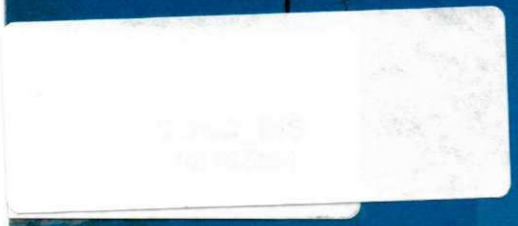
CCT-UFPb

ANÁLISE DE SISTEMAS EM PROJETO DE AUTOMAÇÃO DOCUMENTAL

AGENOR DE SOUSA MARTINS

DEZEMBRO - 1976

UNIVERSIDADE FEDERAL DA PARAIBA
CENTRO DE CIÊNCIAS E TECNOLOGIA
AVENIDA APRIGIO VELOSO, 882 - Cx. POSTAL 518
CAMPINA GRANDE - PB
BRASIL



UNIVERSIDADE FEDERAL DA PARAÍBA

CENTRO DE CIÊNCIAS E TECNOLOGIA - CCT

ANÁLISE DE SISTEMAS EM PROJETO DE AUTOMAÇÃO DOCUMENTAL

AGENOR DE SOUSA MARTINS

TESE SUBMETIDA AO CORPO DOCENTE DA COORDENAÇÃO DOS PROGRAMAS DE PÓS-GRADUAÇÃO DE ENGENHARIA DO CENTRO DE CIÊNCIAS E TECNOLOGIA DA UNIVERSIDADE FEDERAL DA PARAÍBA COMO PARTE DOS REQUISITOS NECESSÁRIOS À OBTENÇÃO DO GRAU DE MESTRE EM CIÊNCIAS - (M.Sc.)

APROVADO POR:

Prof. ORION DE OLIVEIRA SILVA
- Presidente -

COMISSÃO :

Prof. RAIMUNDO HAROLDO DO CARMO CATUNDA

Prof. GOVIND PRASAD GUPTA

Prof. ORION DE OLIVEIRA SILVA
- Orientador -

CAMPINA GRANDE
ESTADO DA PARAÍBA - BRASIL
1 9 7 6

M379a Martins, Agenor de Sousa.
Análise de sistemas em projeto de automação documental / Agenor de Sousa Martins. - Campina Grande, 1976.
[110] f.

Dissertação (Mestrado em Ciências) - Universidade Federal da Paraíba, Centro de Ciências e Tecnologia, 1976.
"Orientação : Prof. Orion de Oliveira Silva".
Referências.

1. Processamento de Informação. 2. Automação Documental - Projeto (Análise de Sistemas). 3. Análise de Sistemas - Automação Documental. 4. Dissertação - Ciências. I. Silva, Orion de Oliveira. II. Universidade Federal da Paraíba - Campina Grande (PB). III. Título

CDU 004.032.2(043)

Aos meus pais

Agradecimentos,

Ao Governo Federal: CAPES/UFPb
Ao Governo do Piauí: SEPLAN/CEPRO
Ao Prof. Orientador: ORION DE O. SILVA

R E S U M O

Este trabalho tem como finalidade proporcionar uma visão geral da Análise de Sistemas aplicada ao controle e ao processamento automáticos de artigos de revista, papers, dossier médico, livros e outros dados impressos.

Apresenta, inicialmente, os conceitos bibliométricos de sistema documental, palavra-chave, frequência de ocorrência e de coocorrência de palavra-chave, coeficiente de similitude, matriz de incidência, matriz de coocorrência e matriz de similitude. Em seguida são analisados os principais processos envolvidos nos subsistemas de aquisição, catalogação e circulação de documentos, como ainda os processos que compõem os subsistemas de recuperação de informação bibliográfica. A exposição é ilustrada com dois exemplos de análise, um retirado da Recuperação de Informação e o outro da gestão documental.

Também são discutidos aspectos gerais relativos à avaliação técnica e à avaliação de custo-benefício dos sistemas documentais.

A B S T R A C T

This work attempts to present an overall view of systems analysis used in the control and automatic processing of articles, magazines, papers, medical dossiers, books and other published material.

Initially presented are bibliographic concepts of documental systems, key-word, occurrence frequency and co-occurrence frequency of key-words, similarity coefficients, incidence matrix, co-occurrency matrix and similarity matrix.

Next, the principal processing involving sub-systems of aquisition, cataloguing and circulation of documentation is analysed, as well as processes that form the sub-system of bibliographic information retrieval.

The work is illustrated by two examples of analysis, one of them extracted from information retrieval, the other from management documentation.

General aspects related to the technical evaluation and cost-benefit from documental systems are also discussed.

S U M Á R I O

I N T R O D U Ç Ã O

No estágio atual de desenvolvimento da Teoria de Recuperação de Informação (RI) seus problemas fundamentais giram em torno dos sistemas que provêm dados - **Management Information Systems** - ou MIS e dos sistemas que providenciam referências ou sistemas documentais. Ambos são sistemas de informação, porém assumem características que os diferenciam. Evidentemente, em se considerando que o campo da Recuperação de Informação apareceu há muito pouco tempo, seu universo de estudo além de não se encontrar ainda perfeitamente delimitado, carece ainda de uma sistematização unificada para o tratamento das questões referentes aos sistemas de informação. Este fato é tanto mais verdadeiro quando se leva em conta a análise e automação dos sistemas de referência onde cada autor ou grupo de pesquisa discorre sobre temas diferentes dando prioridade ora à este, ora àquele assunto. Se a isto for adicionado o fato de que praticamente inexiste uma bibliografia nacional sobre estudos de automação documental (afora algumas dissertações de mestrado) pode-se compreender o grau de dificuldade de quem deseje iniciar-se em RI. Os estudos rarefeitos sobre o assunto, além de se voltarem para questões muito específicas (abstratos, administração de documentos, ou alguma linguagem documental) dentro do universo da Recuperação de Informação, confundem frequentemente o leitor principiante com terminologia não unificada e com a superenfatização de certos aspectos. Assim, por exemplo, grande importância tem sido dada à abordagem dos arquivos sequenciais e invertidos, às estruturas de lista, multilista, listas circulares e de árvores, em detrimento muitas vezes, das demais variáveis do estudo dos sistemas de informação

e, em particular, dos sistemas documentais.

Foi neste contexto que se decidiu pela elaboração deste documento. Ele é um estudo menos atomizado sobre análise de sistemas documentais orientada para o uso computacional. Seu objetivo consiste, assim, em apresentar um **survey** sobre a automação das atividades de bibliotecas e centros de documentação, quer pertençam elas ao nível da gestão de documentos (aquisição, catalogação, circulação) ou ao nível do planejamento de sistemas (avaliação técnica, avaliação econômica) ou, ainda, ao nível das atividades propriamente de recuperação de informação.

A preocupação básica foi oferecer uma visão das atribuições da Análise de Sistemas quando voltada para a automação de bibliotecas, buscando abranger, ao máximo, a temática abordada de uma maneira ordenada e didática. Em contrapartida, seja reconhecido, o preço da abrangência horizontal foi o não aprofundamento vertical (pouca intensidade na abordagem) de determinados assuntos que só ligeiramente foram lembrados.

No primeiro capítulo apresentam-se algumas definições do campo da Bibliometria necessárias para o acompanhamento posterior do raciocínio. Os demais capítulos da Parte I desenvolvem as idéias centrais desta monografia. Partindo-se das características do método sistêmico, é mostrado então seu uso na administração e recuperação da informação documentada. Nos capítulos da Parte II foram destacados dois temas para efeito de exemplificação através de modelos. No que tange à gestão de documentos montou-se um modelo de metodologia que possa ser útil, sobretudo, na automação de bibliotecas descentralizadas de universidade. No último capítulo é mostrado um modelo de análise na área da recuperação de documentos ou da informação

documentada. Após considerar-se a Classificação Decimal Universal (CDU) como uma linguagem de recuperação passível de automação, contudo de pouca vantagem relativa, apresenta-se então, como possível alternativa, um algorítmo de classificação de palavras-chave com razoável nível de detalhamento.

PARTE I: ABRANGÊNCIA DA ANÁLISE DE SISTEMAS EM BIBLIOTECA

CAPÍTULO 1

FUNDAMENTOS DE BIBLIOMETRIA

1.1 Origem da Bibliometria

Em biblioteca o uso computacional requer sempre uma análise de sistemas cujo refinamento e grau de complexidade dependem do subsistema a ser automatizado. Se, por exemplo, o subsistema em foco opera na área da administração documental, então a análise e a automação apresentam menos complexidade do que a exigida por quaisquer subsistemas da área de recuperação de informação. A análise de sistemas em nível de recuperação de informação requer, por exemplo, além do **approach** sistêmico métodos quantitativos oriundos dos mais variados domínios tais como da Estatística, do Cálculo Matricial, da Álgebra Booleana, do Cálculo das Probabilidades e da Teoria dos Grafos.

A partir de 1950, entre os métodos quantitativos que apoiam a análise de sistemas e a Informática Documental* vem tomando lugar um novo ramo de conhecimento. É a Bibliometria. Fundamentalmente a Bibliometria originou-se da interseção de três outros conjuntos de conhecimentos: Estatística, Documentação

---

* Informática Documental é o ramo da Ciência da Computação que trata da especificidade do uso computacional em bibliotecas e centros de análise de informação.

Científica ou Ciência da Informação* e Matemática. O diagrama 1 ilustra esta genealogia.

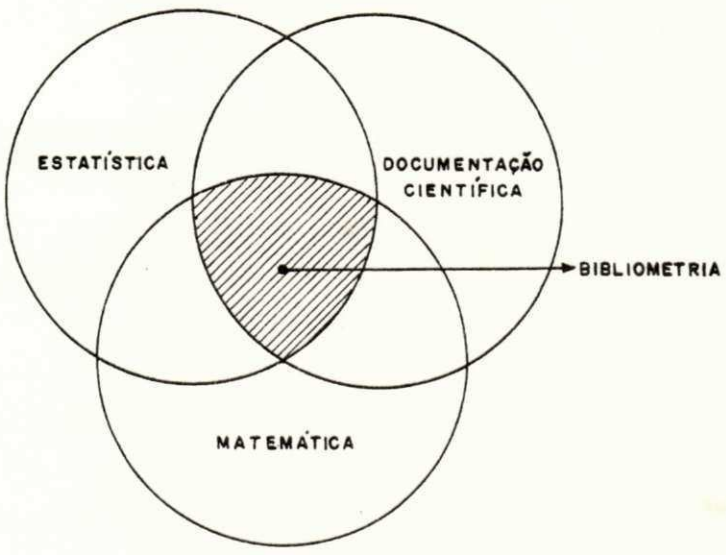


DIAGRAMA 1 - ORIGEM DA BIBLIOMETRIA

O campo principal de pesquisa da Bibliometria é o estudo da biblioteca, da análise documental e dos processos de disseminação da informação utilizando para tal o tratamento quantitativo das propriedades e do comportamento do conhecimento armazenado (ou seja, o comportamento da informação). Note-se que, seu objeto de estudo não se confunde com o da Documentação Científica que se ocupa, antes, com o tratamento da informação não numérica em todas as suas formas. O alvo da Bibliometria, em vez

---

* A diferença entre as duas denominações está no fato da segunda fazer ressaltar o próprio conteúdo do documento - a informação - enquanto que a primeira denominação ressaltava antes o documento como suporte físico da informação.

de considerar as operações de uma biblioteca ou centro de informação do ponto de vista estritamente funcional é examinar, por exemplo, as distribuições estatísticas dos processos relacionados com a dispersão, ou utilização dos itens de informação. Ao descrever a quantidade de empréstimos e as características da circulação de livros numa biblioteca e num determinado período, fazendo uso da distribuição de Poisson a Bibliometria está, de fato, descobrindo suas próprias leis e se firmando como ciência autônoma. Na verdade, é sabido que a maioria das circulações de livros numa biblioteca é realizada dentro de determinados espaços de tempo (circulações rarefeitas) e as características das circulações não são relacionadas umas com as outras. Este fato leva o bibliometrista a adequar a distribuição de Poisson à descrição das estatísticas relacionadas com a circulação de documentos, uma vez que a distribuição de Poisson supõe uma aplicação onde eventos rarefeitos ocorram independentes um do outro e a uma taxa média constante.

Outros fenômenos documentais existem e dão origem à leis empíricas as quais servem de critério de eficiência para o tratamento da informação. Entre os fenômenos que podem ser ajustados, por exemplo, a uma distribuição hiperbólica sejam citados:

(a) a relação hiperbólica, descoberta por Zipf |1|, entre a frequência de ocorrência de uma palavra em dado texto e a linha onde ela aparece.

(b) a relação entre a dispersão de artigos sobre determinado assunto e um conjunto de revista.

(c) a relação entre o número de autores que escrevem sobre determinado assunto e o número de periódicos que cobrem este mesmo assunto.

(d) a relação entre o número de **papers** publicados e o número de autores destes **papers**.

1.2 Definições e conceitos bibliométricos

Todas as distribuições empíricas acima são objeto de estudo da Bibliometria. Neste ponto, contudo, o interesse neste trabalho não é apresentar as leis bibliométricas, mas tão somente extrair da Bibliometria um conjunto de elementos conceituais imprescindíveis, sobretudo, ao desenvolvimento do sexto capítulo. Portanto, as definições e conceitos que seguem serão utilizados para exemplificar a análise de sistemas quando ela abrange aspectos de recuperação documental.

Definição 1.2.1 Coleção documental

Ao conjunto finito $N = (n_1, n_2, \ldots, n_N)$, composto de elementos tais como artigo de revistas, abstrato, livro, jornal, **paper**, dossier médico, etc., dá-se a denominação de coleção documental.

Definição 1.2.2 Vocabulário*

Ao conjunto finito $D = (d_1, d_2, \ldots, d_D)$, composto de palavras-chave ou termos de conteúdo informático que servem para caracterizar os elementos de N, define-se como sendo um vocabulário.

---

* Também chamado de léxico documental, dicionário ou glossário.

Definição 1.2.3 Linguagem documental

É um vocabulário acompanhado de uma estrutura ou sistemática de manipulação deste mesmo vocabulário, com o objetivo de interface (comunicação) entre um usuário e um **data base** documental.

Definição 1.2.4 Frequência de ocorrência de termo

Define-se como

$$F(d) = \sum_{K=1}^{N} f_k^d$$

a frequência absoluta de ocorrência do termo d na coleção documental N.

Definição 1.2.5 Frequência de coocorrência de dois termos

Seja a seguinte matriz:

		Termos associados aos documentos					
		d_1	d_2	d_3	d_4	d_5	d_6
Documentos	n_1	3	0	0	2	0	6
	n_2	0	0	1	3	2	0
	n_3	0	2	3	0	4	0
	n_4	1	2	1	0	3	1

Nesta matriz as palavras-chave d_1 e d_6 estão ambas associadas aos documentos n_1 e n_4 embora com pesos ou frequên

---

* Denominada ainda de **thesaurus** ou linguagem de indexação.

cias diferentes (pesos 3 e 6 no primeiro documento e pesos 1 e 1 no segundo documento). Diz-se, então que os termos (d_1, d_6) coocorrem nos documentos n_1 e n_4.

Define-se, então, a frequência de coocorrência dos termos d_x e d_y como sendo

$F(d_x, d_y)$ = o número de duplas distintas que é possível formar com as palavras-chave (d_x, d_y).

Definição 1.2.6 Esperança de coocorrência de dois termos

Seja F(i) a frequência de ocorrência do termo i. Supondo que o conteúdo semântico de duas palavras-chave não intervenham em suas coocorrências e que todas tenham a mesma probabilidade de representar um documento de N, define-se a esperança de coocorrência de duas palavras-chave como sendo

$$E(d, i) = \frac{F(d)\ F(i)}{N}$$

Definição 1.2.7 Coeficiente de similitude de dois termos

Define-se um coeficiente de similitude S(d, i) entre os termos d e i como sendo uma função simétrica, ou não, da frequência de coocorrência F(d, i). O coeficiente S(d, i) é um valor que serve como fator de associação entre as palavras-chave (d, i).

Definição 1.2.8 Coeficiente simétrico e não simétrico

Se $S(d, i) = S(i, d)$ o coeficiente é simétrico.
Se $S(d, i) \neq S(i, d)$ o coeficiente é não simétrico.

Exemplos de coeficientes simétricos

(i) $S(d, i) = F(d, i)$

(ii) $S(d, i) = \dfrac{F(d, i)}{F(d) + F(i) - F(d, i)}$ → Função de TANIMOTO

Exemplos de coeficientes não simétricos

(iii) $S(d, i) = \dfrac{F(d, i)}{F(i)}$

(iv) $S(d,i) = \dfrac{F(d, i) - E(d, i)}{F(d) + F(i) - F(d, i)}$ → Função de TANIMOTO corrigida

Definição 1.2.9 Relações entre classes de documentos

Se a palavra-chave d_x for associada a alguns dos documentos de N e não a outros, cria-se neste momento duas classes de documentos. A classe X dos documentos associados à palavra-chave d_x e a classe X' dos documentos restantes não associados. Crie-se, agora, a classe Y de documentos indexados pelo termo d_y. Valem, então, para estas classes documentais as seguintes relações formais e operações governadas pela Álgebra Booleana.

(i) Complementaridade de classe

Se $X \cup X' = N$

e

$X \cap X' = \phi$

então X e X' mantêm relação de complementaridade, ou seja, qualquer documento de N pertence ou à classe X, ou à classe X'.

(ii) Soma lógica de classe

Define-se a soma lógica de classes como o conjunto dos documentos que pertencem ou à classe X dos documentos associados a d_x, ou à classe Y dos documentos associados a d_y, ou a X e Y ao mesmo tempo.

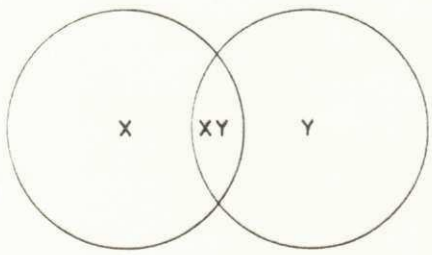
(iii) Produto lógico de classe

Define-se o produto lógico de classes como sendo o conjunto dos documentos que pertencem a mais de uma classe, ou seja, pertencem à classe X e à classe Y. São representados por XY.

Ex.: Sejam,

X = a classe dos documentos que tratam de Documentação Científica.

Y = a classe dos documentos que tratam de Informática. Então



XY = é a classe dos documentos relacionados com Informática Documental*, cujo objeto de estudo é o tratamento não numérico da informação.

Definição 1.2.10 Sistema documental

Dada uma matriz M(D, N) de incidência, ou matriz de termo-documento, que serve para estabelecer uma ligação entre os dois conjuntos N e D, define-se como sistema documental o sistema constituido por

$$< N, D, M(D, N) >$$

Obs.: Os elementos de M(D, N) são representados por valores lógicos (0 ou 1), de tal modo que se o documento n_w não é descrito pela palavra-chave d_x o valor do elemento é 0.

No caso contrário, se o documento n_y é descrito pela palavra-chave d_z o valor do elemento é 1. Aqui M(D, N) é, portanto, uma matriz binária da seguinte forma:

$$\begin{array}{l} \\ \text{DOC. 1} \\ \text{DOC. 2} \\ \text{DOC. N} \end{array} \begin{array}{c} \begin{array}{cccc} \text{TERMO 1} & \text{TERMO 2} & & \text{TERMO D} \end{array} \\ \begin{pmatrix} d_1^1 & d_2^1 & \cdots & d_D^1 \\ d_1^2 & d_2^2 & \cdots & d_D^2 \\ d_1^N & d_2^N & \cdots & d_D^N \end{pmatrix} \end{array}$$

d_z^i = termo z no documento i

---

* Corresponde à Teoria de Recuperação de Informação, na terminologia americana.

Definição 1.2.11 Matriz de coocorrência*

Se teoricamente um sistema documental é constituído pela matriz binária M(D, N), na prática, porém, esta matriz pode ser desenvolvida sob a seguinte organização (organização sobre N):

	TERMO 1	TERMO 2	TERMO D-1
TERMO 1			
TERMO 2		$F(d_x, d_y)$	
TERMO D			

onde, para todo valor de x e de y, o número de duplas** dá a frequência de coocorrência dos conceitos d_x e d_y.

para $\begin{cases} x = 1, 2, \ldots, D-1 \\ y > x \end{cases}$

Uma matriz assim definida é uma matriz de coocorrência.

Definição 1.2.12 Matriz de similitude (ou de associação de termos)

Na definição 1.2.11 fazendo $F(d_x, d_y) = S(d_x, d_y)$ obtem-se a denominada matriz de similitude que especifica, para cada par de palavras-chave, um correspondente fator de

---

* Também denominada de matriz termo/termo.

** Haverá tantas duplas quantas sejam possíveis formar com os conceitos (d_x, d_y).

associação. Os elementos desta matriz dados por $S(d_x, d_y)$ têm as seguintes características:

(1) $0 \leq S(d_x, d_y) \leq 1$

(2) $S(d_x, d_x) = 1$

(3) $S(d_x, d_y) = S(d_y, d_x)$, quando fôr adotado um coeficiente simétrico.

(4) $S(d_x, d_y) \neq S(d_y, d_x)$, quando fôr adotado um coeficiente não simétrico.

CAPÍTULO 2

ANÁLISE APLICADA À GESTÃO DOCUMENTAL

2.1 Características do método sistêmico

O nível de generalização já conseguido pela Teoria Geral de Sistemas (TGS, Ciência de Sistemas) permite definir um sistema S como sendo uma estrutura onde se fazem presentes os seguintes atributos ou componentes:

$$< E, C, T, F : X \rightarrow Y >$$

onde

$E = \{e_1, e_2, \ldots, e_n\}$ → é o conjunto de agentes, processadores ou elementos ativos que integram S.

$C = \{c_1, c_2, \ldots, c_m\}$ → é o conjunto das possíveis configurações ou estados de S.

$T = \{t_1, t_2, \ldots, t_k\}$ → é o conjunto dos instantes de tempo t_i utilizados para definir uma cronologia para a ação de S.

$X = \{x_1, x_2, \ldots, x_p\}$ → é o conjunto dos **inputs*** de S.

$Y = \{y_1, y_2, \ldots, y_q\}$ → é o conjunto dos **outputs**** de S.

F = é uma função definida no conjunto dos **inputs** de S e que os transforma em **outputs** mediante as atividades dos elementos e_i.

Colocada nestes termos, a definição de sistema se torna suficientemente geral, de tal maneira a poder ser utilizada adequadamente pelas diversas ciências, quer tratem elas de sistemas físicos, químicos, biológicos, computacionais, econômico-sociais ou sócio-culturais.

Ao emprestar esta definição, outros conceitos e seu **approach** decorrente, a Teoria Geral de Sistemas contribui e faz o interface com as demais ciências mediante, sobretudo, seus ramos aplicados: a Análise de Sistemas e a Engenharia de Sistemas. Pela Análise, os sistemas particulares são estudados e caracterizados com vista à obtenção de uma solução ótima. Mediante a Engenharia de Sistemas (auxiliada pela Teoria da Regulação, Teoria da Comunicação, Teoria dos Algorítmos, Teoria de Autômata) os sistemas analisados são, então, montados, implementados e controlados.

---

* **Inputs** ou entradas são influências, estímulos (físicos ou informacionais) que proporcionam ao sistema o material de sua operação.

** **Outputs** ou saídas são os resultados processados pelo sistema.

A Análise de Sistemas é, pois, um método que tem como objetivo estudar um sistema ou subsistema como uma unidade de atividade organizada, integrada e funcional, tomando como instrumento, técnicas analíticas e parâmetros tais como **inputs** do sistema, **outputs**, transações, processos, tempo, custo e critérios de otimização*. São objetivos, portanto, bastante amplos que vão além do reconhecimento e ordenação de dados sobre uma estrutura em análise. Evidentemente, os desafios enfrentados pela Análise de Sistemas são tanto maiores quanto mais complexas forem as unidades de atividade onde ela deva ser aplicada. É esta complexidade, inclusive, o que determina a recorrência à técnicas analíticas e teorias de outros campos, sobretudo da Matemática, Pesquisa Operacional e Ciências de Gerência, conforme o quadro de contribuição abaixo**:

ANÁLISE DE SISTEMAS E ÁREAS AFINS

ÁREA CONTRIBUINTE	SUB-ÁREA CONTRIBUINTE
a. Matemática	Teoria dos Grafos Cálculo Matricial Cálculo Relacional Cálculo Informacional ***
b. Pesquisa Operacional (PO)...	Teoria dos Modelos Programação Linear Programação Não-Linear Teoria de Otimização Simulação
c. Ciências de Gerência	Teoria da Decisão Teoria dos Jogos Teoria das Filas Análise Marginal

* A Análise, portanto, pode ou não, ser dirigida para o processamento de dados. Neste trabalho ela se orientará para processos de automação.

** Obviamente foram excluídas do quadro a TGS e a Engenharia de Sistemas.

*** Todas as técnicas matemáticas da Teoria da Informação.

Além destas técnicas a Análise se utiliza ainda de outras, mais ligadas à área onde deva ser aplicada, por exemplo, a Bibliometria se a área é a de Recuperação de Informação.

Quando empregados corretamente pela Análise, estes instrumentos de trabalho são capazes de representar a totalidade das operações de uma unidade de atividade (estrutura), através de um modelo conceitual que identifique subproblemas e suas inter relações. A meta da Análise de Sistemas será atingida quando ela conseguir integrar as soluções individuais encontradas para os vários subproblemas na solução geral do sistema em questão.

As características de uma boa análise que leve em consideração os conceitos da Teoria Geral de Sistemas estão abaixo enumeradas[1].

(1) a abordagem de sistemas exige um nível de detalhamento e de precisão o qual será difícil, ser tido como exagerado.

(2) cada sistema é um subsistema de um sistema mais amplo e cada sistema em si é composto por um número de subsistemas. Portanto, todos os sistemas pertencem tanto a uma micro-hierarquia como a uma macro-hierarquia.

(3) é impossível encontrar a solução de um sistema ou subsistema sem considerar todos os segmentos dos vários problemas. A consideração do todo é que faz evitar o fenômeno de "sub-otimização", que ocorre

---

1 In Robert W.B.Júnior. Biblioteca e Enfoque Sistêmico. R. Esc. Bibliotecon. UFMG, Belo Horizonte, 1(2): 164-83, set. 1972.

quando um subsistema componente opera de modo ótimo, em detrimento do sistema como um todo.

(4) não há nenhuma medida única e definitiva para a eficácia de um sistema, apenas "ótimos" circunstanciais, cada um dos quais devendo ser comparado a todas às outras opções possíveis detectadas pela Análise.

(5) todo enfoque de sistema é por natureza reiterativo; cada repetição sucessiva realiza-se no mesmo nível ou em níveis diferentes.

(6) o controle e o feedback contínuo e de boa qualidade são componentes essenciais da Análise de Sistemas.

(7) por definição, todos os sistemas têm de existir dentro de um meio ambiente. Os fatores ambientais são aqueles que afetam ou estão relacionados com o sistema em discussão não sendo eles, porém, uma parte deste sistema.

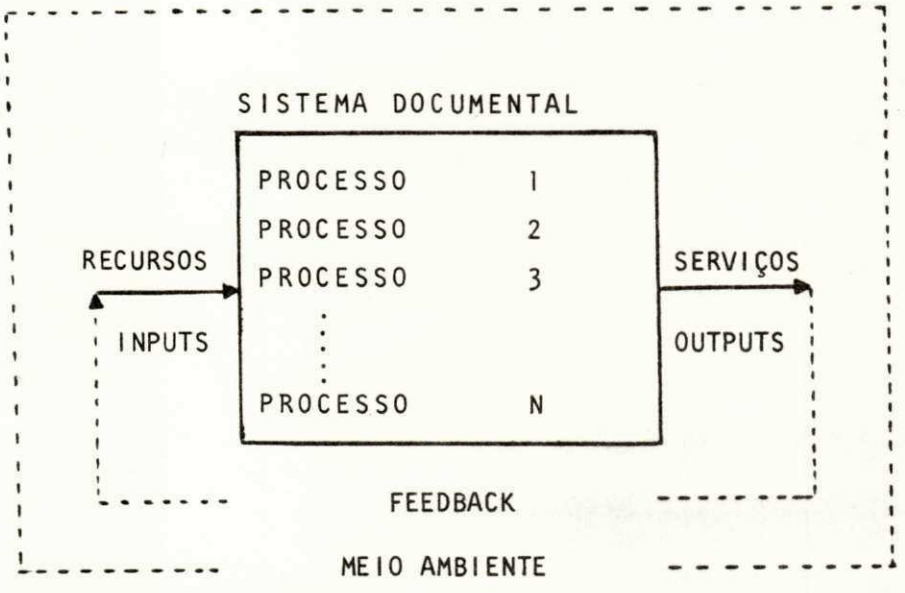
(8) dependendo do sistema particular, o excesso de quantificação pode conduzir a um ponto inconsistente na prática do sistema total. Existem fatores imponderáveis (fatores humanos) que além de não se prestarem à quantificação, podem levar a análise formal e o sistema ao fracasso, ou ao mau funcionamento.

(9) a documentação do ciclo de vida do sistema é uma parte tão essencial na prática de sistema quanto à análise, e ignorar este aspecto é também um convite ao fracasso.

(10) nunca há uma fase final no emprego de métodos sistêmicos, apenas iterações.

2.2 Subsistemas de uma biblioteca

Exposta esta idéia de sistema, como ainda suas implicações, resta agora transportar o conceito para o âmbito da Documentação Científica e da Informática Documental. Já no capítulo sobre Bibliometria foi definido um sistema documental em termos do conjunto N de documentos, do conjunto D de palavras-chave e da matriz M(D,N). No entanto, a linguagem matemática utilizada simplificou, em muito, o significado de um sistema documental. É preciso portanto que, aqui ele seja enriquecido e explicitado em função de seus **inputs**, principais processos e **outputs** resultantes. Ora, este é precisamente o papel da Análise de Sistemas. Supondo um sistema documental* como representado abaixo, à Análise caberá determinar e estudar:



---

* - Fisicamente o sistema é representado por bibliotecas, centros de dados, centros de documentação, bancos de dados, centros de análise de informação. Também, como um documento (no presente caso) é a mesma coisa que uma informação documentada, será empregado neste estudo, indiferentemente, sistema documental ou sistema de informação.

a) os **outputs** totais (objetivos) do sistema documental
b) os **inputs** (recursos econômico-financeiros, recursos humanos, recursos informacionais, recursos tecnológicos)
c) o meio ambiente do sistema
d) os processos documentais* relativos ao controle ou gestão dos documentos
e) os processos documentais relativos à análise e recuperação da informação
f) os processos de planejamento e avaliação determinantes da configuração ou estado do sistema

Por conseguinte, a Análise descreve e representa conceitualmente cada componente de per si, através do detalhamento dos processos envolvidos, da especificação de variáveis e parâmetros e das relações funcionais que possam ser estabelecidas.

Em um sistema como o do diagrama acima, o ponto inicial de suas atividades é a coleta dos dados para a entrada na cadeia de processamentos** . Estes dados são os **inputs** informacionais do sistema. Por simplificação, se forem excluidos os dados oriundos de mapas, partituras, **slides** ,quadros artísticos, restarão como principal entrada os "dados impressos" ou a informação documentada. A população de documentos impressos (informação documentada), ou coleção de registros*** armazenados e destinados

---

* - Todos os processos constituem funções dos elementos ativos do sistema.

** - Cadeia de processamento ou "cadeia documental" é o conjunto das funções documentais que se sucedem numa certa ordem determinada e sempre semelhante em qualquer que seja o sistema documental.

*** - Tais registros quando destinados ao "data base"documental devem ter forma legível pela máquina

a providenciar serviços de informação* é pois, o input fundamental de um sistema de informação (sistema documental). A partir da aquisição destes dados, a cadeia documental em um sistema tipo biblioteca pode ser dada pelas seguintes etapas componentes:

A. Aquisição dos documentos
   1. localização externa/ou geração interna
   2. seleção
   3. ordens de compra
   4. recebimento
   5. divulgação

B. Processamento inicial
   1. padronização ou adaptação dos novos dados ao sistema
   2. criação de registros correspondentes

C. Catalogação/análise
   1. Análise Derivativa (catálogo descritivo, resumo, indexação de palavras)
   2. Análise Associativa (classificação, indexação controlada de assunto

D. Arquivamento/preservação

E. Circulação/Pesquisa retrospectiva

---

* - Informação é qualquer elemento capaz de reduzir a incerteza e de aumentar a precisão em um dado sistema.

Do ponto de vista da análise orientada para a automação, todas estas atividades documentais e administrativas em um sistema-biblioteca podem ser grupadas em cinco categorias básicas de subsistemas conforme a natureza de cada atividade. São eles:

(a) Subsistema de Aquisição
(b) Subsistema de Catalogação
(c) Subsistema de Circulação
(d) Subsistema de Recuperação de Informação
(e) Subsistema de PlanejamentoEAvaliação

Os Subsistemas (a), (b) e (c) congregam as atividades relativas aos processos de gestão e controle dos documentos como seus próprios nomes indicam. Esta administração e controle sobre a coleção de documentos se viabiliza, principalmente, pela elaboração de catálogos, construção e manutenção de listas e registros de vários tipos. Mais precisamente estes subsistemas tratam de:

1. processamento da ordenação dos documentos
2. processamento de aquisição
3. recebimento dos documentos na biblioteca
4. controle da circulação (empréstimo, devolução, reserva)
5. elaboração de listas de aquisição
6. elaboração de índices e listagens de assuntos
7. elaboração de listas de estantes
8. elaboração de boletim de abstrato, etc.

Estas operações são também chamadas de operações de housekeeping.

O subsistema (d)* agrupa e implementa aquele conjunto de atividades já pertencentes à área da recuperação de informação e ao processamento do conteúdo tais como:

1. referência e pesquisa de informação por autor, por título, por assunto, por especificação do conteúdo, etc.
2. disseminação seletiva da informação
3. indexação automática, extração de palavras-chave em textos, resumos e títulos
4. construção de abstratos (ou extratos)
5. construção de thesauri ou classificações (linguagens documentais) hierárquicas ou não.

Quanto ao subsistema (e) de planejamento e avaliação ele se ocupa dos seguintes processos:

(a) Processos operacionais (utilizando o processamento eletrônico)
   1. orçamento do sistema, fundos, verbas para aquisições e convênios
   2. contabilidade do sistema
   3. folha de pagamento do pessoal da biblioteca
   4. contas a pagar
   5. multas sobre usuário, etc.

(b) Processos de avaliação técnica

(c) Processos de avaliação econômica (utilizando a Análise de Custo-Benefício)

Em tese, é este o domínio da Análise de Sistemas no âmbito do processamento de biblioteca ou de outra unidade de ativi

---

* - Os processos deste subsistema além de não serem amplamente implementados, possuem ainda um caráter altamente experimental.

dade que processe documentos. Na parte final deste capítulo,em 2.3, serão desenvolvidos mais alguns aspectos teóricos dos subsistemas de aquisição, catalogação e circulação. Em seguida, nos capítulos 3 e 4 serão considerados respectivamente aspectos relativos à Teoria de Recuperação de Informação (subsistema de recuperação) e a elementos de avaliação técnica e econômica (subsistema de planejamento).

2.3 - Modelos **off-line** e **on-line** em biblioteca

(a) Subsistema de Aquisição - este subsistema cria e desenvolve a população de documentos de um sistema documental, numa forma contínua e segundo uma política de planejamento das aquisições. Estabelecendo critérios para a seleção do material a ser armazenado, é dele que depende (considerando-se,obviamente, o orçamento e as disponibilidades do mercado de livros) a efetividade do sistema em termos de velocidade de crescimento e tamanho da coleção. Suas operações, de maneira simplificada, podem ser visualizadas pelos passos do diagrama 2 seguinte.Inicia-se o funcionamento do subsistema quando à biblioteca é dirigida uma solicitação de documento.

É feita uma pesquisa bibliográfica para se certificar da existência ou não de uma encomenda envolvendo a solicitação inicial. Caso não exista ainda tal solicitação prepara-se uma ordem de compra e o fluxo segue como mostrado no diagrama 2. Se os processos envolvidos forem desdobrados, quatro tipos de operações podem ser originados:operações de processamento de solicitação, operações de preparação de ordem de compra, operações de atualização de arquivo e operações de processamento fiscal. A tabela 01 extraída de [1] ilustra estas operações.

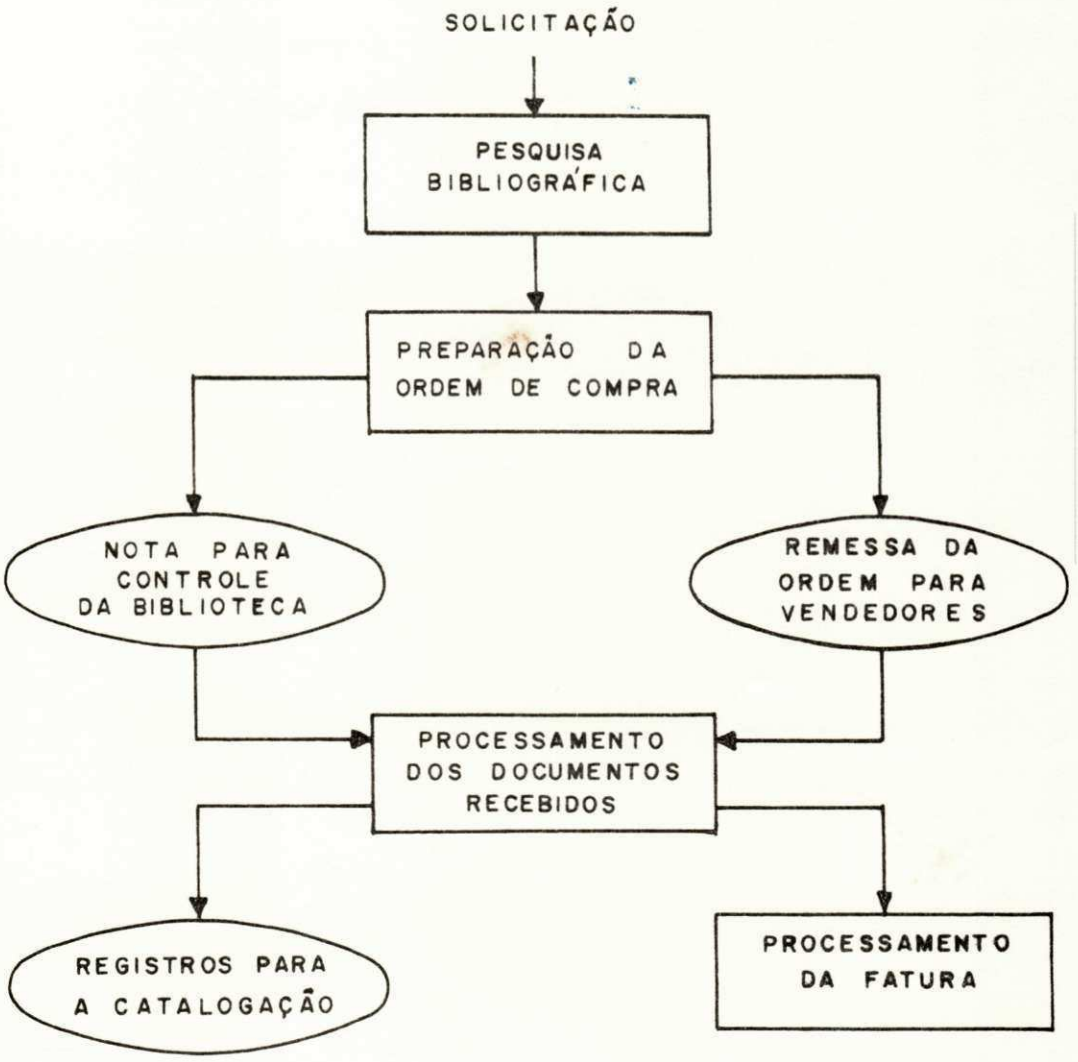


DIAGRAMA 2 — SUBSISTEMA DE AQUISIÇÃO/EXTRAIDO DE [1]

TABELA 01 - Operações de Aquisição de documentos

1 - Processamento de pedidos

- geração da solicitação do livro ou documento
- entrada no arquivo de encomenda de documentos
- pesquisa bibliográfica para examinar a situação do documento solicitado
- pesquisa do arquivo de encomenda para exame da prioridade da solicitação
- decisão se deve ou não ser procedida a ordem de compra
- escolha da verba (fundo) e do vendedor ou editora
- geração da nova ordem de compra e entrada no arquivo de encomenda

2 - Elaboração da ordem de compra

- criação do número de ordem
- checagem da verba e seu correspondente comprometimento
- atualização do arquivo de vendedores pela inclusão do novo item de compra
- preparação de cartões de alterações para criar informes sobre a situação da encomenda no sistema.
- "sorting" das ordens por autor, título e por entrada no arquivo de encomendas

3 - Processamento de arquivo

- processamento periódico dos cartões de alterações
- impressão periódica de listas de encomendas, notícias para usuários solicitantes e registros de catalogação.

4 - Processamento fiscal

- processamento de registros para a fatura, alterações fiscais e pagamento.
- atualização do arquivo de verbas e listagens da situação financeira
- registro das alterações de faturas
- preparação de reclamações e cancelamentos

Portanto, o subsistema de aquisição manipula, pelo menos, quatro arquivos principais:

(a) arquivo de documentos sob encomenda

(b) arquivo de vendedores (editoras)

(c) arquivo de verbas e fundos

(d) arquivo de faturas

Se o sistema tem processamento off-line estes arquivos são processados periodicamente, de acordo com a dinâmica do sistema, e é gerado a partir de então um conjunto de registros de controle. A atualização dos arquivos e a preparação dos outputs estão mostradas no fluxo do diagrama 3 da próxima página.

Num sistema on-line, como todo processamento deste tipo, a principal vantagem seria a atualização contínua dos arquivos. Desta maneira não seria preciso, então o acúmulo de cartões de alterações que teriam de ser processados periodicamente. Todas as alterações e correções nos arquivos on-line são efetuadas no decorrer dos fatos, através do uso de terminais.

(b) Subsistema de Catalogação - o papel deste subsistema é providenciar a extração, organização e manipulação de elementos bibliográficos apropriados para o processo de identificação dos documentos. Os dados que entram para a catalogação são dados chamados derivados (por serem extraídos dos próprios documentos) e a catalogação resultante provém da análise derivativa*. Se no sistema existe integração de arquivos, grande par

---

* *Quanto à análise associativa, ela se confunde com os processos de recuperação de informação que serão posteriormente considerados.*

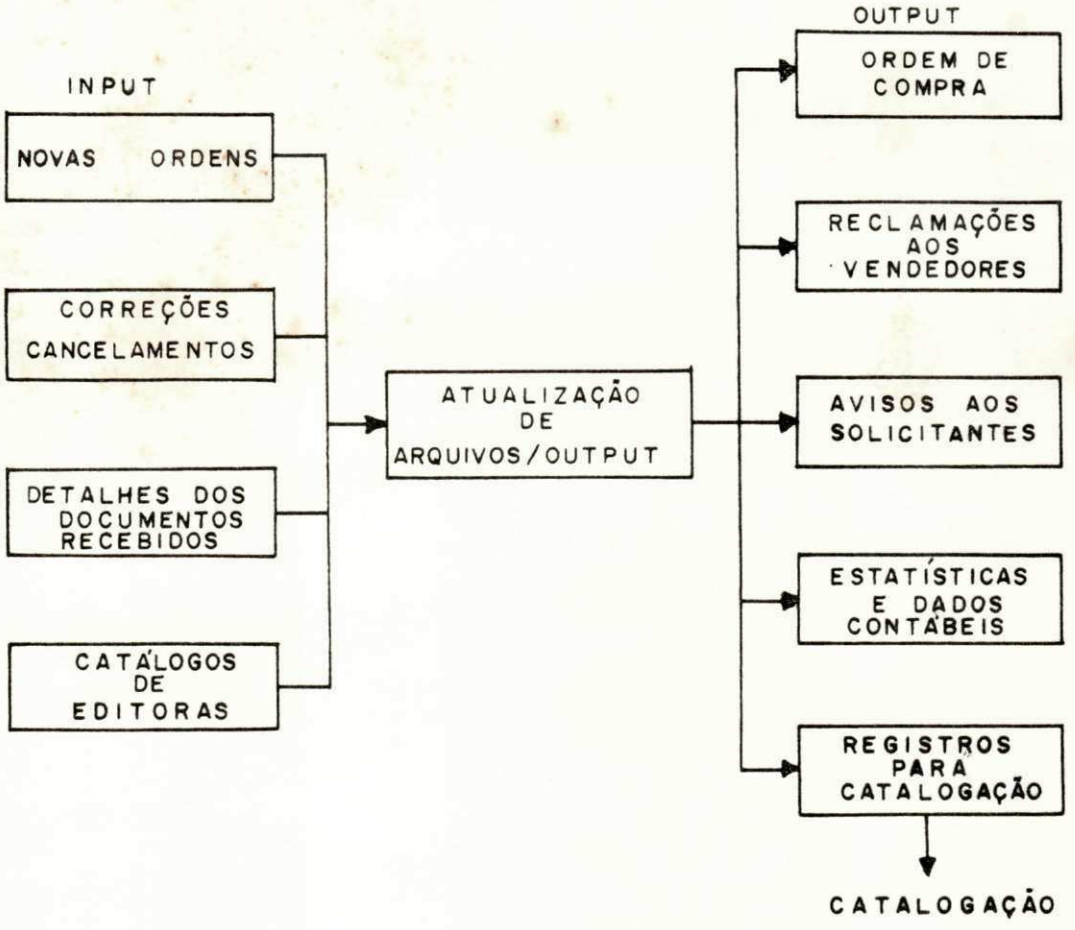


DIAGRAMA . 3 — ATUALIZAÇÃO/ORDEM DE COMPRA

te dos inputs deste subsistema haverá de se originar do subsistema de aquisição. Este fato está visível no diagrama 4, onde os principais processos de catalogação foram simplificados. Geralmente eles são constituídos de operações de edição, merge, sorting e impressão.

Embora um computador tenha outras atribuições neste subsistema, por exemplo, pesquisa bibliográfica e auxílio ao documentalista nas tarefas de alterações e acréscimos de informações, sua principal vantagem reside, contudo na produção de catálogos. Através da combinação dos inputs iniciais este subsistema se encarrega de fornecer uma gama de catálogos dos mais variados tipos. É de costume serem produzidos estes principais catálogos:

1. Catálogo topográfico (informações bibliográficas, nº de exemplares de cada uma das publicações ordenadas pelo nº de localização)
2. Catálogo alfabético (nome do autor, título da publicação, ano da publicação, código da localização)
3. Catálogo de assunto
4. Catálogo de periódico (informações bibliográficas)
5. Catálogos especiais (de congressos, de entidades, de teses, de papers, folhetos)
6. Índice KWIT (Key Word in Title)
7. Índice KWIC (Key Word in Context)
8. Índice KWOC (Key Word out of Context)

Estes índices produzidos com a mesma finalidade dos catálogos (instrumentos de procura e difusão de informação) são originados a partir dos dados memorizados e sem intervenção do documentalista. Construídos sobre palavras-chave seguem, em

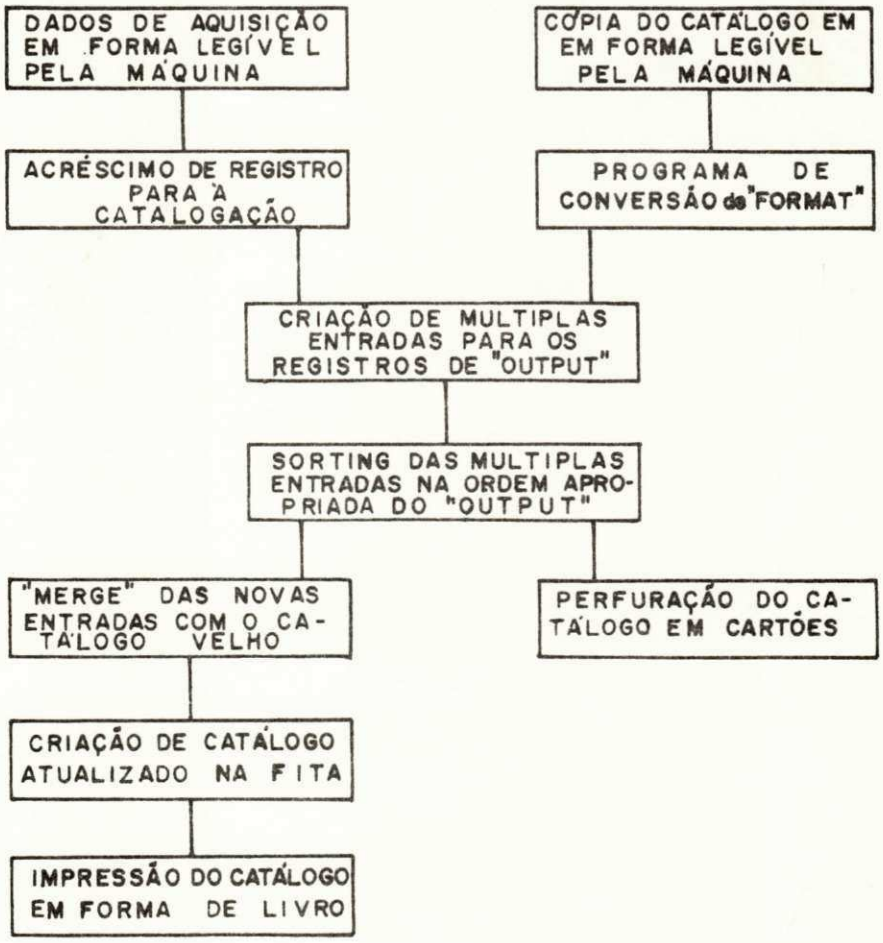


DIAGRAMA 4 — SUBSISTEMA DE CATALOGAÇÃO/
ADAPTADO DE [1]

princípio, dois passos - elaboração de um antidicionário (ou dicionário negativo) para incluir as palavras sem sentido informático e a permuta automática dos termos significativos.

(c) Subsistema de Circulação - as comunicações de um sistema documental com seus usuários são tarefas do subsistema de circulação ou de usuários. Dadas suas características de relacionamento com as pessoas este subsistema incorpora muitos elementos originados do meio ambiente. Por exemplo, o tipo de usuário (que depende do ambiente) determina as características das transações do subsistema. Fundamentalmente, porém, as operações de circulação consistem em criar registros individuais de tal modo a estabelecer relações entre informações sobre usuários e informações sobre documentos emprestados.

Num sistema off-line, com pelo menos dois arquivos - arquivo de usuários e arquivo de documentos - o fluxograma das operações está mostrado no diagrama 5. Já numa configuração on-line os esquemas de circulação podem ter os modelos esboçados nos diagramas 6 e 7. No diagrama 6 tem-se uma simplificação das atividades do subsistema, enquanto no diagrama 7 as operações foram detalhadas dando origem a um modelo conversacional mais realístico.

Em termos de implementação, uma estratégia para o subsistema de circulação consistiria em realizar as operações de indagação e de respostas em tempo real e deixar os demais outputs (reclamações, avisos etc.) para serem processados no modo batch. Como consequência dessa decisão, seria exigido mais espaço de armazenamento para as transações em batch e seria reduzido o tempo das operações on-line.

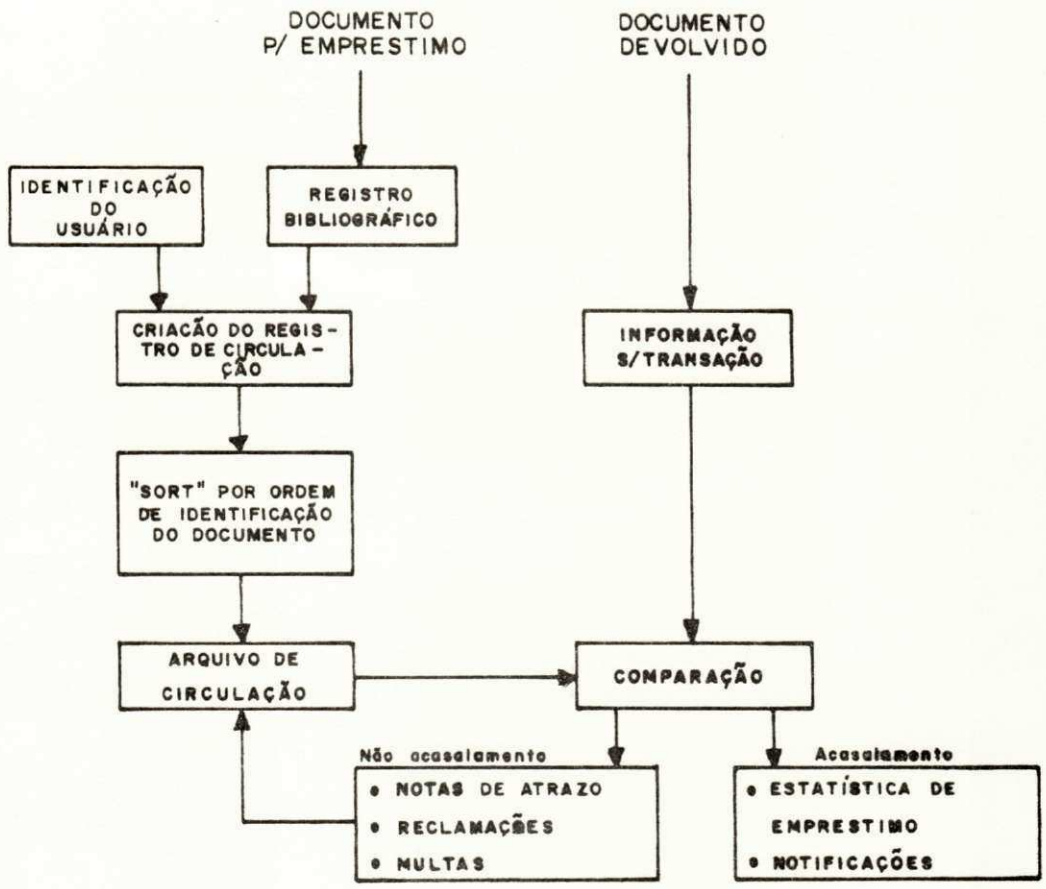


DIAGRAMA 5 - TRANSAÇOES DE CIRCULAÇÃO EXTRAIDO DE [1]

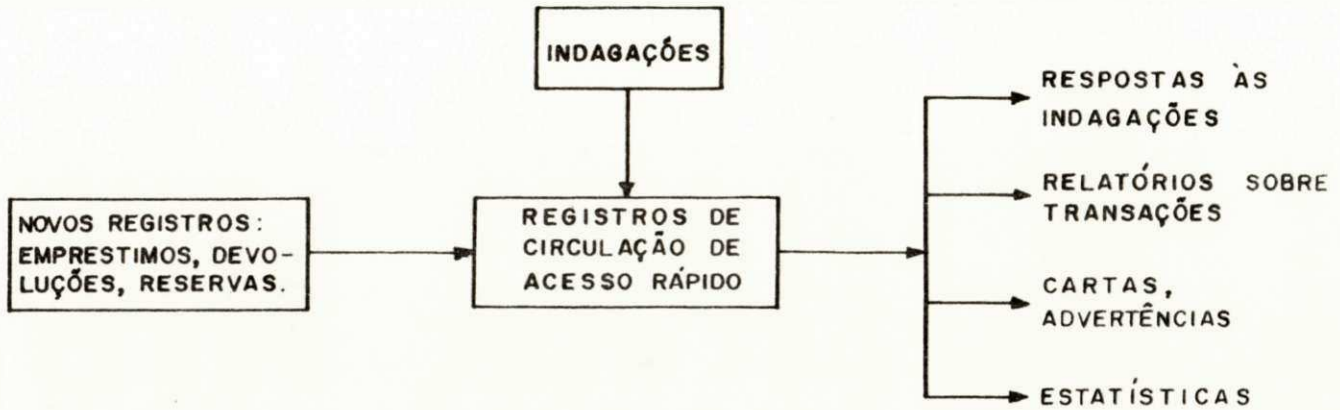


DIAGRAMA 6 — CIRCULAÇÃO SIMPLIFICADA

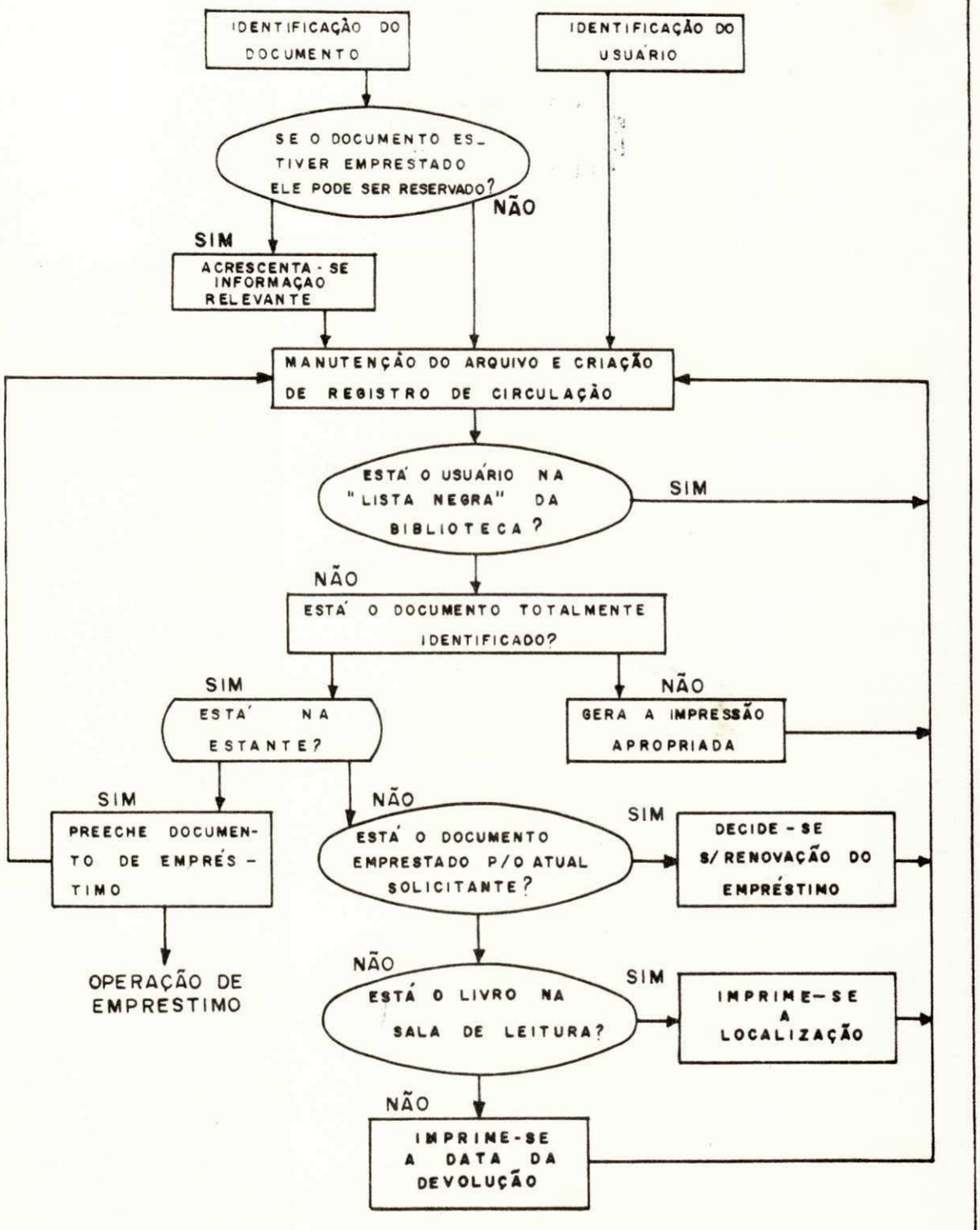


DIAGRAMA 7 - CIRCULAÇÃO DETALHADA EXTRAÍDO DE [1]

CAPÍTULO 3

ANÁLISE APLICADA À RECUPERAÇÃO DOCUMENTAL

3.1 Usuário e recuperação de informação

A Teoria de Recuperação de Informação presentemente distingue dois tipos de sistemas de informação:

(a) sistemas que provêm dados,

(b) sistemas que provêm referências.

Os sistemas que provêm dados se caracterizam pelo fato de responderem à questões sobre dados solicitados da maneira mais específica possível. São os sistemas mais utilizados pela alta administração das empresas e que na prática são chamados de MIS (Management Information Systems). Os sistemas que provêm referências, ao contrário, se caracterizam pelo fato de suas respostas não se constituirem um dado ou um fato específico, e sim uma área específica. Tal como num intervalo de confiança, existe uma folga nas respostas dos sistemas que providenciam referências. A estes sistemas de referência, cujo **output** é uma coleção de documentos, é que vem sendo dado neste trabalho a denominação de sistema documental, em concordância com a corrente francesa de Recuperação de Informação.

Embora exista um relacionamento entre os dois tipos de sistemas eles são, contudo, diferentes:

1. na organização do armazenamento

2. na manipulação dos arquivos

3. nas etapas de processamento

4. no grau de dificuldade para a sofisticação*

5. no manuseio de perguntas e respostas do sistema, através de funções de acasalamento

Sem considerar, porém, estas diferenças o interesse aqui será voltado agora para a análise dos processos de recuperação de um sistema documental, após a abordagem em 2.3 daquelas suas atividades de caráter mais administrativo.

Inicialmente, seja tomado de um sistema documental o seu subsistema de recuperação de referências representado como no diagrama 8 que segue.

Como visto anteriormente, os **inputs** de um subsistema de recuperação de referências são os dados impressos ou a população de documentos sobre a qual atuará o computador e/ou os documentalistas. Pelo Diagrama 8, todos os documentos que entram para a fase de análise de conteúdo deverão ser caracterizados, em termos das palavras-chave contidas neles, ou de seus elementos descritivos. Realizada a extração destes elementos, eles passam a ser organizados sob a forma de índices, vocabulários e thesauri seguindo metodologia própria para cada um destes tipos de representação condensada de documento. A análise é, portanto, a principal etapa de um subsistema de recuperação já que,

---

* Os sistemas documentais possuem técnicas de sofisticação de manuseio mais difícil.

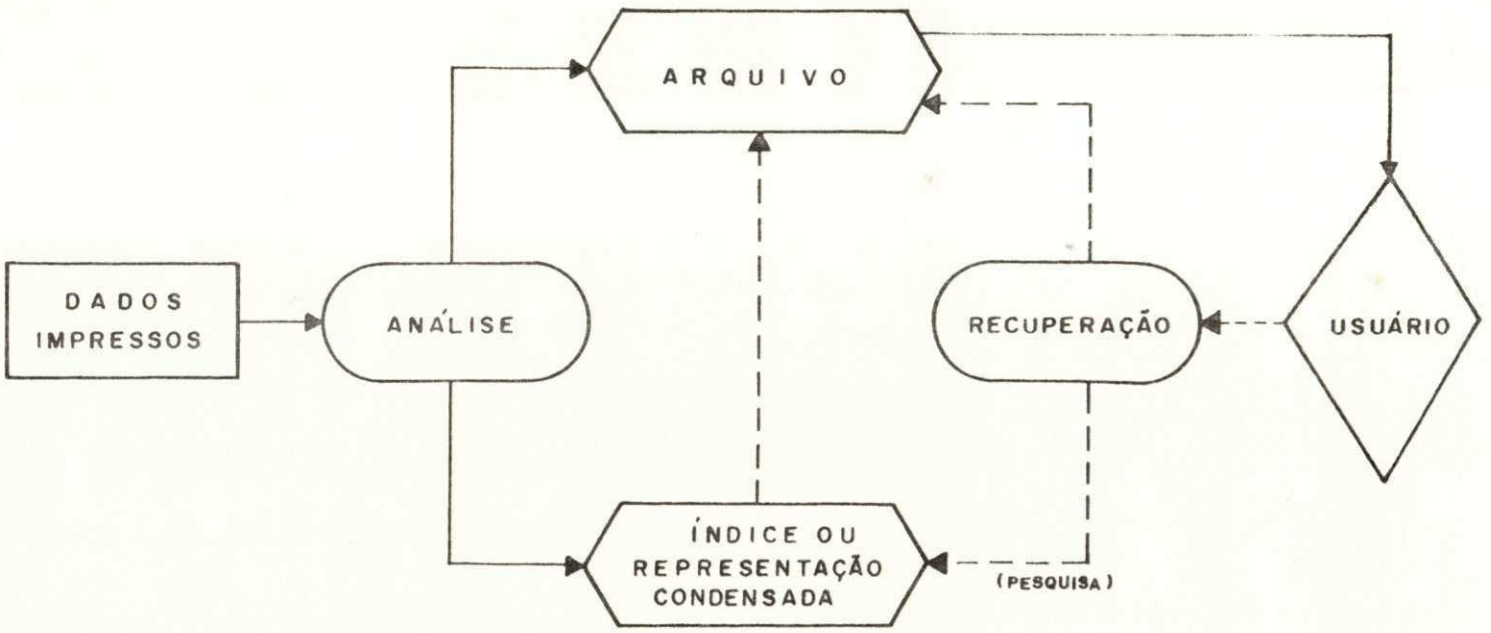


DIAGRAMA 8 – SUBSITEMA DE RECUPERAÇÃO SIMPLIFICADO
EXTRAÍDO DE [2]

dela dependem todas as outras. Os resultados obtidos na análise são então armazenados em arquivos magnéticos*, enquanto os documentos de entrada vão para os arquivos físicos (estantes) de acordo com uma ordem pré-estabelecida. Existem, portanto, em qualquer sistema (subsistema) de recuperação dois tipos de arquivos fundamentais criados após a análise: arquivos físicos e arquivos de representação condensada dos documentos ou arquivos substitutos.

O processo de busca de informação tem início com a ação do usuário do sistema documental. Para recuperar uma informação ele deverá entrar nos índices (thesauri) com aquelas palavras-chave que descrevam suas necessidades de informação afim de que seja realizado o acasalamento entre os elementos que exprimem as necessidades e os elementos constantes dos índices. Caso haja o acasalamento, os documentos que satisfazem ao usuário são então retirados dos arquivos. Nos sistemas simples porém, sem mecanismos de consulta, a busca é feita diretamente aos arquivos físicos. O sexto capítulo exemplificará as chamadas funções de acasalamento existentes nos sistemas de recuperação.

O fluxo mostrado contudo, ao descrever as etapas mais gerais dos processos de recuperação de informação não se aprofunda nesta descrição. Na prática os sistemas, frente às crescentes exigências do usuário, se tornam cada vez mais complexos.

---

* Nos fragmentados estudos brasileiros sobre Recuperação de Informação (geralmente, teses) quase se confunde e se reduz toda a problemática de recuperação ao seu aspecto de organização do armazenamento ou de estrutura de dados.

Kent em 1965 [2] elaborou o Diagrama 9 para mostrar o dilema do usuário diante da pesquisa bibliográfica e a crescente atribuição de funções feita por ele aos sistemas de referência bibliográfica. O esquema da próxima página mostra esta atribuição.

3.2 Operações de recuperação de informação

São, por conseguinte, as exigências do usuário do sistema documental que determinam as características e o nível de refinamento dos sistemas (subsistemas) de recuperação. Assim, por exemplo, levando-se em conta o nível tecnológico dos processos de recuperação desenvolvidos entre 1965-1967, o modelo de recuperação apresentado anteriormente (diagrama 8) pode agora ser detalhado de tal modo a mostrar a complexidade da análise de sistemas que se voltam para a busca documental. O universo das operações envolvidas é constituido de cinco tipos distintos de operações* [10]:

1º Tipo - Operações de análise e controle, incluindo

(1) seleção e aquisição dos documentos

(2) análise descritiva (ou derivativa) e especificações

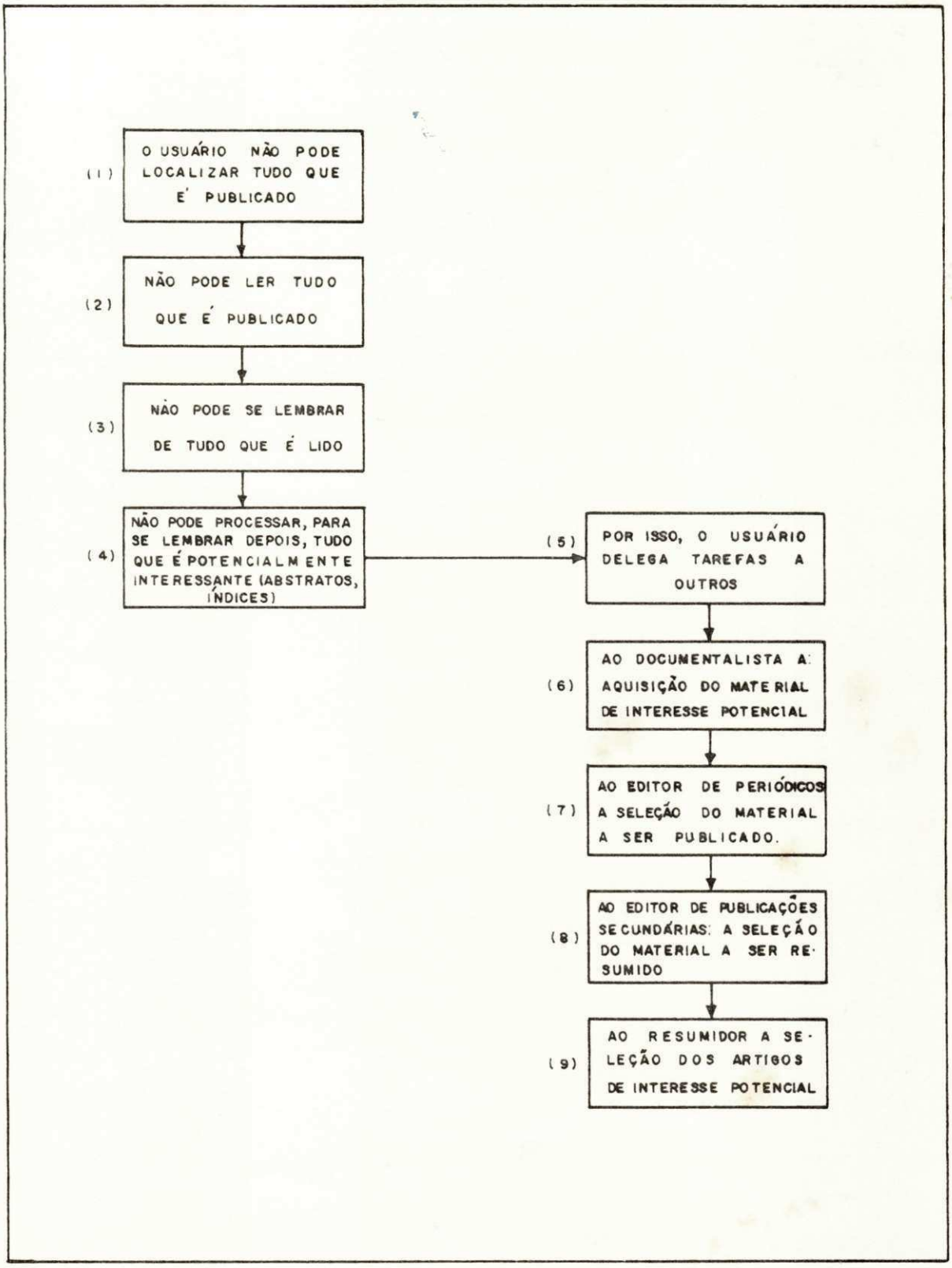
(3) análise de conteúdo e especificações

(4) construção de thesaurus

(5) validação ou verificação das palavras-chave ou descritores mediante uma tabela de verificação

---

* Salton desdobrou estas operações em quatro tipos: análise de informação, armazenamento de informação, operações de pesquisa e processamento da recuperação.


(1) O USUÁRIO NÃO PODE LOCALIZAR TUDO QUE É PUBLICADO
(2) NÃO PODE LER TUDO QUE É PUBLICADO
(3) NÃO PODE SE LEMBRAR DE TUDO QUE É LIDO
(4) NÃO PODE PROCESSAR, PARA SE LEMBRAR DEPOIS, TUDO QUE É POTENCIALMENTE INTERESSANTE (ABSTRATOS, ÍNDICES)
(5) POR ISSO, O USUÁRIO DELEGA TAREFAS A OUTROS
(6) AO DOCUMENTALISTA A AQUISIÇÃO DO MATERIAL DE INTERESSE POTENCIAL
(7) AO EDITOR DE PERIÓDICOS A SELEÇÃO DO MATERIAL A SER PUBLICADO.
(8) AO EDITOR DE PUBLICAÇÕES SECUNDÁRIAS: A SELEÇÃO DO MATERIAL A SER RESUMIDO
(9) AO RESUMIDOR A SELEÇÃO DOS ARTIGOS DE INTERESSE POTENCIAL

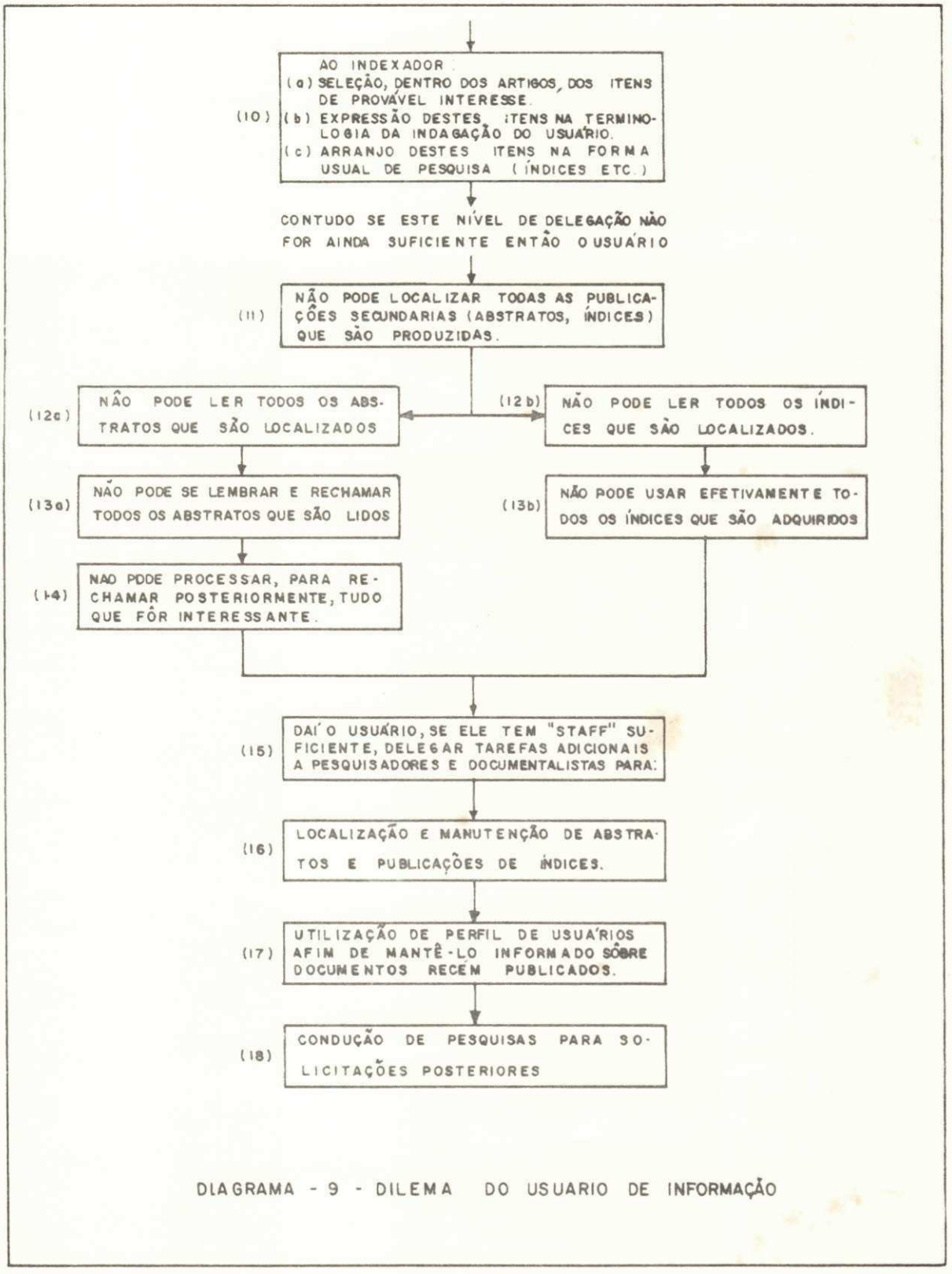


DIAGRAMA - 9 - DILEMA DO USUARIO DE INFORMAÇÃO

(6) acréscimos de registros de indexação, adicionando palavras-chave pré-determinadas que se relacionam

(7) computação e elaboração de relatórios para controle administrativo do sistema de recuperação

(8) administração do sistema

(9) análise do computador a ser utilizado, dos programas necessários, da estrutura de arquivo e do algorítmo de pesquisa

(10) análise dos documentos em relação à indagação

2º Tipo - Operações de substituição

(11) preparação de representações condensadas dos documentos (descritores, referências, abstratos)

(12) conversão dos descritores da linguagem natural para códigos internos

3º Tipo - Operações de transformações físicas

(13) transformação dos documentos originais em microfichas e/ou microfilmes

(14) conversão dos dados para formas legíveis pela máquina

(15) conversão das indagações para formas legíveis pela máquina

4º Tipo - Operações de processamento de arquivos

(16) criação e manutenção de arquivos físicos de documentos

(17) criação e manutenção de arquivos substitutos de documentos

(18) pesquisa a arquivos substitutos, acasalamentos e recuperação

5º Tipo - Operações de visualização das recuperações efetuadas através de listagens, visores, etc.

Em termos de fluxo de operações tal sistema de recuperação é mostrado, passo a passo, através do Diagrama 10.

Deste Diagrama se depreendem logicamente algumas conclusões importantes. Assim,

(a) muitas das atividades antes alocadas no nível da gestão documental (aquisição de documentos, análise descritiva, etc) passaram para o nível da recuperação da informação. Afinal, a gestão de documentos através de computador deve ser considerada como uma etapa preparatória para se atingir o nível mais integrado e complexo da recuperação documental. Quando a recuperação de informação é introduzida em um sistema ela açambarca todas as atividades anteriores.

(b) considerando que em cada bloco do diagrama são supostos ainda rotinas e processos próprios de execução, é errônea a idéia difundida de que é fácil tratar documentos computacionalmente.

(c) contrariamente à idéia também existente em centros universitários brasileiros*, não existe razão lógica

---

* Sejam examinados, por exemplo, os estudos e teses produzidos sobre Teoria de Recuperação de Informação.

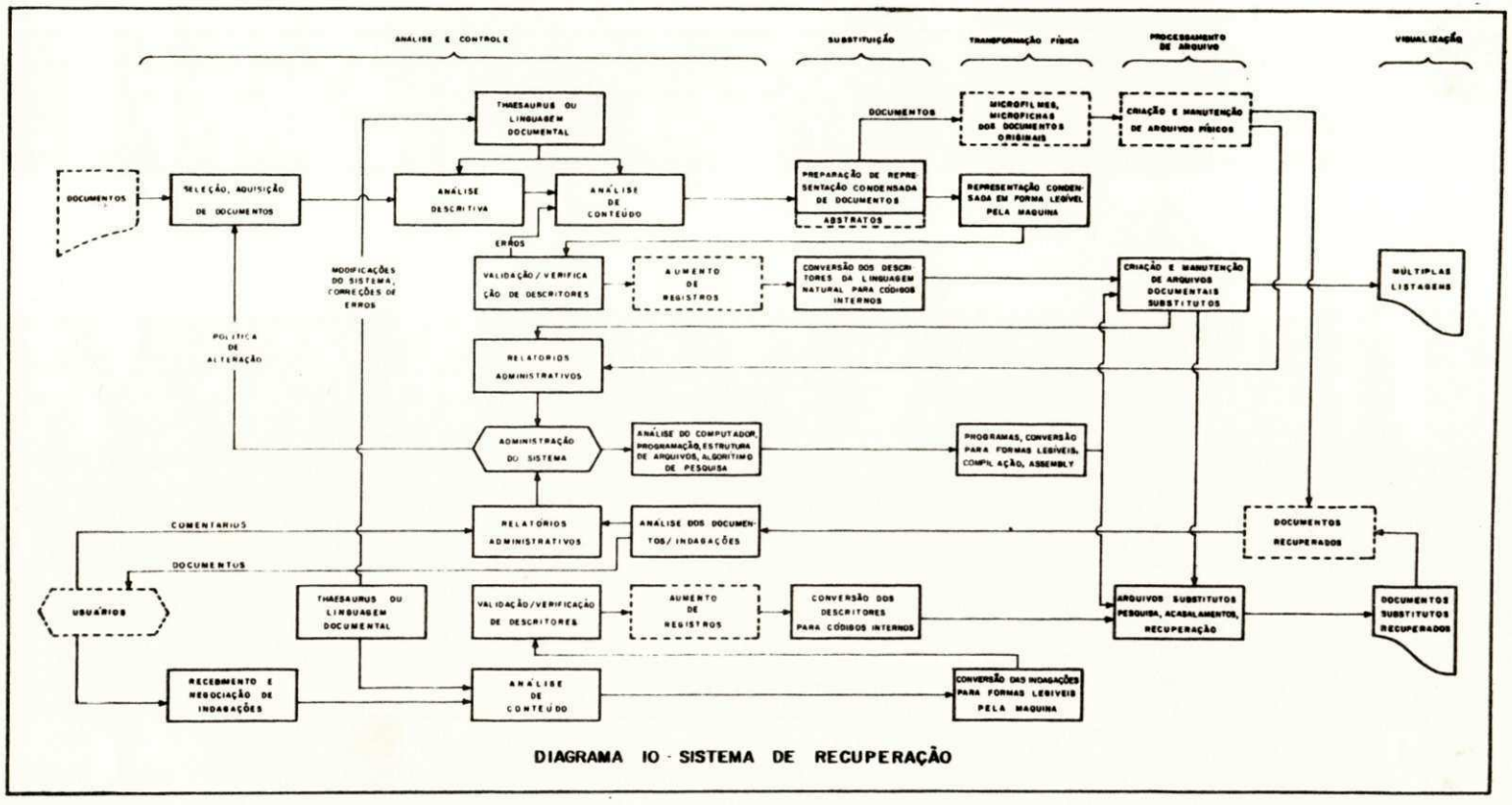


DIAGRAMA IO - SISTEMA DE RECUPERAÇÃO

alguma para superenfatizar, em meio a este conjunto de operações, o tratamento quase exclusivo daquelas questões relativas à organização do armazenamento (estruturas de arquivo) e estruturas de informação.

(d) no bloco relativo à administração do sistema estão subentendidas entre outras, todas as preocupações com a eficiência econômica e técnica dos sistemas de recuperação de informação. Dada a importância da análise destas questões, elas serão brevemente abordadas no próximo capítulo.

CAPÍTULO 4

ANÁLISE APLICADA À AVALIAÇÃO DE SISTEMA DOCUMENTAL

4.1 Avaliação técnica

No quadro complexo do tratamento não numérico da informação (antes visto) a Análise e a Engenharia de Sistemas não somente analisam e projetam os sistemas de informação, como ainda da proporcionam elementos de avaliação para estes mesmos sistemas. No início, a ênfase foi dada à avaliação técnica e ao exame do que era possível ser conseguido por estes sistemas. Era preciso encontrar parâmetros de avaliação que mostrassem não só a viabilidade técnica, como ainda as vantagens da automação documental sobre o trabalho artesanal dos documentalistas. Estes critérios começaram a ser estudados a partir de 1957 com os estudos de Perry e Kent |2|. Os avaliadores de sistemas de recuperação tomaram inicialmente como medidas de efetividade elementos tais como:

N = número de documentos existentes no sistema

L = número de documentos recuperados numa indagação ao sistema

C = número de documentos relevantes

R = número de documentos ao mesmo tempo recuperados e relevantes

Mediante o estudo destas variáveis procuraram descobrir e analisar:

(a) a razão de ser da recuperação de grande quantidade de documentos irrelevantes numa pesquisa empreendida, ou seja, o problema da baixa precisão da recuperação;

(b) o problema do ruído do sistema que ocorre quando documentos que não constam numa solicitação do usuário são, contudo, indevidamente recuperados. Tais documentos são chamados de "documentos parasitas";

(c) o problema do silêncio do sistema que ocorre quando o documento existe na memória do computador (geralmente em forma de palavra-chave), mas não é selecionado quando da interrogação.

Foram estes, basicamente, os problemas que deram origem em Recuperação de Informação aos estudos de avaliação da eficiência dos sistemas. Os parâmetros N, L, C e R no entanto se mostraram insuficientes para aquilatar a complexidade dos sistemas documentais e, a partir deles, foram definidos outros fatores de avaliação técnica mais expressivos tais como os que seguem:

1. Fator de precisão (**precision**)* = R/L
2. Fator de rechamada (**recall**) = R/C
3. Fator de resolução = L/N
4. Fator de eliminação = (N - L)/N
5. Fator de ruído = (L - R)/L
6. Fator de omissão = (C - R)/C

---

* Conhecido ainda como fator de pertinência ou fator de relevância.

Agora, a performance dos sistemas automáticos pode ser descrita com maior segurança mediante estes indicadores, sobretudo o de **precision** e o de **recall**. Os estudos, então, foram intensificados no sentido de descobrir os dispositivos capazes de melhorar a performance destes fatores e de descobrir as relações existentes entre eles. Viu-se, por exemplo, que as percentagens de **recall** e de **precision** num sistema conservam a relação dada pelo gráfico abaixo.

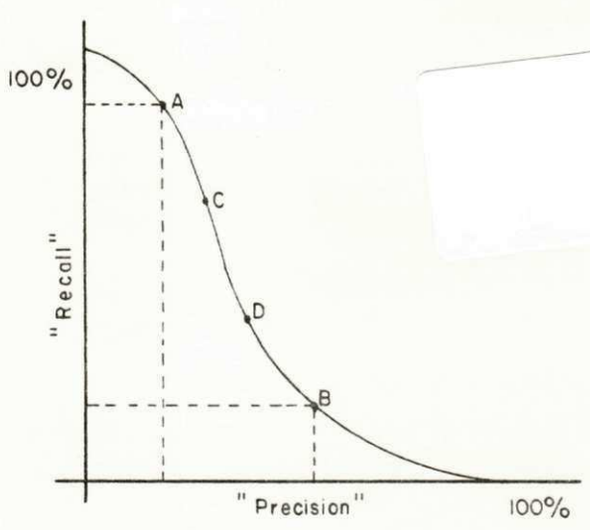


DIAGRAMA - II - GRÁFICO DE RELAÇÃO ENTRE FATORES

O poder do fator de precisão depende basicamente da especificidade do **thesaurus** (abrangência de assuntos) e o fator de rechamada depende do nível de exaustividade da indexação (profundidade dentro do assunto). De acordo com as necessidades do usuário do sistema ele escolherá a distância sobre a curva de performance acima, na qual deverá operar. Se optar por um índice muito específico, com indexação exaustiva poderá o sistema operar entre A e B. Já com um thesaurus menos específico e indexação menos exaustiva o sistema operará sobre um segmento da

curva menor, talvez entre C e D.

Ora, esta decisão pelos níveis de **recall** e de **precision** invariavelmente implicará em problemas de produção e atualização de dicionários e de **thesaurus**. E se mecanismos de melhora de **recall** e de **precision** forem introduzidos no **thesaurus** (**links**, regras, termos com pesos) surge agora a importante questão dos custos de desenvolvimento e dos custos operacionais dos sistemas documentais. Na verdade, só muito recentemente tem sido colocado pelos analistas o problema da avaliação econômica da automação da informação.

4.2 Avaliação de custo-benefício

O desenvolvimento da Economia do Computador como ramo aplicado da Teoria Econômica e da Ciência dos Computadores* é que tem levado a enfocar a massa de informação de um país como "recurso nacional". É um recurso que entra na função de produção** de uma economia ou de uma empresa mediante seu componente tecnológico. A informação é um recurso e um bem que não é livre (no sentido da Economia do Computador) nem do ponto de

---

* A Economia do Computador estuda o computador como um bem produzido pelo sistema econômico e como um fator tecnológico de produção de serviços.

** A função de produção é: $y = \phi(T_i, K_i, MO_i, N_i, I_i)$ onde:

$T_i = (t_1, t_2, \ldots, t_n)$ = Recursos naturais (Terra)

$K_i = (k_1, k_2, \ldots, k_n)$ = Recursos de capital (Capital)

$MO_i = (mo_1, mo_2, \ldots, mo_n)$ = Recursos humanos (Mão-de-obra)

$N_i = (n_1, n_2, \ldots, n_n)$ = Recursos tecnológicos (Tecnologia)

$I_i = (I_1, I_2, \ldots, I_n)$ = Fatores Institucionais

vista macroeconômico, nem do ponto de vista microeconômico. Se, contudo, o mecanismo do mercado de informação não está perfeitamente caracterizado é porque ainda existem restrições a este mercado que teoricamente pode ser descrito pelo gráfico 12.

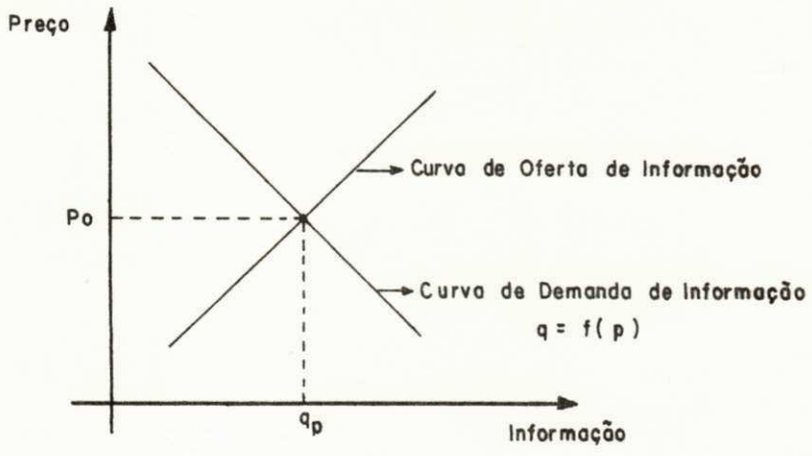


DIAGRAMA - 12 - MERCADO DE INFORMAÇÃO

Um dos entraves ao funcionamento do mercado de informação é o fato de que os consumidores de informação não são também os tomadores de decisão, ou só muito remotamente a eles estão ligados, Ora, como não são os tomadores de decisão quem decidem a quantidade e a qualidade da informação necessária, e como não são os consumidores de informação quem pagam por ela, o mecanismo de mercado não pode, assim, funcionar. O fornecimento da informação passa, então, a ser dado fora do mecanismo de oferta e procura.

Não obstante, a informação tem um valor e um custo econômicos. Considere-se, por exemplo, o seguinte fato concreto re

latado por Jordan em 1970 |2|.

> Quando da instalação de um sistema de disseminação seletiva de informação por uma companhia, constatou-se que seu **staff** técnico e científico estava reservando três (3) horas por semana na atividade de pesquisa bibliográfica e de busca de informação necessária à elaboração dos estudos e projetos da companhia.
> Supondo que um homem/hora custasse à companhia $ 7.00, estima-se que o custo por homem/ano era de $ 1000.00. Enquanto isto, o custo do sistema de disseminação seletiva de informação era de $ 250.00 por homem/ano. Assim, o valor da informação ficou evidenciado.

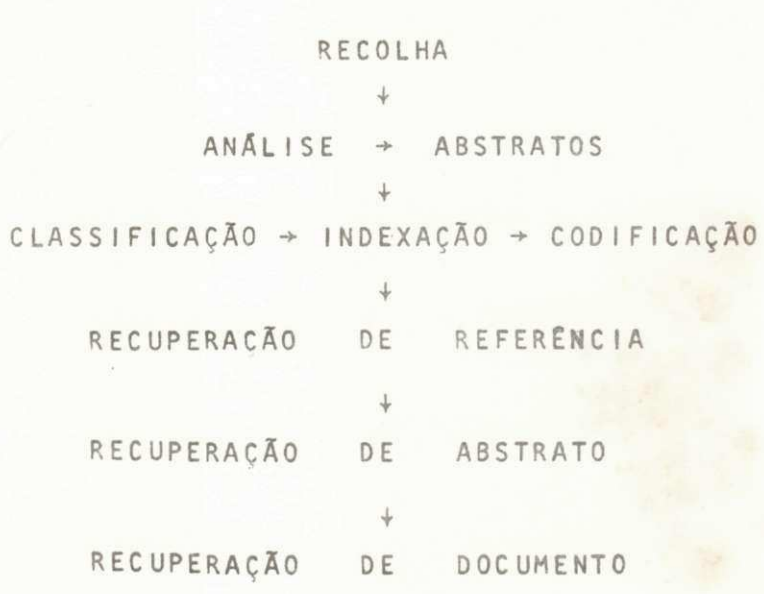
A Análise de Custo-Benefício determina, no entanto, que para haver inversão em qualquer setor é necessário antecipadamente identificar e avaliar todos os custos e benefícios do projeto em consideração. Tanto os custos dados pelos valores dos bens e serviços necessários para que o projeto entre agora em operação, como os custos associados (acrescidos ao atual projeto). A contabilização dos benefícios compreenderia todos os efeitos positivos resultantes da implantação do projeto. Isto posto, determinar-se-ia o quociente dado por,

$$\beta = \frac{\text{BENEFÍCIO}}{\text{CUSTO}}$$

para que a regra de decisão pudesse ser aplicada. Ou seja, o projeto deveria ser implantado quando $\beta > 1$, sendo então os benefícios maiores do que os custos.

Num projeto de sistema de informação, no entanto, se de um lado os custos fixos e variáveis do tratamento de dados são perfeitamente identificáveis e quantificáveis, de outro lado os benefícios já não apresentam a mesma facilidade. Como, por exemplo, mensurar os benefícios da oferta de serviços de informação? Como medir o grau de satisfação de um usuário atendido pelo sistema documental? Os benefícios de um sistema de informação, por conseguinte, são quase todos benefícios ditos "intangíveis" e de quantificação indireta em oposição aos benefícios diretamente mensuráveis.

Quanto aos custos, eles ocorrem em todas as etapas da cadeia documental:



Geralmente eles são classificados em quatro tipos principais:

1. Custo de Análise e Programação

2. Custo de Processamento*

3. Custo de transferência do suporte convencional para o suporte do processamento automático

4. Custo de transferência do suporte automático para o convencional

Outro avanço significativo da Economia do Computador consiste hoje em ela haver identificado os principais elementos, capazes de influir na estrutura de custos do processamento da informação não numérica. Estes elementos determinantes dos custos em um sistema documental podem ser resumidos:

1. processamento da informação em forma codificada (mais barata) ou em linguagem natural (geralmente, de custo proibitivo)

2. tipo de organização do arquivo documental (sequencial, invertido ou a combinação dos dois tipos) levando em consideração:

   2.1. tamanho da coleção de documentos

   2.2. tamanho das palavras-chave e de seus códigos

   2.3. modo de controle de erros

   2.4. numeração da referência bibliográfica

   2.5. frequência de atualização de arquivos

3. tipo de algorítmo ou estratégia de pesquisa determinante do tempo de indagação ao sistema, podendo ser

---

* Os custos de processamento automático algumas vezes são competitivos com os custos de processamento manual.

3.1. ou baseado em funções booleanas (presença ou ausência de descritores)

3.2. ou baseado em processos de atribuição de peso aos descritores

4. características de **thesaurus** do sistema

4.1. treinamento da equipe de construção do **thesaurus**

4.2. volume* e manutenção do dicionário

4.3. seleção das palavras-chave e correção de erros

4.4. nível de detalhamento (exaustividade) para efeito de indagação

4.5. grau de especificidade**

4.6. **links** introduzidos (de equivalência, hierárquicos e associativos)

5. Hardware e software documental utilizados

Finalmente em termos de custo, a eficiência de um sistema de informação pode ser aquilatada ou pela capacidade do sistema em processar o mesmo número de unidades por custo mais baixo, ou pela capacidade de produzir mais unidades pelo mesmo custo. Já os benefícios

---

* Um pequeno dicionário é suficiente para serviços de referência bibliográfica enquanto que um maior será melhor para processamento de indagações mais detalhadas.

** É antieconômico construir **thesauri** mais específicos do que exige a demanda de documentos.

poderão ser aquilatados ou pela capacidade de fornecer mais respostas às necessidades do usuário, mantendo-se constante tempo e custo, ou, mediante a avaliação (não necessariamente monetária) da satisfação das mesmas necessidades com economia de tempo e custo.

PARTE II: MODELOS DE ANÁLISE DE SISTEMAS EM BIBLIOTECA

CAPÍTULO 5

UMA ANÁLISE EM GESTÃO DOCUMENTAL

5.1 Generalidades

Como era de se esperar, a maioria dos centros computacionais das universidades brasileiras, no momento, se mobilizam no sentido de atender não só às necessidades de ensino (graduação e pós-graduação) e pesquisa, como ainda satisfazer à demanda de serviços de modernização administrativa por elas solicitados. Assim é que, vários projetos de modernização se acham em elaboração e execução no âmbito das universidades. Entre eles é comum encontrar os que se ligam ao controle acadêmico, controle de patrimônio e almoxarifado, controle contábil e orçamentário e controle de pessoal. Geralmente a automação nas universidades começa com estes projetos. No entanto, as atividades-meio relativas à documentação e biblioteca estão a exigir cada vez mais sua inclusão nos planos de modernização. Isto, em decorrência, principalmente, do fenômeno de explosão de informações a ocorrer na área científica e tecnológica.

É neste contexto que o presente capítulo tem em vista fornecer um modelo de metodologia para a análise de gestão de documentos em biblioteca universitária. Retomando as idéias discutidas no capítulo 2 mostra-se agora, mais concretamente, como atuar em um sistema documental sem contudo levar em conta ainda os processos de recuperação de informação. Supõe-se, en

tão, ficarem para fases mais adiantadas da automação todas as implementações das atividades ditas "intelectuais" em uma biblioteca. Isto considerando-se o fato de que as tarefas de indexação automática, classificação e elaboração de abstratos, ao envolverem a análise, o processamento de conteúdo e a recuperação de informação, não só demandam acurada tecnologia informacional, tempo e recursos financeiros como possuem, ainda, um caráter altamente experimental.

5.2 Metodologia de top-down/bottom-up

A aplicação da técnica de Análise de Sistemas à quaisquer estruturas (organizações empresariais, governamentais, universitárias, etc.) não tem somente o objetivo de reconhecimento e ordenação de dados. Ao especificar as entradas e as saídas produzidas em determinado processamento, a Análise de Sistemas tem um objetivo mais amplo, qual seja o de projetar um sistema correspondente para a organização em foco. Daí porque a Análise de Sistemas empreendida sobre uma dada estrutura vem sempre seguida do modelo do novo sistema.

Tecnicamente, quando se parte do estudo dos objetivos e problemas de um organismo tendo em vista atingir o projetamento de novo sistema de funcionamento diz-se, então, haver sido empreendida uma Análise de Sistemas de tipo **Top-down**. Ela representa um diagnóstico completo sobre o organismo e ao fruto desta análise - o modelo proposto - dá-se a denominação de Síntese ou Montagem de tipo **Bottom-up**. Ou seja, a análise é feita de cima para baixo e a implementação do novo modelo é feita de baixo para cima, a partir dos resultados detectados pela Análise.

Em automação de biblioteca, é esta a metodologia recomendada por se mostrar a maneira mais científica de descrever e formular corretamente o problema das bibliotecas universitárias. Estuda-se o conjunto de bibliotecas (os problemas, praticamente, são os mesmos) e faz-se resultar como conclusão lógica a montagem do novo sistema ou o projetamento da nova estrutura de funcionamento das atividades documentais na universidade a ser considerada.

A visualização do fluxo de procedimentos a serem obedecidos numa análise típica está contida no lado esquerdo do diagrama 13. No lado direito estão os componentes em cima dos quais deve ser montado o novo sistema. Ele se caracterizará pelo fato de:

(a) redefinir objetivos (objetivo geral, objetivos específicos, justificativas)

(b) pressupor a organização administrativa das bibliotecas

(c) reagrupar as antigas e as novas atividades em subsistemas

(d) redesenhar os fluxos de informação

(e) substituir os arquivos convencionais pelos arquivos magnéticos

Elementos da Análise **Top-down**:

Elemento 1 Objetivos e problemas das bibliotecas

Fundamentalmente o objetivo (ou meta) de

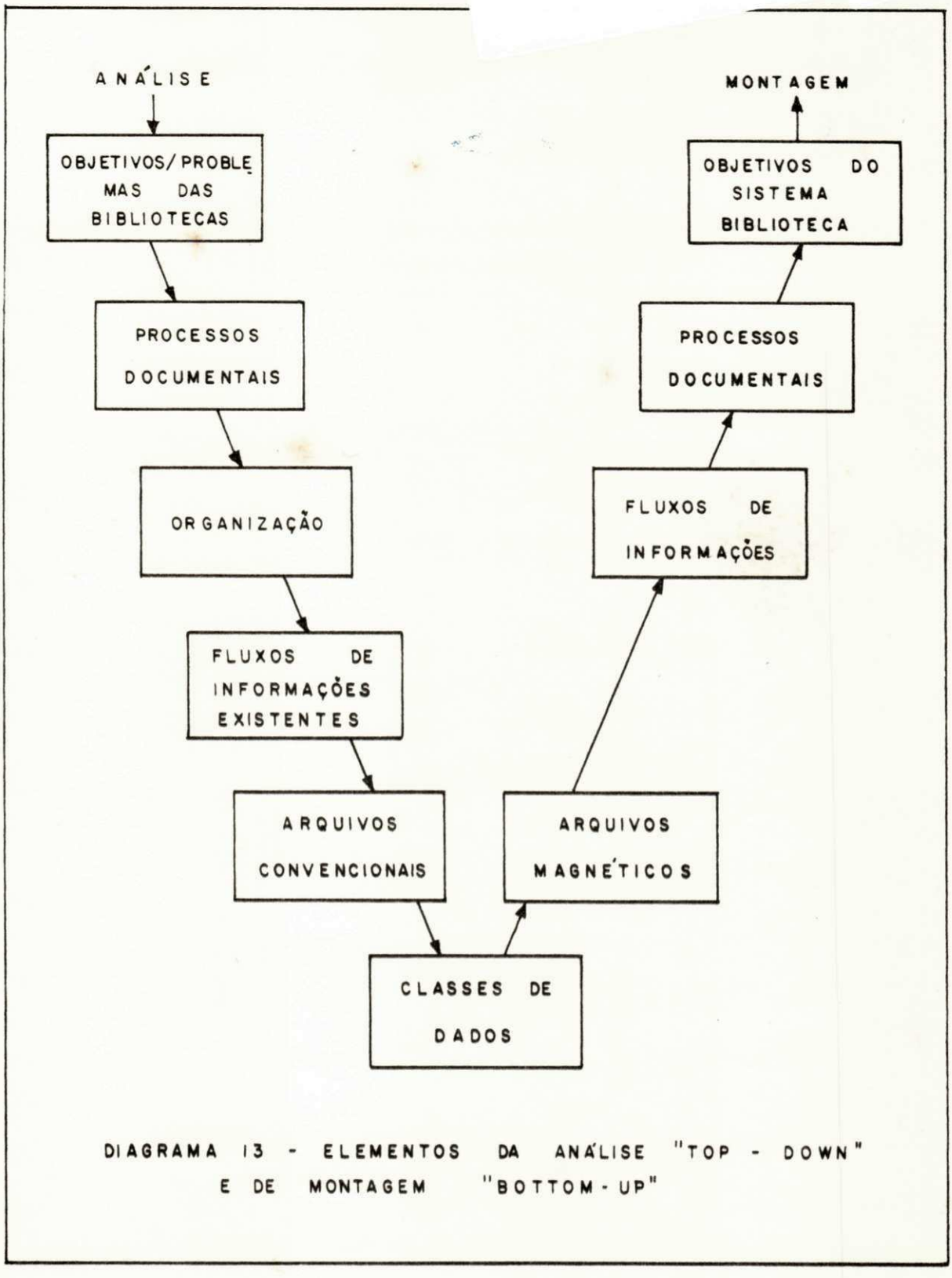


DIAGRAMA 13 - ELEMENTOS DA ANÁLISE "TOP - DOWN" E DE MONTAGEM "BOTTOM - UP"

uma biblioteca em universidade é a prestação de serviços à comunidade universitária*. Um serviço que consiste, sobretudo, em "auxiliar a identificar, prover e usar documento ou informação que melhor ajude o usuário no seu estudo, ensino ou pesquisa, na combinação ótima de custo e tempo"[1] ... Como pode ser inferido, a quantificação em Análise de Sistemas deste tipo de serviço é praticamente impossível. No entanto, avaliar o grau de satisfação das necessidades proporcionado aos usuários, os problemas e o dinamismo de um complexo de bibliotecas (no sentido atribuido por SALTON) isto pode ser conseguido e, com uma gama de informações à respeito deste questionamento podem ser redefinidos os objetivos de uma rede de bibliotecas numa universidade.

Elemento 2 Processos documentais

Do ponto de vista da automação são os processos técnicos de aquisição de documentos, circulação, catalogação/classificação, controle de usuários os aspectos principais. Basicamente eles é que serão o objeto da montagem de tipo **bottom-up**, após a identi

---

* Staff científico, técnicos, professores, corpo discente, funcionários.

1 In LEIMKUHLER, Ferdinand. F. Mathematical models for library systems analysis. School of Industrial Engineering, Pardue University, Sept. 1967. PB 176-113.

ficação de seus pontos de estrangulamento obtida mediante a análise. A análise disseca o funcionamento de cada processo documental.

Elemento 3 Organização das bibliotecas

Considerando-se estritamente a aplicação computacional, é irrelevante o aspecto da Análise de Sistemas relativo à administração das bibliotecas. A razão disto reside no fato dos modelos de automação se endereçarem antes aos processos de gestão dos documentos (processos técnicos) do que aos setores das bibliotecas responsáveis por estes processos*. Esta característica do sistema computacional lhe proporciona a vantagem de quase independência em relação aos aspectos organizacionais (recursos humanos, recursos materiais e financeiros, aspectos de hierarquia, aspectos de espaço físico, aspectos de centralização/descentralização, etc.) A automação de uma biblioteca pressupõe sua organização administrativa. À análise cabe apenas detectar aqueles aspectos organizacionais que mais diretamente afetarão o processamento de dados**.

---

* Por esta razão, do lado da montagem, no diagrama em questão, não aparece mais este elemento.

** Em cada biblioteca, a partir da Biblioteca Central, deve existir institucionalizado um subsistema de apoio logístico que possa avaliar e solucionar os problemas de organização interna em todas suas dimensões.

Elemento 4 Fluxos de informações existentes

Através da análise deste elemento busca-se descobrir os atuais fluxos de informações existentes dentro de cada biblioteca, como ainda os existentes entre elas, os usuários, a Biblioteca Central e o meio ambiente (Centros, Departamentos e outros órgãos da Universidade). Evidentemente como as bibliotecas, em geral, não são estruturadas sob a forma de sistemas, os fluxos de informações costumam não ser significativos e a análise não poderá esperar alto nível de interrelacionamento nas atividades documentais desenvolvidas.

Elemento 5 Arquivos convencionais*

Os arquivos convencionais objeto de estudo são:

a) Arquivos dos documentos originais (estantes)
b) Arquivos das fichas de autores
c) Arquivos das fichas de títulos
d) Arquivos das fichas de assunto
e) Arquivos de periódicos (Kardex)
f) Arquivos de cadastro de usuários
g) Arquivos de palavras-chave (UNITERMOS)**

---

* Todo arquivo não adaptado ao processamento eletrônico.

** A análise destes arquivos é fundamental quando a meta é a recuperação de informação.

Elemento 6 Classes de dados relevantes

Do estudo e triagem realizados por sobre estes elementos surgem as principais classes de informações relevantes para o diagnóstico e para o modelo bottom-up de automação. Estas classes são detectadas pela Análise em cada subsistema das bibliotecas e podem, assim, serem resumidas:

(a) subsistema de aquisição

- política centralizada ou descentralizada de aquisições
- número de livros, coleções e exemplares por biblioteca
- aumento médio anual do nº de livros e periódicos
- área de especialização das bibliotecas
- grau de diversificação do acervo
- relação das editoras e livrarias fornecedoras
- relação dos fornecedores estrangeiros
- relação de entidades convenentes nacionais e estrangeiras

(b) subsistema de catalogação

- tipo de classificação e registro adotados
- modo de colocação dos documentos nas estantes
- catálogos já existentes e processo de elaboração
- catálogos mais utilizados pelos usuários
- tipo de pesquisa bibliográfica comumente efetuada

(c) subsistema de circulação (ou de usuário)

- . rotinas atuais de cadastro, cancelamento de inscrições.
- . rotinas atuais de novos empréstimos, empréstimos especiais e renovações
- . número médio de empréstimos efetuados numa unidade de tempo
- . índice de movimentação do acervo (média de volumes em empréstimo, a cada instante, em relação ao acervo total)
- . multas e controle sobre circulação de documentos
- . número e tipos de usuários da rede de bibliotecas incluindo o corpo técnico-científico da universidade, corpos docente e discente e funcionários

Elementos da montagem **Bottom-up** :

Elemento 1 Arquivos magnéticos integrados

De posse dos dados levantados e estabelecido o alcance da automação (metas) é possível determinar não só o número de arquivos necessários ao redesenho do sistema como os tipos de informações que deverão alimentá-los. Comumente se tem:

(a) arquivos-movimento: aquisição de livros e periódicos, catalogação, empréstimo

(b) arquivos-cadastro: livros em aquisição, usuários, livros em reserva, livros sob empréstimo

(c) arquivos-relatório: livros recebidos , etc.

(d) arquivos-tabela: editoras/fornecedoras, etc.

Elemento 2 Fluxos de informação

Na montagem dos fluxos de informação projetam-se as comunicações entre arquivos que devam intercambiar informações. Como cada arquivo tem origem em dado subsistema, definir comunicações entre arquivos é definir comunicações entre subsistemas de biblioteca. Isto deve ser feito mediante o fluxograma geral do sistema ou dos subsistemas, de suas partes e de suas fases de consistência, classificação, acerto, atualização e relatório.

Elemento 3 Processos documentais

Quanto aos processos técnicos, são reagrupadas suas operações em subsistemas, especificando-se para cada um deles o seguinte:

(a) rotina de funcionamento e objetivos do subsistema
(b) dados e formulários de entrada
(c) dados e formulários de saída
(d) periodicidade de cada operação envolvida
(e) órgãos envolvidos com a operação (biblioteca, centro de computação, órgão financeiro, etc.)

Elemento 4 Objetivos do sistema-biblioteca

Analisados e projetados todos os componen- aqui considerados, tem-se a certeza de que os objetivos do sistema, previamente definidos são agora atingidos.

Na prática são estes os elementos que um projeto de análise e redesenho das operações padronizadas de um sistema-biblioteca leva em consideração. Tudo é feito, contudo, por etapa e dentro de cronogramas (Análise, Montagem e Execução).

O Diagrama 14 da página seguinte mostra as fases de um projeto desta natureza focalizando, inclusive, a orientação sequencial que lhes foi dada. Cada etapa, não sendo estanque nem inflexível, tem a possibilidade de influir uma sobre a outra, até mesmo na redefinição do sistema através do mecanismo de feedback. O conteúdo de cada etapa pode assim ser descrito:

Etapa 1: Diagnóstico ou Análise

Nesta etapa pesquisa-se em profundidade a situação presente das bibliotecas da universidade* a partir dos seus objetivos teoricamente definidos. São também levantados os processos técnicos empregados, as carências estruturais e materiais, como também identificam-se os recursos e potencialidades disponíveis. É uma fase de percepção de necessidades, tendo por base o sistema bibliotecário existente. O diagnóstico representa a própria análise do sistema atual e leva em consideração, pelo menos os seis elementos vistos no lado esquerdo do diagrama 13.

Etapa 2: Definição geral do sistema

Aqui é realizado o estudo de viabilidade técnica e econômica e a determinação dos objetivos

---

* Supõe-se que além da Biblioteca Central existam ainda tantas bibliotecas quantos sejam os Centros da Universidade. Esta pressuposição é válida, sobretudo, em universidade geograficamente descentralizada.

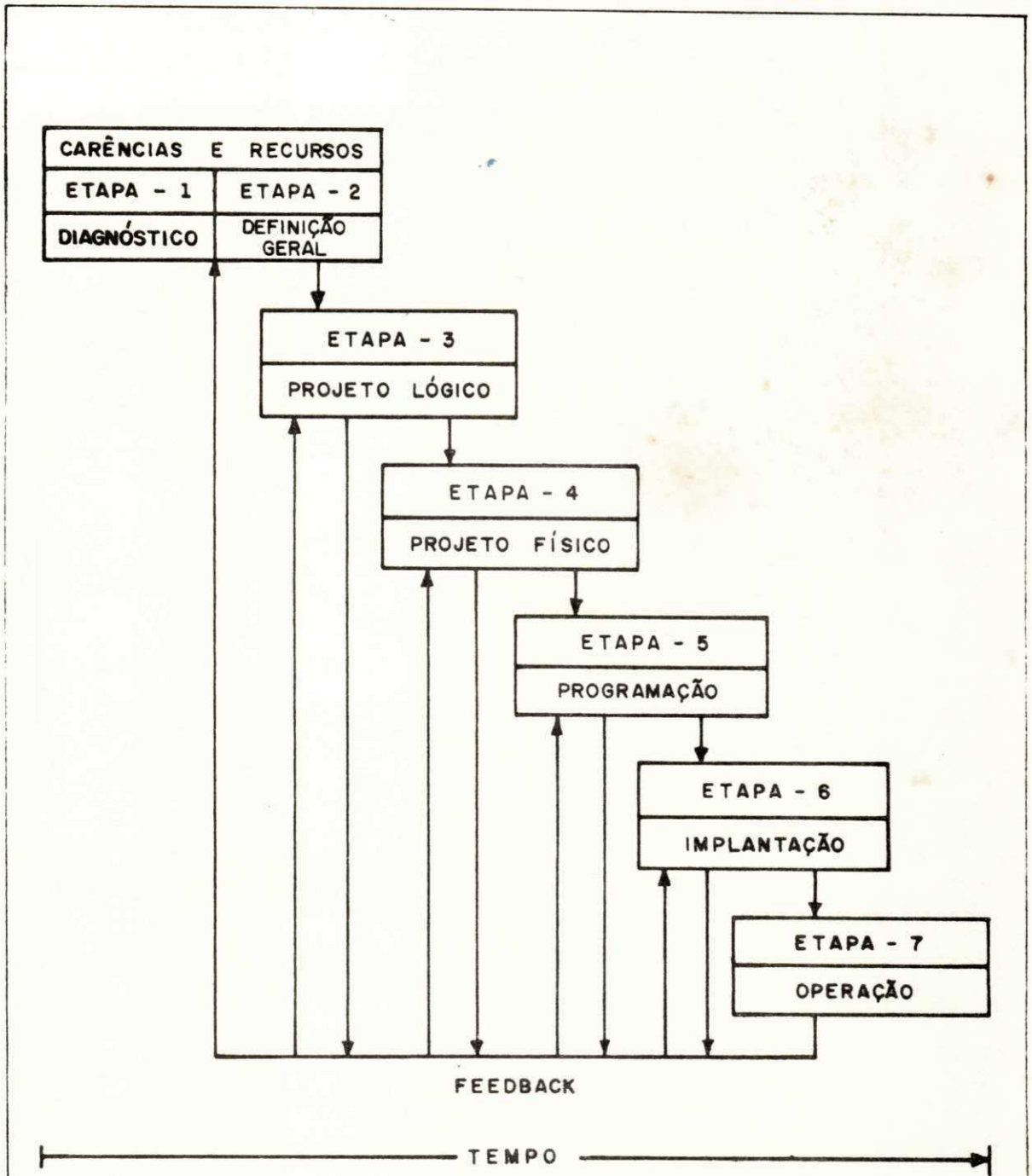


DIAGRAMA 14 - ETAPAS DA AUTOMAÇÃO DOCUMENTAL

da automação do sistema. Apresentam-se as justificativas para os procedimentos a serem adotados. São também estabelecidos outros requisitos para o uso eficiente do computador de tal modo que, a descrição das características mais globais do novo sistema, possa visualizar seu desenho geral ou seu primeiro nível de definição.

Etapa 3: Definição lógica (ou projeto lógico)

O projetamento lógico é o passo responsável pelo desenho detalhado do sistema. Nesta etapa é feita a montagem do processamento do sistema levando em consideração especificações técnica, administrativa e economicamente viáveis.

Etapa 4: Definição física (ou projeto físico)

Define-se nesta etapa a organização do processamento (**on line** ou **batch**), o equipamento e sistema operacional a serem utilizados, os **packages** utilitários e controles do sistema; Determinam-se os tipos de organização dos arquivos, como também são estabelecidas as especificações dos programas.

Etapa 5: Programação

São revisadas nesta fase as especificações dos programas definidos nas etapas 3 e 4 passando-se em seguida, ao desenvolvimento de sua lógica e ao processo de codificação.

Etapa 6: Implantação

É o passo no qual são treinados bibliotecários, auxiliares e operadores do sistema. Engloba ainda a realização dos primeiros testes e a revisão de detalhes que surgem quando da passagem do sistema convencional (manual) para o sistema automatizado; Esta passagem é também chamada de conversão do sistema, quando então os arquivos convencionais são substituidos pelos arquivos magnéticos.

Etapa 7: Operação

Inicia-se, aqui, o processo de produção dos relatórios, catálogos e demais registros e informações já definidos nas etapas anteriores. A partir de então, através do processo de manutenção, apenas são introduzidas no sistema aquelas mudanças ou alterações que objetivem mantê-lo viável.

Estes passos para automação de Bibliotecas, Centros de Documentação, ou Centros de Análise da Informação são válidos para quaisquer que sejam os aspectos da automação, quer pertençam eles ao nível da administração dos documentos (controle de bibliotecа), quer ao nível da recuperação de informação. Cada uma das etapas, evidentemente, deve estar associada a determinado espaço de tempo (como lembra o diagrama 14) dentro do qual ela deve ser concretizada.

Quanto à documentação* do sistema, ela não se constitui propriamente numa fase, já que a atividade de documentar está presente em todas as etapas, registrando informações tais como relatórios sobre o antigo sistema de biblioteca, proposta do novo sistema, fluxograma do novo sistema, documentos de entrada, relatórios de saída, descrição do banco de dados documentais, etc.

A documentação satisfatória em um processo de automação de biblioteca é fundamental não só para registrar o ciclo de vida do sistema, como também para providenciar seu contínuo aperfeiçoamento.

---

* Aqui, documentação tem o sentido utilizado em Processamento de Dados.

CAPÍTULO 6

UMA ANÁLISE EM CLASSIFICAÇÃO AUTOMÁTICA DE PALAVRAS - CHAVE

6.1 - Teoria da Classificação

Da variada gama de operações exigidas pela Recuperação de Informação, conforme apresentadas no capítulo terceiro, escolheu-se investigar aqui, tão somente, alguns aspectos relacionados com as linguagens de recuperação de informação. Em particular serão estudadas as linguagens hierárquicas de recuperação de informação. Ou seja, as linguagens que utilizam a classificação como estrutura de manuseio do vocabulário.

Inicialmente seja retomado o conceito de relação de classes apresentado em 1.2. Uma outra importante relação de classe a ser acrescida agora é a relação de inclusão. A relação de inclusão ocorre quando os membros de uma classe (de documentos,ou de palavras-chave) são também membros de outra classe. Exemplificando, sejam

X = a classe dos documentos sobre sistemas sociais

Y = a classe dos documentos sobre sistemas econômicos

Aqui "X inclui Y" ($Y \cap X = Y$ e $X \cap Y = Y$). Ora, esta relação não é outra senão a conhecida classificação hierárquica onde os itens (documentos, ou palavras-chave) são grupados dentro

duma estrutura de classes* em que as mais amplas ou mais gerais incluem as menos amplas ou mais específicas.

O conceito de classificação se encontra perfeitamente ligado à definição de linguagem documental já vista no primeiro capítulo. Na verdade, se uma linguagem é composta de um vocabulário e uma estrutura que mostre as relações conceituais deste vocabulário, da forma mais evidente possível, então, a classificação pode satisfazer esta exigência estrutural duma linguagem de indexação. A questão fundamental agora posta resume-se em como classificar satisfatoriamente.

Ora, em Recuperação de Informação a Teoria da Classificação** e a Teoria de Indexação se preocupam precisamente com esta questão. Seus problemas básicos são:

(a) a geração de classes usando informações (propriedades ou características) a respeito dos objetos a serem classificados. Surge aqui o problema de criar métodos para obtenção de classes ou de documentos ou de palavras-chave.

(b) a associação dos objetos às classes já existentes ou às classes que estão sendo geradas. É o problema das operações de indexação e procura dos documentos.

(c) a extração das informações úteis à identificação e caracterização das classes de documentos.

(d) a modificação das classes quando ocorre uma expansão da coleção inicial de documentos.

---

* *Vale em Recuperação de Informação o sentido etimológico do verbo classificar. Em Documentação Científica o termo pretende ser mais abrangente.*

** *Não confundí-la com a Taxonomia onde os objetos a serem classificados são bem definidos e de fácil caracterização como, por exemplo, nas classificações botânicas e zoológicas.*

Na parte restante deste capítulo serão focalizados basicamente aspectos ligados aos problemas (a) e (b). Antes, porém, enumeram-se aqui os principais tipos de classificação estabelecidos pela Teoria da Classificação.

1. Quanto ao objeto, a classificação pode ser:

   1.1. Classificação de documento (Tipo 1) - é a classificação que objetiva melhorar e atingir mais rapidamente o resultado duma pesquisa empreendida, através da restrição do campo da pesquisa a somente determinadas partes do arquivo (convencional ou magnético).

   1.2. Classificação de palavras-chave (Tipo 2) - é a que visa grupar as palavras-chave dentro de classes com a finalidade de produzir maiores possibilidades de acasalamento entre as solicitações feitas pelos usuários dos documentos e as palavras-chave deles extraídas.

2. Quanto à metodologia da classificação:

   2.1. Classificação abstrata (Tipo 3) - é aquela classificação obtida mediante processos puramente abstratos, em que procedimentos baseados em critérios bem definidos são aplicados a um conjunto determinado de documentos. Estes critérios são oriundos do Cálculo Relacional, da Taxonomia Numérica e da Análise de Associação.

   2.2. Classificação empírica (Tipo 4) - é a classificação obtida por processos não abstratos cuja qualidade depende dos propósitos da classificação.

3. Quanto ao critério da classificação:

   3.1. Classificação monotética (Tipo 5) - ocorre quando todos

os componentes de uma classe possuem uma propriedade específica em comum

3.2. Classificação politética (Tipo 6) - ocorre quando o critério acima não se verifica.

4. Quanto aos "pontos de vista" do documento:

4.1. Classificação rígida (Tipo 7) - ocorre quando um documento é classificado sob um único "ponto de vista" ou assunto.

4.2. Classificação multidimensional (Tipo 8) - É aquela em que um mesmo documento é classificado sob diferentes "pontos de vista" ou assuntos.

Na prática, no entanto, quase sempre o que existe é a combinação de tipos de classificação de tal modo a tornar mais eficiente o processo de busca documental. Foram omitidas aqui as chamadas classificações especializadas e facetadas.

6.2 - Automação da CDU *

A caracterização da Classificação Decimal Universal (CDU) como uma linguagem de recuperação de informação (thesaurus) é um fato recente que surgiu, sobretudo, em decorrência do reestudo de suas bases metodológicas e dos esforços desenvolvidos para sua automação. De maneira geral ela satisfaz às condições de existência de uma linguagem ou seja, a CDU possui um vocabulário e uma estrutura de manipulação deste vocabulário. Assim, em termos de thesaurus a CDU pode ser definida como:

CDU = VOCABULÁRIO + CLASSIFICAÇÃO TIPO 1

---

* *O que for dito para a CDU será válido para a Classificação Decimal de Dewey e vice-versa.*

onde o vocabulário é a totalidade do conhecimento universal armazenado ou a armazenar e a classificação é a estrutura que permite criar classes hierárquicas de base 10 (dez). São, portanto, características da CDU:

(a) a divisão do conhecimento em dez classes principais (as primeiras classes) e a sucessiva subdivisão em classes secundárias.

Exemplo:

0 Generalidades
1 Filosofia
2 Religião-Teologia
3 Ciências Sociais
4 Filologia-Linguística

5 Ciências Puras...	50 Generalidades	540 Química-Ciências Correlatas
	51 Matemáticas	541 Química Teórica
	52 Astronomia	542 Física-Química Experimental
	53 Física	543 Química Analítica
	54 Química.....	544 Análise Qualitativa
	55 Geologia	545 Análise Quantitativa
	56 Paleontologia	546 Química Inorgânica
	57 Biologia	547 Química Orgânica
	58 Botânica	548 Cristalografia
	59 Zoologia	549 Mineralogia

6 Ciências Aplicadas
7 Belas Artes
8 Literatura

9 História-Geografia...
- 90 Generalidades
- 91 Viagens
- 92 Biografias
- 93 História Antiga
- 94 História da Europa e França
- 95 - 96 História da Ásia e África
- 97 - 99 História da América

(b) como decorrência da característica anterior, a CDU pode ser representada por uma árvore (logo, não existe loop) onde cada nodo da árvore representa um ramo do conhecimento e tem conectado a ele mais dez nodos.

(c) ao pesquisar-se a árvore, à qualquer subconjunto de conhecimento, não importando seu nível dentro da classe, faz-se corresponder um número bem definido que, mediante a quantidade de seus algarismos, determina o nível do subconjunto de conhecimento.

Ex: Ao número 549 sempre corresponde o subconjunto Mineralogia.

(d) outras características de regras de formação e notação podem ser visualizadas pela seguinte indexação utilizando a CDU:

Título do documento: Distribuição e movimento sazonal de peixes nas Baías do Brasil.

Classificação p/CDU:

597:591.9:591.52(285:981) "1964/1966"(047+084.3)

Forma de entrada no indíce:

(047) ..... Relatórios técnicos

(084.3) ..... Mapas

(285) ..... Lagos

(981) ..... Brasil

"1964/1966" ..... Eventos de 1964 à 1966

591.52 ..... Habitats animais e migrações

591.9 ..... Distribuição geográfica de animais

597 ..... Ictiologia. Peixes

EXEMPLO DE INDEXAÇÃO PELA CDU

Quanto à automação da CDU, no entanto, as experiências e estudos têm revelado e levantado graves questões. Umas inerentes à própria constituição da CDU e outras mais ligadas aos aspectos operacionais da automação. Assim, por exemplo:

(a) a utilização da base dez de numeração nos é familiar, mas não apresenta razão lógica nenhuma para que as subdivisões do conhecimento fiquem restritas ao fatídico número dez. Ora, se não há razão lógica para que a classificação deva se limitar à subdivisões em quantidades constantes de subconjuntos seria provavelmente, no plano teórico, o sistema binário o mais recomendado. No mínimo, ofereceria melhor adaptação ao processamento eletrônico.

(b) a classificação não é bem definida, isto é, não é obtido um único resultado para qualquer corpo de dados, de tal modo que dois documentalistas podem classificar um mesmo documento de maneiras diferentes. As regras de indexação, por conseguinte, são ambíguas.

(c) as regras carecem de especificidade, isto é, todas as relações possíveis no corpo de dados devem ser representadas com os poucos dispositivos notacionais existentes.

(d) a CDU possui baixa capacidade de recuperação. Se for tomado o documento Formal Organization, por exemplo, nele serão encontrados assuntos das disciplinas abaixo:

- Teoria da Aprendizagem ..... Psicologia
- Cadeias de Markov ..... Cálculo de Probabilidade
- Organogramas ..... Administração Geral
- Confiabilidade de Sistemas .. Engenharia de Sistemas

Como pela CDU o documento possivelmente será classificado como sendo de Administração Geral*, isto significa que apenas 30% do conteúdo original será recuperado. Os 70% restantes devem ser contabilizados como perda de informação no sistema documental ou como aumento da taxa de "silêncio"**

Não obstante apresentar a CDU este conjunto de restrições, experiências de implementação automática têm sido desenvolvidas sobretudo na Alemanha*** e nos Estados Unidos**** |10| . O fundamento do trabalho consiste em desenvolver uma base racional para o uso da CDU como linguagem de indexação, mesmo que para isto seja feita a combinação com um outro sistema de recuperação já existente*****.

Os algorítmos empregados quase todos obedecem aos seguintes passos gerais.

PASSO 1 : Preparação dos inputs da classificação. Certos caracteres da CDU devem ser eliminados.

PASSO 2 : Padronização na escrita da CDU. São necessários mecanismos para identificar fins de linhas, brancos e outros caracteres especiais.

PASSO 3 : A conversão dos caracteres é muito importante no esta

---

* *O apelo à classificação multidimensional só é possível quando se dispõe de mais de um volume do mesmo documento e quando o nível cultural do documentalista assim o permite.*

** *"Silêncio" é a percentagem de documentos existentes no sistema mas não recuperados por deficiência na indexação.*

*** *Alemanha: Zentralstelle für maschinelle Dokumentation.*

**** *Estados Unidos: American Institute of Physics e School of Library Science*

***** *Por exemplo, CDU/"Sistema Combinado de Pesquisa à Arquivo" implementado no IBM 1401 |10|*

belecimento dos termos do indíce. (Ex: n ← Cn; = n←En; (On) ← En; "n" ← Tn; n'n ← nYn; onde n = 1, ..., 9). Os símbolos à esquerda da seta correspondem à notação genérica da CDU.

PASSO 4 : Criação da tabela de regras de indexação.

PASSO 5 : Criação dos caracteres de controle. Eles permitem o reconhecimento dos termos de índice originais.

PASSO 6 : Sorting dos termos de índice.

PASSO 7 : Os termos solicitados são pesquisados título por título.

PASSO 8 : Depois do sorting, o material é preparado para a impressão do índice. Os termos do índice não desejados podem ser supressos. Caracteres, ou combinações de caracteres são convertidos para sua forma original. Segue a impressão.

Como conclusão destes estudos podem ser registrados os aspectos abaixo:

1) Não há dúvida de que a CDU possa ser usada como linguagem de indexação em sistema automático.

2) Como ela existe no momento, provavelmente não terá a eficiência, em um sistema automatizado, tal como uma linguagem especificamente criada para o processamento em computador. Já foi mostrada, contudo, a possibilidade da CDU ser usada, no processamento em batch, ou no modo interativo não tendo mais sentido, pois, a acusação de Malcolm Rigby: "o computador existe para execu-

tar qualquer operação complicada quando necessária, tanto matemática e estatística, como a de recuperação (seleção, ordenação, combinação e/ou listagens) e não se furtar às tarefas consideradas abaixo de sua capacidade" [8]*.

3) Aplicar ou não computador à CDU também não é uma questão de escala de serviço. Não interessa ao computador a massa de dados a ser processada. É antes uma questão de eficiência, qualidade e consistência dos resultados considerando-se, a priori, a natureza débil da CDU. E isto embora dependa "do campo do conhecimento, vocabulário disponível, hábitos do usuário", depende, principalmente, das características intrínsecas do sistema de classificação a ser automatizado.

4) Portanto, a viabilidade técnica de automação da CDU é relativamente menos vantajosa do que aquela proporcionada pelos sistemas que utilizam palavras-chave extraídas do contexto documental. Isto não só pelo fato destes sistemas superarem as deficiências da CDU, como pelo fato de melhor se adaptarem ao processamento computacional.

6.3 - Algorítmo de Classificação Automática

6.3.1 - Técnica de Trabalho

Não obstante os esforços para atualização e desen -

---

* *Malcolm Rigby desconhece que não são as tarefas de Informática Documental, e sim, as de Inteligência Artificial que se constituem em maior desafio para o computador.*

volvimento da CDU, considera-se altamente significativo o fato de Salton* em 1 não fazer qualquer citação explicita deste sistema convencional. Segundo ele há considerável evidência de que as classificações convencionais em biblioteca não podem satisfazer a todos os requisitos de um sistema documental dinâmico (biblioteca dinâmica). E, de fato, é improvável que as classificações tradicionais possam providenciar soluções para os problemas tais como o da ordem unidimensional das estantes de bibliotecas, bem como o da análise multidimensional do conteúdo dos livros e documentos.

Em vista destes impasses e dos fatores limitantes da CDU, estudos e experiências em torno das palavras-chave têm sido intensificados e implementados, envolvendo basicamente três operações:

OPERAÇÃO 1: Extração automática (ou não) de palavras-chave. É também conhecida como operação de construção de vocabulários. Os inputs da operação são os títulos dos livros, os textos originais dos documentos e os resumos. O output é, obviamente, uma lista ou conjunto finito de palavras-chave conforme a definição dada a vocabulário no subcapítulo 1.2.

OPERAÇÃO 2: Ordenação automática de palavras-chaves, ou processo de formação de classes, ou ainda, classificação automática de palavras-chave. É este o sentido dado pela Recuperação de Informação à noção de classificação. Ela fica restrita, portanto, à obtenção de um conjunto de classes para dada coleção de documento. No entanto, no momento em que é procedida a

---

* *Principal expoente americano na área de Recuperação de Informação*

classificação das palavras-chave (ou construção do thesaurus), também está sendo procedida a classificação dos documentos, pois os documentos não são classificados diretamente por suas palavras-chave, mas, indiretamente pelas classes derivadas das palavras-chave que ocorrem e coocorrem nos documentos. As duas classificações não são, portanto, independentes mas simultâneas como pode ser visto na operação 3. O input na operação 2 é o vocabulário ou parte dele e o output é a classificação obtida (classificação das palavras-chave e consequentemente dos documentos).

OPERAÇÃO 3: Indexação* ou interrogação ao sistema em feedback ou não). É a operação que consiste em realizar o acasalamento entre uma indagação que é feita (uma procura) e um documento que existe no sistema. Este acasalamento é dado por uma função F. Pela operação de indexação verifica-se o dinamismo da classificação de palavra-chave, pois ela não é estanque e se ajusta a cada função de indagação que está sendo aplicada ao sistema. Assim, por exemplo, seja:

F = a função de acasalamento

n = o documento que deve ser acessado como relevante

d_x = a palavra-chave que caracteriza n.

R = a indagação sobre n feita com base na palavra-chave d_x

T = o valor de um certo delimitador (perfil)

---

* *O mesmo que catalogação na biblioteca tradicional*

Supondo que o documento n deva ser acessado como relevante para a indagação R a indexação deve providenciar mecanismos no sentido de ajustar a classificação da palavra-chave d_x de tal modo que $F(R,n) \geq T$.

Por este meio garante-se que o documento n seja realmente recuperado pelo sistema, em resposta à indagação R.

Contrariamente, quando o documento n não é considerado relevante para a indagação R, a classificação é, então, ajustada de tal modo que $F(R,n) < T$.

Na prática, a busca a documentos é feita utilizando-se mais de uma palavra-chave para caracterizar os documentos. Neste caso R deve ser representado como um vetor n-dimensional, o vetor-indagação, e n como o vetor-documento. A função de acasalamento (ou função de similitude) deve refletir o grau de relacionamento (ou de distância) existente no par (indagação, documento). Assim, pode-se definir:

(a) vetor-indagação: $R = (r_1, r_2, \ldots, r_t)$ onde

r_i = é o peso da i-ésima palavra-chave da indagação.

(b) vetor-documento: $n = (n_{j1}, n_{j2}, \ldots n_{jt})$ onde

n_{ji} = é o peso de i-ésima palavra-chave no documento j

(c) função de similitude: $$F(R,n) = \frac{\sum_{i=1}^{t} (r_i \cdot n_{ji})}{\sqrt{\sum_{i=1}^{t} (r_i)^2 \cdot \sum_{i=1}^{t} (n_{ji})^2}}$$

F = função típica de acasalamento sendo

$$0 \leq F \leq 1$$

Numericamente, ao se escolher por exemplo T = 0.75, se $F(R, n) \geq 0.75$ então, todos os documentos que satisfazem esta condição serão recuperados para atenderem ao usuário.

Vários métodos têm sido desenvolvidos visando conduzir logicamente, via computador, as três operações aqui apresentadas. Neste capítulo escolheu-se analisar, tão somente, um dos métodos de classificação automática de tal maneira à servir de modelo para os procedimentos da operação 2 que fornece a estruturação de uma linguagem documental. Após o razoável detalhamento do algorítmo serão então, apresentados os resultados de sua implementação levada a efeito pela equipe que o criou.

A) Objetivo do Algorítmo* - supondo uma primeira etapa de seleção e constituição de um vocabulário, o algorítmo introduz uma estrutura hierárquica no vocabulário total ou em parte dele. Ou mais claramente, dada uma palavra-chave d (ou um perfil que caracterize um grupo

---

* *Encontra-se em "Revue Française d'Automatique, Informatique et Recherche Opérationnelle", (6e année, nº juin 1972.)*

potencial de palavras-chave) o algorítmo deverá, então providenciar o ambiente desta palavra-chave d, ou seja, todas as outras palavras* relacionadas e que devem ser incluídas na classe representada por d.

B) Origem do Algorítmo - nasceu de um convênio firmado entre o Governo Francês (Comitê de Pesquisa em Informática) e o Instituto Gustavo Roussy (Serviço de Documentação Científica), sendo dirigentes da pesquisa M. Wolff-Terroine, D. Rimbert e B. Rouault |6|.

C) Base Teórica e Pressuposições do Algorítmo - o problema principal da construção de uma linguagem é superar as ambiguidades da linguagem natural procurando encontrar métodos estatísticos e informáticos que permitam o tratamento automático dentro de um sistema documental.

Em vista disto, especial atenção deve ser dada ao problema do vocabulário documental a ser submetido à classificação, pois nem todo conjunto D de atributos de documento é, a priori, classificável.

A escolha dos atributos (palavras-chave) duma coleção de documentos constitui, portanto, uma etapa fundamental dentro da cadeia de operações orientadas para a classificação automática, uma vez que esta escolha determina a coerência das partições a serem classificadas. E esta escolha é tanto mais essencial quando se tratar de classificação hierárquica (em razão do aspecto semântico).

---

* *Podem ser sinônimos de d ou relacionados com d pelo sentido.*

A base teórica dos procedimentos de classificação repousa, assim, na normalização inicial do vocabulário e nos parâmetros bibliométricos dele originados. Estes parâmetros foram definidos no capítulo sobre Bibliometria e, em última análise, é deles que depende a eficiência da própria classificação.

São pressuposições do algorítmo aqui analisado:

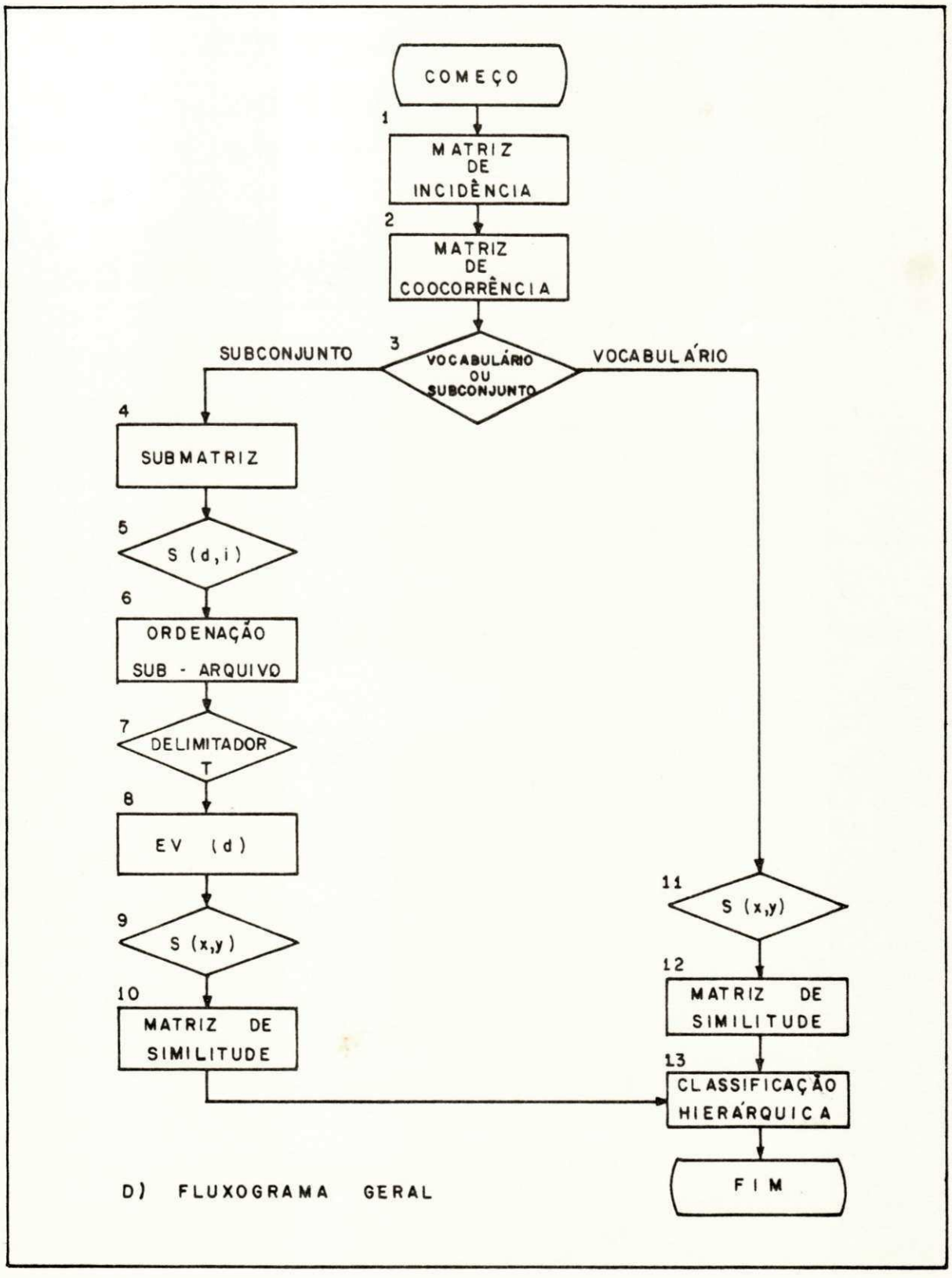
(1) o vocabulário, input da classificação, é escolhido a priori.

(2) a classificação é hierárquica

(3) cada palavra-chave é citada uma única vez por documento.

(4) no processo, perde-se a descrição individual do documento em favor da descrição da coleção de documentos.

(5) no algorítmo não serão consideradas as interrogações realizadas com formulação em feedback.



D) FLUXOGRAMA GERAL

E) Passos do Algorítmo e Análise

PASSO 1: Cálculo da matriz de incidência - determinam-se as frequências absolutas de citações das palavras-chave, montando, assim, a matriz de incidência (ou matriz de termo/documento) que fornece as ocorrências (presença) dos termos por documento. Num processo de indexação probabilística a presença dos termos d_1, d_2, ... d_D no documento n_i seria usada como base para estabelecer que o documento n_i pertence a classe C_k com probabilidade p*.

PASSO 2: Cálculo da matriz de coocorrência - calculam-se as frequências de coocorrência de todos os termos do vocabulário. Este cálculo é a geração da matriz de coocorrência apoiada na formação de um arquivo de todas as duplas possíveis de serem formadas com as palavras-chave (d_x, d_y).

Para todo valor de x e y, o número de duplas distintas (D (D - 1)/2 onde D é o nº de palavras-chave) dá a frequência de coocorrência dos termos d_x e d_y.

PASSO 3: Classificação no vocabulário ou no subconjunto - uma classificação pode ser efetuada sobre todo o vocabulário ou sobre um subconjunto do vocabulário. No primeiro caso o algorítmo segue para o passo 11 onde se inicia a preparação para a classificação. No segundo caso, a escolha do subconjunto precisa ser formalizada e o algorítmo segue para o passo 4.

---

* *Neste algorítmo o processo é, basicamente, de indexação associativa.*

PASSO 4: Determinação da submatriz de coocorrência - determina-se a submatriz a ser trabalhada (ou subconjunto) tomando-se ou uma linha da matriz de coocorrência ou n linhas. No primeiro caso a linha será determinada pela palavra-chave d (palavra indutora ou perfil) na qual se tem interesse. No segundo caso, faz-se a união das palavras-chave nas quais se tem interesse (tumor + seio = tumor do seio = d) e a submatriz (n linhas da matriz) agrupará as palavras-chave que coocorrem com a união encontrada. Este subconjunto E (d) (ou submatriz) determinado de uma destas maneiras é um subconjunto monotético, pois, seus elementos tem uma característica comum - a coocorrência com d. A palavra-chave d passará a caracterizar, a priori, o subconjunto E(d).

PASSO 5: Escolha do coeficiente de similitude S(d, i) - os coeficientes de similitude são funções simétricas ou não, da frequência de coocorrência F(d, i). A esta frequência foi atribuida um sentido especial. Assim, parte-se do princípio de que o conhecimento de uma relação entre dois fenômenos se apoia sobre a observação de um certo tipo de associação fornecida por suas coocorrências. Se esta associação se repete a observação inicial vai se fundamentando. Daqui pode ser inferido o fato de que a frequência de coocorrência é uma noção composta de:

F(d, i) = Parte semântica (dada pela Associação das Coocorrências) mais parte estatística (dada pela esperança E(d, i) de F(d, i) acontecer).

Deste modo, ao se fazer o coeficiente de similitude igual à frequência de coocorrência (S(d,i) = F(d,i)) atri

bui-se, neste momento, ao coeficiente de similitude a capacidade não só de medir o valor estatístico da coocorrência de dois termos, como ainda, avaliar o aspecto semântico desta coocorrência. Se, por exemplo,

$$S(d,i) = \frac{F(d,i) - E(d,i)}{F(i)}$$

pode-se, através de S(d, i), avaliar a proximidade de sentido entre os termos d e i, uma vez que, da frequência de suas coocorrências foi subtraída a parte estatística (E(d,i) do numerador).

Vendo assim estes coeficientes, pode-se neste ponto do algorítmo escolher o coeficiente que melhor se adapte ao problema concreto.

PASSO 6: Ordenação do sub-arquivo - escolhida a fórmula do coeficiente de similitude deve-se, agora, aplicá-lo sobre o subconjunto E(d) determinado no Passo 4, afim de que suas palavras-chaves sejam organizadas segundo o grau de similitude entre elas mesmas.Aqui, o coeficiente de similitude se destina a fornecer uma ordenação no subconjunto de termos E(d). E o interesse pela ordenação neste passo é poder, a posteriori, compará-la com a ordenação ideal do passo 9, depois então, que houver sido aplicado o delimitador T no passo 8.

PASSO 7: Escola do delimitador T - com vistas a estabelecer um grupo EV(d) de elementos com coerência interna (elementos estes, saídos de E(d) escolhe-se um valor arbitrário para um certo delimitador T.

PASSO 8: Formação do Grupo EV(d) - o objetivo deste grupo é conhecer o ambiente semântico da palavra-chave d. Partindo do subconjunto E(d) definido anteriormente, escolhe-se agora neste subconjunto as palavras-chave cujas coocorrências com a palavra-chave indutora d apresente uma significação semântica. Estas palavras irão formar o grupo EV(d) e as condições para que elas pertençam ao grupo EV(d) são as seguintes:

$$EV(d) = \{i \mid i \in E(d) \wedge S(d, i) \geq T\}$$

No grupo EV(d) assim formado, as palavras-chave i coocorrem com d e são significativamente correlacionadas com d. Este grupo é dito de similitude maximal e dele é que será possível posteriormente (na classificação) retirar uma partição que forneça o ambiente de d (ou os "pontos de vista" sobre d).

PASSO 9: Ordenação ideal mediante S(i, j) - a ordenação neste passo é definitiva e é efetuada agora sobre EV(d). A fórmula do coeficiente de similitude deve ser a mesma escolhida no Passo 5, e os valores fornecidos por S(i, j) é que irão constituir os elementos da matriz do passo 10.

PASSO 10: Matriz de similitude - esta matriz representa a ordenação sobre os elementos de EV(d), de tal maneira que EV(d), ordenado e esta matriz são a mesma coisa. Segue para o Passo 13.

PASSO 11, 12: Estes passos são efetuados à maneira dos Passos 9 e 10 com a única diferença de serem realizados sobre o conjunto do vocabulário.

PASSO 13: Classificação hierárquica - O input da classificação será a matriz de similitude (um conjunto de termos e os fatores de associação conceito-conceito). Seja S um delimitador ou perfil que representa e idenfica um grupo a ser submetido à classificação. Ele é escolhido para caracterizar as outras palavras-chave do grupo. Pela classificação hierárquica pode-se definir uma partição única e tal que, se duas palavras-chave estão reunidas dentro de um mesmo grupo de delimitador S, elas estarão ainda para todo $S' < S$. Baseado neste princípio a classificação consistirá em incluir grupos (ou classes potenciais) de palavras-chave em classes reais que estão sendo formadas. Sejam:

a) $G_1, G_2, G_3, \ldots, G_s \rightarrow$ uma partição de EV

onde EV tem coerência maximal S.

b) G_p e $G_q \rightarrow$ Os dois grupos cuja similitude S_{pq} é maximal.

Quando $S = S_{pq}$, G_p e G_q devem ser reunidos para dar $G_r = G_p \cup G_q$. Calcula-se então a similitude S_{ir} do novo grupo G_r com cada uma das outras classes G_i.

Os efetivos ainda não grupados são considerados como classes potenciais.

Pela fórmula de Lance e Williams

$$S_{ir} = \alpha S_{ip} + \beta S_{iq} + \gamma S_{pq} + \delta \mid S_{ip} - S_{iq} \mid$$

$$\alpha, \quad \beta > 0 \quad ; \quad S_{ir} \leq S_{pq}$$

Entre outras maneiras de achar α pode-se fazer:

$$\alpha = \beta = \frac{1 - a}{2} \; ; \quad \gamma = a; \; a = -0.25$$

6.3.2 - Resultados da Implementação

A implementação do algorítmo apresentado foi realizada no âmbito de um sistema com as seguintes características:

(a) Coleção de documento: N = 20.000

(b) Conjunto de palavras-chave: D = 3.700

Estas palavras-chave eram especializadas em Medicina, com área de concentração Cancerologia.

O vocabulário era suficientemente confiável e foi testado para ver se tinha alcançado uma estabilidade suficiente para poder dar origem a medidas estatísticas significativas.

(c) Em média, cada documento foi descrito por sete palavras-chave.

(d) Computador utilizado: UNIVAC 1107

A montagem da matriz de coocorrência levou 2.30 horas, donde se pode ver que o tempo-máquina da experiência é perfeitamente aceitável.

Escolheu-se realizar a classificação sobre um subconjun-

to do vocabulário e não sobre a totalidade do vocabulário. Na seleção deste subconjunto o critério foi a coocorrência com a palavra-chave "TUMOR DO SEIO", ou SEIO(T). Com o perfil SEIO (T) foram realizadas três experiências fazendo variar apenas o coeficiente de similitude e os valores de α, β, γ e δ na fórmula de Lance e Williams

1a EXPERIÊNCIA:

(1) d = perfil = SEIO (T)

(2) S(SEIO(T), i) = coeficiente de similitude onde i foram as palavras-chave que coocorreram com tumor do seio

$$= \frac{F(SEIO(T),\ i)}{F(SEIO(T)) + F(i) - F(SEIO(T),\ i)} =$$

= Função de Tanimoto

(3) Resultado: Com estes parâmetros a primeira experiência foi capaz de agrupar um conjunto de termos onde somente a palavra "Mastectomia" estava mais diretamente ligada ao conteito de "Tumor do Seio", fornecido pelo usuário. Também surgiram conceitos tais como "Classificação Histológica" e "Radiografia pós-operatória".

2a EXPERIÊNCIA:

(1) d = perfil = SEIO(T)

(2) $$S(SEIO(T),\ i) = \frac{F(SEIO(T),\ i) \quad - \quad E(SEIO(T),\ i)}{F(SEIO(T)) + F(i) - F(SEIO(T),i)}$$

(3) Resultado: Aqui o numerador do coeficiente de similitude anterior foi corrigido pela esperança E(SEIO(T),i). O número de palavras mais específicas e correlacionadas com tumor de seio aumentou sensivelmente e além de "Mastectomia" já apareceram, na nova hierarquia conseguida, palavras como "Lactação" e "Menopausa".

3a EXPERIÊNCIA:

(1) d = perfil = SEIO(T)

(2) $$S(SEIO(T), i) = \frac{F(SEIO(T), i) - E(SEIO(T), i)}{F(i)}$$

(3) Resultado: Com este novo coeficiente de similitude o algorítmo grupou entre outros, termos tais como

*
*
*

Pós-menopausa
Pré-menopausa
Ovariectomia
Câncer avançado
Lactação
Mastectomia
Mamografia

*
*
*

Já nesta última tentativa praticamente foram eliminadas do subconjunto submetido à classificação (120 termos) todas as palavras-chave não especificamente relacionadas com "Tumor do Seio".

C O N C L U S Ã O

Evidentemente, tal como apresentado no sexto capítulo, o algorítmo para classificação automática de palavras-chave não pode ainda ser diretamente implementado em computador. Não foi esta a intenção ao introduzí-lo neste capítulo. Seria preciso antes uma pesquisa complementar para efeito de maior detalhamento e conhecimento dos parâmetros e conceitos requeridos pelo algorítmo. Contudo, a finalidade dos dois modelos de metodologias apresentados na Parte II foi apenas ilustrar aspectos teóricos do processamento automático discutidos anteriormente. O objetivo foi, antes de tudo fazer ver os diferentes níveis de dificuldades na automação da gestão e recuperação documentais.

Nos capítulos da Parte I foi estabelecida uma sistematização para o estudo do processamento automático de biblioteca. Sabendo-se da carência de trabalhos nacionais na área do processamento de dados em geral, pode-se constatar a necessidade e a utilidade de estudos particularmente voltados para o processamento de dados documentais. Ao ser mostrado, portanto, o universo da aplicação computacional ao tratamento não numérico buscou-se oferecer subsídios à Analistas de Sistemas, Bibliotecários e Documentalistas que de quaisquer formas venham a se envolver com a automação de biblioteca ou entidade congênere.

Como proposição para futuros estudos, deixa-se a tarefa de aprofundamento de cada aplicação sistêmica aqui abordada, uma vez que, a preocupação maior foi com a abrangência de assuntos. Quando, no entanto, tal iniciativa for tomada não seja supervalorizada a abordagem de itens isolados em detrimento de outros. Pois, será a criação de uma massa de informação qualitativa e

quantitativamente diversificada que enriquecerá a bibliografia nacional sobre Informática Aplicada e Recuperação de Informação.

REFERÊNCIAS BIBLIOGRÁFICAS

| 1 | SALTON, Gerard. Dynamic information and library processing. New Jersey, Prentice-Hall, 1975. 523 p.

| 2 | DOYLE, Lauren B. Information retrieval and processing. California, Melville Publishing Company, 1975. 410 p.

| 3 | LANCASTER, F. Wilfrid. Information retrieval systems, characteristics, testing, and evaluation. New York, John Wiley & Sons, 1968. 222 p.

| 4 | CHAUMIER, Jacques. Les techniques documentaires. Paris, Presses Universitaires de France, 1971. 111 p.

| 5 | DIAS, Donaldo de Souza & GAZZANEO, Giosafatte. Projeto de Sistemas de Processamento de Dados. Rio de Janeiro, Livros Técnicos e Científicos, 1975. 150 p.

| 6 | WOLFF-TERROINE, M. & RIMBERT, D. & ROUAULT, B. La classification automatique: Son utilisation pour la constitution du langage et l'interrogation des systemes documentaires. Revue Française d'Automatique, Informatique et Recherche Opérationnelle, Paris, R.A.I.R.O, 6^{e} année, Juin, 1972.

| 7 | MATTOS, Antônio Carlos Marques. Informática: O sistema de palavras-chave do contexto. Revista de Administração de Empresas, Rio de Janeiro, 12(4):24-39, out/dez.1972.

|8| VICENTINI, Abner Lellis Corrêa et alii. Mecanização da Classificação Decimal Universal: O projeto LEMME. R. Bi-Bibliotecon, Brasília, 1(1), jan./jun. 1973.

|9| UNESCO. FRANÇA. Diretrizes para a elaboração e desenvolvimento de thesauri monolingues destinados à recuperação de informações. Paris, jul. 1970. Tradução.

|10| FREEMAN, R. R. & ATHERTON, P. File organization and search strategy using UDC in mechanized reference retrieval systems. In: Proceedings of the F.I.D/I.F.I.P. Joint Conference, Rome, June 14-17, 1967, p. 122-152.