UMA INTERFACE INTELIGENTE PARA ACESSO DE USUÁRIOS CASUAIS A SISTEMAS DE RECUPERAÇÃO DE INFORMAÇÕES BIBLIOGRÁFICAS

CARLOS HENRIQUE MARCONDES DE ALMEIDA

Comissão Nacional de Energia Nuclear
Centro de Informações Nucleares

Dissertação apresentada ao curso de Mestrado em Ciência da Informação, da Escola de Comunicação da Universidade Federal do Rio de Janeiro e do Instituto Brasileiro de Informação em Ciência e Tecnologia - IBICT - para obtenção do Grau de Mestre em Ciência da Informação.

ORIENTADORA: GILDA MARIA BRAGA, Phd - Instituto Brasileiro de Informação em Ciência e Tecnologia.

RIO DE JANEIRO

1992

RESUMO:

Propõe-se a concepção de um programa, rodando em microcomputador PC-compatível, que se constitua em uma Interface Inteligente para usuários casuais a Sistemas de Recuperação de informações Bibliográficas. Tal sistema funcionaria como um "front-end" a um SRI convencional de modo a aceitar a formulação da consulta do usuário em linguagem natural e incorporar todo o conhecimento não só para compreender a consulta como também para interagir com o sistema, substituindo o intermediário (especialista em informação, bibliotecário de referência) no processo de busca. O programa emprega de técnicas de Inteligência Artificial, tanto para o processamento das consultas em linguagem natural como para dirigir a interação usuário/SRI.

ABSTRACT:

An Inteligent Interface for casual users of bibliografic information retrieval systems is proposed. Such a system would be a "front-end" to a conventional IRS and would accept natural language query formulations and would have the ability to substitute the intermediary in a search session. Artificial Inteligence techniques are used for processing the natural language queries and also to simulate the expertise of the intermediary in interacting with the IRS.

SUMARIO

1. INTRODUÇAO

A Ciência da Informação, no decorrer de seu desenvolvimento, tem se voltado para a pesquisa dos fundamentos teóricos e técnicas de armazenamento e recuperação de informações que dessem conta da explosão informacional que é uma característica dos tempos modernos. O surgimento do computador em fins da década de 40 trouxe o suporte tecnológico capaz de tornar realidade propostas como as do Memmex de Vannevar Bush, de um aparelho capaz de recuperar informações e auxiliar o trabalho do cientista. As primeiras experiências do emprego de computadores no tratamento e recuperação de informações bibliográficas, no sentido que hoje damos aos Sistemas de Recuperação de Informação (SRI) foram realizadas por LUHN (1).

Em nossa sociedade cada vez mais a informação é um insumo essencial para as mais diferentes atividades; a industria de informações em linha é uma realidade: grandes bancos de informação como os sistemas Dialog, Orbit, BRS, Questel, Medline e outros armazenam milhões de registros da literatura científica, tecnológica e de negócios e tornam referências desta literatura disponíveis através do acesso remoto via teleprocessamento. Hoje estes sistemas estão integrados ao complexo editorial de publicções científicas e técnicas e junto com ele compõe um dos canais cada vez mais importantes de disseminação de ciência e tecnologia pelo mundo.

A finalidade de um sistema de recuperação de informações bibliográficas consiste basicamente em "indicar" a existência de literatura que pode ser relevante para as necessidades informacionais de um usuário, seja ele um cientista envolvido numa pesquisa acadêmica ou um técnico voltado para a solução de um problema tecnológico. Um SRI não responde a perguntas de seus usuários como um sistema de consulta a bancos de dados convencional, capaz de responder qual é o saldo de uma conta bancária; apenas indica literatura capaz de responde-las.

Um SRI automatizado funcionalmente consiste basicamente num processo de entrada que realiza o tratamento da literatura que se deseja armazenar no sistema, procurando representar seu conteúdo sob a forma, por exemplo de palavras-chave; este processo é uma antevisão de possíveis consultas que poderão ser feitas ao sistema. Existe também um processo de armazenamento dessas representações da literatura no sistema; e um processo de recuperação das informações armazenadas, que consiste em submeter ao SRI consultas, que são representadas da mesma forma que foram os documentos armazenados no sistema; quando existe uma similaridade entre a representação da consulta formulada e a de um ou mais documentos, estes são recuperados em resposta à mesma.

O computador é o instrumento que viabiliza um sistema de recuperação de informações. Em todo o desenvolvimento da tecnologia da informática, desde seu surgimento até os dias

atuais, uma tendência tem sido constante: a princípio, operadores ou usuários de computadores manipulavam diretamente os circuitos eletrônicos da máquina, alterando sua configuração, de modo a obter diferentes comportamentos e poder realizar diferentes tarefas. Esta forma de "programação" no entanto, era bastante rígida e pouco prática. O computador só avançou no sentido de tornar-se uma tecnologia de largo uso quando formas de programa-lo para a realização de diferentes tarefas tornaram-se mais simples; isto foi obtido através da divisão funcional do computador em "hardware" ou seja, a maquinaria e os circuitos elétricos, que só realizavam um número fixo e imutável de funções, e o "software", que permitia "programar" a seqüência em que as funções básicas do "hardware" eram executadas.

O computador tem assim um componente, a memória, em que um programa é armazenado, instruindo o processador, que é a parte ativa do computador capaz de executar seu elenco de instruções básicas, sobre que instruções básicas executar, em que seqüência, em quantas interações.

Um usuário desta forma utiliza o potencial de um computador intermediado por um "software" ou programa. A evolução desta tecnologia foi crescentemente no sentido de interpor cada vez mais camadas de "software" entre o usuário e a máquina. Historicamente o primeiro tipo de "software" foi a linguagem assembler, que tinha praticamente uma correspondência um a um com o elenco de instruções de máquina, somente facilitando sua memorização pois em vez de o usuário ter que lembrar-se que a instrução de máquina que fazia a soma de A com B era 4524 37 48, que significa somar o conteúdo do endereço de memória 37 com o conteúdo do endereço de memória 48 e armazenar o resultado no endereço de memória 48, ele usava a instrução de assembler ADD A B.

Após o surgimento do assembler, surgiram os Sistemas Operacionais, concebidos como uma camada de "software" que isolava totalmente o usuário e mesmo outros programas, do "hardware"; qualquer solicitação de alguma função realizada pelo "harware", fosse ela feita pelo usuário interagindo diretamente com o computador ou indiretamente, através de algum programa, era intermediada pelo SO.

Esta concepção foi levada as últimas conseqüências no atual estado da arte da tecnologia de software. Quanto mais níveis ou camadas de "software" são interpostas entre a máquina e o usuário, mais fácil e "amigável" se torna a utilização do computador. Toda a tecnologia de "software" tem avançado nesta direção, no sentido de tornar a utilização das facilidades de um computador acessíveis a usuários leigos, não especialistas em computação. A tendência é o usuário solicitar do computador um serviço, não se preocupando no "como" ele vai ser realizado; todo este conhecimento fica embutido no "software", distribuído em suas diferentes camadas especializadas.

Um SRI automatizado é implementado através de um "software". Uma consulta a um SRI tem o aspecto semelhante a seqüência expressões seguintes, tecladas num terminal de vídeo conectado ao sistema Dialog (2):

?B(EGIN) 40

(o usuário se conecta a Base de Dados ENVIROLINE, identificada no sistema pelo número 40 e com cobertura de assuntos na área de meio-ambiente);

?S(ELECT) insecticides **AND** toxicity **AND** wild life

(o usuário usa o comando "SELECT" para selecionar da base de dados referências sobre inseticidas, toxicidade e vida selvagem, simultaneamente);

?S(ELECT) S1 **AND** PY=1985:1990

(o usuário usa o comando "SELECT" para selecionar das referências da consulta anterior (S1) aquelas com ano de publicação entre 1985 e 1990);

?T(YPE)

(o usuário usa o comando "TYPE" para exibir na tela do seu terminal de vídeo algumas referências recuperadas).

No exemplo mostrado, convém destacar algumas características: a interação com o sistema é feita através de expressões ou "sentenças imperativas" que são submetidas ao sistema via terminal de vídeo; esta característica se deve ao alto custo de transmissão via teleprocessamento, que é minimizado submetendo as "sentenças imperativas" ao sistema linha a linha, em oposição à transmissão de uma tela completa; estas "sentenças imperativas" constam de comandos, inteligíveis pelo sistema, como **SELECT** (selecionar) ou **TYPE** (exibir), e de palavras-chave ou descritores, que descrevem o assunto de interesse do usuário, como **insecticides**, **toxicity**, **wild life**; o relacionamento entre os descritores também é de suma importância, sendo utilizados para isso operadores booleanos como E, OU e NAO, realizando respectivamente as operações booleanas de intersecção, união e diferença entre conjuntos de referências indexadas por cada um dos descritores.

Existem grandes dificuldades para um usuário na sua tentativa em obter informações diretamente de um SRI: pelo exemplo pode-se notar, a primeira vista, que esta interação não é trivial, requerendo conhecimento dos comandos do sistema, de sua sintaxe, e da terminologia a ser empregada para descrever o assunto de interesse do usuário; todos este sistemas tem geralmente algum tipo de controle terminológico na escolha dos termos que vão indexar um documento, o que obriga que este controle seja mantido

também para a recuperação dos mesmos. Além disso o usuário tem que escolher o SRI/base de dados adequados a suas necessidades de informação, conectar-se ao sistema, conhecer a linguagem de recuperação, elaborar uma expressão de busca sintaticamente correta nesta linguagem e submete-la ao SRI, alterar a expressão de busca em função de suas necessidades e dos resultados obtidos.

Estas características dos SRI fizeram com que se consagrasse a figura de um especialista em busca de informações como o intermediário entre o usuário final e o SRI e que conhecendo os detalhes não só da linguagem de interação com o sistema, mas também da estruturação de cada Base de Dados (política de coleta, de indexação, vocabulário utilizado, etc.), sirva de intermediário entre o sistema e o usuário final. Estes sistemas são acessados pelo usuário final através de postos de serviço em bibliotecas universitárias, institutos de pesquisa, etc, que sempre contam com intermediários treinados na interação com o sistema e que além disso dispõe do auxílio de instrumentos como manuais do sistema, tabelas de áreas, tesauros, etc.

Esta situação se contrasta com as grandes facilidades de telecomunicação disponíveis mesmo em países como Brasil, como por exemplo a Rede Nacional de Comutação de Pacotes, a RENPAC (3), que oferece facilidades de acesso semelhantes à telefonia, e que permitiriam a um usuário casual consultar um SRI até de sua própria casa, através de um microcomputador conectado ao SRI via teleprocessamento. A necessidade da assistência de um intermediário se torna, desta forma, um obstáculo a acessibilidade dos SRI por usuários casuais.

E justamente o fato de a interação de um usuário com um SRI necessitar de um intermediário, um especialista em busca de informação, que sugere a pesquisa na área de sistemas especialistas como possível solução para implementar um programa de computador que se interponha entre o usuário e o SRI e que seja dotado de "inteligência" capaz de simular as habilidades do intermediário na busca de informações em situações onde o usuário não dispõe do apoio de um especialista.

Um programa convencional de computador, ao incorporar um procedimento, chamado de algorítmo, para resolver algum problema, um cálculo científico ou um procedimento administrativo, era num certo sentido dotado de "inteligência". Mas os problemas resolvidos por programas convencionais, construídos sobre o paradigma algorítmico, eram problemas em que todos os passos para se chegar a solução eram conhecidos de antemão; o sistema não julgava, avaliava entre várias alternativas, planejava os passos da sua solução. Esta já estava previamente definida pelo algorítmo.

Ao contrário do processamento de dados convencional, o paradigma dos sistemas especialistas se baseia não no processamento algorítmico de dados, mas sim no processamento de conhecimentos representados sob alguma forma simbólica. Em (4) são

discutidas as diferenças de paradigma entre as técnicas de IA e o processamento de dados convencional: "A Ciência da Computação clássica preocupa-se em resolver problemas bem definidos, para os quais é possível descobrir os passos da solução antes de começar a resolver o problema. O conjunto destes passos recebe o nome de algoritmo."; enquanto as técnicas de IA são definidas como: "Ela ocupa-se dos problemas para os quais os passos da solução devem ser encontrados durante a resolução. Os problemas com os quais a IA trabalha estão de tal modo cheios de situações inesperadas que a máquina deve descobrir sozinha como resolvê-los. Em suma, a máquina deve improvisar, descobrir os passos que vai seguir, criar a solução".

Nos programas convencionais os conhecimentos acerca do problema que o programa se propõe a resolver, bem como a estrutura que controla a seqüência e a interação do processamento, estão embutidos nas linhas de código do mesmo, constituindo o algorítmo. Num sistema baseado no paradigma dos sistemas especialistas, esse conhecimento, representado em algum sistema simbólico, é transiente, mutável, assim como os dados num sistema convencional. Isto permite que esse conhecimento seja manipulado independentemente do programa, abrindo a possibilidade de que ele seja adquirido, manipulado, armazenado e transportado, sob a forma de Bases de Conhecimento. As Bases de Conhecimento são uma forma nova de armazenar conhecimentos, que se junta aos livros, artigos de periódicos e todo o tipo de meio de armazenamento de conhecimentos convencional. Além de simplesmente armazenar conhecimentos, podem também, associadas a um sistema especialista, tornar este conhecimento manipulável e operacional na resolução de problemas, inclusive aquele conhecimento não formal que consiste em heurísticas, habilidades e conhecimento prático, intuitivo, originário da experiência diária de especialistas em diferentes áreas de atividade.

Os Sistemas Especialistas (SE) têm evoluído nos últimos anos como uma das áreas mais promissoras de pesquisa no chamado campo da Inteligência Artificial (IA). Uma Interface Inteligente, construída como um Sistema Especialista, utilizaria os conhecimentos, heurísticas e habilidades do intermediário humano na interação com SRIs, coletados e armazenados em Bases de Conhecimentos; estes conhecimentos seriam tanto de natureza lingüística, de modo a poder interpretar as consultas formuladas pelo usuário em linguagem natural, como também relacionados às habilidades de interagir com o SRI, de formular consultas usando álgebra booleana, de conhecer a estruturação das bases de dados (linguagem de indexação, estrutura de armazenamento, etc) e o uso de instrumentos de recuperação do tipo tesauros.

A possibilidade de um usuário casual, interessado em consultar um banco de dados de informações remoto, isto é, de acesso público através de teleprocessamento, poder submeter suas consultas a um Sistema de Recuperação de informações (SRI) em linguagem natural, por exemplo, teclando no seu micro a seguinte consulta: "Desejo recuperar artigos que falem dos efeitos dos

inseticidas na extinção de formas de vida animal selvagens", seria uma alternativa que certamente melhoraria a disponibilidade desses sistemas, além de, com certeza, fornecer uma interação "padrão", em uma linguagem conhecido por todos (a LN), com qualquer SRI. O usuário casual ficaria assim dispensado de aprender uma linguagem formal de consulta para poder fazer uso das informações armazenadas no banco de dados gerenciado pelo SRI; pode-se compreender como seria útil que usuários casuais, cientistas, técnicos, pesquisadores, gerentes, pudessem acessar o crescente número de bancos de dados bibliográficos, acessíveis via RENPAC ou disponíveis através de dispositivos leitores de "Compact Disk-Read only Memory" (CD-ROM) a partir de um microcomputador instalado numa biblioteca universitária.

As Bases de Dados brasileiras acessíveis via RENPAC são hoje uma realidade. O anuário "Bases de Dados Nacionais", publicado pela revista INFO e pela EMBRATEL (5) relaciona 285 Bases de Dados, das quais cerca de 57 são acessíveis via RENPAC. Devido ao seu baixo custo, facilidade de operação e grande capacidade de armazenamento, começam a se disseminar pelo país bases de dados residentes em CD-ROM. Esta disponibilidade de informações, no entanto, esbarra com outro tipo de barreira ao acesso às mesmas, que consiste no fato de que essas bases de dados são suportadas por diferentes sistemas, com diferentes tipos de interfaces (menus, comandos), o que obrigaria a um eventual usuário a conhecê-las todas para que pudesse acessar informações armazenadas nos diferentes bancos de dados. No caso das bases de dados bibliográficas, existe um projeto conjunto CIN/IBICT para implementar uma Linguagem de Acesso Comum a Bancos de Dados Bibliográficos, denominada LINCE (6), que se propõe a ser um padrão nacional, inspirada na CCL (Common Comand Language), um padrão proposto pela International Standard Organization (ISO). No entanto, mesmo a LINCE, apesar de significar um padrão, é ainda uma linguagem de acesso convencional a SRIs, calcada na álgebra booleana. São conhecidas na literatura as dificuldades de um usuário casual de interagir com um SRI utilizando como linguagem a álgebra booleana (7).

E nesse sentido que seria de grande valia a um usuário casual, que não conta com a assistência de um especialista para intermediar sua busca, a facilidade de poder submeter suas consultas a um SRI em linguagem natural e esperar que sua IIn se encarregue de funções especializadas como: conectar-se ao computador/SRI mais provável de conter a informação desejada, "compreender" sua consulta em linguagem natural, traduzi-la para uma expressão formal na linguagem de comando do sistema (baseada na álgebra booleana), fornecer auxílio ativo sobre o sistema e suas facilidades, bem como a forma de interagir com o mesmo ("help" online), normalizar a terminologia utilizada (palavras permitidas, palavras proibidas), reconhecer automaticamente variações de grafia de termos ou expressões (microcomputador X micro-computador X micro computador; Fiocruz X Fundação Oswaldo Cruz), ordenar as referências recuperadas em ordem de relevância ("ranking"). Uma interface "amigável" deve suportar também

diferentes graus de precisão na especificação das informações desejadas, como proposto em (8): acesso a um documento específico dado por exemplo seu autor ou título, acesso a grupos de documentos que atendam a um determinado critério (especificado por uma expressão booleana) ou "browsing"/navegação pelos documentos de uma base de dados como num sistema de hipertexto. Estas características apontadas estão de acordo com as últimas tendências em arquitetura de sistemas de processamento distribuído, a arquitetura distribuída "cliente-servidor", que preconizar a um sistema possuir as facilidades de uma interface "user-friendly", rodando no microcomputador, que deixa de ser um mero terminal enviando consultas para o SRI remoto; esta interface cooperaria com o SRI remoto, que ficaria livre das funções de interfaceamento amigável (menus explicativos, "helps", tutoriais, etc), trabalhando somente como um servidor de banco de dados.

Mais do que isso, montar um SRI requer pesados investimentos e um tempo de maturação relativamente longo para alimentar a base de dados com um número significativo de referências. Todos estes sistemas, não somente os existentes no Brasil mas principalmente os internacionais, estão montados sobre uma tecnologia de "software" em muitos aspectos obsoleta, baseadas em arquivos invertidos e usando lógica booleana para realizar as consultas. O uso de técnicas de Inteligência Artificial num programa como o proposto, totalmente construído à parte do SRI, permite incorporar uma evolução tecnológica significativa a estes sistemas, melhorando a acessibilidade e o desempenho dos mesmos, e simultaneamente preservar os investimentos já feitos.

A seguir são apresentados alguns critérios para definir a aplicabilidade do emprego de sistemas especialistas para uma determinada tarefa, tomados de FENLY (9) em um estudo da Library of Congress sobre aplicabilidade dos SEs no ambiente de bibliotecas:

a. A tarefa requer principalmente processamento simbólico, isto é, processamento de conhecimento representado sob algum sistema de símbolos ao invés de processamento algorítmico.

b. A tarefa é tipicamente heurística, dependendo para o seu sucesso do conhecimento de um expert; é não repetitiva, não algorítmica.

c. O domínio em questão é relativamente estreito, especializado.

d. Um certo grau de tolerância a erros, a acertos relativos, a aproximações, pode ser tolerado, ao invés da exigência de respostas tipo absolutamente certo ou totalmente errado.

e. Existem especialistas reconhecidos trabalhando na área.

Estes critérios parecem se aplicar ao problema de interação de um usuário casual com um sistema de recuperação de informações bibliográficas. A intermediação entre um usuário e um SRI requer o

apoio de um especialista, cujos conhecimentos só em parte podem ser objeto de um aprendizado formal, se constituindo em grande parte de conhecimento heurístico adquirido na prática de interação com SRIs, de difícil formalização. A área de apoio à recuperação de informações tem sido inclusive uma das áreas da Ciência da Informação onde este tipo de sistema tem mais sido empregado, conforme atesta a bibliografia relacionada em MARCONDES (10). Outros trabalhos como FIDEL (11) e SARACEVIC (12) tem se centrado também em tentativas de formalizar o conhecimento heurístico do especialista de busca, o que se constitui em outro indicador forte da aplicabilidade das técnicas de IA à situação.

O presente trabalho se divide nas seguintes partes: após a parte 1, introdutória, em que o problema é colocado e seus limites estabelecidos, a parte 2 discute os objetivos e hipóteses de trabalho. A parte 3 faz um apanhado da situação brasileira e relaciona trabalhos brasileiros correlatos, além de discutir problemas técnicos e teóricos envolvidos com a construção da IIn e as soluções propostas na literatura. A parte 4 discute o modelo proposto, dividido em modelo do processo de busca/interação com um SRI, módulo construtor do modelo do usuário/problema, módulo lingüístico capaz de processar consultas ao SRI em linguagem natural num domínio específico, módulo especialista em busca/interação com SRIs e problemas construtivos da IIn; na análise do módulo especialista lingüístico é discutida a importância de um formalismo lingüístico para a compreensão correta das intenções do usuário ao formular a consulta em linguagem natural; é feita uma análise e proposta uma classificação de tipos característicos de consultas encontradas no material analisado e derivada uma estrutura lingüística para as consultas que vai guiar a elaboração do formalismo lingüístico capaz de analizá-las; são discutidos diferentes formalismos lingüísticos e um deles é proposto; são propostas também regras de interpretação semântica para mapear a consulta analisada em expressões de lógica booleana, que é a linguagem-alvo do sistema; na análise do módulo especialista em busca/interação com o SRI, são propostas regras para esta interação, métricas para avaliação dos resultados obtidos, regras para a reformulação e re-submissão de consultas.

A parte 5 apresenta uma avaliação e discussão do modelo proposto e a parte 6 é uma conclusão que esboça ainda possíveis expansões e incrementos na IIn.

Notas e Referências:

1. LUHN, H. P. A Statistical Approach to Mechanised Encoding and Search of Library Information. **IBM Journal of Reseacrh and Development**, 1, p.309-17, 1957.

2. DIALOG INFORMATION SERVICES. **Pocket Guide to DIALOG**. Dialog Information Services, Inc,1987.

3. EMBRATEL. **Introdução à Comutação de Pacotes**. EMBRATEL, 53p.,

4. SIQUEIRA, I. S. P, PEREIRA, A. E. C. Perspectivas de Aplicação da Inteligência Artificial à Biblioteconomia e à Ciência da Informação. **R. Bras. Biblioteconomia e Doc**, 22(1/2), p.39-80, 1989.

5. INFO. **Bases de Dados Nacionais**. Rio de Janeiro, JB/EMBRATEL, 1989. Número especial.

6. CNPq/IBICT/CNEN-CIN. **LINCE**: Linguagem Comum de Recuperação de InformaçÕes em linha (versão preliminar). Rio de Janeiro, CNPq, s.d.

7. HEINE, M. H. A **Logic Assistant for the Database Searcher**. Information Processing & Management, v. 24, n. 3, p.323-329, 1988.

8. FLUHR, C. Information Retrieval. In: **CONSULTIVE MEETING of INIS LIAISION OFFICERS**, 17, Viena, May 1989.

9. FENLY, C. & HARRIS, H. **Expert Systems - Concepts and Applications**. Washington, D. C., Cataloging Distribution Service, Library of Congress, 37p., 1988. (Advances in Library Techology, 1).

10. MARCONDES, C. H. **Interfaces em Linguagem Natural a Sistemas de Recuperação de InformaçÕes: uma Revisão da Literatura**. Trabalho apresentado na disciplina "Processamento da Informação VIIII", curso de mestrado em Ciência da Informação, convênio IBICT/ECO-UFRJ, 1988, 20p.

11. FIDEL, R. **Online searching styles: a case-study-based Model of Searching Behavior**. J. of the American Society for Information Science, v. 35, n. 4, p.211-221, 1984.

12. SARACEVIC, T., KANTOR, P. **A Study of Information Seeking and Retrieval**. JASIS, v. 39, n. 3, part I, p.161-176, part II, p.177-196, part III, p. 197-216, may 1988.

2 OBJETIVOS E PREMISSAS DE TRABALHO

O objetivo principal do trabalho é conceber um programa de computador que se constitua numa interface inteligente (IIn) a sistemas de recuperação de informações bibliográficas, de modo a simular o apoio prestado pelo especialista na intermediação de uma busca de informações. Ao incorporar e manipular vários tipos de conhecimentos especializados, como o conhecimento lingüístico necessário à compreensão de consultas formuladas em linguagem natural, e o conhecimento do especialista intermediário em consultas à SRIs, o sistema proposto se incluiria no paradigma da Inteligência Artificial, dos sistemas baseados em conhecimento, isto é, se constituiria num sistema especialista.

Como objetivos secundários, foi proposta uma classificação das expressões de consultas formuladas em linguagem natural, que serviu de base para a formulação do formalismo lingüístico (gramática) para interpretar estas consultas. Procurou-se também, sempre que possível, generalizar os passos seguidos, no sentido de propor uma metodologia para o desenvolvimento de interfaces em linguagem natural a SRIs bibliográficos. Procurou-se também: sistematizar o conhecimento do especialista em informação que serve de intermediário entre o usuário e o SRI e formaliza-lo sob a forma de regras, que vão se constituir numa das bases de conhecimento empregadas pelo programa; foram investigados também indicadores e métricas (também formalizados sob a forma de regras) para avaliação dos resultados de uma consulta de modo a decidir-se pela sua reformulação ou não e de que forma faze-lo;

A premissa deste trabalho é que o usuário "casual" de Sistemas de Recuperação de informações Bibliográficas interage com o SRI, tendo como cenário provavelmente uma biblioteca de uma universidade ou instituto de pesquisas, ou mesmo uma empresa, que disponha da facilidade de possuir um microcomputador PC-compatível, conectado à RENPAC ou a um leitor de CD-ROM, com capacidade de hospedar a IIn (software) e as bases de conhecimento mínimas necessárias para o seu funcionamento (tesauros, bases de conhecimentos lingüísticos, bases de conhecimento especializado em buscas retrospectivas).

Entende-se por usuário "casual" aquele usuário que faz buscas a SRIs eventualmente, como decorrência de seu próprio trabalho de pesquisa, etc, não sendo porém um especialista em estratégias de busca, manipulação de instrumentos como um tesauro, nem sendo familiarizado com linguagens de consulta, elaboração de estratégias de busca/perfis usando álgebra booleana; esse usuário tem no máximo familiaridade com a operação de pacotes típicos em microcomputadores, como editores de texto, planilhas eletrônicas, gerenciadores de bases de dados, etc; além disso, esse usuário tem uma razoável clareza da terminologia da sua área e principalmente do problema para o qual tenciona que a busca de informações contribua para a solução. Em suma, esse usuário não é um

bibliotecário de referência ou especialista em informação, cujo trabalho é exatamente servir de intermediário à usuários de SRIs.

Num cenário como esse é razoável que um usuário acostumado a operar esses pacotes típicos de microcomputadores espere um mínimo de facilidades de uma IIn, que ela incorpore características amigáveis encontradas em pacotes típicos para microcomputadores e não encontradas em SRIs, que se caracterizam por interfaces bastante pobres, voltadas a linguagens de comando, com características de serem submetidas linha a linha, objetivando reduzir os custos de transmissão. As interfaces dos SRIs convencionais são tipicamente voltadas para um usuário especialista em buscas retrospectivas, este sim um intermediário entre o sistema e o usuário final que necessite das informações armazenadas no banco de dados manipulado pelo SRI.

Algumas dessas características que tornam uma interface "amigável" a um usuário casual, não-especialista em interação com SRIs, são simples de implementar e estão presentes, como já mencionado, na maioria dos pacotes para microcomputadores. Procura-se neste trabalho nos concentrar nas características relativas a interpretação de consultas submetidas em linguagem natural e na simulação do especialista na interação com o SRI como intermediário entre este e o usuário casual.

Os limites do trabalho aqui proposto são os seguintes: a IIn deve se constituir em um programa **independente** do SRI, rodando em micro PC-compatível e acessando um SRI (ou vários SRIs) convencional, baseado na estrutura de arquivos invertidos, cuja linguagem de consulta é calcada na LINCE (3). O micro hospeda todas as bases de conhecimentos e arquivos (tesauros, tabelas de classificação) necessários ao funcionamento do sistema, de forma que não implique em nenhuma alteração no SRI. Supõe também que a principal base de conhecimento, necessária para a geração das expressões booleanas de busca que serão submetidas ao SRI, um tesauro automatizado, existe enquanto um dos módulos da IIn; e principalmente, que esteja em língua portuguesa, o que não é verdade nos exemplos reais coletados, já que as bases utilizadas tem sua indexação e seus instrumentos de recuperação (tesauros) em inglês; após o processamento lingüístico, tomou-se a liberdade de "traduzir" os termos de busca para o inglês.

Notas e Referências:

1. CIN/CNEN significa Centro de Informações Nucleares da Comissão Nacional de Energia Nuclear; IBICT significa Instituto Brasileiro de Informação em Ciência e Tecnologia; BIREME significa Biblioteca Regional de Medicina; EMBRAPA significa Empresa Brasileira de Pesquisas Agropecuarias; FGV significa Fundação Getúlio Vargas.

3 SITUAÇAO ATUAL. O PROBLEMA

Neste capítulo são discutidos as experiências brasileiras com sistemas semelhantes ao proposto e possíveis requisitos para a IIn. A experiência mais conhecida no Brasil de um sistema operativo de consulta a bases de dados em linguagem natural é o sistema FARAO, desenvolvido na EMBRATEL por Clarisse S. de Souza, cujo formalismo lingüístico (gramáticas de determinação), embora empregado no contexto de consulta a uma bases de dados administrativa, possui características que inspiraram algumas soluções no presente trabalho.

Outra experiência brasileira conhecida é o trabalho de Idméa S. P. Siqueira, que trabalhou no desenvolvimento de uma interface em linguagem natural a uma base de conhecimentos, o sistema LINAT, que foi implementado no CTI (Centro Tecnológico para Informática), em Campinas/SP.

E um tanto difícil traçar uma linha de evolução dos SRIs no Brasil; no entanto, um esboço pode ser traçado. O SRI mais antigo em funcionamento no país foi o SUPRIR, operado pelo Centro de Informações Nucleares da Comissão Nacional de Energia Nuclear; o processo que culminou com o desenvolvimento do SUPRIR inciou-se pelos contatos entre a CNEN e a Agência Internacional de Energia Atômica e seu sistema INIS (International Nuclear Information Systems), a qual o CIN se tornou representante no Brasil. Estes contatos com um sistema de informação a nível mundial como o INIS permitiram a equipe do CIN absorver a tecnologia para desenvolver o SUPRIR. Convém ressaltar que, nesta mesma época, o curso de Engenharia de Sistemas do Instituto Militar de Engenharia (IME) possuia uma linha de pesquisa sobre Sistemas de Informação Bibliográfica e mantinha intenso intercâmbio tecnológico com o CIN.

Outra experiência semelhante ao do CIN é a da EMBRAPA, que opera um SRI com a base de dados AGRIS, da FAO, que mantém um sistema similar ao INIS, denominado AGRIS.

A década de 70 viu surgir os primeiros grandes sistemas de recuperação de informações, embora ainda não bibliográficas, quando várias instituições governamentais, detentoras de grandes volumes de informações, iniciam uma política de torna-las acessíveis à toda a sociedade; foi o caso de sistemas de informação como o SIDRA (IBGE), o CIDUL (SERPHAU), etc; a primeira tentativa de um Sistema de Recuperação de informações acessível por teleprocessamento, oferecendo no caso dados sócio-econômicos, foi o sistema ARUANDA, operado pelo SERPRO.

Mais recentemente, com o advento da RENPAC, SRIs em linha, com diversas bases bibliográficas, se tornaram disponíveis em instituições como o IBICT e a BIREME. Pode-se prever um grande incremento da disponibilidade de informações bibliográficas acessíveis em linha com a inauguração próxima (para 1992) do

Sistema Público de Acesso (SPA), uma rede interligando os principais centros de acesso a bases de dados bibliográficas do País, oferecendo ao usuário remoto um acesso transparente a qualquer base em qualquer centro do sistema, com normas comuns de contabilização de custos, oferecendo serviços como correio eletrônico e acesso ao documento final, além de propiciar o acesso através de uma única linguagem comum de acesso, a LINCE, que deverá se tornar um padrão nacional.

No caso do Brasil, um sistema como o proposto se constituiria numa interface a SRIs acessados via RENPAC, como é o caso de bancos de dados operados por instituições como o CIN/CNEN, o IBICT, a BIREME, a EMBRAPA, a FGV, etc. ou um SRI que acesse bases locais, hospedadas num microcomputador que disponha de uma unidade leitora de CD-ROM.

Sistemas de Recuperação de informações Bibliográficas "on-line" são já uma realidade nos países do 1o. mundo há algum tempo e a venda de informações se constitui hoje numa atividade econômica poderosa e rentável.

Estes sistemas tem um componente de entrada, em que são construídas representações dos documentos a serem armazenados no sistema, geralmente constituídas por atributos dos documentos como título, autores, dados do periódico onde o documento foi publicado, dados do congresso onde o documento foi apresentado, data de publicação, resumo e principalmente, descritores que procuram representar o conteúdo do documento.

O outro componente do sistema é o componente de saída, ou de consulta à base de dados. Uma consulta ao sistema é também representada por um conjunto de descritores, tal qual os documentos na base; o que o sistema faz é recuperar descrições de documentos que tenham representações, em termos de conjuntos de descritores, semelhantes a consulta formulada.

Para obter este resultado, em termos de tecnologia de software, estes sistemas se baseiam em uma estrutura de índices e listas invertidas contendo, cada entrada da lista, o endereço em disco do registro que contém a representação de um documento. O sistema tem ainda a capacidade de implementar as operações booleanas de união (E), intersecção (OU) e diferença (E-NÃO) sobre listas de ponteiros de documentos.

A tecnologia em que os SRIs estão baseados, associada as suas características "on-line", ou seja, a necessidade de reduzir volumes de dados transmitidos e portanto custos, praticamente consagrou para estes sistemas um tipo de interface baseada em comandos submetidos linha a linha (ao invés de transmitir uma tela inteira) constituindo uma linguagem de consulta baseada na álgebra booleana; uma interação típica com um SRI já foi vista no capítulo 1.

Esta tecnologia, principalmente o formalismo da álgebra booleana como linguagem de consulta, tem tido sua eficácia bastante questionada na literatura, por não prover uma adequada e precisa descrição das necessidades de informação de um usuário: em uma consulta em que vários descritores são unidos pelo conectivo E, não existe forma de dar peso maior para um ou vários dos descritores, para refletir sua importância para o usuário; também uma expressão de consulta com vários descritores unidos pelo conectivo E, não recupera um documento que seja indexado por todos os descritores menos por um deles e que poderia, eventualmente, ser relevante para o usuário; esse formalismo também é incapaz, por si só, de ordenar documentos recuperados em ordem de sua relevância em relação a consulta formulada. Trabalhos como os de FOX, MARON, TAHANI, RADECKI, LOSEE, BLAIR, GORDON e muitos outros contém críticas ao formalismo de álgebra booleana e aos SRIs convencionais em geral e trazem várias propostas para incrementar sua eficiência e eficácia.

No entanto, os pesados investimentos já efetuados no tratamento de documentos pelos SRIs convencionais tornam-se um obstáculo aos avanços tecnológicos nesta área, por implicar em um re-tratamento de milhões e milhões de documentos, o que se torna praticamente inviável. Uma alternativa que vem sendo investigada é propor melhoramentos somente na interface com o usuário destes sistemas, sem interferir no SRI convencional. É nesta linha que se situa o presente trabalho.

3.1 ETAPAS DO PROCESSO DE BUSCA/INTERAÇAO SRI/USUARIO

No processo intermediado por um especialista humano, a interação se inicia por uma etapa descrita na literatura como "entrevista pré-busca", onde o intermediário procura traçar um quadro da situação-problema com a qual se depara o usuário, seu objetivo com a busca de informações, o tipo e o grau de generalidade desejado das mesmas, etc. Segue-se a formulação da consulta propriamente dita, onde o usuário EXPOE suas necessidades de informação, enfoques específicos sobre o assunto desejado, restrições sobre as características e o assunto das informações desejadas, etc.

A consulta a um moderno SRI bibliográfico com informações em linha se faz geralmente através de um terminal de vídeo, diante do qual ficam o usuário e o especialista. De posse dos dados da entrevista, o intermediário formula uma primeira estratégia de busca, consulta instrumentos como um tesauro ou uma tabela de classificação e finalmente materializa esta estratégia sob forma de uma expressão booleana, que é então submetida ao sistema. Um vez recuperado um conjunto de documentos, o intermediário submete este primeiro conjunto à avaliação do usuário. Em função da avaliação feita pelo usuário e do número de documentos recuperados, o intermediário reformula a expressão booleana original, utilizando-se de diversas técnicas que, dependendo da intenção, ora favorecem a revocação, ora favorecem a precisão,

como truncagem, navegação na hierarquia do tesauro para termos mais genéricos, mais específicos, ou termos relacionados, inclusão ou exclusão de termos na expressão booleana de busca, etc.

3.2 PAPEL DO ESPECIALISTA EM INFORMAÇÃO COMO INTERMEDIARIO NO PROCESSO DE BUSCA

Um SRI convencional não pode prescindir da figura do especialista em informação como intermediário do processo de busca entre o usuário e o SRI. Essa intermediação tornou-se essencial, à medida que crescia a quantidade de informações disponíveis em bancos de dados bibliográficos, juntamente com o conhecimento técnico demandado no processo de busca/interação com SRIs, do tipo: escopo de cada base de dados, terminologia, política de coleta/indexação de cada base, uso de instrumentos como tesauros, dicionários e tabelas de classificação, sintaxe da linguagem de consulta e sua maior ou menor potencialidade para recuperar informações.

Cabe ao intermediário basicamente, durante o processo de busca/interação com um SRI: obter do usuário descrição de suas situação-problema/necessidades de informação, elaborar/negociar uma estratégia para a obtenção das informações que presumivelmente possam ser relevantes para a situação/problema, traduzir esta estratégia numa expressão de consulta formal sintaticamente correta utilizando uma linguagem de consulta baseada na álgebra booleana e submetê-la ao SRI, avaliar os resultados com o usuário e possivelmente reformular a estratégia de busca/consulta para re-submete-la ao SRI.

E amplamente reconhecido na literatura o papel da etapa do processo de busca conhecida como "entrevista pré-busca" (pre-search interview). As técnicas da entrevista pré-busca, bem como sua importância para o sucesso das etapas posteriores foram estudados em FIDEL (10) e em SARACEVIC (11). Estes autores destacam a importância, para uma correta formulação da busca, de que o intermediário consiga obter do usuário informações sobre o motivo ou objetivo da busca e para que as informações obtidas vão ser usadas.

Nesta linha situa-se o trabalho de BELKIN que afirma que o usuário, ao buscar um serviço de informações, estaria num estado mental denominado pelos autores de ASK, de "anomalous state of knowledge" - estado anômalo de conhecimento, que se caracterizaria pela carência de conhecimentos ou falta de um quadro conceitual coerente a respeito de uma situação-problema, que ele pretende superar com as informações obtidas. Seria então incorreto solicitar do usuário que formulasse suas necessidades de informação uma vez que seu "estado anômalo de conhecimento" a respeito de uma situação-problema ou tópico impediria que o mesmo soubesse que informações necessitaria para superar o ASK. Os autores propõe reconstruir este estado anômalo de conhecimento, determinar suas características estruturais, basicamente entrevistando o usuário, não sobre suas necessidades de informação

que, supõe-se, o usuário desconheça, mas sobre a **situação-problema** com a qual o mesmo se defronta. A técnica de entrevista empregada resulta em uma rede de conceitos interrelacionados por técnicas estatísticas de co-ocorrência, que seria a representação do ASK; a representação, um verdadeiro mapa conceitual da situação problema, serviria de base para a formulação da busca.

A hipótese de BELKIN - ASK, coloca uma nova perspectiva na interação usuário/SRI, passando a considerar também seu aspecto cognitivo, embora ela não possa ser generalizada para toda a situação de busca por informações, que em muitos casos que podem se dar também segundo as hipóteses propostas por FLUHR, isto é: busca a um documento preciso, formulação exata das necessidades de informações (principalmente na área tecnológica, quando um usuário é um técnico altamente especializado que se defronta com um problema específico e sabe exatamente de que informações necessita) e "browse" pelos documentos de uma base. Inegavelmente a situação-problema se constitui em um insumo da maior relevância para elaboração da estratégia de busca e para o sucesso da interação usuário-sistema de informação; no entanto, a técnica usada por BELKIN para determinar as características estruturais dos ASKs nos parece inadequada, um tanto quanto tosca, como também parece desconhecer os avanços já obtidos na área de inteligência artificial em modelagem de usuários e de situações.

SCHANK propõe um formalismo para modelar a compreensão da linguagem, baseado na maneira como os seres humanos realizam o processo de compreensão. Segundo este formalismo, compreender seria mais que identificar o significado, mesmo que de todas as palavras de uma expressão, mas situar o que foi dito dentro de uma "visão do mundo", que torne coerente a expressão falada e que permitiria a quem ouviu realizar uma série de inferências sobre a situação exposta, tirando conclusões que em nenhum momento foram verbalizadas; segundo SCHANK "We view the process of understanding as the fitting in of new information into a previously organized view of the world"; ou então: "The basis of understanding is the assignment of new inputs to previously stored episodes in memory that wil make sense of then". Acreditamos só ser possível modelar uma situação-problema de um usuário na sua busca de informações se o intermediário tiver uma razoável visão da problemática do usuário, que lhe permita completar (inferir) as lacunas da descrição feita pelo mesmo de modo a obter um quadro conceitual coerente que permita inferir que informações seriam importantes para soluciona-lo.

Por exemplo, se um usuário tem um OBJETIVO/PROBLEMA que consiste em se livrar dos rejeitos radioativos da usina de Angra I, deduz-se que as suas NECESSIDADES DE INFORMAÇAO vão consistir de referências sobre rejeitos radioativos, materiais para acondiciona-los, como transporta-los, sua disposição final, normas de segurança e legislação específica, nacional e internacional, etc. Estas inferências só são possíveis com base em um conhecimento prévio do problema proposto e do tipo de informações que possa contribuir para a solução do mesmo. Um tipo de

conceitualização desta situação-problema poderia ser o que é sugerido na FIG.01.

Dentro da visão de que compreender seria um processo mais complexo do que simplesmente juntar os significados de diversas palavras, seria interessante que a IIn tivesse um conhecimento das "intenções" do usuário ao se dirigir a um sistema de informação ou seja, a IIn deve ter um "comportamento cooperativo", baseado em um "script"/"roteiro" do processo de interação com um usuário na busca de informações.

ex: num sistema inteligente que responda consultas em Linguagem Natural sobre os horários de partida e chegada em uma estação de trem, a resposta a pergunta formulada por um usuário:

- "A que horas chega o trem de Montreal?"

o sistema responderia:

- "As 3:15 Hs., na plataforma no. 7."

ex. de: ALLEN, J. F., PERROULT, C. R.

Ou seja, o sistema responde mais do que o solicitado (a resposta: "às 3:15 Hs." seria uma resposta correta à pergunta formulada), pois possui um modelo do usuário (suas INTENÇOES e necessidades) e portanto tenta se adiantar às necessidades do usuário, informando também em que plataforma o trem chegará.

3.3 ALTERNATIVAS PARA O PROCESSAMENTO LINGUISTICO DE CONSULTAS A BAMCOS DE DADOS (porque utilizar técnicas lingüísticas).

Os diferentes sistemas especialistas voltados para a recuperação de informações bibliográficas não tem implementadas interfaces em linguagem natural ou as tem empregando técnicas rudimentares do ponto de vista lingüístico. E no campo mais amplo de sistemas ou interfaces de consultas a bancos de dados comerciais (não bibliográficos) que pode-se encontrar precedentes e modelos para a proposta deste trabalho.

Na verdade a linguagem enquanto fenômeno cognitivo ainda é pouco conhecida: nas palavras de RUWET : "vê-se imediatamente que o fato central que a Lingüística sincrônica deve dar conta, é o seguinte: todo indivíduo adulto que fala uma determinada língua é, em qualquer momento, capaz de emitir espontaneamente, ou de perceber e compreender, um número indefinido de frases, que ele jamais pronunciou nem ouviu antes". Este fenômeno, a competência Lingüística dos seres humanos, tem sido objeto de intensas pesquisas multidisciplinares nos campos da Lingüística, psicologia, inteligência artificial, ciência cognitiva e mesmo na Ciência da Informação.

Nesta seção serão discutidas e analisada alternativas para o processamento lingüístico de consultas em linguagem natural como o proposto neste trabalho.

As técnicas automáticas (algorítmicas) de geração de expressões booleanas de consultas em Sistemas de Recuperação de informações a partir do texto das consultas formuladas diretamente pelo usuário final do sistema em linguagem natural, tem sido utilizadas como uma das primeiras formas de processamento lingüístico, ainda que primitivo, da expressão de consulta formulada por um usuário . Além de permitir a interação direta de um usuário com o sistema em "linguagem natural", estas técnicas são de fácil implementação. Elas se baseiam num procedimento de busca de palavras ou expressões no texto da consulta (string-search) cotejando-as com um dicionário. No entanto, por não trabalharem a compreensão do texto da consulta, estas técnicas podem resultar em equívocos sérios entre as intenções do usuário e a expressão booleana gerada, a qual será submetida ao SRI. Propõe-se no presente trabalho como alternativa o emprego de técnicas de Inteligência Artificial, especificamente técnicas de Processamento (Compreensão) de Linguagem Natural, como forma de processar uma consulta a SRIs formulada em linguagem natural. As citações seguintes mostram que grande parte dos sistemas similares ao proposto não tem qualquer tratamento lingüístico da expressão de consulta, limitando-se ao emprego da técnica de "string-search/busca a dicionário".

> "Outro aperfeiçoamento pode ser feito no estágio de entrada do sistema. Atualmente os usuários entram com um conjunto de termos um a um. Uma forma mais natural de fazer isto seria permitir que os usuários entrassem com uma consulta em linguagem natural e fazer o sistema identificar os termos significativos (utilizando, por exemplo, um algorítmo de extração de raízes de palavras e uma lista de palavras não significativas) e então realizar a busca por termos melhores." (SHOVAL).

> "... no sistema PLEXUS nós optamos por não usar um analisador e escolhemos o caminho oposto, usando raízes de palavras e uma lista de palavras não significativas... O processo de formar uma expressão de busca e em essência o seguinte: cada quadro é percorrido, palavras são extraídas e suas raízes são pesquisadas no dicionário e nas redes de categorias e de sinônimos. Sinônimos são unidos pelo conectivo OU. Grupos de sinônimos e termos únicos são unidos pelo conectivo E." (VICKERY).

> "1 . Palavras são primeiramente extraídas da formulação da consulta em linguagem natural, usando uma técnica comum de análise automática e indexação de textos. Assumindo que os termos T1, T2 ... Tn foram extraídos da formulação original da consulta em linguagem natural, uma expressão booleana inicial pode ser formulada na forma de (T1 or T2 or ... or Tn).

2 . correlações entre os termos da consulta podem ser computadas baseadas na co-ocorrência de pares ou triplas de termos da consulta nos documentos da coleção. Pares ou triplas de termos que co-ocorrem um certo número de vezes superior que um certo nível podem se adicionados a expressão booleana da consulta e unidos pelo conectivo E. Assim, se a correlação entre termos co-ocorrentes Ti, Tj e Tm, Tn exceder um certo nível, a expressão booleana da consulta pode ser incrementada da seguinte maneira: (T1 OU T2 OU ... OU Tn) OU (Ti E TJ) OU (TM E TN).

3 . Refinamento adicionais podem ser introduzidos permitindo que o usuário estipule um certo número máximo m de documentos que ele gostaria que fossem recuperados na consulta. Estimando o número de documentos recuperados por cada termo ou grupo de termos da expressão booleana de consulta, suficiente termos de busca podem ser incluídos na expressão de consulta, de modo a atingir-se o número de documentos desejado". (SALTON).

Examinemos as seguintes consultas hipotéticas, formuladas a um sistema:

a) "Desejo informações sobre inteligência artificial, especificamente os tópicos de representação do conhecimento e reconhecimento de padrões."

- caso esta consulta fosse interpretada utilizando uma técnica de string search/busca a vocabulário, resultaria na seguinte expressão booleana:

"inteligência artificial OU representação do conhecimento OU reconhecimento de padrões".

obs: uma regra "cega" para a formulação de booleanas estabelece que se os termos envolvidos tem alguma identidade semântica (no caso, os três termos pertencem a mesma hierarquia), devem ser unidos pelo conectivo OU, pois representam diferentes graus de especificidade do mesmo conceito; termos que não pertençam a mesma hierarquia nem sejam de alguma forma relacionados, devem ser unidos pelo conectivo E.

Um especialista humano no entanto, ao interpretar esta consulta, formularia a seguinte expressão booleana:

"representação do conhecimento OU reconhecimento de padrões".

b) "Desejo informações sobre inteligência artificial, especificamente os tópicos de representação de conhecimento para reconhecimento de padrões."

- caso esta consulta fosse interpretada utilizando uma técnica de string search/busca a vocabulário, resultaria na seguinte expressão booleana:

"inteligência artificial OU representação de conhecimento OU reconhecimento de padrões."

Um especialista humano no entanto, ao interpretar esta consulta, formularia a seguinte expressão booleana:

"representação do conhecimento E reconhecimento de padrões".

c) "Desejo informações sobre construção de compiladores exceto para as linguagens COBOL e RPG."

- caso esta consulta fosse interpretada utilizando uma técnica de string search/busca a vocabulário, resultaria na seguinte expressão booleana:

"construção de compiladores E COBOL OU RPG".

Um especialista humano no entanto, ao interpretar esta consulta, formularia a seguinte expressão booleana:

"construção de compiladores E-NAO (COBOL OU RPG)".

Dos exemplos levantados conclui-se que a técnica de "string-search" possui qualidades de robustez e facilidade de implementação, mais é lingüisticamente tosca; se a expressão de consulta consiste exclusivamente da Especificação Semântica do assunto desejado, ela mostra um desempenho "razoável", como por exemplo, na consulta "técnicas de **isolamento** de **cabos elétricos**"; seu desempenho no entanto piora, a medida que a consulta cresce em complexidade lingüística, por exemplo, em "necessito para o desenvolvimento do meu trabalho, saber tudo acerca do uso de radiações ionizantes, do tipo alfa e beta, para a preservação de alimentos de origem vegetal".

A tese aqui exposta é de que a compreensão de consultas a SRIs formuladas em Linguagem Natural requer um formalismo lingüístico mais sofisticado que simplesmente o método "string-search"/busca a dicionário.

Existem diferentes enfoques e soluções para a análise de expressões em linguagem natural por sistemas computadores. Estes enfoques variam de acordo com a natureza do problema a ser enfrentado. As diferentes soluções encontradas na literatura se agrupam em torno das respostas dadas a questão do papel da sintaxe e da semântica para a compreensão de expressões. A análise sintática consiste em, mais do reconhecer palavras de uma sublinguagem, reconhecer e ratificar se as mesmas estão estruturadas corretamente, dentro de um padrão aceitável. As diferentes experiências tem posições distintas sobre a real importância da análise sintática para a compreensão de uma expressão em linguagem natural submetida a um sistema computador.

Em WOODS, por exemplo, o sistema analisador de consultas se estrutura segundo um rígido formalismo de análise sintática, uma

gramática representada por um autómata a estados finitos, formalismo este proposto inicialmente por CHOMSKY para representar uma gramática, e relativamente fácil de ser implementado em um programa de computador. Um autômata a estados finitos se constitui em um grafo em que nodos representam estados, um deles sendo considerado o estado inicial e um ou mais como sendo estados terminais, e de transições, onde a passagem de um nodo ou estado a outro se dá pela aceitação de uma palavra ou uma estrutura sintática mais complexa como um predicado (formado, por exemplo por um verbo seguido de um complemento nominal) ou um complemento nominal (formado, por exemplo, por um determinante seguido de um substantivo comum).

Uma certa seqüência de palavras permite que o grafo seja percorrido de um estado inicial até um estado final; se tal fato se ocorre, aquela seqüência de palavras (uma sentença) é dita "reconhecida", ou sintaticamente correta para a gramática representada pór aquele autômata.

A representação computacional deste formalismo é conhecido como ATN (Augmented Transition Network), sendo já bastante conhecido e empregado em diversos sistemas. O mais conhecido deles, LUNAR, se constitui numa interface a uma base de dados com informações sobre geologia lunar. Este formalismo tem a vantagem de já ser bastante conhecido e testado, além de poder ser mais facilmente implementado com ferramentas padrão de IA, como Prolog ou Lisp. No entanto poderia ser muito rígido para a aplicação em questão, pois exigiria que todas as palavras que constarem da consulta formulada em Linguagem Natural sejam conhecidas (constem nos Léxicos do sistema). Essa característica provavelmente resultaria numa gramática pouco "robusta" para a situação. Segue-se um esboço de uma gramática, cujas categorias sintáticas foram criadas livremente e que poderia ser implementada no formalismo ATN; note-se ao estabelecer-se este esboço, já foram levadas em conta as características estruturais das expressões de consulta, conforme estabelecido em 4.3.2.

CONSULTA -> (ESP-NEC), (ESP-CAR), ESP-ASS, (ESP-DET-RES).

ESP-NEC -> (AGENTE), SINT-VERB.

AGENTE -> eu.

SINT-VERB -> (LOC-VERBAL)/VERBO.

LOC-VERBAL -> VERBO, VERBO-INF.

VERBO -> preciso, necessito, gostaria.

VERBO-INF -> obter, saber, recuperar.

ESP-CAR -> (PREP), TIPO-INF.

PREP -> de.

TIPO-INF -> TIPO-SIMPL / TIPO-COMP.

TIPO-SIMPL -> informaçOes, referências.

TIPO-COMP -> TIPO, PREP, ADJETIVADOR, QUALIF-TEMP.

TIPO -> artigo(s), anais.

ADJETIVADOR -> periódico(s), conferência(s).

QUALIF-TEMPO -> IND-PERIODO / IND-INTERVALO.

IND-PERIODO -> recentes, novos ,atuais.

IND-INTERVALO -> entre DATA1 e DATA2, anteriores a DATA1, posteriores a DATA1.

ESP-ASS -> PREP-INTRODUTORIA, (PROCESSO), COORDENAÇAO.

PREP-INTRODUTORIA -> sobre, acerca (de), atinentes, versando sobre, a respeito de.

PROCESSO -> uso(s), processo(s), utilização(Oes), emprego, produção, análise, síntese, aproveitamento, obtenção.

COORDENAÇAO -> TERMO-QUAL < COORD, TERMO-QUAL>.

COORD -> de, para, por, sobre, em.

TERMO-QUAL -> (ADJETIVADOR), TERMO.

ADJETIVADOR -> menos, exceto, extritamente, especificamente, principalmente.

TERMO -> palavra do VOCABULORIO ESPECõFICO DO DOMõNIO ou da LISTA DE AUTORIDADES.

A Gramática descrita permitiria analisar consultas como as seguintes:

- "Necessito de referências sobre uso de radioisótopos na industria."

- "Eu necessito de artigos de periódico acerca de métodos de prevenção de acidentes nucleares."

Em SCHANK, por outro lado, o analisador de consultas está estruturado de forma a dar grande ênfase à analise semântica, ou seja, gerar uma representação semântica da sentença analisada num formalismo denominado Dependência Conceitual (CD) que tem como objetivo guardar o significado conceitual universal de um sentença, independente do idioma e das palavras utilizadas para expressar seu sentido; a representação obtida poderia ser

utilizada por exemplo para gerar paráfrases da sentença original, isto é, sentenças diferentes mas que guardem o mesmo significado da sentença original. Não se objetiva aqui gerar uma árvore com a estrutura sintática da sentença, mas empregar fontes de conhecimento mais amplas, conhecimento extralingüístico, como por exemplo um modelo do mundo em que o contexto da sentença se aplique e que auxilie na sua compreensão; o objetivo é descobrir uma representação semântica da sentença. Isso é possível no sistema de SCHANK, devido ao papel central que desempenha aí o dicionário, que incluí grande quantidade de informação semântica, grande parte dela de natureza não-lingüística, sobre o sentido das palavras e dos seus possíveis papeis numa sentença e sobre as expectativas que o reconhecimento de uma palavra desperta, que permitem antever e restringir o significado da sentença como um todo. Seu analisador, ao reconhecer uma palavra, lhe atribui um papel semântico (embora não definitivo) e dispara uma série de expectativas para as próximas palavras a serem analisadas, expectativas essas obtidas do dicionário, ao permitir mais que o reconhecimento da própria palavra e seu papel sintático, mas também dos seus possíveis papeis semânticos no contexto da sentença analisada e das expectativas de palavras que se seguirão e completarão a conceitualização de uma dada situação.

Outra importante experiência é a proposta de DE SOUZA, denominada de "Gramáticas de Determinação". Sua utilização em aplicações semelhantes a IIn, isto é, como "front-end" de um sistema gerência de bases de dados, utilizado no caso para gerenciar uma base de dados comercial, faz com que esta experiência deva ser considerada com bastante atenção. Nesta proposta a estrutura de uma consulta é motivada pelo Princípio da Determinação: existe sempre numa consulta um elemento determinante ou qualificador das informações a serem recuperadas e um elemento determinado, que corresponde a que informações se pretende recuperar. A estrutura de uma consulta é composta desses dois elementos: informações "pedidas" ou solicitação de recuperação, que seriam as informações determinantes; e informações "fornecidas" ou critérios de seleção, que seriam as informações determinadas; eventualmente pode surgir um terceiro elemento, as informações relativas ao modo de publicação ou exibição das informações solicitadas. Por sua vez cada um destes componentes pode ser analisado como composto também de uma parte determinante e uma parte determinada. O reconhecimento de que palavras ou expressões constituíam a estrutura essencial da consulta em termos de seus elementos determinantes e determinados, torna fácil o seu mapeamento para a linguagem de consulta ao banco de dados, neste caso o Adascript; a identificação dessas palavras ou expressões essenciais permite que o formalismo lingüístico desconsidere ou despreze muitas palavras, concentrando-se nos elementos semanticamente significativos dentro da estrutura da consulta, que tornam possível o mapeamento da mesma da linguagem natural para a linguagem de consulta do banco de dados.

Das experiências anteriores percebe-se claramente a complexidade do problema de fazer um sistema em computador

"compreender" consultas em linguagem natural. A linguagem, como todos os fenômenos cognitivos, ainda é um fenômeno para o qual a ciência ainda não conseguiu uma interpretação satisfatória.

Notas e Referências:

EMBRATEL significa Empresa Brasileira de TELECOMUNICAÇOES, a estatal brasileira responsável pelo setor de TELECOMUNICAÇOES.

1. DE SOUZA, C. S. **Gramáticas de Determinação: uma proposta metodológica.** In: Jornadas Argentinas de Informática e Investigación Operativa, v. 2, 1988.

2. SIQUEIRA, I. S. P, PEREIRA, A. E. C. **Perspectivas de Aplicação da Inteligência Artificial à Biblioteconomia e à Ciência da Informação.** R. Bras. Bibliotec. e Doc., v. 22, n, 1/2, p.39-80, 1989.

3. FOX, E. A. **Practical Enhanced Boolean Retrieval: Experiences with the SMART and SIRE Systems.** Information Processing & Managenment, v.24, n. 3, p.257-67, 1988.

4. MARON, M. E. **Probabilistic Design Principles for Conventional and Full-Text Retrieval Systems.** Information Processing & Managenment, v.24, n. 3, p.249-55, 1988.

5. TAHANI, V. **A Fuzzy Model of Document Retrieval Systems.** Information Processing & Managenment, v.12, p.177-87, 1976.

6. RADECKI, T. **Probabilistic Methods for ranking Output Documents in Conventional Boolean Retrieval Systems.** Information Processing & Managenment, v.24, n. 3, p.281-302, 1988.

7. LOSEE, R. M. **Intergrating Boolean Queries in Conjunctive Normal Form with Probabilistic Retrieval Models.** Information Processing & Managenment, v.24, n. 3, p.315-21, 1988.

8. BLAIR, D. C. **An Extended Relational Document Retrieval Model.** Information Processing & Managenment, v.24, n. 3, p.349-71, 1988.

9. GORDON, M. D. **The Necessity for Adaptation in Modified Booean Document Retrieval Systems.** Information Processing & Managenment, v.24, n. 3, p.339-47, 1988.

10. FIDEL, R. **Online searching styles: a case-study-based Model of Searching Behavior.** J. of the American Society for Information Science, v. 35, n. 4, p.211-221, 1984.

11. SARACEVIC, T., KANTOR, P. A Study of Information Seeking and Retrieval. JASIS, v. 39, n. 3, part I, p.161-176, part II, p.177-196, part III, p. 197-216, may 1988.

12. BELKIN, N. J., ODDY, R. N. BROOKS, H. M. **ASK for Information Retrieval.** Journal of Documentation, v. 38, n. 2 p.61-71, 1982. Part I.

13. FLUHR, C. **Information Retrieval.** In: CONSULTIVE MEETING of INIS LIAISION OFFICERS, 17, Viena, May 1989.

14. SCHANK, R. **Representing and Understanding of Text.**

15. ALLEN, J. F., PERROULT, C. R. **Analyzing Intention in Utterances.** Artificial Intelligence, v. 15, p.143-178, 1980.

16. RUWET, N. **Introdução à Gramática Gerativa.** São Paulo, Ed. perspectiva, Ed. da Universidade de São Paulo, 357p., 1975.

17. SHOVAL, P. **Principles, Procedures and Rules in an Expert System for Information Retrieval.** Information Processing & Management, v. 21, n.6, p.475-87, 1985.

18. VICKERY, A. BROOKS, H. e ROBINSON, B. **A Reference and Referral System using Expert System Techniques.** Journal of Documentation, v.43, n.1,p.1-23,march 1987.

19. SALTON, G. **A Simple Blueprint for Automatic Boolean Query Processing.** Information Processing & Management, v. 24, n. 3, p.269-80, 1988.

É interessante notar que no trabalho de Salton, o exemplo sobre o qual o autor desenvolve seu método, "Excretion of phosphate or pyrophosphate in urine", é desdobrado pelo seu procedimento de geração automática de booleanas em diferentes expressões de consulta, exceto na mais natural e imediata para uma pessoa que entendesse a formulação dessa consulta em Linguagem Natural:

"Excretion AND (phosphate OR pyrophosphate) AND urine".

20. WOODS, W. A. **Transition Network Grammars for Natural Language Analysis.** Communications of the ACM, v. 13, n. 10, octuber, p.591-606, 1970.

21. SCHANK, R. **Conceptual Information Processing.** New York, North-Holland/American Elsevier, 374p. 1975.

4 MODELO PROPOSTO

O material básico para a pesquisa foi um conjunto de formulários de busca retrospectiva utilizados no CIN para registrar os pedidos de buscas retrospectivas de usuários que solicitam diretamente o serviço (atendimento tipo posto de serviço) ou de usuários que solicitam o serviço via correio; o formulário em questão, denominado SUPRIR - SOLICITAÇAO DE BUSCA RETROSPECTIVA, contém a formulação da consulta do usuário em linguagem natural, além de uma ou mais expressões booleanas elaboradas pelo intermediário, juntamente com os resultados (número de documentos) obtido. Para formulação do modelo lingüístico capaz de interpretar as expressões de consulta em linguagem natural, foram examinadas cerca de 60 (sessenta) formulários, com buscas referentes aos anos de 1986 e 1989, contendo buscas retrospectivas feitas sobre diferentes bases de dados das existentes no CIN.

Tomou-se como base que a IIn trabalharia sobre bases de dados que possuam instrumentos típicos de recuperação como tesauros, tabelas de categorias, manuais de indexação, uma vez que os conhecimentos contidos nestes instrumentos vão se constituir nas bases de conhecimento com as quais a IIn irá trabalhar (segundo as palavras de MINSKY, um dos precursores da Inteligência Artificial, um sistema baseado em conhecimento é tanto mais poderoso quanto mais conhecimento tenha armazenado).

No modelo proposto, o processamento de consultas através da IIn será feito segundo três etapas, conforme esquematizado na FIG.02: a) - a entrevista pré-busca, onde é elaborado o modelo do usuário; b) - o processamento lingüístico da expressão de consulta, resultando na formulação de uma expressão booleana representando a consulta do usuário; e c) - a interação do módulo especialista em busca com o SRI, a partir da expressão de consulta original, com o SRI propriamente dito para a recuperação de informações e com o usuário final para avaliá-las e prover o necessário "feedback". O modelo proposto tem portanto três componentes: primeiro o módulo construtor do modelo do usuário, de seu problema e de suas necessidades de informação; em seguida, o módulo de processamento lingüístico, que executa a análise de expressões de consultas formuladas em linguagem natural às bases de dados do CIN e realiza o mapeamento das mesmas para a linguagem de consulta do SRI, baseada na álgebra booleana; e, por fim, o módulo especialista em avaliação/reformulação de expressões de busca, que interage com o SRI para a recuperação de referências e com o usuário para a avaliação das mesmas. Em seguida são apresentados em detalhe os três módulos, precedidos de uma introdução geral sobre a interação entre eles.

4.1 Processo de busca e interação com um SRI via IIn.

Nesta seção será descrito em linhas gerais o processo de busca e interação entre um usuário e um SRI e paralelamente será

formulado o modelo proposto para esta interação através da IIn. O processo de busca e interação entre um usuário e o SRI, intermediado pela IIn, segue aproximadamente as mesmas etapas de uma interação intermediada por um especialista humano. A FIG.03 procura mostrar um paralelo entre os dois processos, destacando a correspondência entre eles.

Na interação com a IIn existe também a necessidade formal de uma etapa correspondente à entrevista pré-busca, que tem como objetivo fixar os parâmetros iniciais do "modelo do usuário"; no caso da IIn, o modelo do usuário servirá para sintonizar a operação do sistema, principalmente a parte interativa que se constitui na interface com o usuário, com as necessidades, o grau de perícia e de experiência e a própria clareza que possua o usuário do objetivo/assunto de sua busca.

Após esta etapa e fixados estes parâmetros iniciais, o sistema passa para a segunda etapa onde o usuário formula sua consulta em linguagem natural. A IIn procura então utilizar seu formalismo lingüístico para interpretá-la; a IIn pode interrogar o usuário sobre a interpretação correta de palavras ou expressões e solicitar a substituição de palavras desconhecidas por outras conhecidas. Como última alternativa será tentado um procedimento de "string-search".

Se a IIn não conseguir interpretar satisfatoriamente a consulta formulada pelo usuário, será considerado que a formulação da consulta do usuário é pouco clara, ou ele não está afeito à área de cobertura da base de dados nem a sua terminologia, ou seu interesse está fora da área de cobertura da mesma; o modelo do usuário será então alterado e a IIn mudará a forma de interação com o usuário, passando a exibir-lhe a tabela de classificação da base de dados, no sentido de mostrar ao usuário a gama de assuntos e seus enfoques coberta pela base e em seguida passará a exibir o tesauro da base para que o usuário possa escolher termos adequados. Esta fase corresponde à proposta de FLUHR (1) e os detalhes desta interação são secundários para o presente trabalho e portanto não serão aqui desenvolvidos.

Se a formulação da consulta do usuário em linguagem natural foi compreendida pela IIn, a próxima etapa será o mapeamento da estrutura da consulta com seus atributos, para a linguagem-alvo do sistema, uma expressão booleana contendo descritores e outros itens de busca, unidos pelos conectivos lógicos **E**, **OU** e **E-NÃO**. Esta expressão booleana é o ponto de partida do módulo especialista em busca, responsável pela interação da IIn com o SRI; uma vez recuperados alguns documentos, também aqui eles serão submetidos à avaliação do usuário e, em função desta avaliação e dos parâmetros estabelecidos no modelo do usuário, e empregando heurísticas de reformulação, a expressão de consulta original será reformulada e re-submetida ao SRI; este processo continuará até que o usuário se considere satisfeito com o conjunto de documentos recuperados.

Para maior clareza, em seguida é apresentado um quadro das etapas do processo de interação usuário/SRI através da IIn, destacando que etapas são realizadas pelo SRI convencional, e que etapas demandam conhecimento e perícia que devem ser simulados pela IIn:

- Entrevista pré-busca; construção do modelo do usuário/problema.
- Formulação dos objetivos/necessidades de informação por parte do usuário em LN.
- Compreensão da consulta pela Interface Inteligente (tarefa inteligente).
- Formulação da estratégia de busca (tarefa inteligente).
- Recuperação das informações (a cargo de um SRI convencional).
- Avaliação/Feedback/Reformulação (tarefa inteligente).

4.2 A entrevista pré-busca; elaboração do modelo do usuário.

Entende-se aqui por modelo cognitivo ou modelo do usuário todo um componente/processo da IIn que tem como objetivo obter e manter informações relacionadas com o usuário em aspectos relativos ao problema que motivou sua busca de informações, a natureza das informações demandadas pelo usuário, as expectativas do mesmo em relação ao SRI, seu grau de perícia em interagir com um SRI, seu conhecimento do assunto que é objeto da busca. O modelo cognitivo, num processo de busca intermediado por um especialista humano, é construído por este durante a entrevista pré-busca. No caso da IIn, as informações obtidas e mantidas no modelo do usuário permitirão graduar a interação da IIn com o mesmo, melhorando a performance e aceitabilidade do sistema. A este respeito algumas perguntas devem ser respondidas: em primeiro lugar, que informações deve conter o modelo do usuário; em segundo lugar, como obtê-las; e em terceiro lugar, como a IIn deve usá-las como intermediária na interação entre o usuário casual e o SRI?

As informações para a formulação do modelo do usuário foram portanto, também no caso da IIn, obtidas na etapa de entrevista pré-busca. Nesta etapa, o sistema levou a cabo um diálogo com o usuário no sentido de descobrir se o usuário está interessado em revocação (recall) ou precisão (precision), além de solicitar ao usuário o número máximo de documentos que interessava examinar, tipos de material de interesse (artigos de periódico, teses, relatórios, etc). Esses parâmetros iniciais formarão o modelo do usuário para o sistema e graduarão o funcionamento da IIn nas etapas de formulação de expressão booleana de busca e na sua posterior reformulação, de acordo com a avaliação do usuário. Este diálogo, como recomenda a literatura (DANIELS), deve ser conciso e poderia ser executado através de menus (3).

4.3 PROCESSAMENTO LINGUISTICO DA EXPRESSAO DE CONSULTA.

O papel de um formalismo lingüístico no contexto da IIn, que seja capaz de processar expressões de consultas formuladas em linguagem natural com o objetivo de recuperar informações de um banco de dados gerenciado por um SRI, praticamente se confunde com o papel mesmo destinado a IIn. A pesquisa em processamento de linguagem natural tem como objetivo proporcionar uma interação o mais confortável possível entre um usuário humano e um sistema computador, dispensando-o de ter que aprender uma linguagem específica (comandos, por exemplo) para poder interagir com o computador.

As pesquisas nesta área e os sistemas desenvolvidos que obtiveram um certo grau de efetividade só o conseguiram às custas de **restringirem** seu potencial de processamento lingüístico, geralmente limitando o seu vocabulário, isto é, o domínio sobre o qual exercem sua competência lingüística, ou só processando expressões que se limitem a uma sentença, já que a complexidade de processar um texto composto de várias sentenças cresce exponencialmente com o número de sentenças.

Levando estas questões em consideração, além dos limites intrínsecos do presente trabalho, procurou-se desenvolver um formalismo lingüístico que possa ser adaptado a outros SRIs/bases de dados. As restrições impostas a sistemas de obtiveram algum sucesso também terão que ser impostas neste caso; no entanto, procurou-se levar em consideração que talvez a restrição de domínio imposta a esses sistemas, no caso da IIn, pudesse ser considerada não em relação a um assunto ou cobertura específicos, alguma coisa como o escopo de uma base de dados sobre a qual a IIn atuaria. Seria desejável que domínio, no contexto da IIn, pudesse ser considerado com "recuperação de informações bibliográficas" e que a potencialidade da IIn não se limitasse a consultas a uma base de um domínio específico, por exemplo, às bases que serviram de suporte a este trabalho.

Considere-se a consulta: "Preciso de artigos recentes, em português ou inglês, sobre esterilização de pragas na agricultura por irradiação"; ou a consulta: "Desejo informações as mais recentes, nos idiomas português e inglês, a respeito de normas de segurança para o transporte de materiais nucleares". Verifica-se claramente que as palavras empregadas numa expressão de consulta em linguagem natural podem ser consideradas como pertencendo a dois grandes domínios: algumas palavras compõe o jargão de "recuperação de informações bibliográficas", como por exemplo "referências", "artigos de periódicos", "informações", "recentes", "a partir de 1985", etc, enquanto outras palavras são diretamente pertencentes ao domínio ou escopo da base de dados pesquisada e que normalmente compõe o seu vocabulário controlado ou tesauro.

Ver-se-a adiante como esta hipótese se confirma, a partir da análise estrutural das expressões de consulta, feitas predominantemente em consultas sobre a base INIS, mas abrangendo

também uma amostra de consultas sobre outras bases. Isso permite supor que o esquema proposto é suficientemente genérico, que o corte de domínio efetuado pode ser melhor caracterizado de fato como "recuperação de informações bibliográficas" ao invés de restringir-se ao escopo de uma ou outra base de dados. Daí poder-se tirar como conclusão, a ser incorporada no formalismo lingüístico a ser desenvolvido, a generalidade desta característica estrutural de uma expressão de consulta que consiste no fato de que as palavras que a compõe poderem ser classificadas como pertencentes a dois "vocabulários" distintos.

A este esboço de requisitos para o modulo lingüístico da IIn agrega-se um outro: robustez; o sentido dado a este requisito aqui é o seguinte; veja-se por exemplo a seguinte consulta:

- "Obter informações sobre critérios de dimensionamento, especificações de materiais, restrições, critérios para qualificação de equipamentos e materiais elétricos, garantia de qualidade de equipamentos e componentes elétricos, sistemas de segurança, on site power, off, site power, de reatores de pesquisa";

foi transformada na seguinte expressão booleana:

-"(electric measuring instruments OU electric equipment) E research reactors E (dimensions OU specifications).

Ou seja, somente algumas palavras da expressão original da consulta em LN tem relevância para a formulação da expressão booleana, ao mesmo tempo que servem de base para que sejam feitas diversas inferências, com base num tesauro, no sentido de obter descritores que possam descrever da melhor forma possível as necessidades de informação da consulta formulada em linguagem natural. O formalismo lingüístico deve possuir características semelhantes, no sentido de obter da expressão da consulta original somente palavras que possam ser traduzidas em descritores do tesauro ou possam conduzir, através de inferências/navegação pela estrutura do tesauro, em descritores do mesmo; a compreensão/tratamento lingüístico de todas as palavras que aparecem na formulação da consulta seria provavelmente excessivamente complicada e dispendiosa, sem representar um aporte significativo na qualidade da expressão booleana gerada, que deve ser, em última instância, o resultado pretendido por todo o processamento lingüístico da expressão de consulta em linguagem natural. Esta característica obriga que se examine a princípio como uma das alternativas de tratamento lingüístico para as expressões de consulta uma técnica de simular compreensão de expressões em linguagem natural largamente usada no tratamento automático de textos, em indexação automática e mesmo, como veremos adiante, em interfaces a SRIs: a técnica "string-search", ou seja, de busca de palavras-chaves constantes de um dicionário no texto de uma expressão em linguagem natural, um título de um artigo, ou a formulação de uma consulta a um SRI, e a geração de expressões booleanas a partir dessas palavras-chaves ou

descritores recuperados. No entanto algumas outras questões devem ser antes resolvidas.

4.3.1 O que significa "compreender" uma consulta em linguagem natural. Características da linguagem-alvo (álgebra booleana).

No contexto deste trabalho, a compreensão da linguagem natural por um sistema (programa de computador) consiste em passar para este sistema instruções em linguagem natural, ou seja, conseguir do sistema um comportamento útil, que seria solicitado através de sentenças em linguagem natural. O comportamento útil que se deseja de um SRI é a recuperação de informações bibliográficas relevantes para a situação/problema do usuário. Os comandos que são inteligíveis a um SRI convencional são consultas formuladas unindo palavras-chave com os conectivos booleanos (E, OU e E-NÃO) representando as operações de intersecção, união e diferença de conjuntos de documentos representados por descritores.

Portanto, para que se possa submeter a um SRI convencional consultas em linguagem natural e obter um desempenho útil para um usuário que o esteja consultando, é necessário que uma Interface Inteligente anteposta ao SRI consiga mapear as consultas do usuário, formuladas em linguagem natural, em expressões de álgebra booleana, que possam resultar em ações úteis por parte do SRI. A álgebra booleana portanto é a representação-alvo na qual se pretende mapear as consultas formuladas pelo usuário em linguagem natural.

A representação-alvo em questão implica, naturalmente, como já se viu, numa simplificação da expressão original da consulta formulada em linguagem natural. A álgebra booleana como representação-alvo é muito mais pobre expressivamente que a linguagem natural; por exemplo, as seguintes consultas em LN:

- "Preciso de referências a respeito de segurança de instalações nucleares."

- "Estou procurando informações sobre segurança de instalações nucleares";

- "segurança de instalações nucleares";

- "Eu gostaria de obter o máximo de informações sobre o tópico segurança de instalações nucleares";

seriam todas mapeadas, por qualquer intermediário humano, na seguinte expressão booleana:

- segurança E instalações nucleares.

Ou seja, o mapeamento de expressões de consulta em linguagem natural para expressões em álgebra booleana implica certamente numa **simplificação** e conseqüente perda de conteúdo semântico da

expressão original. Na coordenação de dois ou mais conceitos, gerando um outro conceito (técnica bastante conhecida e discutida em Ciência da Informação), a forma ("faceta", "papel", ou "caso semântico") com que cada um dos conceitos componentes participa na formulação do conceito resultante fica obscurecida pela pobreza expressiva do conectivo booleano "E". Veja-se os seguintes exemplos:

- "Gás natural (A) como subproduto da prospecção de petróleo (B)".

- "Uso de eletricidade (A) para produção de alumínio (B)".

No primeiro caso, tanto A como B tem o papel semântico de produtos, enquanto no segundo caso, A é um instrumento para a obtenção do produto B.

Em ambas as formulações, a expressão booleana resultante seria provavelmente: **A E B**; no entanto, fica claro que nas duas consultas, os componentes A e B tem papeis semânticos distintos. A pobreza expressiva da álgebra booleana como linguagem-alvo não consegue captar essa distinção. Dessa forma, pode-se esperar do formalismo lingüístico aqui proposto que consiga, no máximo, ser tão efetivo quanto a álgebra boolena; este é o seu limite. Poder-se-ia pensar numa interpretação semânticamente mais rica da consulta, baseada talvez nos "casos", como o proposto por CUNHA e empregado para indexação; o formalismo lingüístico aqui proposto poderia facilmente ser adaptado para obter esta representação. No entanto isso só teria alguma efetividade se a mesma sistemática fosse também empregado na entrada de informações no sistema, na indexação das referências, o que foge à proposta do presente trabalho.

Muitas vezes, esta perda de conteúdo não é devida somente às características da representação-alvo, mas também devido as características da própria linguagem de indexação, que pode ser mais orientada à revocação ou mais orientada à precisão; por exemplo, a consulta:

> - "Processamento de imagens a partir de dados obtidos por tomografia ou outro método de diagnóstico usado em medicina" (RS311/86);

seria mapeada em:

> - **Image processing**;

ou a consulta:

> - "Necessita-se informações sobre trabalhos experimentais e modelos teóricos sobre os processos de absorção de gases, particularmente em colunas recheadas. São importantes não só informações sobre absorção física como também para o caso em que ocorre reação química, para que se estabeleça uma análise

do mecanismo predominante no processo, se é a difusão do Gás ou a cinética da reação" (RS217/89 - Base INSPEC);

seria mapeada em:

- **Gás absoption.**

O processamento de linguagem natural num sistema como o proposto teria então dois sentidos: em primeiro lugar parece inquestionável que a possibilidade formular consultas em linguagem natural se torna sem dúvida num fator facilitador do acesso ao SRI; nada parece melhor ou mais "humano", até mesmo do ponto de vista psicológico/epistemológico, que interagir com um sistema em linguagem natural. Em segundo lugar, devido à pobreza expressiva da álgebra booleana como representação-alvo, conforme já foi discutido anteriormente, a formulação de uma consulta em linguagem natural não seria mais que o ponto de partida para um processo interativo, de refinamentos sucessivos, que o módulo especialista em busca/interação levaria a efeito, empregando outras técnicas já experimentadas em sistemas automáticos, como "relevance feedback" - ver DILLON, ou formulação da expressão de busca a partir de um ou mais documentos previamente conhecidos pelo usuário e julgados relevantes - ver VERNIMB.

4.3.2 Tipologia das consultas.

A elaboração do formalismo lingüístico baseou-se num estudo e classificação das expressões de consultas investigadas segundo suas características estruturais. As expressões de consultas examinadas sofreram um primeiro agrupamento segundo suas características estruturais enquanto sentenças da língua portuguesa. Procedeu-se a uma análise sintática do que parecia ser um modelo canônico de uma expressão de consulta; após esta análise sintática inicial procedeu-se a uma análise estrutural, de modo a obter-se que sintagmas componentes da estrutura sintática da consulta eram mais relevantes para a formulação da expressão booleana a ser submetida ao SRI.

Os diferentes tipos foram obtidos pela maior ou menor proximidade deste modelo canônico, considerando também o seu maior ou menor grau de (a)gramaticidade, conforme a estrutura sintática do português.

Os tipos de consultas dirigidas ao SUPRIR, conforme atestam os formulários SUPRIR - SOLICITAÇAO DE BUSCA RETROSPECTIVA, são as mais diversificadas possíveis; bàsicamente, temos os seguintes tipos, que seria necessário que o formalismo lingüístico pudesse processar:

. Consultas bem estruturadas, gramaticais, não ambíguas e com a terminologia adequada, como por exemplo:

- "Deseja-se obter trabalhos que discutam a variação da viscosidade de óleos usados em sistemas hidráulicos com a temperatura" (RS 319/86, Base INSPEC).

"Deseja-se obter o maior número possível de referências bibliográficas acerca da modelagem e simulação do controle de robôs hidráulicos, com servoatuadores lineares ou rotativos" (RS 402/86, Base INSPEC).

- "O objetivo desta busca é obter informações sobre simulação e modelagem de processo de extração de urânio por solvente" (RS 251/86, Base INIS).

- "O objetivo da busca é a recuperação de documentos atinentes à teoria de controle de servomecanismos aplicados aos atuadores na robótica" (RS 401/86, Base INSPEC).

. Estruturada, com especificação:

- "Documentos teóricos e experimentais sobre a análise do espalhamento de partículas alpha de fontes naturais ou artificiais por átomos de gás hélio (espalhamento de Coulomb); interessam as medidas para partículas alpha de energia até 10 MEV" (RS 241/86, Base INIS).

- "Busco informações sobre o comportamento de plantas PWR p/ pequenos e grandes transientes; não desejo informações de acidentes ou LOCA" (RS 321/89, Base INSPEC); obs: LOCA significa "Loss of coolant accident".

. Consultas "enumerativas":

- "Fontes alternativas de energia: biomassa, maré, solar, geotérmica, eólica, etc." (RS 268/86, Base FONTE).

- "Obter referências sobre: aplicação da teoria de sistemas à biologia e medicina; biofísica teórica e matemática; biofísica de processos neurofisiológicos; redes neurais." (RS 16/89, Base INSPEC).

. Consultas "telegráficas": estas limitam-se a especificar o assunto em questão; geralmente são a-gramaticais.

- "Tratamento de rejeitos nucleares".

- "Extração de ouro por colunas de troca iônica" (RS 198/86, Base METADEX).

obs: note-se que este tipo de consulta é agramatical, considerando-se a gramática do português.

. Consultas usando a linguagem final do sistema.

- "Programação lógica $ Sistemas Especialista" (RS431, Base INSPEC).

- "Solar Energy $ (Residential Buildings + Residential Sector)" (RS 5/86, Base FONTE).

- "Idiun Processes $ Galliun Processes" (RS 310/86).

obs: (o caráter "$" significa, na linguagem de interação do Suprir o conectivo "E"; o caráter "+" significa o conectivo "OU").

. **Consultas tendo como argumento de busca um atributo OBJETIVO das referências:**

- "Autor: Eduardo Pena Franca" (RS 210/89, Base INIS).

4.3.3 Características estruturais das expressões de consulta em linguagem natural.

Uma referência bibliográfica, como todo objeto representado e armazenado em um sistema computador, é descrito através de seus atributos. Atributos típicos de uma referência bibliográfica, encontrados em praticamente todas as bases de dados das mais diversas coberturas, são título, autor, periódico, imprenta, data de publicação, além de descritores ou palavras-chaves que procuram descrever o conteúdo do documento, responder a questão "sobre o que" trata o documento. Os atributos de uma referência bibliográfica podem ser divididos em dois tipos:

- atributos OBJETIVOS: o título, o nome de um autor, o tipo (se um periódico, um livro, uma norma, um artigo), o nome de um periódico, a língua em que o documento original está escrito..

- atributos SEMANTICOS: bàsicamente o assunto ou conteúdo de uma referência: os descritores, a classificação de um documento.

As consultas ora se referem a atributos OBJETIVOS, ora a atributos SEMANTICOS, ora a ambos.

As consultas relacionadas em 4.3.2 sugerem uma estrutura aproximada como a da FIG.04 (baseada na proposta de "indicador sintagmático" - phrase-marker, de Noan Chomsky):

ex: em uma consulta hipotética, apliquemos este esquema:

- "Preciso obter artigos de periódicos, em inglês ou alemão, sobre prevenção de acidentes em reatores tipo PWR; preciso especificamente de artigos que tratem de exemplos práticos".

- Esp.Neces.Infor: "Preciso obter..."

- Esp.Carac.Obj: "...artigos de periódicos, em inglês ou alemão..."

- Esp.Semântica: "...sobre prevenção de acidentes em reatores tipo
PWR..."

- Esp.Detalhe/Restr.: "...preciso especificamente de artigos que tratem de exemplos práticos."

4.3.4 Modelo proposto: um conjunto gramática/vocabulários, e regras de compreensão e mapeamento.

O enfoque aqui utilizado para a estruturação de um formalismo lingüístico para a análise de consultas formuladas em linguagem natural a um SRI deve consistir, então, de dois elementos básicos: um Vocabulário (conjunto de palavras reconhecido) e uma Gramática (as estruturas permitidas para o encadeamento das palavras). A Gramática, entendida aqui como o reconhecimento da estrutura de uma expressão de consulta que permita a compreensão de consultas formulada à IIn, é inspirado no formalismo conhecido em PLN como "Gramática Semântica"; consiste em mapear a estrutura de uma sentença, não em estruturas sintáticas para depois mapear essas estruturas em estruturas semânticas, mas, neste caso, tem o papel de guiar o reconhecimento de traços semânticos e expectativas relevantes para o posterior mapeamento da expressão de consulta para a linguagem-alvo. Como já foi visto, muitas consultas são claramente agramaticais e a estrutura sintática canônica do português auxilia pouco na compreensão da expressão de consulta e no seu posterior mapeamento para a linguagem-alvo. Nesse sentido, a análise estrutural da consulta tem menos um papel "sintático" e predominantemente um semântico.

Note-se que estas opções se dão fundamentalmente condicionadas pelo que seria "compreender" uma consulta no âmbito da IIn, ou seja, mapear a expressão de consulta formulada pelo usuário em linguagem natural numa expressão da linguagem de consulta do sistema, baseada na álgebra booleana. Este enfoque aproveita a robustez da técnica de "string-search" e busca ao Dicionário que criticamos anteriormente, superando a fragilidade lingüística desse método, que poderia levar a erros de interpretação como os apontados. O método de "string-seacrh" é robusto porque pode desprezar as palavras desconhecidas, construindo a expressão booleana somente com as palavras identificadas no dicionário; como já foi analisado, seu desempenho é razoável na parte expressão de consulta referente à Especificação Semântica (ESP-ASS). O Analisador Sintático-Semântico proposto estender esta técnica; ele baseia sua análise na estrutura de uma consulta como a da FIG.05; a estrutura de uma consulta se divide aí em duas grandes partes, a ESPECIFICAÇAO-DAS CARACTEISTICAS-OBJETIVAS (ESP-CAR) de uma consulta e a ESPECIFICAÇAO-DO-ASSUNTO (ESP-ASS), esta introduzida por uma expressão equivalente a preposição "sobre", introduzindo exatamente o assunto de interesse do usuário. O Analisador Sintático-Semântico somente "encherga" ou reconhece do texto de

uma consulta palavras que indicam alguma subestrutura significativa, como por exemplo "artigos", que indica uma subestrutura do tipo TIPO-INF (tipo de informação), um TIPO-SIMPL (tipo simples) como "artigos" sozinho, ou um TIPO-COMP (tipo composto) com "artigos-de-periódico". O Analisador Sintático-Semântico trata o texto de uma consulta da esquerda para a direita, palavra por palavra e procede uma análise de baixo para cima (bottom-up). O reconhecimento de termos (palavras ou expressões) se faz através de três vocabulários e é orientado pelo reconhecimento "sintático" da estrutura da consulta. Os três vocabulários são os seguintes:

- VOCABULARIO ESPECIFICO DO DOMINIO: é o elemento de um SRI conhecido como Tesauro; no modelo de Sistema Especialista proposto para a IIn, o Tesauro constitui-se numa rede semântica que é a principal Base de Conhecimento do sistema.

  - FRAGMENTOS: decomposição do Vocabulário específico em unitermos.

- VOCABULARIO GENERICO DA APLICAÇAO: todos os terminais que compõe a parte da estrutura da consulta que consiste em ESP-INF.

- LISTA DE AUTORIDADES:

Este Procedimento será tanto mais eficiente quanto consiga instanciar (reconhecer) a expressão da consulta na estrutura básica proposta, Especificação das Caraterísticas Objetivas ligada à Especificação do Assunto por uma expressão equivalente a preposição "sobre". Esta performance está portanto condicionada à identificação da palavra ou expressão com a função sintático-semântica de PREP-INTRO, o que torna válido colocar no Dicionário do sistema o maior número de termos ou expressões equivalentes a PREP-INTRO.

Este procedimento funcionaria da seguinte maneira, tendo como exemplo a seguinte consulta:

> "Eu desejo recuperar artigos de periódico os mais recentes possíveis acerca de prevenção de acidentes em instalações nucleares".

PROCEDIMENTO (esboço):

1a. fase . Seria feita a separação de cada palavra da expressão da consulta, colocando-as em uma Tabela em que para cada palavra seria assinalada a sua posição no string da consulta e seu tamanho; simultaneamente o Analisador tentaria indentificar uma PREP-INTR.

```
TABELA:01.EU
       02.DESEJO
       03.RECUPARAR
       04.ARTIGOS
```

05.DE
06.PERIODICO
07.OS
08.MAIS
09.RECENTES
10.POSSõVEIS
-> 11.ACERCA
-> 12.DE
13.PREVEÇAO
14.DE
15.ACIDENTES
16.EM
17.INSTALAÇOES
18.NUCLEARES

2a. fase. Se PREP-INTR não fosse encontrada, o Analisador assumiria que a consulta consta somente da parte referente a ESP-ASS e passaria a procurar os Descritores no Tesauro, tendo como auxiliar o Dicionário de Palavras do Tesauro (unitermos) para facilitar a identificação se um descritor é formado por um termo simples (unitermo) ou se composto por mais de um termo; estes seriam movidos para a mesma entrada da Tabela.

3a. fase . Se PREP-INTR fosse encontrada, o Analisador instanciaria uma estrutura de quadro para a consulta como a da fig.04 e tentaria preencher seus atributos; as palavras desconhecidas até PREP-INTR seriam desprezadas (como EU, DESEJO, RECUPERAR, OS, MAIS, POSSIVEIS); o Analisador iniciaria uma análise "bottom-up" da parte da consulta referente a ESP-INF, usando o Vocabulário Específico. Além de tentar ir preenchendo os atributos da estrutura da consulta, o Analisador teria também um mecanismo de "back-tracking" para permitir o retrocesso e nova tentativa de instanciar a estrutura da consulta com outras palavras do Vocabulário Específico, à medida que houvesse uma "falha" na identificação de algum componente; os terminais da Gramática formadas por mais de um termo no Vocabulário Específico, teriam seus termos componentes agregados na mesma entrada da Tabela. Ex:

- despreza EU
- despreza DESEJO
- despreza RECUPERAR
- identifica o termo ARTIGOS e o instancia como TIPO-INF
- identifica o termo DE e o instancia como o inicio de uma expressão de um PERIODO; pela Gramática, um PERIODO inicia-se pela preposição DE seguida de uma DATA ou um intervalo (DATA1 a DATA2)
- falha ao tentar instanciar o termo PERIODICO como uma DATA
- retrocede (back-tracking) até o ponto onde haviam dois caminhos a seguir, que é o ponto onde foi identificado o termo ARTIGOS como TIPO-INF (o termos ARTIGOS no Vocabulário Especifico corresponde a mais de uma entrada, ARTIGOS simplesmente e ARTIGOS DE PERIODICO)
- identifica o termo ARTIGOS e o instancia como o início de uma expressão de um TIPO-INF

- identifica o termo DE e o instancia como parte de uma expressão de um TIPO-INF
- identifica o termo PERIODICO e o instancia como parte de uma expressão de um TIPO-INF, completando a identificação do TIPO-INF "artigos de periódico"
- identifica o termo RECENTES e o instancia como QUALIFICADOR de um PERIODO.

```
TABELA: 04.ARTIGOS-DE-PERIODICO
        09.RECENTES
    ->  11.ACERCA-DE
        13.PREVEÇAO
        14.DE
        15.ACIDENTES
        16.EM
        17.INSTALAÇOES
        18.NUCLEARES
```

4a. fase . Ao terminar a análise da parte da expressão de consulta referente a ESP-INF, o Analisador passaria a analisar a parte da consulta referente a ESP-ASS, conforme especificado no passo 2.

```
TABELA: 04.ARTIGOS-DE-PERIODICO
        09.RECENTES
    ->  11.ACERCA-DE
        13.PREVEÇAO
        14.DE
        15.ACIDENTES
        16.EM
        17.INSTALAÇOES-NUCLEARES
```

5a. fase. A estrutura da ESP-ASS passa então pelo crivo de Regras de Compreensão, com o objetivo de resolver omissões de termos e mesmo identificar Descritores que a primeira vista, através dos procedimentos de busca ao Tesauro e/ou dicionário de Fragmentos, não foram identificados. O fenômeno da omissão de termos ocorre, por exemplo, no seguinte trecho de uma expressão de consulta:

"...produção e lavra de urânio..."

Na verdade, nesse trecho, os Descritores seriam: PRODUÇAO DE URANIO e LAVRA DE URANIO. A tabela com as palavras da consulta conteria PRODUÇAO, E, LAVRA-DE-URANIO. Eventualmente a palavra PRODUÇAO poderia mesmo ser uma palavra desconhecida; mas na configuração dada, o reconhecimento da omissão de termos poderia se dar pela aplicação da seguinte regra:

R1 . SE, na ESP-ASS, existem dois tokens, token1 e token2, unidos pelo conectivo E, a primeiro deles é desconhecido ou só existe no dicionário de fragmentos e se o outro token é parte de um Descritor composto, obedecendo a seguinte configuração: <token2> DE <token3> e se existe o Descritor formado por <token1> DE <token3>

ENTAO, instanciar a entrada da Tabela correspondente a palavra desconhecida com o Descritor composto TOKEN1-DE-TOKEN3.

Ou, a regra contrária:

R2 . SE um token identificado pelo Analisador sintático/semântico é desconhecido mas se constitui em um fragmento de um ou mais termos (descritores do Tesauro ou Autoridades)
ENTAO, mostrar ao usuário este(s) termo(s) e solicitar ao usuário que escolha um ou mais para a substituição na expressão de busca.

ex: ...metodologia de PROJETO de um sistema...

PROJETO não existe no Tesauro, mas existem:

ENGENHARIA DE PROJETO e

AMBIENTE DE SUPORTE A PROJETOS

O usuário poderia escolher um dois descritores, ambos ou nenhum em substituição ao termo PROJETO (ver ex.2)

As Regras de Compreensão muitas vezes implicam em uma interação com o usuário de modo a esclarecer a compreensão de algum termo ou solicitar a substituição de outros que sejam desconhecidos pelo sistema.

. outras REGRAS DE COMPREENSAO:

R3 - SE um descritor é desconhecido a menos de um dos seus fragmentos
ENTAO mostrar ao usuário os descritores onde apareça o fragmento conhecido (na mesma área ou classificação que o descritor formado pelos fragmentos conhecidos) e solicitar a sua substituição na expressão de busca.

ex: BANCOS DE DADOS INTELIGENTES (INTELIGENTES é o fragmento desconhecido).

BANCOS DE DADOS -> conhecido, área "X".

mostrar:

* SISTEMAS INTELIGENTES
PROGRAMAS INTELIGENTES
INTERFACES INTELIGENTES -> todos conhecidos, área "X".

expressão reformulada:

...BANCOS DE DADOS E SISTEMAS INTELIGENTES...

R4 - SE um descritor é desconhecido a menos de algum(ns) de seu(s) fragmento(s)

ENTAO mostrar ao usuário os TEs dos descritor formado pelo(s) fragmento(s) conhecido(s) e solicitar se o usuário quer substituí-lo por um dos TEs.

ex: BANCOS DE DADOS INTELIGENTES (INTELIGENTES é o fragmento desconhecido).

BANCOS DE DADOS -> conhecido.

TE: BANCOS DE DADOS TEMPORAIS
BANCOS DE DADOS ORIENTADOS A OBJETO
* BANCOS DE DADOS DEDUTIVOS

expressão reformulada:

...BANCOS DE DADOS DEDUTIVOS...

No exemplo citado, a Tabela ficaria com as seguintes entradas: PRODUÇAO-DE-URANIO, E, LAVRA-DE-URANIO.

6a. fase. Uma estrutura como a da FIG.06 com seus atributos instanciados seria a estrutura final, entregue então ao módulo MAPEADOR, que a converteria em uma expressão Booleana.

A Gramática que suportaria esta segunda opção de análise é apresentada a seguir. Note-se o enfoque "botton-up" considerado para a estratégia de análise de uma consulta, exatamente o oposto da proposta de uma gramática ATN:

Termo do VOCABULARIO ESPECIFICO DO DOMINIO ou da LISTA

DE AUTORIDADES -> TERMO.

ex: Radiação, Reatores PWR, AIEA, etc.

Menos, Exceto, Estritamente, Especificamente, Principalmente -> ADJETIVADOR.

de (do, da, dos, das), para, por, sobre, em (no, na, nos, nas), com, e, assim como, como também -> COORDENADOR.

(ADJETIVADOR), TERMO -> TERMO-QUALIFICADO.

sobre, acerca (de), que digam respeito a, a respeito de, atinentes, versando sobre, que mencionem -> PREP-INTRODUTORIA.

PREP-INTRODUTORIA, TERMO-QUALIFICADO < (COORDENADOR), TERMO-QUALIFICADO> -> ESP-ASSUNTO.

de DATA1, de DATA1 a DATA2, entre DATA1 e DATA2 -> INTERVALO.

em diante -> DURAÇAO.

anteriores, após, posteriores -> MODIFICADOR.

(MODIFICADOR), DATA, (DURAÇAO)/INTERVALO -> PERIODO.

artigos de periódicos, anais de conferências -> TIPO-COMPLEXO.

referências, artigos, patentes, informações -> TIPO-SIMPLES.

TIPO-SIMPLES/TIPO-COMPLEXO -> TIPO-INFORMAÇAO.

recentes, atuais, novos -> QUALIFICADOR.

tudo, todas, qualquer -> QUANTIFICADOR-TOTALIZADOR.

algumas, alguns -> QUANTIFICADOR-RESTRITOR.

(QUANTIFICADOR-RESTRITOR), (TIPO-INFORMAÇAO),(QUALIFICADOR), (IDIOMA), (PERIODO), (LUGAR-GEO) -> ESPECIFICAÇAO-INFO.

QUANTIFICADOR-UNIVERSAL, (TIPO-INFORMAÇAO) -> ESPECIFICAÇAO-INFO.

(ESPECIFICAÇAO-INFO), (ESPECIFICAÇAO-ASSUNTO) -> CONSULTA .

obs1: os parêntese indicam elementos opcionais (não-obrigatórios) na formação de um elemento mais complexo; a barra inclinada (/) indica elementos excludentes (alternativos - "ou exclusivo") na formação de um elemento mais complexo.

obs2: Na dicionarização dos terminais que compõe ESPECIFICAÇAO-INFO, cada entrada (terminal) aponta para o atributo que pode preencher, de modo a facilitar a análise "botton-up".

ex: artigos-de-periódico -> TIPO-INFORMAÇAO.

Um Analisador Sintático/Semântico como o descrito poderia analisar consultas como as seguintes:

- "Para mim seria muito útil obter artigos de periódicos recentes sobre prevenção de acidentes em instalações nucleares".

- "Necessito de artigos de periódicos bem recentes que digam respeito a prevenção de acidentes em instalações nucleares.

- "Artigos de periódicos recentes acerca do tema prevenção de acidentes em instalações nucleares".

- "Quero tudo sobre o acidente de Chernobyl".

4.3.5 Regras de Mapeamento.

Trata-se agora de mapear a Representação semântica na Representação alvo, que será finalmente submetida ao SRI convencional. Uma vez obtida a estruturada instanciada com a da FIG.06, esta estrutura teria que ser convertida na representação-alvo desejada, a linguagem de interação com o sistema, no nosso

caso a linguagem LINCE; isto é feito através das Regras de Mapeamento.

As Regras de Mapeamento se baseiam na compreensão semântica do papel de fragmentos lingüísticos como conjunções, preposições, etc. como coordenadores de termos ou conceitos, formando conceitos semanticamente diferentes dos originais. Esta noção, de coordenação entre conceitos, já é bastante conhecida da Ciência da Informação. Na gramática proposta, esses fragmentos pertencem à categoria sintático-semântica denominada COORDENADORES (COORD); pertencem a essa categoria preposições como "de", "para", "por", "sobre", "em" (no, na, nos, nas), "com"; além dessas preposições a categoria COORD contém também a conjunção coordenativa "e". Em CUNHA (14) destaca-se a função das preposições como "vocábulos gramaticais invariáveis que relacionam dois termos de uma oração, de modo que o sentido do primeiro (antecedente) é explicado ou complementado pelo sentido do segundo (conseqüente)". Diante do exposto parece clara a opção de traduzir as preposições pelo conectivo booleano E; a única exceção que implica em alguma ambiguidade de interpretação é justamente a conjunção coordenativa "e", que será objeto de regras específicas.

.(descritor) EM (descritor) -> descritor E descritor

ex: ...acidentes em reatores nucleares ... -> acidentes E reatores nucleares (RS 278)

.(descritor) DE (descritor) -> descritor E descritor

ex: ...simulação de pressurizadores de centrais nucleares... -> simulação E pressurização E centrais nucleares (RS 240)

.(descritor) POR (descritor) -> descritor E descritor

ex: ...análise química de impurezas em compostos de urânio por espectrometria de emissão atômica... -> compostos de urânio E espectrometria atômica (RS 159 - 09/03/89)

.(descritor) SOBRE (descritor) -> descritor E descritor

ex: ...influência da temperatura sobre propagação de trincas em metais... -> efeitos da temperatura E propagação de trincas (RS 200/86)

.(descritor) COM (descritor) -> descritor E descritor

ex:...simulação do controle de robôs hidráulicos com servoatuadores lineares... -> robôs E servomecanismos (RS 402)

.(descritor) NÃO (descritor) -> descritor E-NÃO descritor

ex: ...plantas PWR para pequenos e grandes transientes; não desejo informações de acidentes ou LOCA... -> reatores PWR E análise de transientes E-NAO (acidentes OU loca) (RS 321/89)

Conjunção coordenativa "e" - neste caso, o tipo de coordenação entre os descritores unidos pela conjunção coordenativa "e" é ambíguo; existem algumas heurísticas propostas na literatura para se levantar essa ambiguidade - ver DAS-GUPTA, baseadas na proximidade semântica entre os termos coordenados, que no ambiente do presente trabalho pode ser obtida do tesauro. Por exemplo, entre descritores do tesauro que guardam entre si uma relação hierárquica (TG, TE) ou são relacionados (TR) pode-se afirmar que são semanticamente relacionados ou guardam uma proximidade semântica.

a- Um descritor como antecedente e um descritor como conseqüente:

caso1: os descritores possuem proximidade semântica:

(descritor) e (descritor) -> descritor OU descritor.

ex:...uso de microcomputadores e minicomputadores em bibliotecas... -> (microcomputadores OU minicomputadores) E bibliotecas. Neste caso os dois descritores tem um termo genérico em comum, "computadores".

caso 2. os descritores não possuem proximidade semântica:

(descritor) e (descritor) -> descritor E descritor

ex: influência do fumo e stress como indutores do câncer -> fumo E stress E câncer. Neste caso os descritores coordenados pelo "e" não pertencem a mesma hierarquia do Tesauro nem são termos relacionados.

(descritor) E (descritor) -> descritor OU descritor

ex:...análise e síntese de sistemas de controle automático de reatores nucleares...-> (análise OU síntese) E reatores nucleares (RS 93/89)

b- Mais de um descritor como antecedente e um descritor como conseqüente:

caso 1. os descritores possuem alguma identidade semântica:

(descritor), (descritor) E (descritor) -> descritor OU descritor OU descritor

ex: ...normas e processos para demonstração de impacto ambiental... -> (standardization OU regulations OU regulatory guides) E environment impact (RS 325/86)

obs: descritores coordenados pelo operador OU formam um único conceito quando coordenados com um único descritor (conceito) através de um operador E;

ex: ...produtos agrícolas (frutas, legumes, cereais, etc.) através de irradiação com raios gama... -> (food OU vegetables OU food processing OU cereals OU fruits) E gamma radiation (RS 414/86)

c- Descritores separados por vírgulas -> caso eles tenham identidade semântica o conectivo deve ser "OU"; caso contrário, o conectivo deve ser "E".

d- Elementos que estão preenchendo atributos: ESP-INF -> etiqueta do atributo mais o conectivo "E".

atributo "autor" preenchido com o valor "Vickery":

AU=Vickery E...

atributo "data de publicação" preenchido com "1988":

DT=1988 E...

atributo MODIFICADOR, alterando o sentido de um PERIODO: se o MODIFICADOR é "recentes" ou "novos" então a identificação do campo ANO vem preenchida com "> ano atual menos dois anos"; se o MODIFICADOR é "antigos" então a etiqueta ANO vem preenchida com "< ano atual menos dez anos"; ex: atributo MODIFICADO preenchido com "recentes":

AN>1988 E...

e- Descritor ou autoridade antecedido de um ADJETIVADOR: ADJETIVADOR descritor -> E-NÃO descritor

ex: ... exceto reatores PWR -> E-NÃO reatores PWR.

Na elaboração do modelo proposto foram selecionados cerca de 30 (trinta) formulários de consulta representativos das inúmeras variações da estrutura proposta. Após essa fase foram então selecionados a esmo cerca de outros 30 (trinta) formulários de consulta, de modo a testar o modelo; os resultados deste teste estão apurados no capítulo 5.

4.4 INTERAÇÃO DO MODULO ESPECIALISTA EM BUSCA COM O SRI

A busca de informações por um usuário em um sistema de Recuperação de Informações Bibliográficas é essencialmente uma atividade interativa. Sendo assim, a avaliação dos resultados de uma busca, a reformulação da estratégia de busca e sua posterior submissão ao sistema são uma prática comum. A IIn deve portanto incorporar estas características.

Ao submeter uma consulta ao SRI, os resultados podem se classificar em três tipos, quanto ao número de documentos recuperados:

. Nenhum documento foi recuperado.
. Alguns documentos foram recuperados.
. Um número excessivo de documentos foi recuperado, ou seja um número de documentos intratável pelo usuário (o usuário é incapaz de fazer "browse" devido ao grande número, ou o usuário não se interessa por imprimir as referências examiná-las).

4.4.1 Heurísticas gerais de avaliação dos resultados de uma busca.

Parâmetros de avaliação:

. número de documentos que o usuário deseja.
. se o usuário está interessado em Revocação ou Precisão.
. número de documentos recuperado.

Feedback: alterando REVOCAÇAO e PRECISAO de uma consulta através da alteração da EXAUSTIVIDADE e da PROFUNDIDADE.

As principais medidas da performance de um SRI são a Revocação e a Precisão. Os principais fatores que afetam estas duas medidas são a Exaustividade de uma consulta e a Especificidade da mesma. Para fins deste trabalho trabalharemos com as seguintes definições desses dois conceitos:

Exaustividade: é uma medida do grau em que diferentes assuntos discutidos em um particular documento são reconhecidos/traduzidos pela linguagem de indexação - LANCASTER; em termos práticos, para a IIn, consistirá no número de conceitos diferentes, materializados sob a forma de descritores únicos ou grupos de descritores unidos pelo conectivo OU, extraídos de um tesauro, estão contidos na formulação da consulta.

Especificidade: é a capacidade da linguagem de indexação em descrever um tópico precisamente - LANCASTER (9); em termos da IIn será a profundidade de cada conceito contido na formulação da consulta, em relação ao Top-term da hierarquia a que ele pertence.

"...pode-se dizer que a exaustividade da linguagem de indexação controla a capacidade de revocação de um sistema, enquanto a especificidade da linguagem de indexação controla a capacidade de precisão." - LANCASTER (9).

Revocação e Precisão de uma consulta variam de acordo com os fatores de Exaustividade e Especificidade segundo a FIG.07.

4.4.2 Regras de reformulação.

Em função de uma avaliação do usuário, a IIn poderia alterar a formulação da consulta no sentido de obter maior Revocação ou maior Precisão. Um exemplo de como o conhecimento do especialista

em informação, que o sistema pretende simular, poderia ser representado, está nas técnicas de variação da Revocação e da Precisão de uma consulta em função da alteração, pelo especialista em informação, de parâmetros como a Exaustividade e da Especificidade dos termos componentes da expressão booleana da consulta, da seguinte forma:

Se o usuário deseja maior Revocação, então trabalhar (acrescentar) com mais conceitos.

Se o usuário deseja maior Precisão, então trabalhar com menos conceitos.

Se o usuário deseja maior Revocação, então trabalhar com conceitos mais gerais.

Se o usuário deseja maior Precisão, então trabalhar com conceitos mais específicos.

Essas heurísticas e outras mais sugerem como forma de representação do conhecimento para este componente da IIn, as Regras de Produção, com algumas características específicas, como o fato de que o Conseqüente de uma Regra de Produção poder ser a identificação de um Procedimento a ser evocado.

São amplamente citadas na literatura exemplos de procedimentos que serviram de base para a elaboração das regras citadas e que foram durante muito tempo investigadas na literatura de recuperação automática de informações sob o nome de "automatic feedback"; essas técnicas se baseiam no uso dos próprios documentos recuperados ou de um documento que o usuário reconheça previamente como relevante, juntamente com a avaliação do usuário, para reformular a expressão de busca original. Entre os trabalhos nessa área podem ser citados os de BARKER, de DILLON, de GRZELAK e de VERNIMB. O artigo de BATES sobre táticas para a recuperação de informações em sistemas automatizados, também cita várias heurísticas para a reformulação de expressões de busca.

R5 - SE a busca não recuperou nenhum documento
ENTAO ir reformulando a expressão de busca, substituindo conceitos representados por um único termo por termos mais genéricos, mudar STATUS para "reformular" e ESTRATEGIA para "R5".

R5.1-SE STATUS = "reformular" e ESTRATEGIA = "R5"
ENTAO subir uma grau na hierarquia do tesauro do próximo conceito representado por um único termo e STATUS = "submeter consulta".

R6 - SE a busca não recuperou nenhum documento
ENTAO ir reformulando a expressão de busca, substituindo conceitos representados por um único termo por um conjunto de seus termos relacionados unidos pelo conectivo "OU", mudar STATUS para "reformular" e ESTRATEGIA para "R6".

R6.1-SE STATUS = "reformular" e ESTRATEGIA = "R6" e a busca não recuperou nenhum documento
ENTAO substituir o próximo conceito representado por um único termo pelo termo genérico seguinte na sua hierarquia e STATUS = "submeter consulta".

R6.2-SE STATUS = "reformular" E estratégia = "R6" e a busca não recuperou nenhum documento e a consulta já foi reformulada pela regra R6.1
ENTAO ir excluindo conceitos representados por um único termo da expressão de busca e STATUS = "submeter consulta".

R7 - SE a busca recuperou mais documentos que o especificado pelo usuário
ENTAO ir reformulando a expressão de busca, substituindo conceitos representados por um único termo por termos mais especificos, mudar STATUS para "reformular" e ESTRATEGIA para "R7".

R7.1-SE STATUS = "reformular" e ESTRATEGIA = "R7"
ENTAO descer uma grau na hierarquia do tesauro do próximo conceito representado por um único termo e STATUS = "submeter consulta".

R8 - SE a busca recuperou mais documentos que o especificado pelo usuário
ENTAO solicitar ao usuário algum outro termo/faceta.

R9 - SE o usuário avaliou as referências como muito genéricas
ENTAO apresentar ao usuário uma lista de descritores presentes nas referências julgadas relevantes mas ausentes da expressão de busca e interrogar o usuário sobre quais deles seriam relevantes para serem incluídos na expressão de busca.

R10 - SE STATUS = "avaliar pelo usuário"
ENTAO mostrar os documentos ao usuário e solicitar sua avaliação.

AVALIAÇAO:

. avaliação por documento: INTERESSA/NAO INTERESSA.

. avaliação de um conjunto de documentos: MUITO GENERICO/SATISFATORIO/MUITO ESPECIFICO.

R11 - SE a busca não recuperou nenhum documento e a expressão de busca contém um mais conceitos representados por um único descritor e unidos pelos conectivos "E" ou "E-NAO"
ENTAO vá truncando cada um desses descritores

ex: COMPUTADORES E ...

expressão reformulada:

...COMPUT$ E ...

R12- SE a busca é julgada satisfatória mas o usuário está interessado em Revocação e o número de documentos recuperados é pequeno
ENTÃO reformular a expressão de busca, substituindo um a um cada conceito representado por um único descritor por um conjunto de seus RTs unidos pelo conectivo OU

ex: SISTEMAS ESPECIALISTAS E BIBLIOTECAS

expressão reformulada:

...(SISTEMAS ESPECIALISTAS) E BIBLIOTECAS...
OU
PROCESSAMENTO DE LINGUAGEM NATURAL
OU
ENGENHARIA DO CONHECIMENTO
OU
INTELIGENCIA ARTIFICIAL)

R13 - SE STATUS = "submeter consulta"
ENTÃO submeter a consulta atual ao SRI e STATUS = "avaliar pela IIn".

4.5 PROBLEMAS CONSTRUTIVOS

Os sistemas especialistas descritos na literatura comportam-se como sistemas de inferência e recuperação de conhecimentos armazenados em uma Base de Conhecimentos. Sua saída portanto é conhecimento, sob a forma de um diagnóstico, um aconselhamento, uma resposta, recuperado ou inferido da Base de Conhecimentos. Assim são os Sistemas Especialistas clássicos como o Mycyn, dedicado ao diagnóstico de doenças infecciosas, ou o Prospector, cuja a finalidade é analisar perfis geológicos para descobrir depósitos de minerais. A entrada para o Mycyn por exemplo, são sintomas e a saída é um diagnóstico; estes são sistemas do tipo "stand-alone". A busca pela solução de um problema neste sistemas se dá através da caracterização de um ou mais estados iniciais, ou pontos de partida, e um ou mais estados finais (fatos), a que se aplicam as operações (geralmente regras que cuja aplicação cause uma mudança de estado, ou seja, cujo antecedente seja um estado e o conseqüente um novo estado); busca-se atravessar um grafo que ligue o(s) estado(s) inicial(ais) ao(s) estado(s) final(ais); pode-se aplicar as regras tanto do(s) estado(s) inicial(ais) para o(s) estado(s) final(ais) (raciocínio para frente) como ao contrário (raciocínio para trás); existe sempre um mecanismo de controle para este processo de aplicação de regras e verificação se uma meta (estado inicial ou estado final) foi atingida.

A IIn pretendida neste trabalho é de natureza diferente; não se trata simplesmente de fornecer as condições iniciais de um

problema e buscar uma solução. Sua saída não pode ser um diagnóstico sobre a situação de uma interação usuário/SRI, ou um aconselhamento sobre que ação tomar se o usuário não está satisfeito com os documentos recuperados; também não é possível que a IIn "reconheça" se atingiu ou não a solução do seu "problema", pois isso é em última instância uma atribuição do usuário; o que se pretende é que a IIn exerça uma ação direta sobre o SRI convencional, não somente monitorando a interação de um usuário casual com o SRI, mas também intermediando esta interação, agindo/comandando o SRI de modo a obter um resultado ótimo para o usuário.

A IIn não somente dirá o que deve ser feito (o que implica no tipo de conhecimento declarativo característico dos SEs convencionais) mas efetivamente agirá sobre os parâmetros da busca e comandará o SRI. A natureza do conhecimento requerido pela IIn de modo a conseguir este efeito será não somente de natureza declarativa, mas também de natureza procedimental. Este último fato terá conseqüências profundas na arquitetura da IIn. Vamos exemplificar a diferença entre os dois tipos de conhecimentos:

- Se temperatura acima de 38o e
  inflamação das vias respiratórias
  então
  diagnóstico é resfriado.

- Se numero de documentos recuperados = 0
  então
  diminuir a especificidade dos termos da expressão
  de busca e re-submete-la ao SRI.

A conseqüência da regra de produção é uma ação ou seqüência de ações sobre o SRI e os parâmetros da busca, e não um fato, ou diagnóstico. A avaliação de uma regra ou seqüência de regras tem que ser seguida de uma ação, e a IIn tem que poder avaliar os resultados dessas ações sobre os resultados apresentados pelo SRI. Isso sugere uma arquitetura em que o motor de inferência (avaliador de regras) esteja acoplado ao módulo executor de ações sobre o SRI.

A IIn é simplesmente uma interface a um SRI convencional; pressupOe-se que esse SRI seja um sistema basedo numa estrutura de arquivos invertidos, capaz de responder a consultas formuladas em álgebra booleana, tendo todas as potencialidades para responder a consultas formuladas segundo a LINCE; existirá também, no âmbito do SRI, um Tesauro, armazenando a terminologia da área específica a ser consultada e as relações normalmente existentes em um Tesauro.

Nesse contexto, a IIn consitiria de um módulo especialista como coordenador, avaliando metaregras (estratégias) e supervisionando outros especialistas mais específicos (especialista lingüístico, por exemplo) e módulos executores. Os resultados da avaliação de qualquer módulo especialista

subordinado ou módulo executor e mesmo do próprio modulo coordenador seriam registrados numa estrutura comum, visível por todos eles, conhecida na arquitetura de sistemas especialistas como "Quadro Negro". Um esboço da arquitetura proposta está na FIG.08.

No Quadro Negro estariam parâmetros como:

. STATUS - a tarefa imediata a realizar.

. HISTORICO DAS REGRAS AVALIADAS.

. ESTRATEGIA EM ANDAMENTO (METAREGRA) - seqüência de tarefas a realizar visando um objetivo.

. MODELO DO USUARIO - sua experiência, seu interesse em REVOCAÇAO ou PRECISÃO, o número máximo de documentos que está interessado em examinar.

. CONSULTA ORIGINAL.

. NÚMERO DE DOCUMENTOS RECUPERADOS PELA ULTIMA CONSULTA.

. AVALIAÇAO DO USUARIO EM RELAÇAO A ULTIMA CONSULTA E EVENTUALMENTE A ALGUNS OU TODOS OS DOCUMENTOS DESSA CONSULTA.

Em síntese, pode-se resumir os passos metodológicos seguidos para a elaboração do modelo proposto para a IIn e de suas funcionalidades, no seguinte:

- escolha dos parâmetros que irão compor o modelo do usuário: os parâmetros estabelecidos como operacionais e passíveis de serem manipulados pela IIn com a finalidade de graduar o seu funcionamento foram: interesse do usuário por revocação ou precisão e número máximo de documentos que o usuário gostaria de examinar. Estes parâmetros seriam obtidos interrogando o usuário, através de um diálogo/entrevista estruturado, e de expressões características (estruturas sintático-semânticas) obtidas da própria consulta formulada pelo usuário em LN.

- elaboração do formalismo lingüístico: para isto, em primeiro lugar foi examinado um determinado número de consultas em linguagem natural (sessenta consultas); foi efetuada uma análise sintática, de acordo com as categorias do português, de modo a reconhecer os componentes sintáticos básicos de uma consulta formulada em linguagem natural; analisou-se primeiramente quais destes componentes sintáticos seriam relevantes para a formulação da expressão booleana, obtendo-se assim uma estrutura sintático-semântica inicial para uma consulta em linguagem natural, a exemplo de formalismos lingüísticos utilizados em sistemas similares (gramática semantica); foi realizada uma categorização tentativa inicial dos diferentes tipos de consulta, de acordo a existência ou não de padrões gramaticais e de componentes

relevantes para a formulação da expressão booleana; estes componentes foram estabelecidos então como categorias gramaticais e baseado neles foi então proposta uma gramática sintático-semântica, capaz de analisar as consultas em linguagem natural.

Esta gramática está orientada para categorias sintático-semânticas estabelecidas em função de sua relevância para a formulação final da expressão booleana. Em função dessas categorias sintático-semânticas, reclassificou-se definitivamente as consultas da massa de teste.

Se o formalismo proposto não reconhecer uma consulta, a IIn apelará para um formalismo lingüístico mais tosco, a técnica "string-search/busca ao dicionário".

Com este formalismo procurou-se obter uma expressão booleana inicial tão boa quanto possível, que servisse de ponto de partida inicial para a interação com o SRI; o passo lingüístico é de importância fundamental para o desempenho da IIn, uma vez que, partindo de uma expressão booleana o mais próxima possível das intenções do usuário, formuladas através de sua consulta em linguagem natural, a etapa subseqüente, a interação usuário/SRI intermediada pela IIn, já pode contar com o "feed-back" da própria base de conhecimentos especializado em buscas retrospectivas que faz parte da IIn, como também do próprio "feedback" do usuário ao examinar os resultados iniciais obtidos.

- interação com o SRI a partir da expressão booleana inicial obtida no passo anterior; para isso, reformular a expressão booleana corrente em função dos parâmetros estabelecidos no modelo do usuário e da própria avaliação dos documentos recuperados feita pelo usuário; para esta reformulação, é utilizada a base de regras contendo conhecimento especialista em buscas retrospectivas.

- finalmente, avaliação do modelo proposto, sobre uma massa de teste de trinta consultas.

Notas e Referências:

1. FLUHR, C. Information Retrieval. In: CONSULTIVE MEETING of INIS LIAISION OFFICERS, 17, Viena, May 1989.

2. DANIELS, P. J. **Cognitive Models in Information Retrieval - An Evaluative Review.** Journal of Documentation, v. 42, n. 4, p.272-304, December, 1986.

3. Posteriormente, o potencial do formalismo lingüístico poderia ser utilizado reformular estes parâmetros, como, por exemplo, na seguinte situação: "quero saber tudo sobre o acidente de Chernobyl"; o quantificador "tudo" parece indicar claramente, no caso desta consulta, um interesse marcadamente por revocação da parte do usuário.

4. CUNHA, I. M. R. **Estruturação de Vocabulário.** In: Análise Documentária: a Análise da Síntese. Brasília, Instituto Brasileiro de Informação em Ciência e Tecnologia - IBICT, p.62-85, 1987.

5. DILLON, M., DESPER, J. **The Use of Automatic Relevance Feedback in Boolean Retrieval Systems.** Journal of Documentation, v. 36, n. 3, p. 197-208, September 1980.

6. VERNIMB, C. **Automatic Query Adjustment in Document Retrieval.** Information Processing & Management, v. 13, p.339-53, 1977.

7. CUNHA, C. F. da. **Gramática de Base.** Rio de Janeiro, Ministério da Educação e Cultura - Fundação Nacional de Material Escolar, 1979.

8. DAS-GUPTA, P. **Boolean Interpretation of Conjuctions for Document Retrieval.** JASIS, v. 38, n. 4, p.245-54, 1987.

9 LANCASTER, F. W. **Information Retrieval Systems.** John Wiley & Sons, New York, 222p, 1968.

10. BARKER, F. H., VEAL, D. C., WYATT, B. K. **Towards Automatic Profile Construction. Journal of Documentation,** v. 28, n. 1, March 1972.

11. GRZELAK, HANNA, KOWASKY, K. **Automatic Construction of Information Queries.** Information Processing & Management, v. 19, n. 6, p.381-9, 1983.

12. BATES, M. J. **Information Search Tatics.** JASIS, v. 4, n. 30, p.205-14, 1979.

5 . AVALIAÇAO E DISCUSSAO DO MODELO PROPOSTO

Do modelo proposto é apresentado a seguir uma avaliação somente do formalismo lingüístico, devido a facilidade de simular a IIn. Como resultado espera-se que o formalismo lingüístico obtenha, a partir das expressões de consulta formuladas em linguagem natural, uma expressão booleana inicial razoavelmente coincidente com as intenções do usuário, de modo a servir de ponto de partida para uma interação usuário/SRI. Para avaliação do modelo proposto foram examinadas ao acaso cerca de 30 (trinta) formulários de consulta, dos anos 1986 à 1989, abrangendo diferentes bases de dados. Os resultados obtidos foram os seguintes:

- em 16 (dezesseis) consultas o modelo conseguiu reconhecer a estrutura proposta e com ela guiar a interpretação da consulta, obtendo uma expressão booleana próxima ou eventualmente mais rica que a obtida pelo especialista humano; nesses casos a expressão booleana obtida fornece um ponto de partida seguro para iniciar a interação com o SRI, resultado que portanto foi considerado satisfatório;

- em 12 (doze) consultas o modelo não conseguiu reconhecer a estrutura proposta e analisou a expressão de consulta como se ela constasse somente da ESPECIFICAÇAO DO ASSUNTO, utilizando a técnica de "string-search" e busca ao tesauro da base, o que pode ser considerado lingüisticamente sofrível;

- em 2 (duas) consultas o modelo falhou em reconhecer a estrutura proposta e ao analisar a expressão de consulta segundo a técnica de "string-search"/busca no tesauro, obteve interpretações incorretas das intenções do usuário.

A seguir, seguem-se alguns exemplos de consultas processadas pelos formalismos lingüísticos sugeridos:

ex.1 - "O objetivo da busca é fazer um trabalho sobre metodologias de projeto de um sistema de controle digital enfocando as etapas do projeto que são necessárias e os documentos de projeto que devem ser gerados em cada etapa" (RS25/89 base INSPEC).

```
1a. fase - TABELA: 01.O
                   02.OBJETIVO
                   03.DA
                   04.BUSCA
                   05.E
                   06.FAZER
                   07.UM
                   08.TRABALHO
                ->  09.SOBRE
                   10.METODOLOGIAS
                   11.DE
                   12.PROJETO
```

13.DE
14.CONTROLE
15.DIGITAL
16.ENFOCANDO
17.AS
18.ETAPAS
19.DE
20.PROJETO
21.QUE
22.SAO
23.NECESSORIAS
24.E
25.OS
26.DOCUMENTOS
27.DE
28.PROJETO
29.QUE
30.DEVEM
31.SER
32.GERADOS
33.EM
34.CADA
35.ETAPA.

4a.fase - TABELA: 10.METODOLOGIAS (desconhecido - despreza)
11.DE
12.PROJETO (fragmento conhecido - usar R2)
13.DE
14.CONTROLE-DIGITAL
16.ENFOCANDO (desconhecido - despreza)
17.AS (desconhecido - despreza)
18.ETAPAS (desconhecido - despreza)
19.DE
20.PROJETO (fragmento conhecido - usar R2)
21.QUE (desconhecido - despreza)
22.SAO (desconhecido - despreza)
23.NECESSORIAS (desconhecido - despreza)
24.E
25.OS (desconhecido - despreza)
26.DOCUMENTOS (use DOCUMENTACAO-DE-SISTEMAS)
27.DE
28.PROJETO
29.QUE (desconhecido - despreza)
30.DEVEM (desconhecido - despreza)
31.SER (desconhecido - despreza)
32.GERADOS (desconhecido - despreza)
33.EM
34.CADA (desconhecido - despreza)
35.ETAPA. (desconhecido - despreza)

5a. fase - TABELA:
11.DE
12.ENGENHARIA-DE-PROJETO
13.DE

14.CONTROLE-DIGITAL
19.DE
24.E
26.DOCUMENTAÇAO-DE-SISTEMAS
27.DE
28.AMBIENTE-DE-SUPORTE-A-PROJETO

. expressão booleana gerada:

"ENGENHARIA-DE-PROJETOS E CONTROLE-DIGITAL E DOCUMENTAÇAO-DE-

SISTEMAS E AMBIENTE-DE-SUPORTE-A-PROJETO"

. expressão booleana formulada pelo especialista humano:

"(DIGITAL CONTROL OU DIGITAL SYSTEMS) E (HARDWARE OU DESIGN ENGENEERING)"

ex.2 - "Visa obter referências sobre o estado de desenvolvimento alcançado no projeto e análise de modelos matemáticos utilizados na simulação de pressurizadores de centrais nucleares". RS 240/86 - Base INIS.

1a. fase - TABELA: 01.VISA
02.OBTER
03.REFERENCIAS
-> 04.SOBRE
05.O
06.ESTADO
07.DE
08.DESENVOLVIMENTO
09.ALCANÇADO
10.NO
11.PROJETO
12.E
13.ANALISE
14.DE
15.MODELOS
16.MATEMATICOS
17.UTILIZADOS
18.NA
19.SIMULAÇAO
20.DE
21.PRESSURIZADORES
22.DE
23.CENTRAIS
24.NUCLEARES.

3a.fase - TABELA:
03.REFERENCIAS
-> 04.SOBRE
05.O
06.ESTADO
07.DE
08.DESENVOLVIMENTO

09.ALCANÇADO
10.NO
11.PROJETO
12.E
13.ANALISE
14.DE
15.MODELOS
16.MATEMATICOS
17.UTILIZADOS
18.NA
19.SIMULAÇAO
20.DE
21.PRESSURIZADORES
22.DE
23.CENTRAIS
24.NUCLEARES.

4a.fase - TABELA:

03.REFERONCIAS
-> 04.SOBRE
05.O (desconhecido - despreza)
06.ESTADO (desconhecido - despreza)
07.DE (desconhecido - despreza)
08.DESENVOLVIMENTO (desconhecido - despreza)
09.ALCANÇADO (desconhecido - despreza)
10.NO (desconhecido - despreza)
11.PROJETO
12.E
13.ANOLISE
14.DE
15.MODELOS
16.MATEMATICOS
17.UTILIZADOS (desconhecido - despreza)
18.NA
19.SIMULAÇAO
20.DE
21.PRESSURIZADORES
22.DE
23.CENTRAIS
24.NUCLEARES

5a. fase - TABELA:

03.REFERONCIAS
-> 04.SOBRE
11.PROJETO (Regra R2 - nenhum foi escolhido)
12.E
13.ANOLISE (Regra R2 - nenhum foi escolhido)
14.DE
15.MODELOS-MATEMOTICOS
18.NA
19.SIMULAÇAO
20.DE
21.PRESSURIZADORES

22.DE
23.CENTRAIS-NUCLEARES

. expressão booleana gerada:

"MODELOS-MATEMATICOS E SIMULAÇÃO E PRESSURIZADORES E CENTRAIS-NUCLEARES"

. expressão formulada pelo especialista humano:

"PRESSURIZES E SIMULATION E MATHEMATICAL MODELS"

ex3 - "Obter uma literatura suficiente para entender o processo de refino por zona, em tarugos de silício, relação da potência necessária com o diâmetro do tarugo, frequência, atmosfera e velocidade do processo". RS142/89 - Base METADEX

Ao não identificar um sintagma equivalente a PREP-INTRO (sobre, acerca de, etc), o procedimento limita-se a recorrer ao técnica de "string-search" das palavras do texto da consulta e pesquisa ao Tesauro da base, obtendo as seguintes palavras:

TABELA: 01.REFINO-POR-ZONA
02.SILÕCIO

. expressão booleana gerada:

"REFINO-POR-ZONA E SILICIO"

. expressão booleana formulada pelo especialista humano:

"ZONE MELTING E SYLICON"

ex4 - "Levantamento do estado atual da tecnologia disponível, equipamento, parâmetros e exigências necessárias para subsidiar projetos e execuções de instalações de aterramento de centros de processamento de dados". RS195/89 - Base INSPEC

Ao não identificar um sintagma equivalente a PREP-INTRO (sobre, acerca de, etc), o procedimento limita-se a recorrer ao técnica de "string-search" das palavras do texto da consulta e pesquisa ao Tesauro da base, obtendo as seguintes palavras:

TABELA: 01.PROJETOS
02.ATERRAMENTO
03.CENTROS-DE-PROCESSAMENTO-DE-DADOS

. expressão boolena gerada:

"PROJETO E ATERRAMENTO E CENTROS-DE-PROCESSAMENTO-DE-DADOS"

. .expressão booleana formulada pelo especialista humano:

"(EARTHING OU EARTHING SYSTEMS OU GROUNDING) E COMPUTING INSTALAÇOES"

ex5 - "Obter especificações funcionais sobre os seguintes elementos de uma rede local: servidos de arquivo, servidor de impressão, interconexões de rede, serviços e gateways, protocolos". RS342 - Base INSPEC

1a. fase - TABELA: 01.SOBRE
02.SERVIDOR
03.DE
04.ARQUIVOS
05.SERVIDOR
06.DE
07.IMPRESSAO
08.INTERCONEXOES
09.DE
10.REDES
11.SERVIÇOS
12.E
13.GATEWAYS
14.PROTOCOLOS

Ao localizar a preposição "sobre", o modelo reconhece a estrutura proposta e considera a parte da expressão de consulta até "sobre" como pertencendo a subestrutura sintático-semântica ESPECIFICAÇAO DAS CARATERISTICAS OBJETIVAS das informações, deixando de reconhecer uma faceta de grande interesse para usuário, as expressões "especificações funcionais" e "rede local".

. expressão booleana gerada:

"SERVIDOR-DE-ARQUIVOS OU SERVIDOR-DE-IMPRESSAO OU INTERCONEXAO-DE-REDES-DE-SERVIÇOS OU PROTOCOLOS OU GATEWAYS"

. expressão booleana formulada pelo especialista humano:

1 . "LOCAL-AREA-NETWORK E (FILE SERVER OU FILE SERVERS)"

2 . "expressão1 E (PERFORMANCE OU EVALUATION)"

3 . "expressão2 E (GATEWAY OU GATEWAYS)"

ex6 - " Resinas de troca iônica são bastante utilizadas no tratamento de água de reatores de potência e também na descontaminação de determinados fluxos de rejeitos radioativos líquidos e após sua utilização esses rejeitos podem ser considerados rejeito radioativo.

Uma das técnicas de gerenciamento utilizada para este tipo de rejeito é sua imobilização ou condicionamento em matrizes do tipo

cimento-polímero, comummente denominadas "composites, o qual é o principal objetivo da busca". RS 181/89 - Base INIS

Ao não identificar um sintagma equivalente a PREP-INTRO (sobre, acerca de, etc), o procedimento limita-se a recorrer ao técnica de "string-search" das palavras do texto da consulta e pesquisa ao Tesauro da base, obtendo as seguintes palavras:

TABELA: 01.RESINAS
02.IONS
03.REATORES-DE-POTENCIA
04.REJEITOS
05.COMPOSITE

. expressão booleana gerada:

"RESINAS E IONS E REATORES-DE-POTENCIA E REJEITOS E COMPOSITE"

. expressão booleana formulada pelo especialista humano:

"COMPOSITE MATERIAL E RESINS E RADIOATIVE WASTE

Nos exemplos descritos pode-se verificar que uma fase crítica para o desempenho da IIn é a obtenção da expressão booleana de consulta inicial, uma vez que daí em diante a IIn pode recorrer permanentemente interação com o usuário no sentido de obter "feedback" para seu funcionamento; daí a importância do formalismo lingüístico para o desempenho da IIn.

Pode-se avaliar também a importância de um sistema como a IIn no contexto de um sistema como o SPA (Sistema Público de Acesso), que interligará os maiores centros de acesso à informações do país e permitirá o acesso aos mesmos via RENPAC; devido a escassez e ao longo de tempo de maturação necessário para formar profissionais com experiência em acesso a bases de dados que serviam de intermediários nos SRIs tradicionais, a IIn torna-se mais um fator facilitador para o acesso e uso intensivos das informações armazenadas nestes bancos de dados.

Outro fator a ser considerado ao avaliar a IIn é o fato de o usuário poder graduar e possuir todo o controle de sua interação com o SRI via IIn; deve-se lembrar que a IIn realiza parte de seus procedimentos ainda sem conectar-se ao SRI (análise lingüística, pesquisa aos Tesauros) e que só se conecta com o SRI para o envio de uma consulta já suficientemente validada pelo usuário ou para receber seus resultados, o que é um fator redutor dos custos de comunicação. A presença do intermediário humano, que eventualmente pode ser considerado um fator inibidor para o usuário, provocando a submissão de consultas antes que elas estejam suficientemente avaliadas pelo usuário ou obrigando-o a avaliações mais longas do que seria necessário, também fica superada na interação via IIn, pelo fato do próprio usuário graduar sua interação.

Vê-se assim que um instrumento como a IIn, somando-se a outros como bases de dados disponíveis via RENPAC, Catálogos Coletivos e sistemas de comutação bibliográfica, pode ser um fator facilitador do acesso a documentos primários, de compartilhamento de recursos informacionais, enfim, um fator de otimização do uso da informação em um país como o Brasil

6 . CONCLUSOES.

O modelo que permeia um Sistema de Recuperação de Informações bibliográficas convencional consiste basicamente em cotejar uma consulta que expressa as necessidades de informação de um usuário, sob a forma de uma lista de termos, com as representações de diferentes documentos presentes numa base de dados e também representados cada um por pelo menos uma lista de termos, que expressariam seu conteúdo. Embora este modelo seja muito simplório tanto do ponto de vista dos aspectos cognitivos envolvidos no processo de busca de informações por um usuários como nos aspectos semânticos da representação do conteúdo dos documentos, é um modelo consagrado; implementações práticas deste modelo, baseados na estrutura de arquivos e listas invertidos, constituem hoje os grandes Sistemas de Recuperação de Informações em uso comercial por todo o mundo. Tal modelo remota a LUHN, em suas experiências com SRIs de fins da década de 50.

Embora ao longo destes anos tenham se avolumado críticas ao modelo e proposta várias alternativas, com diferentes suportes teóricos, como o modelo de espaços vetoriais ou o modelo probabilístico, ou modelos implementando lógicas nebulosas (fuzzy sets) (1), (2), (3), resultando mesmo em algumas experiências práticas, como os sistemas SMART e SIRE, propostos por SALTON (4), o modelo original permanece como base da maioria dos sistemas de uso corrente.

Ao longo deste trabalho pudemos perceber que as deficiências mais fortes do modelo tradicional prendem-se a dois aspectos: o modelo cognitivo envolvido nas motivações que levam um usuário a buscar um Sistema de Recuperação de Informações e a interagir com ele, e a própria pobreza da representação de documentos em um SRI.

Em relação a essas duas questões, a IIn poderia trabalhar também no pressuposto de que uma das bases para o seu comportamento inteligente repousa na premissa de que:

OBJETIVO/PROBLEMA -> NECESSIDADES DE INFORMAÇAO -> INFORMAÇAO.
(para que?) (o que?)
(natureza)
(grau de
(especificidade)

Ou seja, os objetivos de um usuário na busca de informações (ao procurar um SRI) devem implicar ou condicionar a formulação das necessidades de informação; estas devem por sua vez implicar nas informações (referências) necessárias à satisfação dessas necessidades, bem como no grau de especificidade das mesmas; essas cadeias de inferências são realizadas num sistema convencional pelo especialista/intermediário - ver (5) e (6).

A cadeia de inferências exposta acima poderia ser mais facilmente realizada por um sistema se houvesse uma maneira

uniforme de se representar a semântica tanto do OBJETIVO/PROBLEMA do usuário, das NECESSIDADES DE INFORMAÇÃO que podem contribuir para a solução do problema proposto, quanto das próprias informações armazenadas no SRI. Tentativas de se caminhar na direção de uma representação semântica uniforme para alguns ou mesmo esses três elementos são descritas em (7), em (8), em (9) e em (10), que trabalham em representações semânticas dos textos armazenados num banco de dados baseadas em lógica de primeira ordem ou em estruturas de quadros ("frames"). No entanto, provavelmente, essas experiências vão além do escopo do presente trabalho, por implicarem numa outra estruturação/representação das referências armazenadas num banco de dados, assumindo este cada vez mais o caráter de uma base de conhecimento propriamente dita, que permitisse uma recuperação "semântica" das referências ou mesmo do texto completo de documentos, o que significaria certamente a obsolescência dos bancos de dados explorados comercialmente, tipo Orbit, Dialog, BRS, Questel, etc.

Embora o presente trabalho objetive a construção de uma Interface Inteligente a Sistemas de Recuperação de Informações bibliográficas convencionais, ou seja, só trabalhar na recuperação dos dados de um sistema convencional, onde o conteúdo de uma referência bibliográfica é descrita no máximo pela atribuição de descritores à mesma como nos sistemas comerciais citados, acreditamos que a linha delineada nos trabalhos citados seja a linha de evolução dos SRIs do futuro.

Todo este esquema sugere empregar o formalismo lingüístico não somente para captar as necessidades de informação do usuário, mas sim, numa etapa anterior, na entrevista pré-busca, para processar a formulação das situação problema/objetivos do usuário na sua busca de informações, como sugere a hipótese ASK (11), (12). No entanto, embora esta hipótese pareça mais correta e consistente por levar em consideração os aspectos cognitivos levantados por BELKIN, não foi utilizada no presente trabalho porque o material disponível, os formulários de SOLICITAÇÃO DE BUSCA RETROSPECTIVA do CIN somente continham diretamente a formulação das necessidades de informação do usuário em linguagem natural.

Em relação aos objetivos iniciais pode-se considerar que o presente trabalho cobre teoricamente e propõe soluções práticas para os aspectos mais relevantes para especificação/construção de um programa que se constitua numa Interface Inteligente a SRI bibliográficos.

Outro objetivo inicial proposto, a obtenção de uma metodologia para a construção de interfaces em linguagem natural para outros sistemas de informação, foi também atingido, uma vez que a metodologia de coleta de consultas, sua classificação em padrões e a construção de uma gramática semântica para analisa-las pode ser generalizada.

Um subsídio importante, a compilação do conhecimento do especialista intermediário em buscas retrospectivas proposta, estabelece uma base inicial para uso em outros sistemas, chama a atenção para a necessidade de um trabalho coletivo de sistematização deste e de outros tipos de conhecimento na Ciência da Informação além de se constituir numa possibilidade de incremento do desempenho da própria IIn, uma vez que novas regras poderão ser agregadas à Base de Conhecimento inicial proposta, a partir de contribuições de profissionais motivados por este trabalho.

Não se deve, num país como Brasil, em que as precariedades em matéria de informação são grandes, achar que devemos passar obrigatoriamente por todos os estágios tecnológicos dos países mais adiantados. A tecnologia avança globalmente e ela deve ser encarada como um fator de alavancagem para a superação de nossas deficiências; estas só poderão ser superadas, ao contrário, pelo uso intensivo das tecnologias mais avançadas.

A proposta da IIn procura superar a contradição existente, talvez por um planejamento inadequado, entre as facilidades de comunicação já disponíveis no país, como foi destacado no capítulo 1, e as dificuldades de acesso à informação através dessas facilidades para o usuário final. No caso dos Sistemas de Recuperação de Informações Bibliográficas, a necessidade da assistência de um profissional com um perfil extremamente raro, de difícil formação, que tem que conhecer informação, sistemas de recuperação de informações, linguagens de acesso, álgebra booleana, além da área de assunto específica, com sua terminologia e suas particularidades, torna-se um grande obstáculo para o usuário final.

A proposta da IIn é justamente no sentido de se tornar um fator facilitador do acesso as informações armazenadas nos bancos de dados bibliográficos do país e acessíveis via RENPAC. Por funcionar como uma interface a estes sistemas, facilitando este acesso e dispensando a intermediação do especialista humano, por sua própria natureza de qualidade e desempenho variáveis, a IIn pode tornar corriqueiro, "fácil" e otimizado o acesso a informações como insumo a atividades técnicas e científicas.

Para a área de Ciência da Informação o trabalho, ao propor a "automação" de uma atividade antes realizado por um "expert" humano, traz também, necessariamente, a contribuição de sistematizar e formalizar este conhecimento. Ao dispensar o especialista humano da função de intermediário entre o usuário e o sistema, na verdade uma posição de atravessador, a IIn oferece facilidades ao próprio usuário final para que ele possa realizar por seus próprios meios, sua pesquisa retrospectiva em SRIs bibliográficos, tornando mais próximo da realidade o sonho de Vannevar Bush, de um trabalho científico solidamente apoiado no insumo informacional de fácil acesso.

Notas e Referências.

1. FOX, E. **Practical Enhaced Boolean Retrieval: Experiences with Smart and Sire Systems.** Information Processing & Management, v. 24, n. 3, p.257-267, 1988.

2. MARON, M. E. **Probabilistic Design Principles for Conventional and Full-Text Retrieval Systems.** Information Processing & Management, v. 24, n. 3, p.249-255, 1988.

3. TAHANI, V. **A Fuzzy Model of Documento Retrieval Systems.** Information Processing & Management, v. 12, p.177-87.

4. SALTON, G. **A Simple Blueprint for Automatic Boolean Query Processing.** Information Processing & Management, v. 24, n. 3, p.269-80, 1988.

5. LANCASTER, F. W. **Information Retrieval Systems.** John Wiley & Sons, New York, 222p, 1968.

6. BELKIN, N. J., VICKERY, A. **Interaction in Information Systems.** London, British Library, 1985. (Library and Information Research Report, 35).

7. TONG, R. M., APPELBAUM, L. A. **Conceptual Information Retrieval from Full-Text.** RIAO, p.899-909, 1988.

8. WATTERS, C. R. e SHEPHERD, M. A. **A Logic Basis for Information Retrieval.** Information Processing & Management, v. 23, n. 5, p. 433-45, 1987.

9. LEBOWITZ, M. **Intelligent Informations Systems.** ACM SIGIR, 1983.

10. DEJONG, G. **Artificial Intelligence Aplications for Information Retrieval.** Communications of ACM, v. 17, n. 4, 1983.

11. BELKIN, N. J., ODDY, R. N. BROOKS, H. M. **ASK for Information Retrieval.** Journal of Documentation, v. 38, n. 2 p.61-71, 1982. Part I.

12. _____. ASK for Information Retrieval: Journal of Documentation, v. 38, n. 3 p.145-65, 1982. Part II.

7 BIBLIOGRAFIA:

ALLEN, J. F., PERROULT, C. R. **Analyzing Intetion in Utterances.** Artificial Intelligence, v. 15, p.143-178, 1980.

BARKER, F. H., VEAL, D. C., WYATT, B. K. **Towards Automatic Profile Construction. Journal of Documentation,** v. 28, n. 1, March 1972.

BATES, M. J. **Information Search Tatics.** JASIS, v. 4, n. 30, p.205-14, 1979.

BELKIN, N. J., ODDY, R. N. BROOKS, H. M. **ASK for Information Retrieval.** Journal of Documentation, v. 38, n. 2 p.61-71, 1982. Part I.

_____. ASK for Information Retrieval: Journal of Documentation, v. 38, n. 3 p.145-65, 1982. Part II.

BELKIN, N. J., VICKERY, A. **Interaction in Information Systems.** London, British Library, 1985. (Library and Information Research Report, 35).

BRUCE, B. **Case Systems for Natural Language.** Artificial Intelligence, v. 6 p.327-360, 1975.

CNPq/IBICT/CNEN-CIN. **LINCE:** Linguagem Comum de Recuperação de Informações em linha (versão preliminar). Rio de Janeiro, CNPq, s.d.

CUNHA, C. F. da. **Gramática de Base.** Rio de Janeiro, Ministério da Educação e Cultura - Fundação Nacional de Material Escolar, 1979.

CUNHA, I. M. R. **Estruturação de Vocabulário.** In: Análise Documentária: a Análise da Síntese. Brasília, Instituto Brasileiro de Informação em Ciência e Tecnologia - IBICT, p.62-85, 1987.

DANIELS, P. J. **Cognitive Models in Information Retrieval - An Evaluative Review. Journal of Documentation,** v. 42, n. 4, p.272-304, December, 1986.

DAS-GUPTA, P. **Boolean Interpretation of Conjuctions for Document Retrieval.** JASIS, v. 38, n. 4, p.245-54, 1987.

DEJONG, G. **Artificial Intelligence Aplications for Information Retrieval.** Communications of ACM, v. 17, n. 4, 1983.

DE SOUZA, C. S. **Gramáticas de Determinação: uma proposta metodologica.** In: Jornadas Argentinas de Informática e Investigación Operativa, vol.2, 1988.

DIALOG INFORMATION SERVICES. **Pocket Guide to DIALOG.** Dialog Information Services, Inc,1987.

DILLON, M., DESPER, J. **The Use of Automatic Relevance Feedback in Boolean Retrieval Systems.** Journal of Documentation, v. 36, n. 3, p. 197-208, September 1980.

EMBRATEL. **Introdução à Comutação de Pacotes.** EMBRATEL, 53p.,

FENLY, C. & HARRIS, H. **Expert Systems - Concepts and Aplications.** Washington, D. C., Cataloging Distribution Service, Library of Congress, 37p., 1988. (Advances in Library Techology, 1).

FIDEL, R. **Online searching styles: a case-study-based Model of Searching Behavior.** JASIS, v. 35, n. 4, p.211-221, 1984.

FLUHR, C. Information Retrieval. In: CONSULTIVE MEETING of INIS LIAISION OFFICERS, 17, Viena, May 1989.

FOX, E. **Practical Enhaced Boolean Retrieval: Experiences with Smart and Sire Systems.** Information Processing & Management, v. 24, n. 3, p.257-267, 1988.

GILCHRIST, A. **The Thesaurus in Retrieval.** London, ASLIB, 1971.

GRZELAK, HANNA, KOWASKY, K. **Automatic Construction of Information Queries.** Information Processing & Management, v. 19, n. 6, p.381-9, 1983.

HARTNER, S. P. & PETERS, A. R. **Heuristics for Online Information Retrieval: a Typology and Preliminary Listing.** Online Review, v. 9, n. 5, 1985.

HEINE, M. H. **A Logic Assistant for the Database Searcher.** Information Processing & Management, v. 24, n. 3, p.323-329, 1988.

INFO. **Bases de Dados Nacionais.** Rio de Janeiro, JB/EMBRATEL, 1989. Número especial.

INTERNATIONAL ATOMIC ENERGY AGENCY. **Inis: Subject Categories and Scope Descriptions.** International Atomic Energy Agency - IAEA, Viena, 1983.

INTERNATIONAL ATOMIC ENERGY AGENCY. **Inis: Thesaurus.** International Atomic Energy Agency - IAEA, Viena, 1984.

INTERNATIONAL STANDARD ORGANIZATION. **Working Draft for Search and Operational Support Protocols (Command Language for Interrogation of Information Retrieval).** ISO/TC 46/SC 4/WO 5. Paris, may, 1980.

LEBOWITZ, M. **Intelligent Informations Systems.** ACM SIGIR, 1983.

LANCASTER, F. W. **Information Retrieval Systems.** John Wiley & Sons, New York, 222p, 1968.

LUHN, H. P. A Statistical Approach to Mechanised Encoding and Search of Library Information. **IBM Journal of Reseacrh and Development**, 1, p.309-17, 1957.

MARON, M. E. **Probabilistic Design Principles for Conventional and Full-Text Retrieval Systems.** Information Processing & Management, v. 24, n. 3, p.249-255, 1988.

METZIER, D. P. at alli. **Constituent Object Parsing for Information Retrieval and Similar Text Processing Problems.** Jasis, v. 40, n. 6, p. 398-423, 1989.

NILSSON, N. J. **Principles of Artificial Intelligence.** California, Tioga Publishing Co., 476p, 1980.

RUWET, N. **Introdução à Gramática Gerativa.** São Paulo, Ed. perspectiva, Ed. da Universidade de São Paulo, 357p., 1975.

SALTON, G. **A Simple Blueprint for Automatic Boolean Query Processing.** Information Processing & Management, v. 24, n. 3, p.269-80, 1988.

SARACEVIC, T., KANTOR, P. **A Study of Information Seeking and Retrieval.** JASIS, v. 39, n. 3, part I, p.161-176, part II, p.177-196, part III, p. 197-216, may 1988.

SCHANK, R. **Conceptual Information Processing.** New York, North-Holland/American Elsevier, 374p. 1975.

_____ **Representing and Understanding of Text.**

SHOVAL, P. **Principles, Procedures and Rules in an Expert System for Information Retrieval.** Information Processing & Management, v. 21, n. 6, p.475-87, 1985.

SIQUEIRA, I. S. P, PEREIRA, A. E. C. **Perspectivas de Aplicação da Inteligência Artificial à Biblioteconomia e à Ciência da Informação. R. Bras. Biblioteconomia e Doc**, v. 22, n. 1/2, p.39-80, 1989.

SPIEGLER, I., ELATA, S. **A Priori Analysis of Natural Language Queries.** Information Processing & Management, v. 24, n. 6, p.619-31, 1988.

TAHANI, V. **A Fuzzy Model of Documento Retrieval Systems.** Information Processing & Management, v. 12, p.177-87.

TODESCHINI, C. **INIS: Manual for Indexing.** Viena, IAEA, 106p, 1974.

TONG, R. M., APPELBAUM, L. A. **Conceptual Information Retrieval from Full-Text.** RIAO, p.899-909, 1988.

VERNIMB, C. **Automatic Query Adjustment in Document Retrieval.** Information Processing & Management, v. 13, p.339-53, 1977.

VICKERY, A., BROOKS, H. e ROBINSON, B. **A Reference and Referral System using Expert System Techniques.** Journal of Documentation, v. 43, n. 1, p.1-23, march 1987.

WATTERS, C. R. e SHEPHERD, M. A. **A Logic Basis for Information Retrieval.** Information Processing & Management, v. 23, n. 5, p. 433-45, 1987.

WOODS, W. A. **Transition Network Grammars for Natural Language Analysis.** Communication of the ACM, v. 13, n. 10, octuber, p.591-606, 1970.

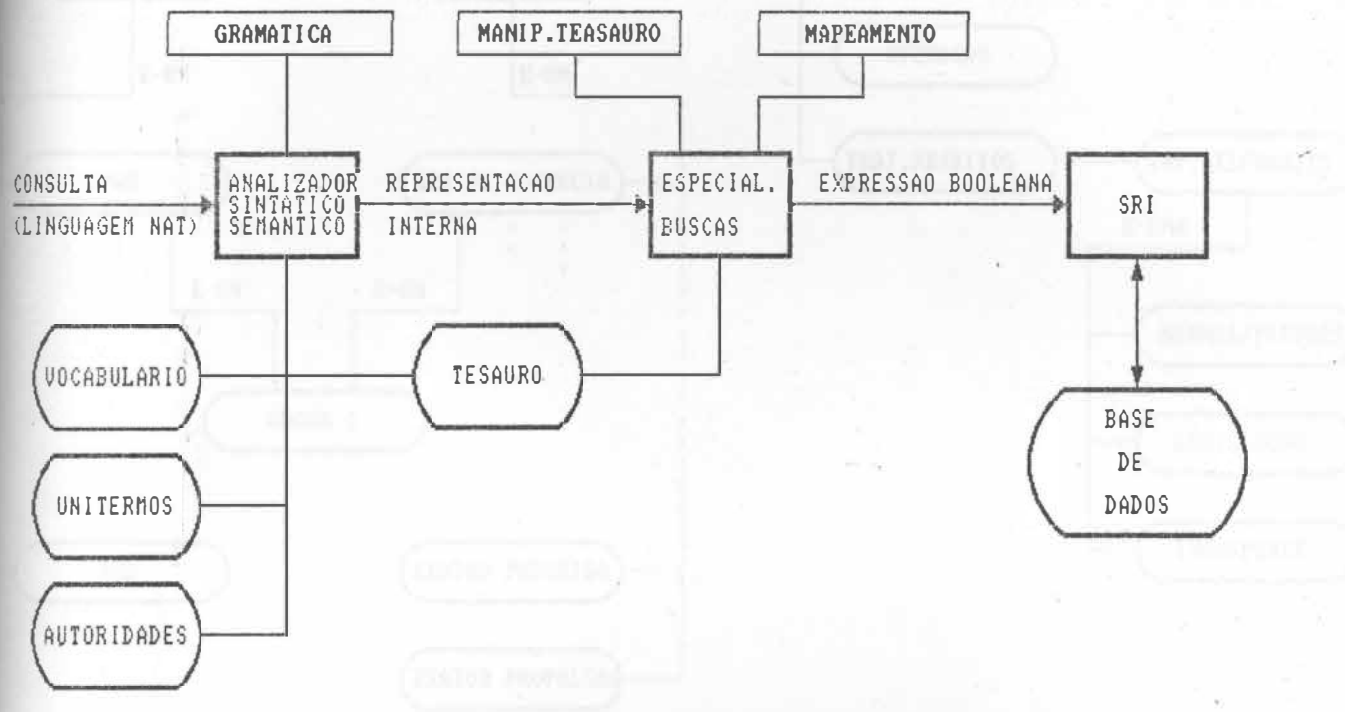

FIG.01 - SEQUENCIA DE ETAPAS DE PROCESSAMENTO DA IIn

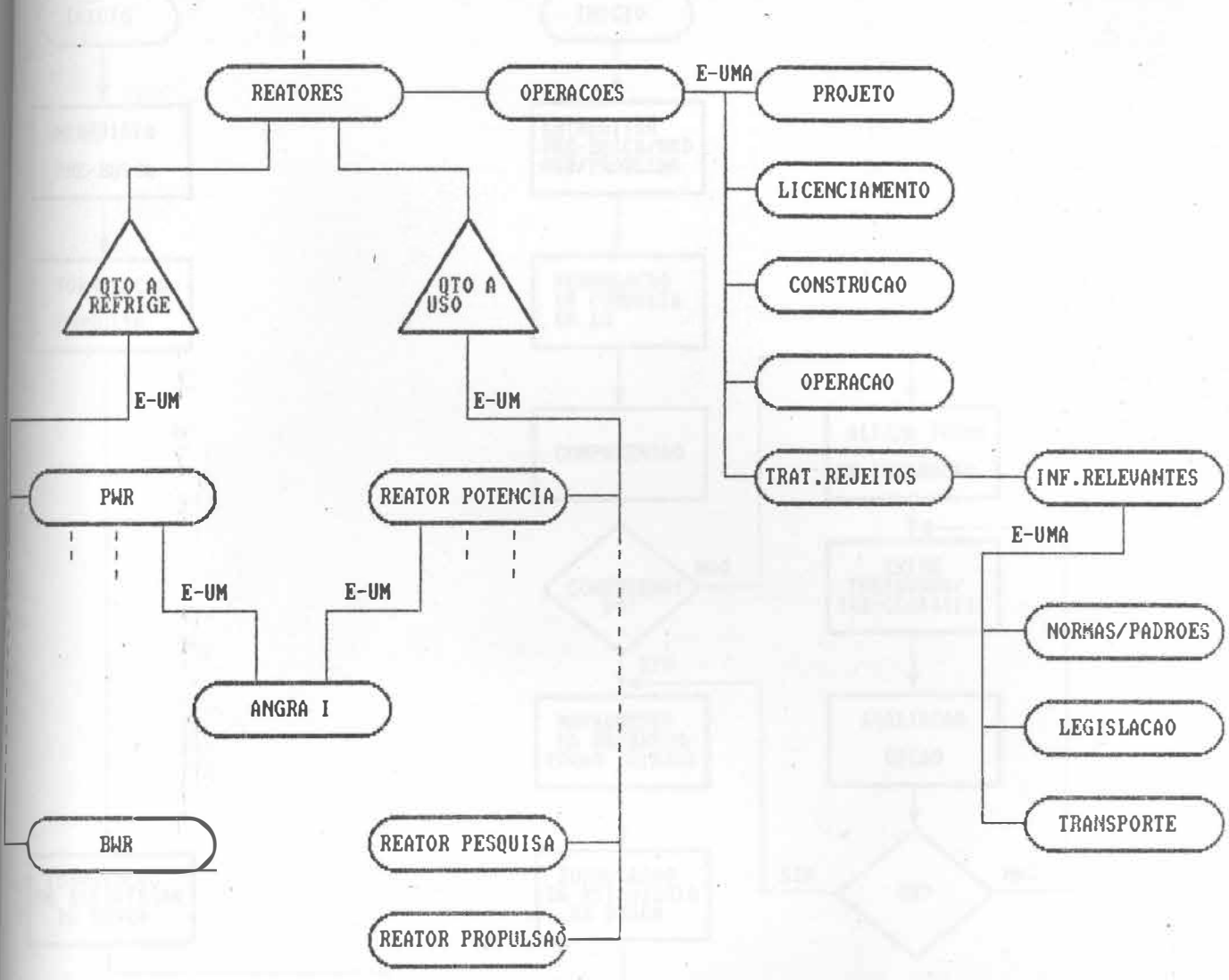

FIG.02 - REDE SEMANTICA REPRESENTANDO "REATOR" E CONCEITOS RELACIONADOS

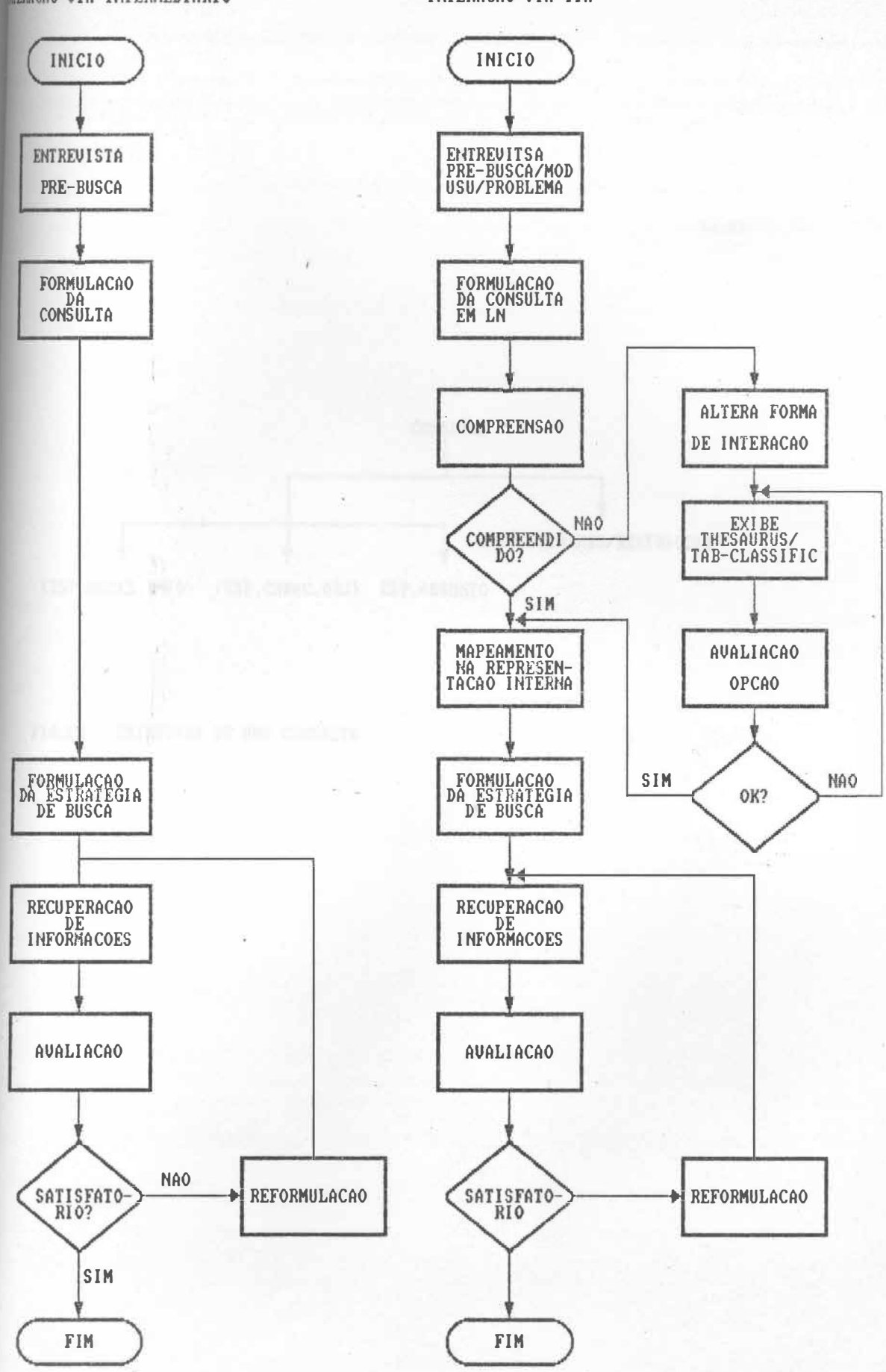

FIG.03 - PROCESSO DE INTERACAO COM UM SRI

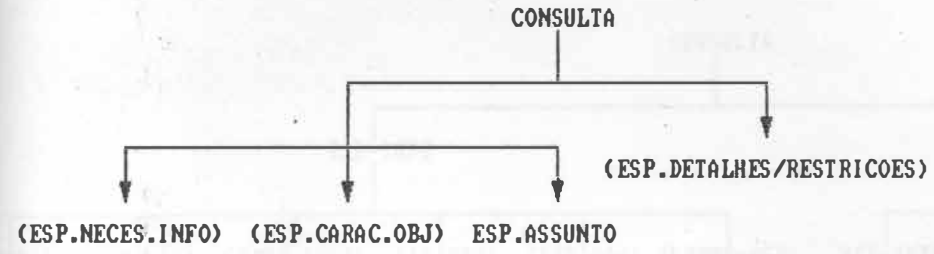

FIG.04 - ESTRUTURA DE UMA CONSULTA

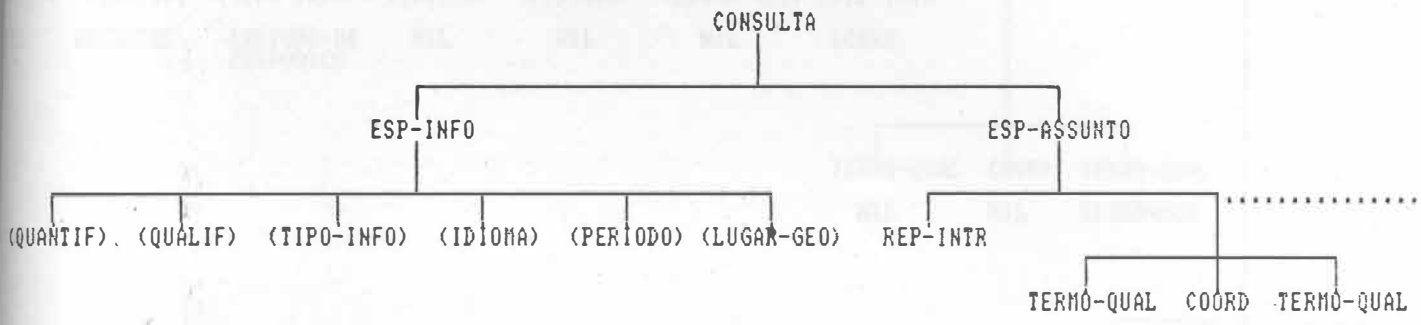

FIG.05 - ESTRUTURA DE QUADROS E ATRIBUTOS PARA UMA EXPRESSAO DE CONSULTA

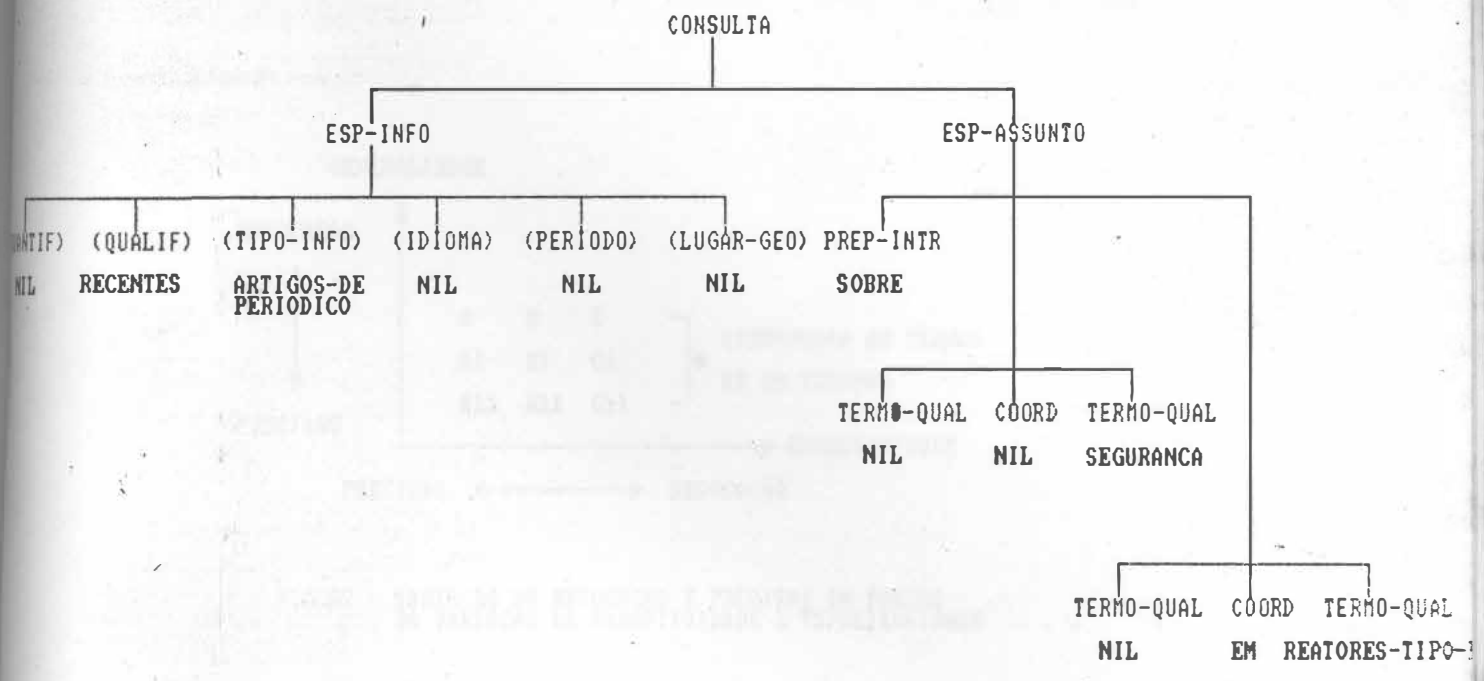

FIG.06 - ESTRUTURA DE QUADRO E ATRIBUTOS PREENCHIDOS PARA UMA CONSULTA

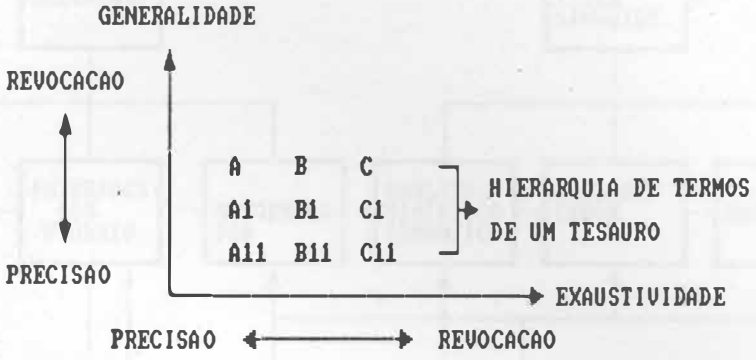

FIG.07 - VARIACAO DA REVOCACAO E PRECISAO EM FUNCAO DA VARIACAO DA EXAUSTIVIDADE E ESPECIFICIDADE

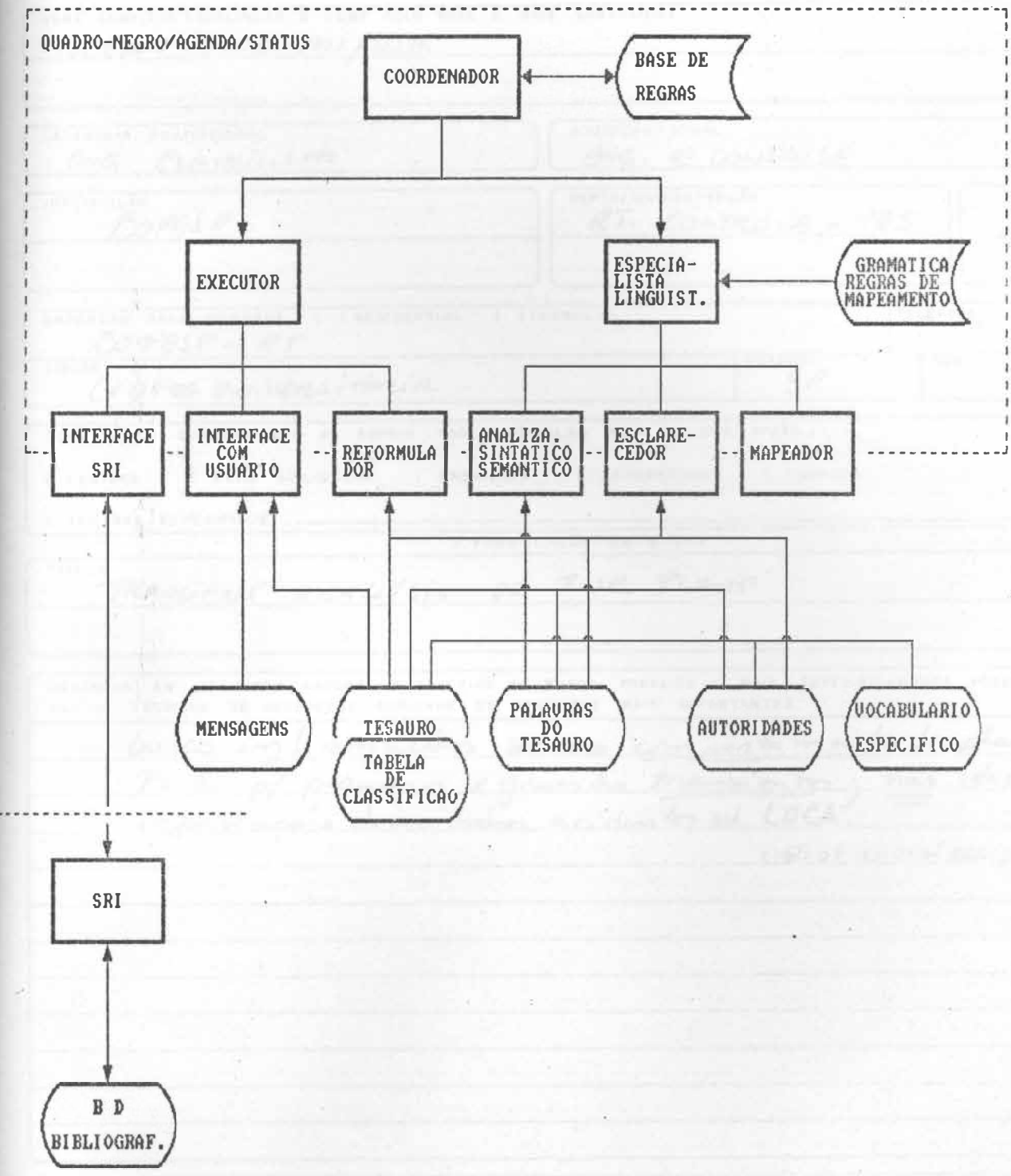

FIG.08 - ARQUITETURA E MODULOS FUNCIONAIS DA IIn

SERVIÇO PÚBLICO FEDERAL

COMISSÃO NACIONAL DE ENERGIA NUCLEAR
CENTRO DE INFORMAÇÕES NUCLEARES
RUA GAL. SEVERIANO Nº 90 - BOTAFOGO - RJ BRASIL TEL.295 2232 R301/2958545
CEP 22294 - TELEX 021-21280 CNEN - BR

INPE 312

VOCÊ É USUÁRIO DO SONAR? IWIS ☒ SIM ☐ NÃO

SUPRIR - SOLICITAÇÃO DE BUSCA RETROSPECTIVA

NOME COMPLETO (SUBLINHAR O NOME PELO QUAL É MAIS CONHECIDO): NELSON FERNO MIYOSHI

SEXO: M

CATEGORIA PROFISSIONAL: ENG. ELETRICISTA

OCUPAÇÃO ATUAL: ENG. DE CONTROLE

INSTITUIÇÃO: COPESP

DEPTº/DIVISÃO/SEÇÃO: RT - CONTROLE - 485

SIGLA: RT. 485

ENDEREÇO PARA REMESSA ( ) RESIDENCIAL ( ) TRABALHO: COPESP - RT

TELEFONE DO TRABALHO:

CIDADE: CIDADE UNIVERSITÁRIA

ESTADO: SP

CEP:

COMO TOMOU CONHECIMENTO DO SUPRIR (PODE ASSINALAR MAIS DE UMA OPÇÃO):

( ) AMIGOS ( ) SUA BIBLIOTECA ( ) REVISTAS ( ) CONGRESSOS ( ) CURSOS

( ) OUTRAS (ESPECIFIQUE):

2. FORMULAÇÃO DA BUSCA

TÍTULO: TRANSIENT ANALYSIS OF PWR PLANT

DESCREVA EM LINGUAGEM NATURAL O OBJETIVO DA BUSCA. FOCALIZE, O MAIS ESPECIFICAMENTE POSSÍVEL, OS ASPECTOS TÉCNICOS DE INTERESSE. SUBLINHE OS CONCEITOS MAIS IMPORTANTES.

- busco informações sobre o comportamento de plantas PWR p/ pequenos e grandes transientes; não desejo informações de acidentes ou LOCA.

LOSS OF COOLANT ACCIDENT

PALAVRAS - CHAVE SUGERIDAS (LIVRES)

TRANSIENTS analysis — treat reactor

PWR

SE DESEJAR, INDIQUE O(S) IDIOMA(S) E OU O(S) ANO(S) DE PUBLICAÇÃO E/OU A(S) ÁREA(S) ESPECÍFICA(S) E CLASSIFICAÇÃO DOS DOCUMENTOS A SEREM SELECIONADOS.

IDIOMA(S)	INGLÊS
ANO(S)	1970 - 1989

RELACIONE ENFOQUES, TÓPICOS OU APLICAÇÕES SOBRE OS QUAIS NÃO DESEJE QUE O ASSUNTO SEJA ABORDADO

RELACIONE TÍTULOS DE PUBLICAÇÕES, AUTORES, SIGLAS OU ORGANIZAÇÕES CUJOS TRABALHOS SÃO RELEVANTES PARA O ASSUNTO A SER PESQUISADO. SE POSSÍVEL, ANEXE A 1ª PÁGINA OU O RESUMO DOS TRABALHOS

QUANTIDADE CONVENIENTE DE REFERÊNCIAS (COBRA-SE POR CADA REFERÊNCIA IMPRESSA)

☐ SEM RESTRIÇÕES ☐ 1 - 10 ☐ 11 - 30 ☐ 31 - 60 ☐ 61 - 100

3. FORMULAÇÃO DAS CONSULTAS

1ª BOOLEANA - BASE DE DADOS UTILIZADA

(POWER^{P} PLANTS + INDUSTRIAL^{D} PLANTS) $ (PWR^{2} TYPE REACTORS + TREAT^{3} REACTOR) $ TRANSIENTS^{3} : ACCIDENTS = 45

RI: 5863

30 DOC's

20'

RESTRIÇÕES	
ÁREA	
IDIOMA	
ANO	
PAÍS	

2ª BOOLEANA - BASE DE DADOS UTILIZADA

RESTRIÇÕES	
ÁREA	
IDIOMA	
ANO	
PAÍS	

3ª BOOLEANA - BASE DE DADOS UTILIZADA

RESTRIÇÕES	
ÁREA	
IDIOMA	
ANO	