CDU 002.6:681.3

INTERACÇÃO DA ECONOMIA DA AUTOMATIZAÇÃO NUM SISTEMA DE RECUPERAÇÃO DE GRANDE DIMENSÃO

L. Rolling e J. Piette
Euratom, C. I. D., Bruxelas

Tradução de:
Maria Cecília Barbosa de Melo

1. INTRODUÇÃO

Este folheto tem como finalidade ilustrar o impacto dos critérios económicos na concepção e operação de vários componentes de um sistema de recuperação da informação de grande dimensão, em especial de componentes que envolvem um certo grau de mecanização. Apresentamos como exemplo o sistema Euratom.

2. ECONOMIA DA CONCEPÇÃO DO SISTEMA

2.1 PORQUÊ A AUTOMATIZAÇÃO? PORQUÊ A ECONOMIA?

É quase lugar-comum justificar com a chamada *"explosão literária"* a introdução de computadores no tratamento da informação não numérica.

Contribui também para esta tendência para a automatização o facto de a nossa sociedade tecnológica precisar, hoje mais do que nunca, da informação e de os requisitos de qualidade exigidos num serviço de documentação, em termos de rapidez, exaustividade e precisão serem cada vez maiores[1].

A automatização é a meta, pois se espera que o computador trate maior quantidade de informação de uma maneira mais rápida e melhor do que os ficheiros e catálogos do documentalista, já fora de moda.

O documentalista argumenta que a automatização é ainda demasiado cara para ser aplicada aos problemas complexos da documentação.

Qual será realmente o custo da automatização, aplicada à documentação? E serão os critérios económicos aplicáveis à documentação, já que ninguém consegue medir o valor real da informação, nem mesmo aqueles que a recebem e a utilizam?

Há quem pense que a informação é tão necessária à riqueza de um país que devia ser gratuita, como um serviço público; que pelo menos o ónus do investimento para a criação de um serviço de documentação não devia pesar sobre os utilizadores. Esta opinião, aliada à ideia de que um sistema deve ser tanto melhor quanto mais dinheiro for gasto no seu desenvolvimento, levaram à tendência para elaborar sistemas de documentação sofisticadas que são caros, tanto na concepção como na operação. Parece haver uma tendência especial para sistemas completamente automatizados mesmo nos casos em que uma operação parcialmente automatizada poderá prestar serviços comparáveis e menos dispendiosos.

Quando se consegue o apoio do governo para actividades de *"Investigação e Desenvolvimento"*, esse apoio em geral desaparece no momento em que o sistema se está a tornar operacional!

2.2 AUTOMATIZAÇÃO COMPLETA OU PARCIAL?

Como a automatização é, hoje em dia, sinónimo de progresso, quase se tem vergonha de conceber um sistema que tenha componentes não automáticos. Por isso se criou o hábito de ignorar ou de negar qualquer espécie de importância àqueles componentes de sistemas que não possam ser automatizados. Parte-se do princípio que estes componentes não fazem parte do sistema mas são em parte complementares dum sistema que é *"completamente"* automatizado.

A falta de um documentalista que seja parte-do-sistema e que possa dedicar uma hora do seu tempo a cada uma das questões postas pelo utilizador, obriga muitas vezes este a gastar mais algumas horas para poder aproveitar o serviço prestado pelo sistema *"completamente automatizado"* mas apesar disso incompleto.

Parece-nos que a afirmação de A.Weinberg: *"o cientista que trabalha deve...partilhar muitas das tarefas tradicionalmente desempenhadas pelo documentalista profissional"* [2] deve ser interpretada com certa cautela.

Pelo contrário, a automatização deve ajudar o documentalista profissional a aliviar o cientis-

ta no trabalho de reunir a sua documentação própria com base em catálogos e no seu próprio sistema de classificação. O tempo do cientista activo é demasiado precioso para ser gasto com a aquisição dos conhecimentos práticos necessários par fazer o computador responder àquilo que ele deseja. Além disso, o tempo do computador é também demasiado caro para ser gasto na afinação de métodos.

Também por estas razões é pouco provável que alguma vez venhamos a ter uma documentação completamente computorizada e, apesar disso, competitiva. Algumas operações componentes continuarão a ser realizadas de uma maneira mais barata, usando outras técnicas, e o emprego de documentalistas especializados, qualquer que seja a sua remuneração, ficará menos dispendioso em muitas operações, especialmente se se pretende atingir um alto grau de eficiência e precisão.

2.3 *A ECONOMIA DE COMPONENTES DO SISTEMA*

O progresso no tratamento da informação consistiu sempre na divisão do processo num número cada vez maior de componentes de sistema, cada vez com maior número de intermediários entre o produtor e o utilizador da informação.

A primeira fase começou muito antes da nossa era, com a criação da Biblioteca de Alexandria; a segunda, com a criação da *"Chemisches Zentralblatt"*, por volta de 1830; a terceira fase teve como consequência o desenvolvimento de sistemas de classificação e índices de assunto.

A fig.1 ilustra a situação actual, que consiste numa cadeia de sete operações que incluem:

1) Escrever, publicar, colocar
2) Dar a notícia bibliográfica, fazer o *"abstract"*, catalogar
3) Classificar, indexar, codificar
4) Recuperar pela referência
5) Recuperar pelo *"abstract"*
6) Recuperar pelo documento
7) Ler

Note-se que as fases 1,2 e 3 envolvem a compressão de toda a informação em formas mais condensadas, ao passo que as fases 5,6 e 7 envolvem a expansão da informação seleccionada, endossando-a para a forma original.

Há para cada uma destas operações constituintes vários métodos de implementação manual e mecânica.

Não vamos considerar aqui nem a alínea 1, mecanização do processo de publicação e do tratamento na biblioteca, nem a alínea 7, transferência da informação do documento para o leitor.

Cada uma das outras fases apresenta uma alternativa manual e uma ou mais alternativas mecanizadas. Temos de considerar em cada caso não só o custo do trabalho mas também o custo da transferência da informação do meio manual para o meio usado pelo processamento automático e da sua recuperação.

A fase 2, fazer os *"abstracts"*, por exemplo, oferece como alternativas o *"abstract"* feito por um documentalista que leu todo o documento e o *"auto-abstracting"*, método proposto pela primeira vez por H.P. Luhn [3] . Enquanto o custo de um *"abstract"* pode ser da ordem dos $10.00 em qualquer dos casos, o custo do armazenamento do texto do documento (e da impressão do *"abstract"* resultante) é de cerca de $50.00. Isto, aliado ao facto de o *"abstract"* feito pelo documentalista ser hoje em dia de muito melhor qualidade, impediu que o *"auto-abstracting"* se tornasse um processo aceitável. A fase 3, indexação (ou codificação) de um *"abstract"*, pode realizar-se manualemente, atribuindo códigos de uma tabela de classificação ou descritores de um vocabulário mais ou menos controlado. Pode também ser feita por um computador, fazendo a correspondência de cada uma das palavras do *"abstract"* com um dicionário de termos admissíveis e não admissíveis, armazenados em fitas magnéticas ou discos, e seleccionando um certo número de termos de acordo com um conjunto de regras representadas no seu *"software"* [4, 5, 6, 7, 8] .

O custo da indexação normal e da indexação mecânica é da ordem dos $1.00, de tal modo que o homem e a máquina seriam hoje competitivos se não fosse o enorme investimento exigido para criar dicionários da máquina eficientes (ver o item 3.4 e ref. [9]) e *"software"*.

A recuperação pela referência (fase 4) pode ser feita manualmente se a colecção for pequena ou se a pesquisa puder ser feita superficialemnte. Mas para uma colecção de grande dimensão com um alto padrão de rapidez e exaustividade, não há outro modo senão a automatização.

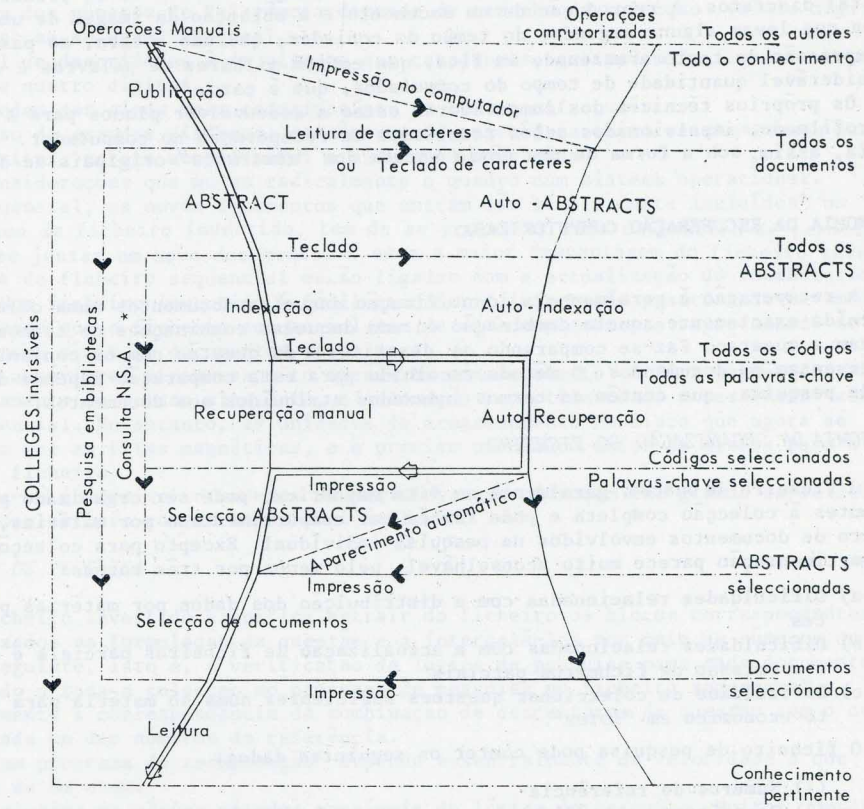
As fases 5 e 6 são as que conduzem desde a referência recuperada até ao texto do *"abstract"* e do documento. Se o documento completo (fase 2) ou o *"abstract"* (fase 3) foram armazenados em fita magnética, o custo da impressão no computador não é muito mais elevado do que a recuperação manual do *"abstract"* no ficheiro ou do documento na estante mais a subsequente fotocópia. Mas se o registo dos dados foi demasiado caro exclui-se a alternativa da impressão no computador.

Enquanto todas as fases componentes de um sistema podem em teoria ser mecanizadas, é evidente, pelas considerações anteriores, que a mecanização é relativamente barata para processos que envolvam informação altamente *condensada*, ou seja, informação codificada, ao passo que o seu custo é proibitivo para o tratamento de informação *expandida*, isto é, informação em linguagem natural.

2.4 *CONCEPÇÃO DO SISTEMA ÓPTIMO*

De acordo com o que foi exposto, o sistema mais económico seria aquele que está representado em

linhas duplas na fig.1. Foi esse o escolhido pela Euratom para o seu Sistema de Documentação Nuclear, de que vamos a seguir fazer um ligeiro esboço:



(*N.T. - N.S.A. Nuclear Science Abstracts)

Fig. 1. Economia da automatização
Operações caras — Soluções económicas
Em desenvolvimento

(2) A Euratom colecciona *"abstracts"* a partir de periódicos de *"abstracts"* e também de documentos primários; criou ou apoiou a criação de três periódicos de *"abstracts"* em assuntos insuficientemente cobertos. A verificação por duplicação é automatizada.

(3) A indexação é baseada em *"abstracts"*. É feita por especialistas as matérias em alguns institutos científicos.

(3/4) Os termos indexados são armazenados em fita magnética. O contrôle dos erros é automático.

(4) Usa-se um computador IBM 360/40 para a recuperação. A organização do ficheiro e a lógica da pesquisa estão descritos no capítulo 3.

(4/5) A saída do computador faz-se em números de referência, ou números de referência e descritores.

(5) Os *"abstracts"* são fotocopiados dos ficheiros para o utilizador e faz-se o enquadramento manual.

(6) Podem fornecer-se, a pedido do utilizador, documentos pertinentes, que não estejam sob direitos de autor, ou em tamanho natural ou em microcópias.

Vários dos outros princípais centros de informação seguem o mesmo processo, com pequenas modificações devidas a particularidades das suas colecções de documentos, a actividades editoriais e ao tipo de utilizadores [10, 11, 12].

2.5 *INFLUÊNCIA DO "HARDWARE" DISPONÍVEL*

A economia do tratamento da informação depende em larga medida das características do *"hardware"* disponível [13].

Utilizar um computador de grande capacidade da última geração representa um aumento da capacidade de cálculo.

Mas há ainda muito trabalho a fazer para construir o necessário *"solftware"* note-se que o *"sol*

tware" que as operações de tratamento da informação de grandes dimensões requerem tem de ser completamente diferente das condensações clássicas para gestão e processamento de dados.

Os computadores electrónicos não são a única solução para o problema do tratamento da informação 14. A indústria óptica e fotográfica também desenvolveu *"Hardware"* adaptável à prdução e reprodução de documentos e à recuperação da informação (Equipamento Filmorex e Filesearch).

Os sistemas ópticos têm a vantagem de a informação não ter de se processar na forma de quanta (bits) discretos. A reprodução de um documento e a obtenção da imagem de uma microficha são operações que levam alguns segundos do tempo do copiador, que não é caro, ao passo que o armazenamento e a impressão do texto armazenado em fita, que contem milhares de palavras e caracteres, gasta uma considerável quantidade de tempo do computador, que é caro.

Os próprios técnicos dos computadores estão a desenvolver planos para a obtenção de material microfilmado, impulsionados pelos resultados da recuperação no computador. A saída de resultados seria, assim, sob a forma de uma cópia barata dos *"abstracts"* originais de documentos.

3. ECONOMIA DA RECUPERAÇÃO COMPUTORIZADA

A recuperação é geralmente a identificação daqueles documentos numa colecção aos quais foi atribuída exactamente aquela combinação ou uma daquelas combinações de termos indexados que representam a questão. Faz-se comparando os descritores da questão com os conjuntos de descritores que representam os documentos. O método escolhido para esta comparação depende da organização do ficheiro de pesquisa, que contém os termos indexados atribuídos aos documentos.

3.1 ECONOMIA DA ORGANIZAÇÃO DO FICHEIRO

O ficheiro de buscas, geralmente em fita magnética, pode ser organizado para conter os dados pertinentes à colecção completa e pode também ser compartimentado por matérias, de modo a reduzir o número de documentos envolvidos na pesquisa indivídual. Excepto para colecções muito grandes esta compartimentação parece muito aconselhável, pelo menos por três razões:

a) Dificuldades relacionadas com a distribuição dos dados por matérias parcialmente coincidentes
b) Dificuldades relacionadas com a actualização de ficheiros parciais e o tratamento separado e a omissão de ficheiros parciais
c) Dificuldade de coleccionar questões suficientes numa só matéria para garantir o processamento económico em *"lotes"*.

O ficheiro de pesquisa pode conter os seguintes dados:

(1) Números de referência
(2) Descritores ou códigos
(3) Dados de catalogação, tais como título, autor e indicações da fonte
(4) *"Abstracts"*.

(1) e (2) são indispensáveis, ao passo que o autor e os dados da fonte só deverão estar na fita de pesquisa se as pesquisas por autor ou fonte forem necessárias. Os *"abstracts"* não devem estar numa fita de pesquisa.

Se os resultados vão ter a forma de dados bibliográficos impressos e *"abstracts"*, é preferível proceder em duas fases:

(a) Recuperação de números de referência a partir de um ficheiro de pesquisa
(b) Extracção e impressão dos dados a partir de um ficheiro mestre (ficheiro bibliográfico).

A organização do ficheiro de pesquisa pode ser ou sequencial (ficheiro linear, ficheiro serial, ficheiro directo) ou invertida.

No ficheiro sequencial, cada número de documento é seguido pelos termos indexados atribuídos; no ficheiro invertido, os descritores (ou códigos), em ordem alfabética (ou numérica), são seguidos pelos números do documento a que foram atribuídos.

Cada uma destas duas fases, ou até uma combinação das duas, pode ser mais económica; depende de:

- O tamanho da colecção
- O volume do thesaurus
- A extensão dos descritores da sua representação em código
- O modo de contrôle dos erros
- A numeração de referência
- A frequência da actualização.

O custo da recuperação é proporcional ao tempo do computador requerido, o qual depende em larga medida da extensão da fita magnética que tem de ser processada.

Se a extensão normal dos descritores ou códigos é da mesma ordem que a extensão dos números de

referência, é fácil determinar a influência do tamanho da colecção: a ordenação sequencial é preferível para colecções muito pequenas, ou seja, colecções em que o número de documentos é mais pequeno que o número de descritores no *thesaurus;* a ordenação invertida é de certo modo preferível no caso de colecções maiores.

A extensão relativa dos números de referência e descritores desempenha um papelsignificativo.

A extensão mínima dos números de referência depende do tamanho da colecção (cinco dígitos até 99.999 items, ou seis, se se usam ligações).

A extensão normal de descritores é de oito ou dez caracteres, que podem facilmente ser substituídos por códigos de quatro dígitos.

Ambos os dados podem ser ainda mais reduzidos por um código alfa-numérico.

Desde que, no caso de grandes sistemas, os números de referência sejam mais extensos que as representações do descritor, parece aconselhável a ordenação invertida.

Mas há outras considerações que mudam radicalmente o quadro num sistema operacional.

Num ficheiro sequencial, os novos documentos que entram são simplesmente incluídos no fim do ficheiro. Num sistema de ficheiro invertido, tem de se processar todo o ficheiro de pesquisa de todas as vezes que se juntar um novo documento. É esta a maior desvantagem do ficheiro invertido.

As maiores falhas do ficheiro sequencial estão ligadas com a actualização do vocabulario e o contrôle dos erros. Num ficheiro invertido é fácil transferir informação posicionada de um descritor para outro, ao passo que o ficheiro sequencial tem de ser completamente reprocessado se se mudar um só descritor numa série de documentos.

As considerações anteriores dizem respeito apenas ao armazenamento em fita magnética. No caso de se usar o armazenamento de acesso aleatório, a ordenação invertida tende a ser mais económica que a ordenação sequencial. No entanto, as unidades de armazenamento em disco que agora se utilizam são mais caras do que as fitas magnéticas, e é preciso um número maior de discos para o processamento de um grande ficheiro.

O quadro modifica-se quando se introduzem os Registos, que oferecem uma grande capacidade de armazenamento de acesso aleatório com um custo aceitável.

3.2 ECONOMIA DO PROCESSO DE PESQUISA

Pesquisar num ficheiro invertido significa extrair do ficheiro os blocos correspondentes a todos os descritores usados na formulação da questão e a intercalá-los por meio de números de referência. O processo seguinte, isto é, a verificação da lógica de pesquisa para cada documento, é o mesmo que foi aplicado a toda a colecção no processo de pesquisa no ficheiro sequencial.

Faz-se sucessivamente a correspondência da combinação de descritores da questão com o conjunto de descritores sob cada um dos números de referência.

A realização de um programa de recuperação depende essencialmente da velocidade a que se faz esta correspondência de um a um.

Mas examinemos primeiro os vários métodos possíveis de lógica de pesquisa. Muitos descritores são relevantes para as necessidades do utilizador; correspondem a um número mais limitado de conceitos básicos (8 descritores para 3 conceitos no exemplo da fig.2/3). Há estratégias de pesquisa com base em descritores e com base em conceitos; cada uma delas pode ser melhorada por prioridades estatísticas e probabilísticas e por processos de ordenação. Vejamos o seguinte exemplo de uma estratégia com base em descritores: seleccione todos os documentos representados (entre outros) pelos descritores A, B, C, D, E, e F. Como a muito poucos documentos serão atribuídos todos os seis descritores, uma condição estatística mudaria a questão para: seleccione todos os documentos que contenham quatro de entre os descritores A, B, C, D, E e F.

A lógica Booleana é uma estratégia com base em conceitos. Os descritores que representam o mesmo conceito sao ligados como alternativas pela operação disjuntiva OR; os grupos de descritores que representam conceitos a combinar (*"coordenação de conceitos"* equivale a *"indexação coordenada"*) são conectados pelo operador conjuntivo AND. Os descritores podem ser excluídos da formulação da questão pelo operador (AND) NOT.

Um exemplo da lógica de Boole: seleccione todos os documentos que contenham um dos descritores A ou B ou C, e um dos descritores D ou E ou F, e um dos descritores G ou H.

Uma estratégia baseada em descritores melhorada por prioridades levaria a:

A—4, B—4, C—2, D—2, E—1, F—1, $Q = 12$.

O peso de corte 12 seria atingido por qualquer uma das combinações ABCDEF, ABCDE, ABCDF, ABCEF e ABCD, mas já não seria atingido, por exemplo, pelas combinações ABCE, ACDEF, ou BCDEF. A saída de uma pesquisa pode ser ordenada de acordo com os pesos cumulativos.

Uma formulaçao Booleana também pode ser melhorada por prioridades, dando maior importância, na formulação de uma questão, a um só descritor dentro de um grupo de conceitos ou a um só grupo de conceitos dentro da formulação da questão [18]. Uma questão de prioridades Booleana seria assim: (A—3 ou B—2 ou C—1) e (D—3 ou E—2 ou F—1) e (G—2 ou H—1). Um corte de 6 limitaria os resultados das combinações ADG; BDG; ADH; CDG; AFG; BDH; AEH; e excluiria as combinações AFH, CDH, BEH, BFG, CEG; BFH, CEH, CDF; CFH. Na falta de um corte, as referências resultantes podem ser ordenadas de acordo com o total dos valores mais altos para cada conceito.

Quanto mais complicada for a lógica de pesquisa, menor é a velocidade de adequação documento/questão. Adicionar pesos e adequar o total com um valor de corte leva evidentemente mais tempo do que o simples testar da presença ou ausência de um descritor. É também considerável o tempo de computador que a ordenação dos resultados de pesquisa requer.

Em ordem a um prcesso de recuperação rápido e por consequência económico, a lógica de Boole, sem rotinas de prioridades e ordenação, foi escolhido como a estratégia de pesquisa para o Sistema de Documentação Nuclear Euratom.

3.3 A SOLUÇÃO EURATOM

As considerações seguintes levaram a uma rotina ainda mais rápida e económica:

A maioria dos documentos numa grande colecção não são pertinentes à questão que está a processar-se. Assim, no decurso do processo de correspondência (descritores da questão com descritores dos documentos) num determinado momento há uma decisão negativa: um dos descritores da questão não se encontra entre os descritores do documento. Esta decisão põe fim à correspondência para este documento.

Quanto mais cedo se puder tomar esta decisão mais rápido será o processo. É por isso aconselhável ordenar os descritores da questão por crescente frequência de uso. Se o descritor menos usado tem uma frequência de uso de 500 numa colecção de 500.000 documentos, isso significa que para 999 documentos em 1.000 a correspondência envolve só um descritor da questão.

Na formulação de uma questão Booleana que envolva ANDs e ORs as frequências de descritores alternativa num só grupo de conceitos devem juntar-se, para que esta regra se mantenha válida. Por isso é também aconselhável ordenar os grupos de conceitos por números crescentes de descritores alternativa.

O tempo preciso para a correspondência de descritores é proporcional ao número de caracteres, ou melhor, o número de *bits* (ou *bytes*) usados para a representar na fita magnética.

Um descritor em linguagem natural, com uma média de nove caracteres, leva ± 72 bits (numa extensão variável). Se os descritores forem substituídos por códigos de quatro dígitos, o número de bits reduz-se para 32 (numa extensão fixa).

Se o número de bits puder reduzir-se a um, o tempo de correspondência podia ser reduzido outras tantas 32 vezes. Isto seria , se todos os descritores tivessem de ser idênticos, mas é realmente possível se os bits que representam os descritores, em número limitado, puderem ser identificados pela sua *posição* no local de armazenamento.

No sistema Euratom, que tem um *thesaurus* [15] de 4665 descritores, cada documento pode ser representado por uma matriz binária de 4665 posiçoes, em que cada posição corresponde a um só descritor. Os descritores realmente atribuídos (uma média de 15) têm um 1 nas posições correspondentes; todas as outras (4650) posições são preenchidas com zeros.

Se o descritor ZIRCONIUM for o primeiro mencionado na questão, a posição 4622 é testada primeiro; se contém um bit zero, está acabado o processo de correspondência para o documento correspondente.

A fig. 2 representa a rotina de correspondência na formulação de uma questão Booleana complexa. A fig. 3 é a representação em fluxograma do programa de pesquisa.

O tamanho considerável da matriz binária que representa cada documento cria uma dificuldade; 4665 bits ocupam uma grande extensão de fita e, em consequência disso, o tempo de pesquisa seria consideravelmente aumentado pelo tempo de a fita.

Ultrapassou-se esta dificuldade condens... a matriz binária em forma codificada para armazenamento em fita e revelando-a na sua forma expandida para recuperação só no local de armazenamento do computador.

O tempo de pesquisa é realmente reduzido a uma pequena proporção do tempo de processamento total, e o tempo de rodar a fita passa a ser o maior factor de custo. Por isso é económico processar as questões em lotes.

A experiência demonstrou que não há diferença sensível no tempo requerido para o processamento de uma ou de trinta questões, devido à sobreposição do canal de entrada.

No sistema Euratom não se usam as prioridades pois isso iria ocupar uma parte considerável do tempo do computador. Porém toda a selecção que podia ser feita com as prioridades pode também fazer-se a partir de uma formulação Booleana.

Por outro lado o método usado no sistema Euratom envolve a formulação de várias *"sub-questões"* para a mesma questão.

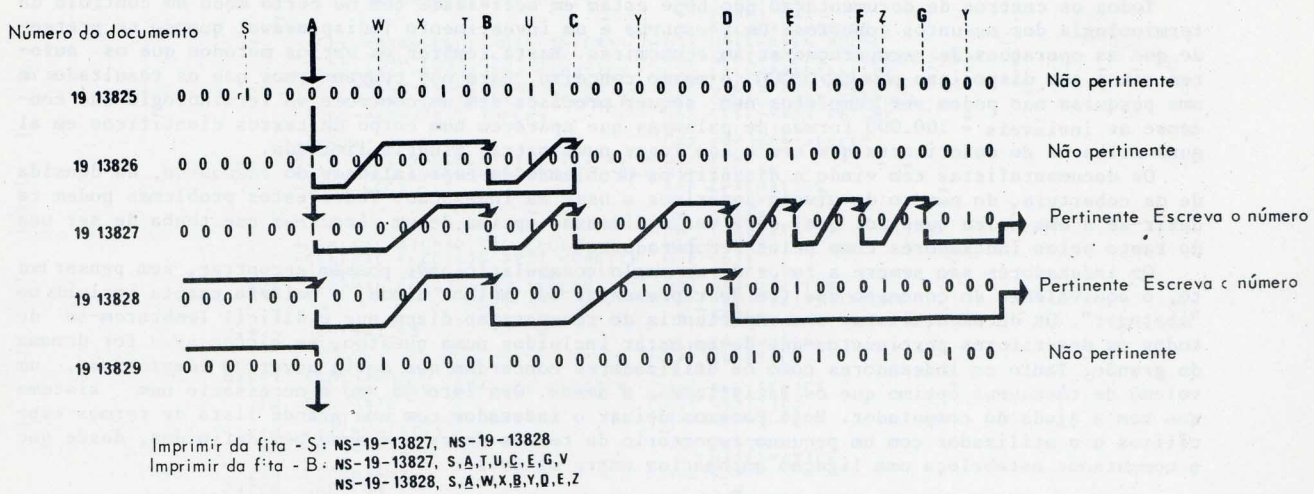
As *"sub-questões"* podem ser *"redundantes"*; podem ser também *"inclusas"*, isto é, os resultados da questão *"perdida"* (nº i+1) incluem os resultados da questão *"de continuação"* (no. i). Este processo produz uma ordenação das referências recuperadas em que as respostas às subquestões *"de continuação"* são as mais pertinentes.

A possibilidade que o documentalista tem de processar várias pesquisas paralelas sem um custo adicional dá-lhe uma flexibilidade sem precedentes. Nem ele nem o computador perdem tempo em testes.

É mesmo possível melhorar a ideia do utilizador determinando antecipadamente o número aproximado de referências que serão recuperadas pela formulação de uma dada questão. A fórmula

$$R_n = \frac{f_1 f_2 \cdots f_n}{y^{n-1}} k^{n-1}$$

em que f_i é a frequeência acumulada de uso dos descritores num grupo, n é o número de grupos de des-



Questão típica:
A e (B ou C) e (D ou (E e (F ou G) ou H).
CORROSÃO e (ZIRCÓNIO ou ZIRCALOY) e (VAPOR ou EVAPORAÇÃO) ou HUMIDADE).

Fig. 2 . CONSULTA DE MATRIZES BINÁRIAS

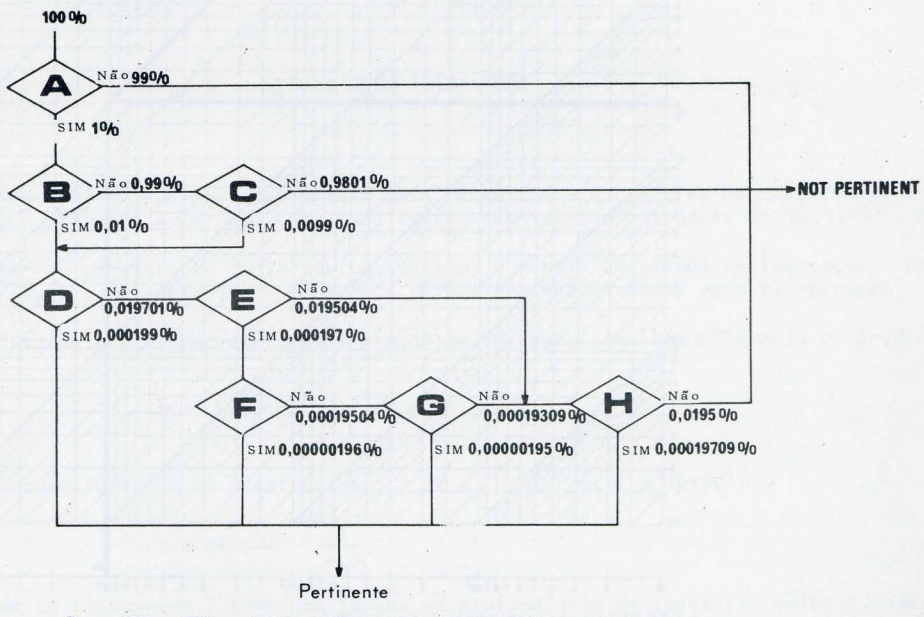


Fig. 3 - PROCESSAMENTO LÓGICO DE UMA QUESTÃO BOOLEANA TÍPICA

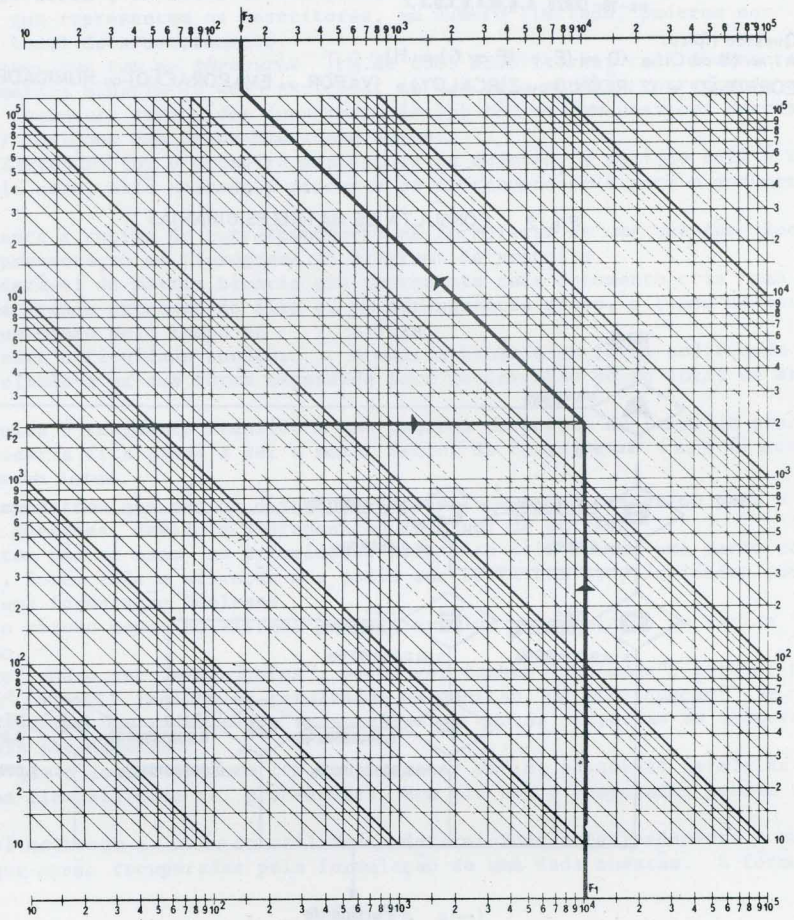
critores, *K* o factor de associação e *V* o tamanho da colecção, tem sido aplicada com êxito a mais de 1.000 questões. Na intenção de evitar cálculos cansativos com números de frequência muito grandes, G. Romerio da Euratom/CID desenvolveu um nomograma logarítmo (fig. 4) que substitui os cálculos por uma elementar operação com régua de cálculo.

3.4 ECONOMIA DO CONTROLE DA TERMINOLOGIA

Todos os centros de documentação que hoje estão em actividade têm de certo modo um contrôle da terminologia dos assuntos cobertos. Um *thesaurus* é um investimento indispensável quando se pretende que as operações de recuperação sejam económicas. Basta lembrar os vários métodos que os autores têm à sua disposição para expressar o mesmo conceito, para nos convencermos que os resultados de uma pesquisa não podem ser completos nem sequer precisos sem um controle da terminologia que condense as inviáveis ± 100.000 formas de palavras que aparecem num corpo de textos científicos em alguns milhares de descritores que têm o seu lugar numa matriz binária limitada.

Os documentalistas têm vindo a discutir os problemas da especialidade do *thesaurus*, da densidade da cobertura, do número de níveis genéricos a usar na indexação. Todos estes problemas podem reduzir-se a uma única questão: qual deve ser a dimensão óptima de um *thesaurus* que tenha de ser usado tanto pelos indexadores como pelos recuperadores?

Os indexadores são sempre a favor de um amplo vocabulário onde possam encontrar, sem pensar muito, o equivalente do conceito que tem de representar ou, melhor ainda, a palavra exacta incluída no *"abstract"*. Os documentalistas com experiência de recuperação dizem que é difícil lembrarem-se de todos os descritores pertinentes que devem estar incluídos numa questão, se o *thesaurus* for demasiado grande. Tanto os indexadores como os utilizadores concordam que devia haver um compromisso, um volume de *thesaurus* óptimo que os satisfizesse a ambos. Ora isto já não é necessário num sistema que tem a ajuda do computador. Hoje podemos deixar o indexador com uma grande lista de termos específicos e o utilizador com um pequeno repertório de termos de recuperação bem definidos, desde que o computador estabeleça uma ligação automática entre os dois.



EXEMPLO: $F_1 = 10.000$; $F_2 = 2.000$
RESULTADO: $F_3 = 160$

Fig. 4 - MONOGRAMA QUE DA O NÚMERO DE REFERÊNCIAS ESPERADAS

```
      17  GELL-MANN MODEL                    USE NUCLEONS
                                             +  PIONS
                                             +  SCATTERING
      84  GELL-MANN THEORY                   USE QUANTUM NUMBER
                                             +  STRANGENESS
      12  GELL-MANN-BRUECKNER MODEL          USE ELECTRONS
                                             +  FERMI GAS
                                             +  NUCLEAR MODELS
          -GELL-MANN-GOLDBERGER THEORY       USE GELL-MANN MODEL
          -GELL-MANN-LEVY RELATION           USE GOLDBERGER-TREIMAN RELATION
          -GELL-MANN-NISHIJIMA RELATION      USE GELL-MANN THEORY
          -GELL-MANN-NISHIJIMA SCHEME        USE GELL-MANN THEORY
          -GELL-MANN-OKUBO MASS FORMULA      USE OKUBO MASS FORMULA
          -GELL-MANN-PAIS THEORY             USE GELL-MANN THEORY
          -GELS                              USE COLLOIDS
          -GEMINI PROJECT                    USE PROJECT APOLLO
          -GEN. EL. VALLECITOS BOIL. W. R    USE VBWR
          -GEN. NUCL. ENGINEERING CO. BWR    USE BWR-M1
          -GENE LOCI                         SEE CHROMATIDS
                                             OR  CHROMOSOMES
                                             OR  GENES
                                             OR  GENETICS
          -GENERAL ATOMIC GAS-COOLED R.      USE GACOR
          -GENERAL ELECTRIC TEST REACTOR     USE GETR
          -GENERAL ELECTRIC-SIEMEN CORRES    USE GESCR
     123  GENERAL RELATIVITY THEORY          USE GRAVITATION
                                             +  RELATIVITY THEORY
                                             +  SPACE
   3,424  GENERATORS
     213  GENES                              USE CHROMOSOMES
                                             +  GENETICS
   4,004  GENETICS
          -GENEVA UNIV. AGN-201 REACTOR      USE AGN SERIES
      14  GENOTYPE                           USE GENETICS
       1  GENTILLY POWER REACTOR             USE REACTORS
     468  GEOCHEMISTRY
      23  GEODESY                            USE EARTH
                                             +  MATHEMATICS
     188  GEOGRAPHY                          USE EARTH
   2,860  GEOLOGY
      84  GEOMAGNETIC COORDINATES            USE GEOPHYSICS
                                             +  MAGNETIC FIELDS
                                             +  ZONES
          GEOMAGNETIC CUT-OFF RIGIDITY       USE COSMIC RADIATION
                                             +  EARTH
                                             +  ENERGY RANGE
                                             +  MAGNETIC FIELDS
       7  GEOMAGNETIC EQUATOR                USE EQUATOR
                                             +  GEOPHYSICS
                                             +  MAGNETIC FIELDS
       3  GEOMAGNETIC POLES                  USE GEOMAGNETIC COORDINATES
      48  GEOMAGNETIC STORMS                 USE GEOPHYSICS
                                             +  MAGNETIC FIELDS
                                             +  TURBULENCE
     180  GEOMAGNETISM                       USE GEOPHYSICS
                                             +  MAGNETIC FIELDS
     232  GEOMETRY                           USE CONFIGURATION
   1,785  GEOPHYSICS
          -GEORGE WASHINGTON SUBMARINE       USE SSW SERIES
      64  GEORGIA                            USE USA
          -GEORGIA TECH. RESEARCH REACTOR    USE GTRR
          -GEOSTROPHIC FORCE                 USE CORIOLIS FORCE
```

Fig. 5 - EXTRACTO DE THESAURUS

O Thesaurus Euratom [15] é o primeiro a incluir mais termos não-descritores específicos do que descritores; é também o primeiro a ter uma estrutura de referências que faz parte do *"software"* do computador.

A fig. 5 apresenta-nos um extracto do Thesaurus Euratom. O número de termos de indexação foi estimado em cerca de 10.000, enquanto que o número de termos realmente usados para recuperação é muito inferior.

C. Vernimb, encarregado das operações da recuperação no Euratom, combinou a "formula de prognóstico da recuperação"

$$R_n = \frac{f1f2...fn}{V^{n-1}} \; k^{n-1}$$

com a definição da frequência média de descritores $f = aV / T$ para obter a fórmula

$$T = ak \frac{n}{R_n k} \frac{V}{}$$

que apresenta o volume do *thesaurus* T como uma função da profundidade de indexação média a, o factor de associação (ou redundância) k, o tamanho da colecção V, e Rn que é o número de documentos recuperados numa pesquisa que combine n, grupos de descritores. Esta equação mostra que:

(1) O volume do *thesaurus* tem de ser maior se os documentos tiverem de ser indexados com mais pormenor (a);

(2) Se o *thesaurus* aumentar de volume, haverá maior redundância de indexação (k);

(3) Uma colecção maior exige um *thesaurus* maior mas o aumento está longe de ser proporcional (V);

(4) A dimensão do *thesaurus* depende daquilo que o utilizador pretende: um *thesaurus* pequeno é

suficiente para a compilaçao de grandes bibliografias, ao passo que um *thesaurus* grande é melhor para o processamento de questões precisas (*Rn*).

(5) A dimensão do *thesaurus* depende da política de pesquisas: pode ser relativamente pequeno para indexação coordenada e pesquisa Booleana, mas teria de ser muito grande para a pesquisa de índices ($n=1$).

Um outro problema de terminologia importante — a formulação da questão — podia aproveitar de um conceito que tem estado a ser testado em vários lugares: o conceito de semelhança semântica.

Se fosse possível definir numericamente o grau de semelhança entre pares de descritores relacionados semanticamente, bastaria ao documentalista alimentar o computador com dois ou três descritores que melhor expressassem os conceitos centrais da sua questão, e confiar ao computador a tarefa de juntar, como descritores alternativa, aqueles que tivessem o maior grau de semelhança com os conceitos centrais.

3.5 ECONOMIA DO CONTROLE DE ERROS

O problema do controle de erros não pode ser resolvido por aqueles que concebem os sistemas, pois eles não sabem onde os erros vão surgir. Estes erros, porém, em especial na fase de entrada, constituem um factor da maior importância para a economia do sistema de documentação.

Os erros que mais frequentemente afectam a entrada são erros de indexação sistemáticos, cometidos pelos indexadores, erros de grafia e numeração também dos indexadores, e erros de perfuração tanto nos descritores (ou códigos) como nos números de referência.

Os erros na escrita do código e na perfuração do código não podem ser detectados, porque um erro de codificação produz muitas vezes outro código. É por isso que os descritores não são codificados nesta fase do trabalho do Euratom.

A percentagem de erros de perfuração não pode ser reduzida drasticamente pela perfuração *dupla;* é um processo dispendioso, indispensável no tratamento de problemas de gestão e cálculos científicos, mas que pode evitar-se no processamento do texto.

Os erros de perfuração e de grafia podem ser facilmente detectados fazendo a comparação de todos os termos armazenados com o *thesaurus*, pelo custo de duas operações de ordenação, antes e depois do controle.

A operação cara, porém, não é a detecção destes erros mas a sua *correcção*. A correção manual não só exige mão-de-obra como leva um certo tempo, durante o qual os documentos correspondentes não podem ser introduzidos no sistema. Por isso é importante que a maior parte os erros sejam removidos imediatamente através de uma rotina de correcção *automática*.

Alguns dos termos desconhecidos detectados pela correspondência do *thesaurus* correspondem a erros muito elementares, causados pela adição ou omissão de um caracter, substituição de um caracter por outro ou inversão de dois caracteres adjacentes. Podem ser tratados por um programa de correcção automática, mudando, por exemplo, REATORS, REACATORS, REALTORS e REATCORS em REACTORS. O erro de ortografia mais complicado (RCACTOR) não pode ser corrigido por um programa deste tipo.

Os termos de indexação mais curtos (quatro letras ou menos) não devem ser processados por um programa como este porque há uma probabilidade relativamente grande de que uma combinação desconhecida de caracteres possa resultar da ortografia errada de dois termos de indexação diferentes: BONY pode ser a grafia errada de BONE ou BODY. Erros de ortografia complexos ou pequenos devem portanto ser processados manualmente, juntamente com os termos recentemente introduzidos, que podem ser incorporados ou nos termos aceites ou nos termos rejeitados.

As correcções individuais podem ser armazenadas numa fita separada e usadas como dicionário complementar no tratamento de subsequentes lotes de material de indexação.

Pareceria lógico, no processamento de lotes ulteriores, aplicar este *"dicionário de erros"* aos termos desconhecidos que ficam *depois* da aplicação do programa automático de correcção da grafia. Mas verificou-se que o programa automático consome mais tempo de computador por termo corrigido do que a correspondência ao dicionário. Como se vê na fig.6, a correspondência ao dicionário precede o processamento automático e os erros corrigidos pela rotina automática são acrescentados ao dicionário de erros.

A rotina de correcção automática podia ser posta de parte, se todos as possíveis adições, omissões, substituições e inversões fossem incluídas no dicionário de erros; mas o número de erros possíveis é tão grande que também este método não seria económico.

Todos estes processamentos tratam de erros acidentais resultantes da introdução de termos desconhecidos. Mas há vários outros tipos de erros além dos termos ou números mal escritos. Um erro na numeração feito pelo indexador ou um perfurador, por exemplo, terá como resultado a colocação dos descritores atribuídos num número correspondente a outro documento, e esses descritores vão juntar-se num conjunto de descritores duplos e enquanto o número correcto não vai aparecer na colecção. As Rotinas automáticas podem identificar os conjuntos de descritores duplos e as falhas.

A maneira mais económica de corrigir este tipo de erro é anular os conjuntos de descritores duplos e preencher ambos os tipos de lacuunas perfurando de novo os conjuntos de descritores.

Um outro tipo de erro de que não resultam termos mal ortografados é a indexação incorrecta pela substituição de um descritor por outro. Em vez de testar os termos atribuídos um por um, uma simples rotina automática aplicável a homógrafos e outros termos ambíguos pode assegurar que o uso errado destes termos é imediatamente detectado.

O programa baseia-se na verificação do contexto semântico dos termos individuais. O uso errado do descritor PLASMA (que no Sistema Euratom representa o gás ionizado) no sentido de plasma sanguíneo será detectado pela co-ocorrência de PLASMA com outros descritores que digam respeito ao campo da biologia. O uso errado de MERCÚRIO (o metal) para o planeta Mercúrio será detectado procurando a co-ocorrência de MERCÚRIO com termos respeitantes ao campo da astronomia.

A má qualidade da indexação pode ser remediada pela indexação paralela feita por duas pessoas [16], mas geralmente é mais económico modificar a estratégia de pesquisa para explicar uma certa inconsistência na indexação.

A necessidade de algumas rotinas de correcção de erros faz aumentar o custo da automatização aplicado à documentação, de tal modo que o volume de colecção crítico (o limite acima do qual a automatização passa a ser competitiva) se aproxima mais da marca de documentos 100.000 do que do nível de item 10.000.

4. ECONOMIA DA UTILIZAÇÃO DO SISTEMA

Os serviços de informação tradicionais, não mecanizados, serviam um número limitado de utilizadores que tinham de estar próximo do local em que a colecção, os ficheiros e os catálogos estavam instalados.

A mecanização, na medida em que fez aumentar a capacidade de processamento desses centros, alargou também o número de utilizadores que podem ser servidos a partir de uma localização central. A velocidade das comunicações postais não aumentou com a mesma dimensão. Por isso um dilema se nos depara: por um lado, a capacidade crescente e a diversidade do *hardware* impulsionam a criação de grandes centros de informação centralizada; por outro lado, a necessidade de serviços rápidos gera a tendência para a descentralização dos serviços. O telefone e o teleprocesso são meio de comunicação excelentes para a transmissão dos pedidos de informação ou questões e até para pequenas quantidades de dados registados, mas já não podem cobrir grandes quantidades de "*stock*" impresso. Estas considerações, aplicadas à situação actual da documentação relativa à ciência e à tecnologia, levam às seguintes conclusões:

Uma capacidade aumentada de *hardware* convida a um máximo de centralização.

O conhecimento acumulado de cada um dos ramos da ciência (medicina, química, metalurgia, física, engenharia) pode hoje ser armazenado e processado nas memórias de um dos maiores computadores [17].

No entanto, a centralização do processamento de todo o conhecimento científico e técnico a nível regional ou nacional não é aconselhável, visto que as várias matérias requerem diferentes métodos de processamento e vocabulários diferentes.

Isto não deve impedir as autoridades nacionais de agrupar no mesmo local as instituições nacionais que contribuem para os diferentes centros de documentação internacionais. Claro que não é fácil conseguir a concordância internacional para a criação de um centro de informação. É por isso que na prática, os centros de documentação internacionais orientados por assuntos se vão desenvolver através de maiores facilidades nacionais ou regionais. A NASA foi a primeira a dar maior extensão às suas actividades de recuperação da informação, cobrindo a aviação e a tecnologia do espaço, na Organização Europeia de Pesquisa (ESRO) seguindo-se-lhe a Biblioteca Nacional de Medicina com a criação das ramificações MEDLARS na Grã-Bretanha e na Escandinávia (a que se seguiram outras). A física, a química, e o campo nuclear estão prestes a seguir estes exemplos. O sistema desenvolvido pela EURATOM não requer quaisquer mudanças para servir de base a uma cooperação mais ampla.

Se a tendência para a centralização se deve principalmente ao aumento de capacidade de *hardware*, ela já não engloba necessariamente os componentes manuais. O "*abstract*" e a indexação, que, são ainda melhores e mais baratos feitos manualmente, podem e devem continuar descentralizados. Além disso, os centros nacionais estão, pelo menos em teoria, em melhor posição para coligir e explorar a literatura dos seus respectivos países do que os organismos internacionais. Isto também se aplica aos componentes não-computorizados da saída do sistema, tais como a formulação da questão, o enquadramento dos resultados e a reprodução de "*abstracts*" e documentos para o utilizador.

Há dois requisitos prévios para a saída descentralizada: um é que o pessoal em serviço deve ser muito bem treinado a usar o sistema, o outro é que deve poder dispor-se de colecções de documentos completas e correntemente actualizadas.

5. CONCLUSÃO

Há um compromisso entre as operações manuais e as operações mecanizadas que permite conseguir um óptimo económico. De todas as vezes que um "*hardware*" melhorado aparece no mercado, o óptimo dá um passo na direcção da automatização. Mas o óptimo da economia operacional não pode muitas vezes atingir-se porque a transição do "*hardware*" velho para o novo iria desencadear operações, requerer um investimento em custo de máquinas, horas de programador e transferência e modificação da informação, ou criar imcompatibilidades com outros utilizadores do sistema "*descentralizado*".

A economia é como um traço vermelho que percorre todas as partes de um sistema. Todos os critérios de qualidade podem realmente ser expressos em termos de custo. Na fig.7 o custo por item armazenado e o custo de uma questão processada são planeados pelos parâmetros usados para a avaliação de rapidez, perfeição e precisão. É de notar que estas curvas são todas idênticas. O custo sobe exponencialmente quando os requisitos de qualidade aumentam. Passar de 50% para 90% de eficiên-

cia fica mais barato do que cobrir os últimos 10%. Note-se que as curvas representam um sistema óptimo, que é o limite de sistemas reais representados pela área que está sobre as curvas. Podíamos resumir o significado destes gráficos por estas palavras: *"A perfeição é dispendiosa mas também se pode operar um sistema caro sem perfeição"*.

* A fig. 7, realmente, não representa um conjunto ordenado de valores determinados no EURATOM mas tão somente uma indicação das ordens de grandeza e tendências que se encontram.

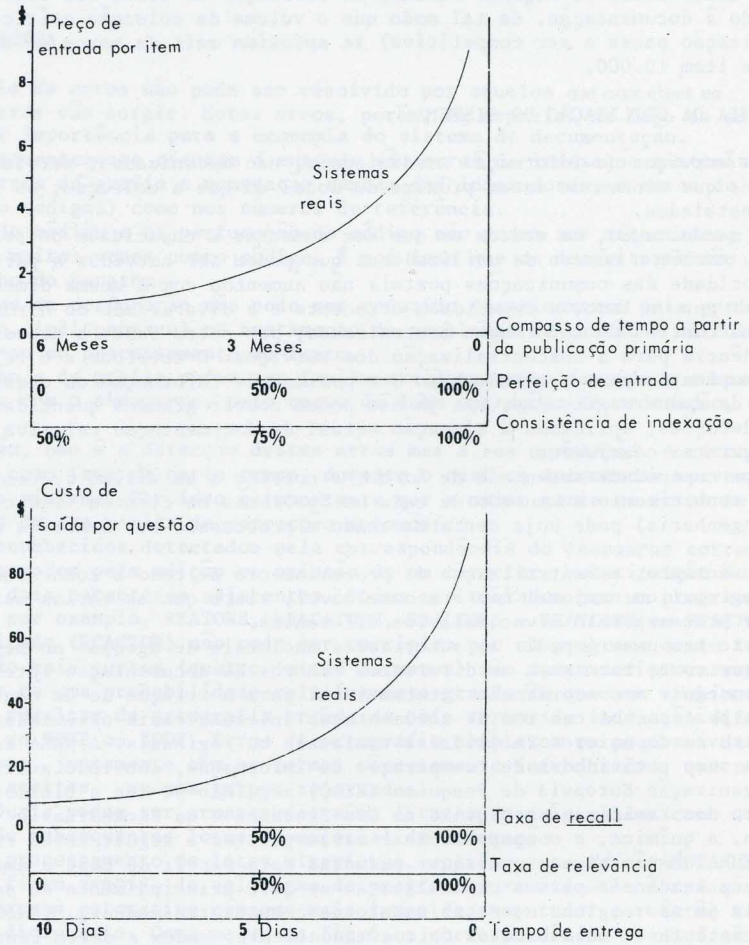
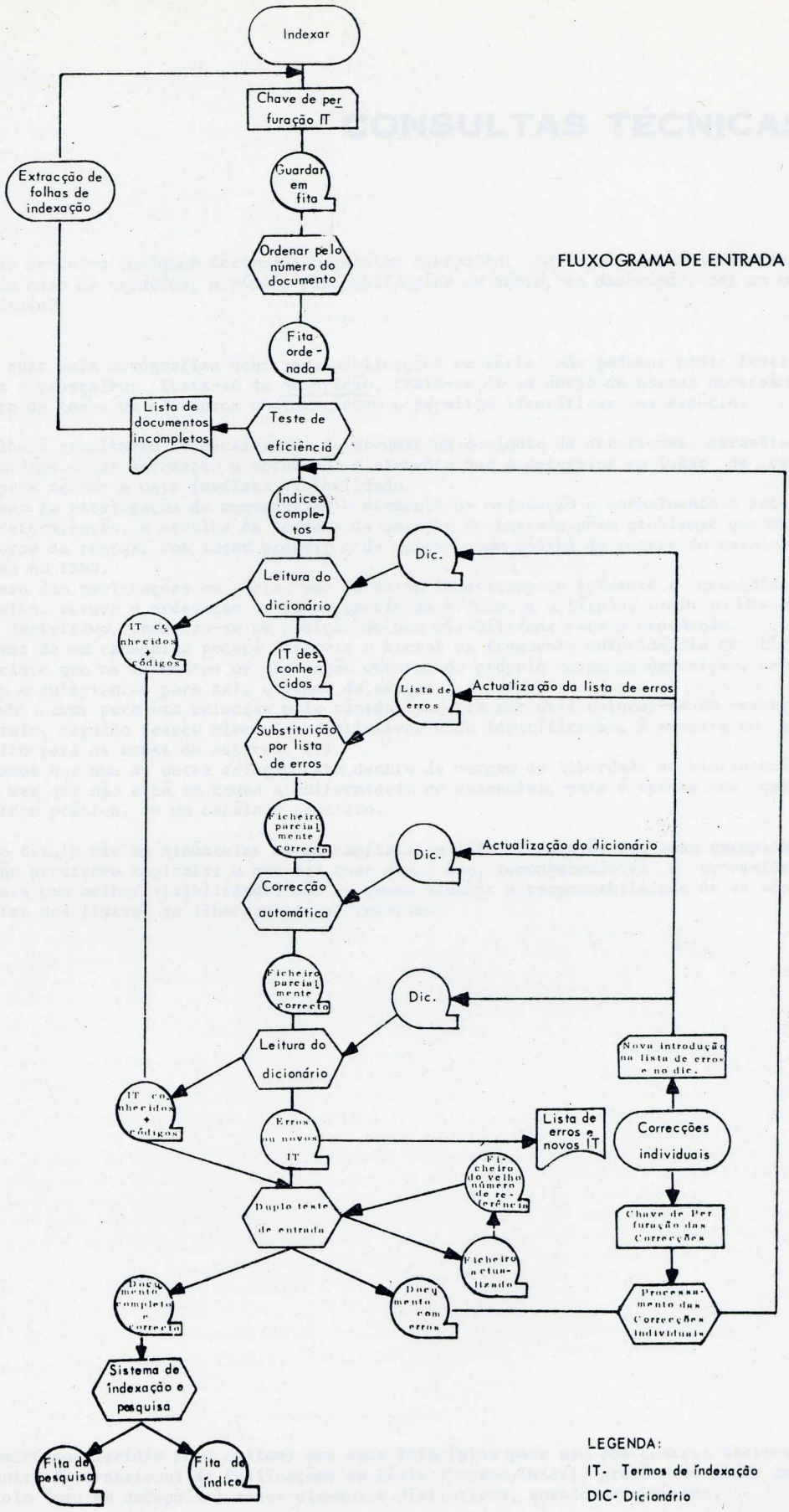


Fig. 7 - O CUSTO DO APERFEIÇOAMENTO



FLUXOGRAMA DE ENTRADA

LEGENDA:
IT- Termos de indexação
DIC- Dicionário

Fig. 6 - FLUXOGRAMA DE ENTRADA