CDU 002.6:80

COMUNICAÇÃO CIENTÍFICA E TECNOLÓGICA: A DISSEMINAÇÃO SELETIVA DE INFORMAÇÕES

JORGE EDUARDO FREUND

Assessor Técnico no Centro de Pesquisas Informáticas do IPT.

MARI TOMITA

Analista de Sistemas no Centro de Pesquisas Informáticas do IPT.

Para qualquer tipo de atividade na área de pesquisa e desenvolvimento em ciência e tecnologia, a informação é um insumo dos mais importantes. Poucos são, no Brasil, os serviços destinados a atender às necessidades de informação de técnicos e cientistas. O Instituto de Pesquisas Tecnológicas do Estado de São Paulo S/A – IPT desenvolveu um sistema de Disseminação Seletiva de Informações (DSI) para atender a seus técnicos e a interessados de outras instituições brasileiras.

O presente trabalho conceitua de maneira simples um sistema DSI e descreve as soluções adotadas para a implantação de tal sistema no IPT. É comentada a filosofia que orientou o desenvolvimento do software para o sistema implantado em computador de pequeno porte do Instituto.

INTRODUÇÃO

Mais de um milhão de artigos ("papers") em ciência e tecnologia são produzidos anualmente, e entregues à comunidade através de publicações especializadas. Mesmo numa área específica, é muito grande o número de novas publicações (Fig. 1).

Informação precisa e atualizada pode ser tão importante para o trabalho de um técnico ou pesquisador quanto boas instalações e bons equipamentos. Impossível porém, a qualquer um, ler, ou ao menos tomar conhecimento, de tudo o que é publicado; o problema é de seleção do que realmente interessa. É necessário auxiliar o leitor a selecionar o que realmente deve ser lido, de acordo com sua particular área de atuação ou interesse.

Diversas técnicas de tratamento de informação surgiram recentemente, acompanhando c crescimento exponencial do volume de publicações.

Disseminação Seletiva de Informações (DSI) é uma destas técnicas e foi definida por Lunh em 1961 como "o serviço dentro da organização que se encarrega da canalização de novas informações, qualquer que seja a fonte, para aqueles locais na organização onde é alta a probabilidade destas informações serem úteis". Em outras palavras, DSI é o serviço que, conhecendo os interesses específicos de cada usuário, procura selecionar (em geral com auxílio de computador), e encaminhar-lhe as informações mais relevantes à sua área de atuação. O propósito do sistema não é encontrar um determinado documento que atenda a uma necessidade específica; é uma sistemática de processamento e encaminhamento seletivo de novas informações recebidas pela organização, de uma forma rotineira e periódica.

DISSEMINAÇÃO SELETIVA DE INFORMAÇÕES

Disseminação Seletiva de Informações baseia-se no processamento em computador de perfís de usuários contra "data-base" de referências bibliográficas. Um sistema DSI consiste de:

i – "data-base" de referências bibliográficas
ii – perfil de interesse de cada usuário
iii – um sistema para processamento dos perfís contra o "data-base"
iv – saída de referências selecionadas em forma de relatório ou de fichas catalográficas
v – possibilidade de ajustes dos perfís de interesse
vi – possibilidade de fornecimento de documentos referenciados
vii – realimentação e determinação de estatísticas.

Em sua grande maioria, os sistemas DSI se utilizam de fitas de referencias bibliográficas produzidas por entidades especializadas em tratamento de informaçoes. As mais utilizadas são MEDLARS (Ciências biomédicas), CHEMICAL ABSTRACTS (química), COMPENDEX (engenharia e tecnologia), INSPEC (física, eletronica. computadores e controle e outros).

As referências bibliográficas são originadas no tratamento das principais revistas técnicas publicadas no mundo todo, além de anais de congressos, livros e outros, passando cada artigo por um processo de indexação e resumo. As referências consistem de indicação de título, autor (es), palavras-chave e resumo ("abstract") sendo outras informações incluídas de acordo com a particular fonte de informações e com a área de atuação (p. ex. CODEN, fórmula-molecular, etc.).

Especial atenção deve ser dada, num sistema DSI, para a elaboração dos perfís e para seu processamento, de modo a só fornecer ao usuário as informações que realmente lhe interessam, não omitindo informações de interesse (fig. 2). O sucesso do sistema depende fundamentalmente de sua precisão.

O Instituto de Pesquisas Tecnológicas do Estado de São Paulo S.A. – IPT está im-

plantando um sistema DSI processando inicialmente as fitas COMPENDEX produzidas pela Engineering Index Inc. e fornecidas mensalmente. O sistema é descrito a seguir.

O SISTEMA DSI

O sistema implantado no IPT consta dos seguintes módulos (fig. 3):

a. processamento e cadastramento dos perfís de interesse
b. reformatação da fita COMPENDEX, formando o "data-base" com formato adequado ao processamento
c. pesquisa no "data-base" e recuperação das referências relevantes a cada perfíl
d. formatação de saída, e emissão das listagens personalizadas, utilizando o Sistema Etiquetas de Endereçamento (SEND) desenvolvido no IPT.

As seguintes considerações orientaram a elaboração do sofware do sistema DSI:

1. modularização dos programs e arquivos permitindo a implantação do sistema no computador de pequeno porte (B-1726) do IPT
2. programação em COBOL prevendo uma possível conversão para outros sistemas
3. conversão das fitas COMPENDEX para formato próprio do IPT, permitindo no futuro o processamento de outras fitas e a formação de um banco de dados para numa 2ª fase implantar um sistema de busca retrospectiva.

ELABORAÇÃO DE PERFÍS DE INTERESSE

Um perfil é formado por uma série de termos que representam o interesse do usuário, e interelacionados numa expressão lógica. Por ser a fita COMPENDEX produzida nos Estados Unidos, a indexação das referências é feita em inglês, e por este motivo os perfís são elaborados também em inglês. Para minimizar os problemas decorrentes deste fato, é utilizado o "thesaurus" do Engineering Index como referência. Não é necessário porém se ater somente aos termos deste thesaurus visto que a indexação é feita com vocabulário livre.

Aos termos que descrevem um determinado perfil são designados rótulos (letras de A a Z) e podem indicar nomes de autores (tipo A) palavras de texto (tipo T), CODEN (tipo K) ou classificação Card-A-Lert do Engineering Index (tipo C). Cada rótulo pode ser formado por um ou mais termos que serão interpretados como sinônimos, sem que seja necessário especificá-los individualmente na expressão lógica.

A expressão lógica, que relaciona entre si os diversos termos do perfil, é formada pelos diferentes rótulos, por parentesis, e pelos operadores $\wedge$ (e), $\vee$ (ou) e $\neg$ (não).

Cada termo, que pode se constituir de uma ou mais palavras é limitado em 40 caracteres. Permite-se, para maior flexibilidade, o truncamento de palavras à direita e/ou esquerda.

No verso da planilha são prenchidas as informações acerca do usuário necessárias à utilização do sistema de endereçamento. A mesma planilha é utilizada também para alteração ou cancelamento de perfís de interesse.

Na fig. 4 o exemplo de um perfil para um usuário interessado em obter referências relativas ao **código 723** do Card-A-Lert (Sofware) relativas a **Sofware de Computador** e especificamente em **técnicas Recursivo-Descendentes** ou **compiladores** (ou **interpretadores**). O usuário deseja ainda obter referências das publicações de **Gries**, porém não deseja saber das publicações de **Hopgood**.

PROCESSAMENTO DE PERFÍS

Os perfís de insteresse dos usuários do sistema são prefurados em cartões de 96 colunas e após um processamento de consistência são armazenados num arquivo em disco, ficando então prontos para o processamento de pesquisa no "data-base". O programa utilizado tanto na criação do arquivo de perfís quanto em sua atualização (inclusão, modificação ou exclusão de perfís), compõe-se de:

a) crítica de consistência: verificação de campos alfabéticos e numéricos, da validade dos rótulos da expressão lógica contra os rótulos dos termos do perfil, da validade dos tipos de termos, da validade dos operadores e da formação da expressão lógica. Não são permitidos operadorea $\wedge$ e $\vee$ consecutivos. Os perfís com erro são rejeitados e os erros indicados em listagem.

b) expansão da expressão lógica: a expressão lógica do perfil correto é expandida com a introdução de sub-rótulos ligados pelo operador $\vee$ para os diferentes termos que apareçam em um mesmo rótulo (sinônimos).

Ex: No perfil do exemplo da fig. 4 as palavras COMPIL * e INTERPRET * são utilizadas como sinônimos no mesmo rótulo D. As palavras COMPLIT * e INTERPRET * são truncadas à direita de modo a permitir recuperação de referências indexadas com COMPILER, COMPILERS, COMPILATION.

INTERPRETER, INTERPRETERS, INTERPRETATION, etc. São criados dois sub-rótulos D1 e D2 e a expressão lógica é expandida para:

$$A \wedge (B \wedge C \wedge (D1 \vee D2) \vee G) \wedge \neg F \vee E)$$

c) transformação em notação polonesa: a expressao lógica expandida é transformada para notação polonesa (pós-fixa) simplificando com isso seu processamento

posterior, pois os operadores ficam relacionaods na ordem em que as operações devem ser executadas. No algoritmo para transformação em notação polonesa é usada uma estrutura de pilha e são atribuídos níveis de prioridade aos operadores, a saber:

¬ : prioridade 4
∧ : prioridade 3
∨ : prioridade 2
⊂ : prioridade 1
⊃ : prioridade 0

Operandos são transcritos para a expressão pós-fixa e os operadores colocados na pilha até que surja operador de menor prioridade, quando estão são transcritos para a expressão pós-fixa.

Ex: No perfil do exemplo da figura 4 temos:
expressão lógica expandida:

A ∧ (B ∧ C ∧ ((D1 ∨ D2) ∨ G) ∧ ¬ F ∨ E)

expressão lógica expandida em notação polonesa:

A B C ∧ D1 D2 ∨ G ∨ ∧ F ¬ E ∨ ∧

d) geração de arquivos: dois arquivos são utilizados pelo programa para armazenamento dos perfís:

– o arquivo PERFIL guarda o código do usuário, e a expressão lógica em notação polonesa, onde os operandos (termos) são substituídos pelos seus endereços no arquivo OPERANDO.

– o arquivo OPERANDO guarda os termos relacionados no perfil juntamente com o respectivo tipo e comprimento.

Ex: para o perfil do exemplo da Fig. 4, na página seguinte.

Para indicar truncamento ou não, são inseridos brancos apropriadamente antes e/ou depois dos termos.

Dependendo de indicação na planilha o programa fará a inclusão, exclusão ou modificação de perfil no arquivo.

Para inclusão de determinado perfil não deve haver outro já cadastrado com o mesmo código de usuário.

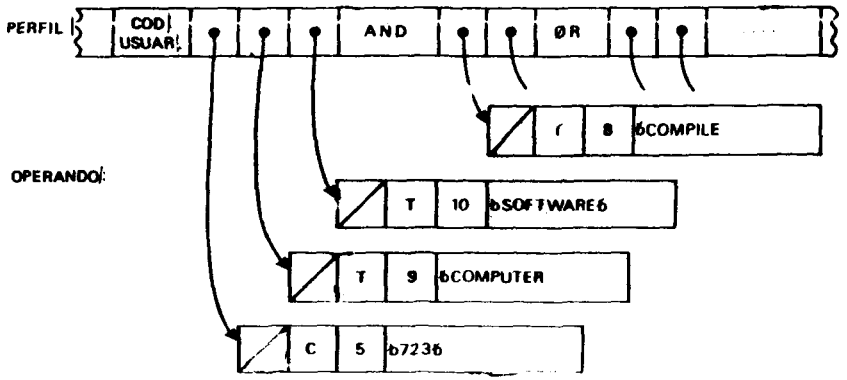


Fig. 4

Na exclusão o sistema localiza o perfil em questão também pelo código do usuário, efetuando então a exclusão.

Numa modificação de perfil são efetuadas sucessivamente as operações de exclusão e inclusão. Desta forma não se permite alteração parcial de perfil. O perfil deve sempre ser reescrito totalmente qualquer que seja a alteração a ser efetuada.

Qualquer anormalidade no processo de Inclusão, Exclusão ou Modificação é acusada pelo programa, não sendo nestes casos realizada a operação em questão.

Será implantada numa segunda fase uma rotina para simplificação das expressões lógicas. A implantação desta rotina é justificada pelo fato de que, uma vez cadastrados, os perfis serão utilizados inúmeras vezes no processamento de "data-bases" e portanto a simplificação efetuada uma única vez poderá acelerar os processamentos mensais.

PESQUISA E RECUPERAÇÃO

As fitas COMPENDEX recebidas mensalmente são processadas contra o arquivo de perfis. Os sistema de processamento consta de:

a) Geração do data-base:

A fita Compendex (mensal) é composta de aproximadamente 6000 referências gravadas em registros de tamanho variável sendo que cada registro contém no máximo 8000 bytes e a média gira em torno de 1200 bytes. Com o intuito de aumentar a eficiência do sistema, ao invés de manipular os registros no formato original, as informações foram estruturadas de modo a obter arquivos inversos em disco, em forma de lista ligada para títulos, autores, etc. Um arquivo mestre contém para cada registro (referência bibliográfica) os apontadores para os diversos arquivos inversos (fig. 5).

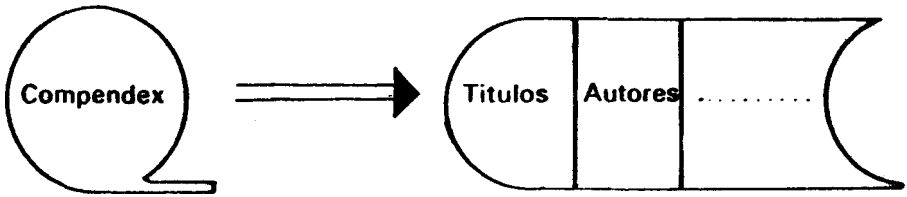


Fig. 5

A partir desta estrutura, a pesquisa poderia ser feita de uma única vez considerando-se que cada arquivo inverso (títulos, autores) teriam em média 6000 registros no mês. Porém, se ocorrer alguma anormalidade, precisar-se-ia manipular todos os arquivos desde o início.

Para acelerar o processo de recuperação em caso de anormalidade no processamento, dividimos cada um desses arquivos em n arquivos distintos formando módulos menores. Cada módulo é constituído de m registros do arquivo de títulos, m registros do arquivo de autores etx., sendo a pesquisa então feita por módulos.

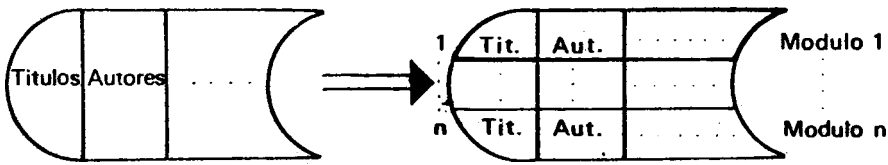


Fig. 6

n é um número escolhido pelo próprio analista e a escolha deve ser feita levando-se em consideração a capacidade do disco e o tempo de processamento de cada módulo.

b) Seleção de perfís e Pré-processamento:

Utilizamos como entrada o arquivo "PERFIL" que contém expressões lógicas dos perfís de usuários em notação polonesa. Caso se deseje a recuperação de referências para um sub-conjunto de perfís existe um programa que seleciona os perfís que devem ser pesquisados. Para o programa de recuperação é indiferente se o arquivo "PERFIL" contém um conjunto ou sub-conjunto de perfís. Ainda, caso uma empresa especializada em uma determinada área queira cadastrar seus técnicos ou cientistas para se manterem atualizados, o sistema fornece uma opção para que os perfís desses funcionários sejam pesquisados não em todo o data-base, mas somente naquelas referências recuperadas para o perfíl da empresa em questão.

c) Pesquisa de perfís no "data-base":

Os termos (palavras) podem ser procurados nos campos de título, autor, "headings", "sub-headings", "free-language terms", "cross-reference terms" e códigos Card-A-Lert de acordo com a especificação de cada um e conforme formatação das fitas COMPENDEX (fig. 6).

Nesta fase não é feita pesquisa no campo de "abstract". Esta pesquisa além de excessivamente demorada não é relevante na grande maioria dos casos.

A avaliação das expressões lógicas é feita utilizando-se uma pilha em que são armazenados passo a passo os resultados de cada pesquisa (verdadeiro ou falso) e o resultado de avaliação de expressões lógicas parciais (verdadeiro ou falso).

O número do documento é o número do perfil que se relacionam são gravados em um arquivo de trabalho "WKELEOK".

Para o número do documento que não se relaciona com o perfil gravamos seu número do perfil e o estado da expressão que forma o perfil em um outro arquivo de trabalho "ESTADO".

Com isto permitimos em fase posterior tentar recuperar documentos procurando os termos nos "abstracts" (se a especificação do termo pede a procura no "abstract") usando os arquivos "ESTADO" e "ABSTRACT".

Os documentos recuperados nesta fase serão adicionados ao arquivo "WKELEOK".

Note que com a modularização dos arquivos e seleção de perfís permitimos recuperação de um determinado perfil nos documentos referentes a mes (es) anterior (es) sem a necessidade de processar todos os arquivos para todos os perfís.

d) Emissão de listagens:

Para cada elemento do "WKELEOK" procuramos as informações nos arquivos a fim de montar a saída em forma de listagem.

Porém, antes de imprimi-las diretamente na impressora gravamos essas informações em um outro arquivo de trabalho "WKOBRAOK".

Classificamos esse arquivo por ordem numérica de perfil e listamos os documentos por usuário. Nesta fase procuramos informações do usuário no arquivo "ETIQUE-

TAS" e imprimimos também o nome na listagem. Através desse arquivo emitimos também etiquetas de endereçamento para remeter as listagens pelo correio (usuário externo). São sempre impressas as referências completas, inclusive "abstract", para cada pesquisa bem sucedida.

CONCLUSÕES

O sistema Disseminação Seletiva de Informações tem como objetivo o fornecimento de informações atualizadas aos técnicos do IPT e clientes externos. Através de contatos com o IBICT – Instituto Brasileiro de Informação em Ciência e Tecnologia – procurou-se elaborar um sistema compatível com outro já existente e criar um padrão a ser adotado em todo país.

Outros "Data-Bases" poderão ser processados em função das necessidades dos usuários e do meio técnico e científico em geral.

O técnico do IPT, de posse da referência bibliográfica produzida pelo sistema DSI poderá, através da biblioteca do Instituto, solicitar o encaminhamento do documento em questão.

Depois de um período de operação normal do sistema, serão feitas estatísticas de sua utilização, com o objetivo de verificar quais as áreas mais solicitadas, número de referências apontadas, etc.

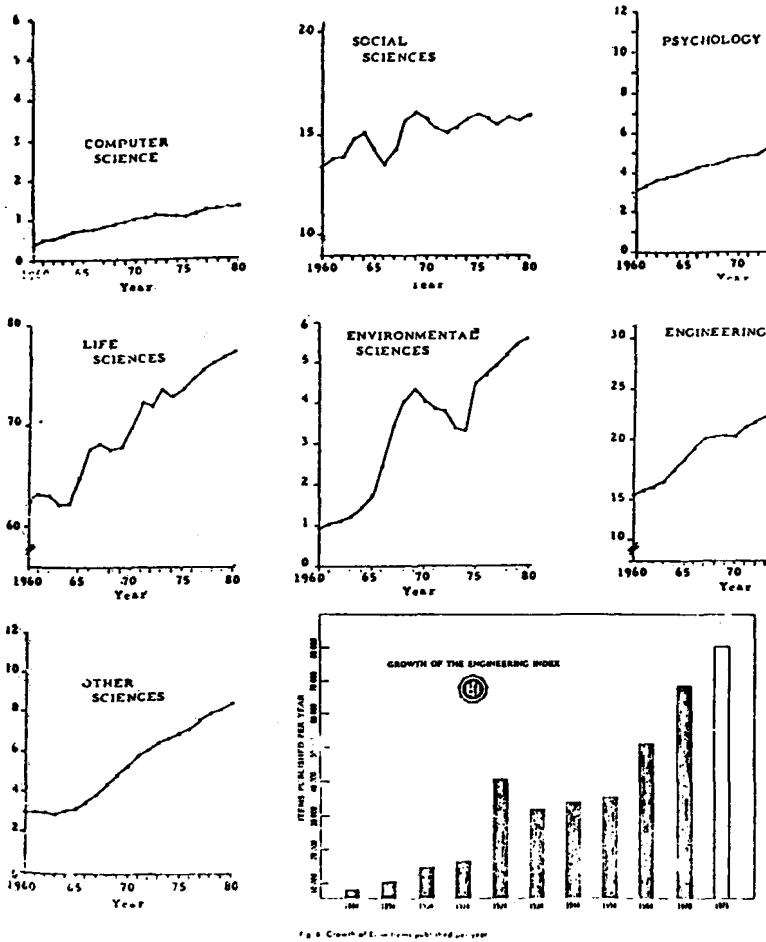
Da maior importância para um sistema DSI é a realimentação por parte do usuário. A partir das informações provenientes do usuário, acerca de relevância, ou não, das referências emitidas e da própria solicitação de documentos, será feita a análise da acuracidade dos perfís de interesse e seu possível ajuste.

Um perfil de interesse não é limitado a um indivíduo. Pode sim, indicar as áreas e assuntos de interesse de um agrupamento, de um projeto ou mesmo de uma empresa.

Os primeiros perfís processados no IPT dizem respeito a agrupamentos e projetos. Com isso pode-se atender a todas as áreas do Instituto e, paulatinamente, individualizar os perfís.

Usuários externos ao IPT serão atendidos após o sexto mês de operação do sistema, isto é, após superados os problemas naturais da implantação do sistema e feitos os ajustes necessários.

NUMBER OF U. S. S&T SCHOLARLY JOURNAL ARTICLES BY FIELD OF SCIENCE: 1960-1980



Fontes: King Research Inc
Engineering Index Inc

Fig. 1

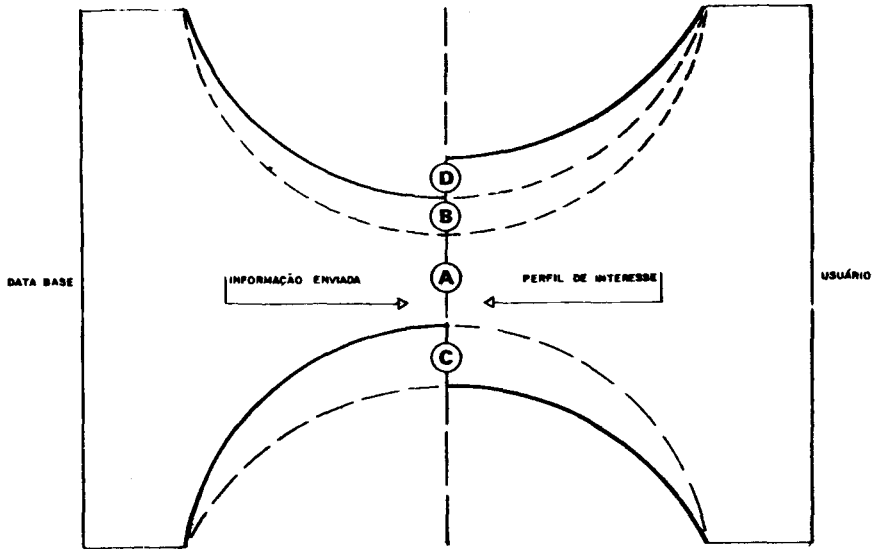


Fig. 2

INTERFACE USUÁRIO/SISTEMA DSI

A – INFORMAÇÃO DE INTERESSE ENVIADA AO USUÁRIO

B – INFORMAÇÃO SEM INTERESSE ENVIADA AO USUÁRIO – PERFIL MUITO AMPLO, DEVE SER REFEITO

C – INFORMAÇÃO DE INTERESSE NÃO ENVIADA AO USUÁRIO – PERFIL MUITO RESTRITO, DEVE SER REFEITO

D – INFORMAÇÃO DE INTERESSE NÃO DISPONÍVEL NO D. B. – NECESSÁRIA PESQUISA COMPLEMENTAR

Fonte: SDI – Georg R. Mauerhoff

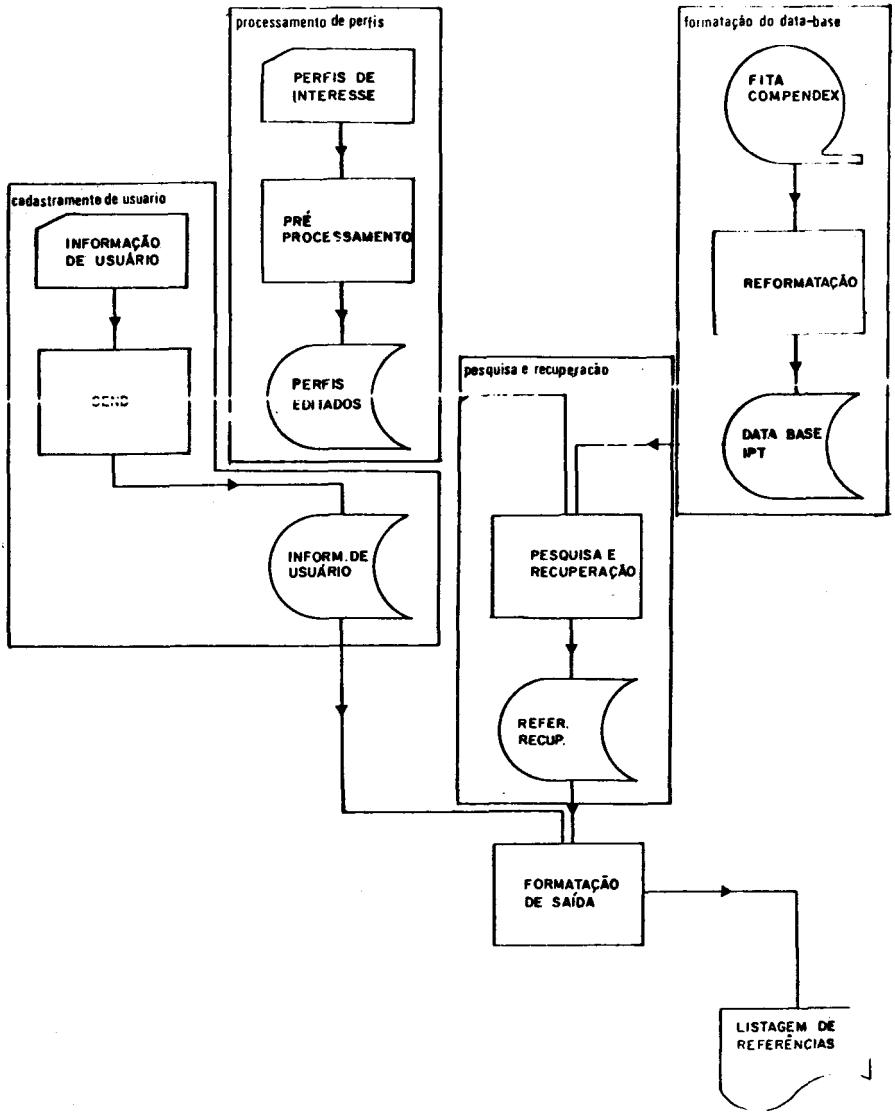

processamento de perfis
PERFIS DE INTERESSE
PRÉ PROCESSAMENTO
PERFIS EDITADOS
formatação do data-base
FITA COMPENDEX
REFORMATAÇÃO
DATA BASE IPT
cadastramento de usuario
INFORMAÇÃO DE USUÁRIO
CENO
INFORM. DE USUÁRIO
pesquisa e recuperação
PESQUISA E RECUPERAÇÃO
REFER. RECUP.
FORMATAÇÃO DE SAÍDA
LISTAGEM DE REFERÊNCIAS

Fig. 3

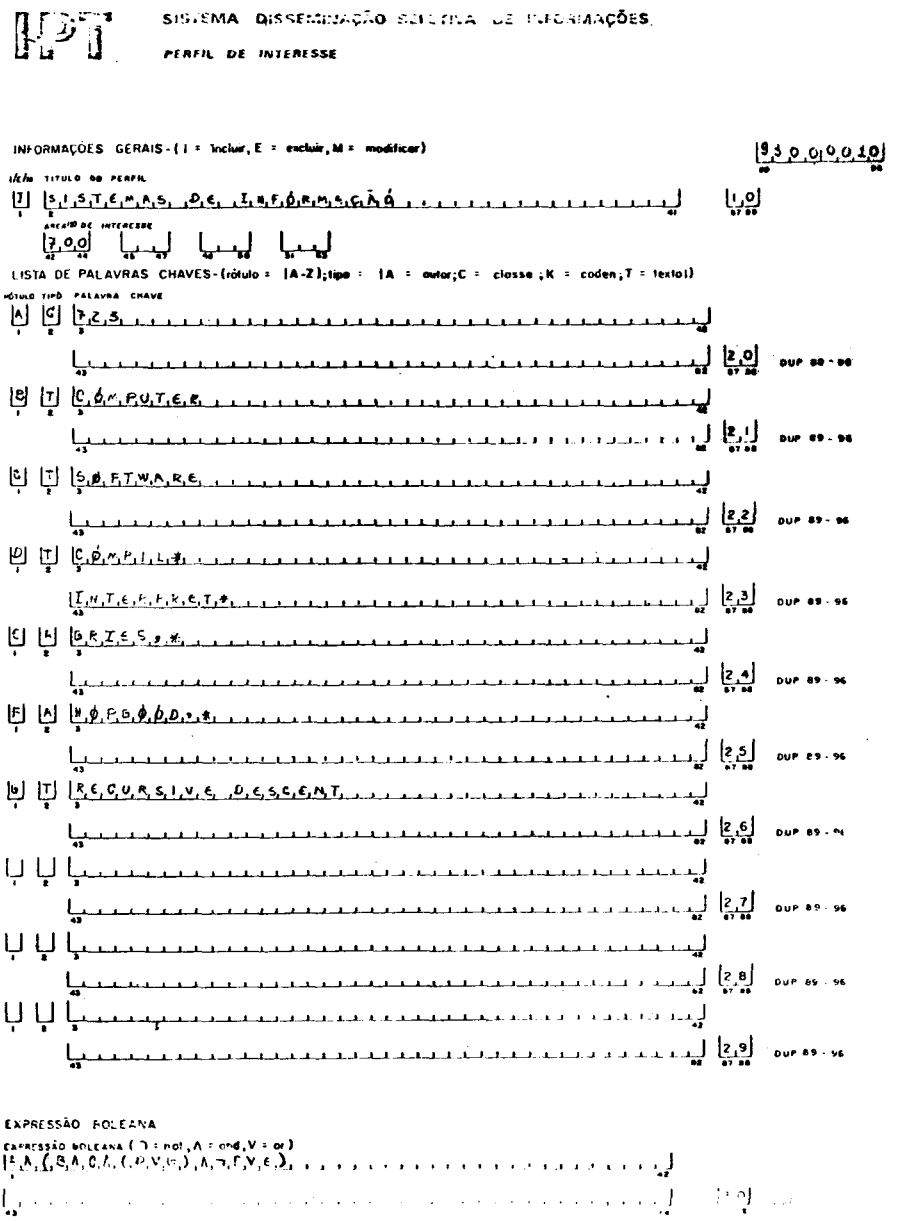

SISTEMA DISSEMINAÇÃO SELETIVA DE INFORMAÇÕES
PERFIL DE INTERESSE
INFORMAÇÕES GERAIS - (I = incluir, E = excluir, M = modificar)
LÍNGUA
TÍTULO DO PERFIL
ANO/MÊS DE INTERESSE
LISTA DE PALAVRAS CHAVES - (rótulo = |A-Z|; tipo = |A = autor; C = classe; K = coden; T = texto|)
RÓTULO
TIPO
PALAVRA CHAVE
DUP 89 - 96
EXPRESSÃO BOLEANA
EXPRESSÃO BOLEANA ( ) = not, A = and, V = or)

Fig. 4

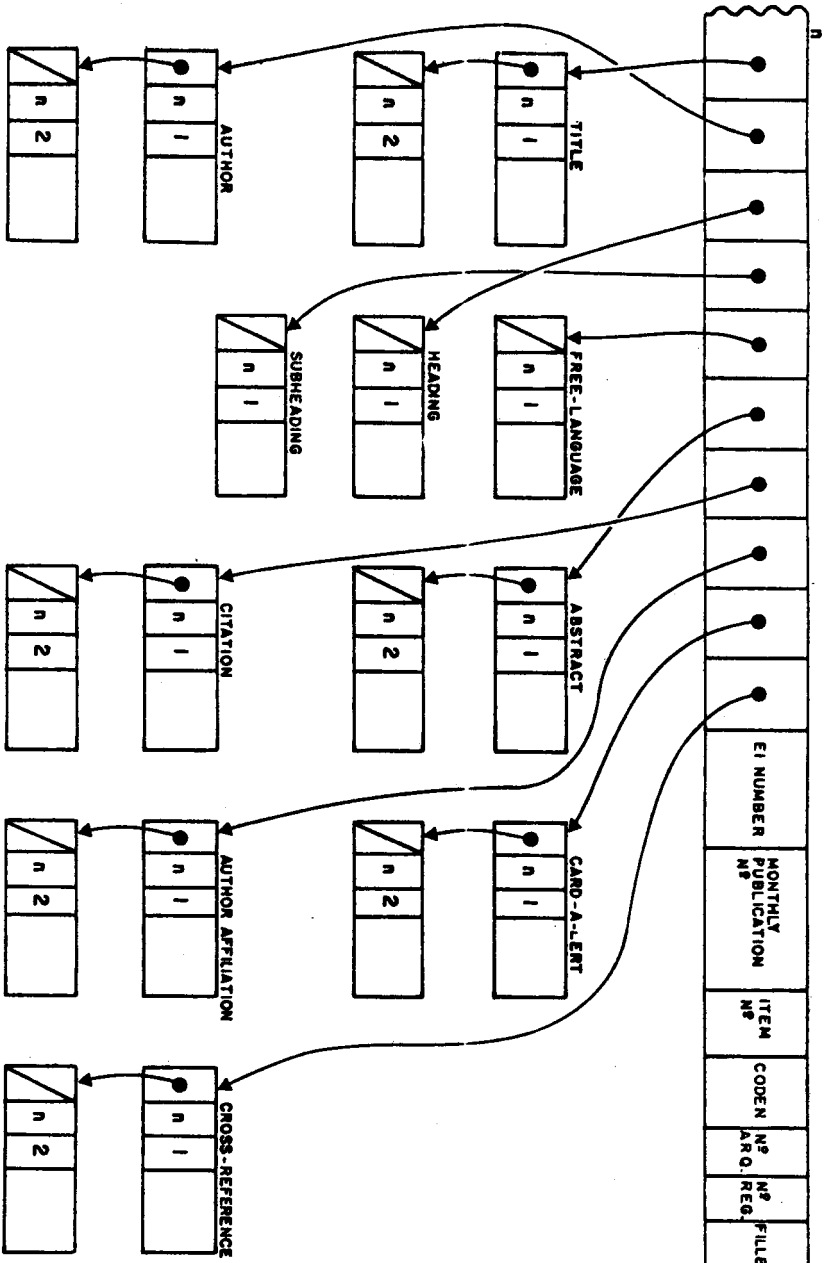
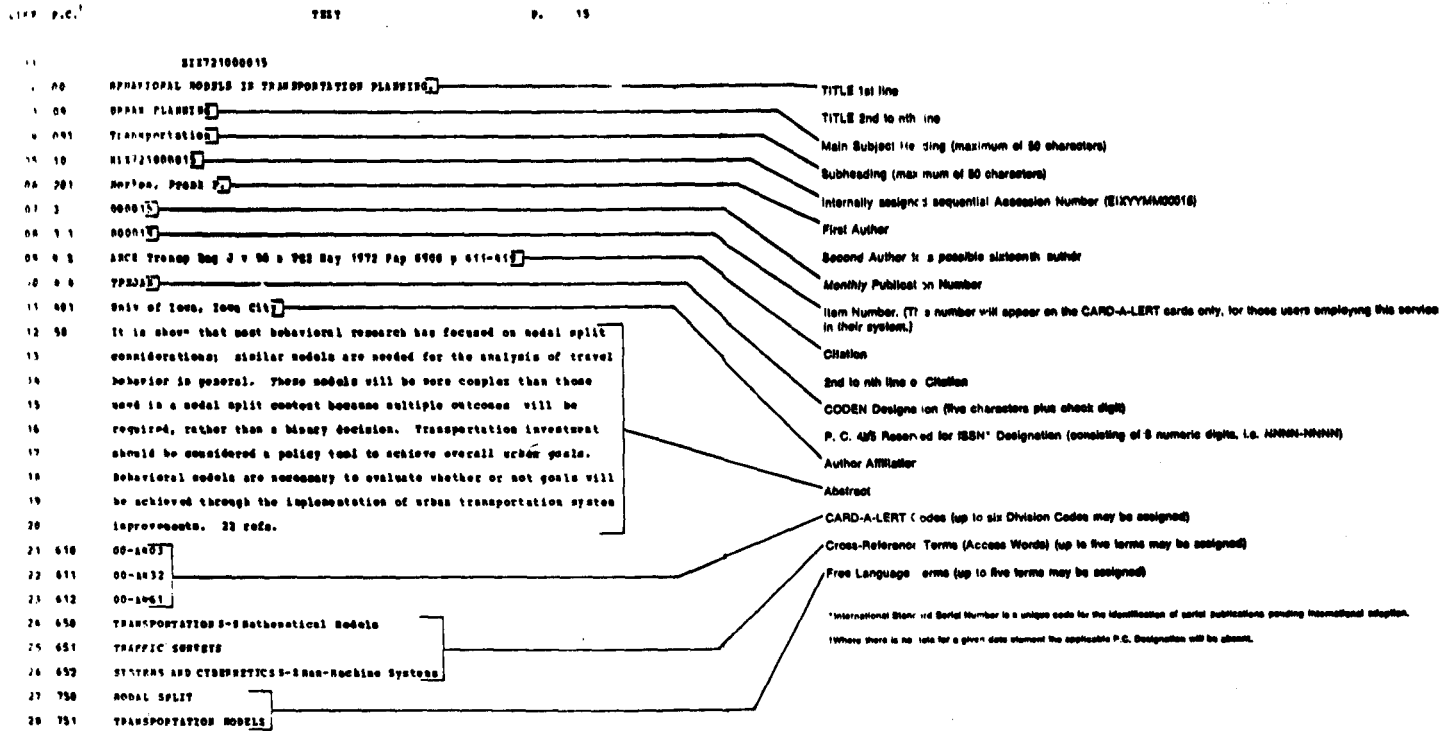

n
EI NUMBER
MONTHLY PUBLICATION Nº
ITEM Nº
CODEN
Nº ARQ.
Nº REG.
FILLER
TITLE
AUTHOR
FREE-LANGUAGE
HEADING
SUBHEADING
ABSTRACT
CARD-A-LERT
CITATION
AUTHOR AFFILIATION
CROSS-REFERENCE
n
1
2

Fig. 5



Fonte Engeneering Index, Inc.

Fig. 6

ABSTRACT

Information is of an atmost importance in the field of research & development as far as science & technology is concerned. A very few good offices in Brazil are afforded to technicians & scholars. The "Instituto de Pesquisas Tecnológicas do Estado de São Paulo S/A" developed a system of selective Dissemination of information to be ready at call of technicians & other Brazilian Institutions as well. The present paper forms an opinion about a very simple system of SDI & describes the solutions taken to introduce such a system in APT.

It's also discussed the rational explanation that guided the development of the software to the implanted system into a small computer of the institute.

REFERÊNCIAS

1. WILLIAMS, Martha E. Provision of Information to the Rescarch Staff. IIT Rescarch Institute. The American Institute of Chemical Engineers. 63rd Annual Meeting, 1970.

2. COMPENDEX. Technical description and specification manual. Engineering Index Inc., 1976.

3. MAUERHOFF Georg R. Selective dissemination of information. In: ADVANCES IN LIBRARIANSHIP. Canada, National Science Library, 1974 (4) : 25-62.

4. BUTTERLY, E. The evolution and underlying problem of the SDI system. Pretoria Atomic Energy Board, 1974.