject headings*. 7. ed., Washington, D.C., 1966.
5. Lancaster, F.W. *Information retrieval systems: characteristics, testing and evaluation*. New York, Wiley, 1968.
6. Knuth, Donald E. *The art of computer programming*. Mass., Addison-Wesley Publ. Co., 1969.
7. Lucena, Carlos J. P. *Introdução às estruturas de informação*. Rio de Janeiro, GB, Ao Livro Técnico, 1970.
8. Foskett, D. Jl. *Serviço de informação em bibliotecas*. São Paulo, Polígono, 1969.
9. *ROGET'S international thaesaurus of english words and phrases*. Thomas Y. Crowell Co., USA, 1970.

Artigos

10. Luhn, H.P. *Key word in context index for technical literature (K.W.I.C.)*. Yorktown Heights, N.Y., International Business Machines Corp., Advanced Development Division, 1959.
11. Samuelson, Kjell. *Proceeding of the FID-FIF conference on mechanical information storage, retrieval and dissemination*. North Holland, 1967.
12. Alvarez, José Cesário R. *Informática, conceitos* gerais. São Paulo, Fundação Getulio Vargas, 1971.
13. Kraft, A.H. Comparison of key word in context (K.W.I.C.) indexing of titles with a subject reading classification system. *American Documentation*, v. 15, p. 48-52, 1964.
14. Rodrigues, Eduardo Celestino. Centro de Informações facilita a atualização. Jornal O *Estado de São Paulo*, Atualidade Científica, 24-10-1971.
15. IBM, Manual Form C17-0012, Técnicas para Fluxogramas e Diagramas de Blocos.

[1] Ver referências bibliográficas 2, onde vários critérios são estudados.