Proposta de um Sistema de Recuperação de Informação Assistido por Computador - SRIAC

Hélio Kuramoto

Os problemas de utilização dos Sistemas de Recuperação de Informação textual tradicionais (SRI) são discutidos com base em alguns princípios como: leveza, rapidez, exatidão, visibilidade, multiplicidade e interatividade, inspirados no livro de Ítalo Calvino, Lições americanas: 6 proposições para o terceiro milênio.

Palavras-chave: Sistemas de Recuperação de Informação. Sintagmas nominais. Descritores. Indexação Automática. Interface de Recuperação de Informação.

1 INTRODUÇÃO

Com o aparecimento dos computadores, milhões de informações foram registradas em várias bases de dados, em diversos domínios de conhecimento e sob diversas formas como: numéricas, textuais, de imagens e outras. Se hoje os recursos informacionais se tornaram disponíveis também aos usuários pessoais, por outro lado, no entanto, continuam os problemas de como fornecer-lhes as informações de que necessitam.

Com a inserção das Novas Tecnologias da Informação e da Comunicação (NTIC), impulsionadas pela rede de redes,Internet, e também pelos Estados Unidos com o anúncio de seu programa *National Information Infrastructure* (NII), constatou-se um rápido e incrível crescimento do volume de informação, como também das ferramentas de processamento e de busca da informação.

Segundo Pierre Lévy (1997, p. 17) “a prosperidade das nações, das regiões, das empresas e dos indivíduos depende de suas capacidades de navegar no espaço do saber. A potência ou força é conferida doravante pela gestão ótima do conhecimento, que elas sejam de ordem tecnológicas, científicas, da comunicação ou que elas resultem da relação ‘ética’ com o outro”.

Os contextos apresentados constituem, na realidade, um único contexto. De um lado, percebe-se a oferta crescente de informação, como consequência — além da produção crescente de informação — da baixa de preços dos recursos de computação,

e da inserção das NTIC. Do outro lado, tem-se o discurso da importância do conhecimento e do saber para as nações, as regiões, as empresas e os povos (a utilização da informação). Pierre Lévy faz uma constatação daquilo que se vê hoje em dia — a modificação de paradigmas — tendo em vista a chegada das NTIC. *A* gestão ótima dos conhecimentos passa necessariamente pela aquisição e produção da informação. É nesse aspecto que se percebe a importância dos Sistemas de Recuperação da informação (SRI), dado que são estas as ferramentas que se utiliza para a busca e acesso à informação.

Vários tipos de SRI podem ser encontrados no mercado, segundo as características da informação ou da aplicação. Nesse ponto, faz-se necessário distinguir entre o que seja um SRI e uma Interface de Recuperação de Informação (IRI). O primeiro, é um sistema completo que compreende a interface de recuperação de informação, o tratamento da informação, seu armazenamento e sua organização. Concorda-se, então, com Gerald Salton e Michael McGill (1983), quando afirmam que um SRI trata da representação, do armazenamento, da organização e do acesso aos itens de informação.

A IRI faz parte de um sistema qualquer; ela faz apenas a intermediação entre o usuário e as bases de dados. Segundo Harter*, uma IRI é um dispositivo que se interpõe entre o usuário potencial e a coleção de informações (Harter, l986). É ela quem recebe as solicitações de busca de informação, sob a forma de uma expressão de busca, do usuário. É fácil observar que uma IRI faz parte também de um SRI.

Existem autores que usam o termo Sistemas de Recuperação de Informação (SRI), tanto para denotar um sistema completo de recuperação de informação, quanto para uma interface de recuperação de informação. Utilizaremos a notação SRI neste documento, tendo em vista que serão discutidos aqui, tanto os aspectos da recuperação da informação, quanto aqueles do tratamento e da indexação automática da informação.

Neste documento vamos propor um novo sistema de recuperação de informação, tendo como base alguns princípios inspirados na obra de Ítalo Calvino. Assim, apresentaremos, inicialmente, alguns dos problemas encontrados na utilização de um SRI e em seguida discutiremos os princípios de base para a concepção do novo SRI.

2 OS PROBLEMAS DE UTILIZAÇÃO DE UM SRI

Frente à proliferação de fontes de informação e à utilização generalizada de ferramentas de informática cada vez mais evoluídas, o usuário final busca dispor de sistemas de utilização simples e amigável. Ou seja, os usuários buscam aqueles

sistemas que lhes permitem ter uma visão sintética das informações, mas também que lhes forneçam respostas mais precisas. Segundo Polity (1994), os SRI tradicionais possuem interfaces orientadas à "linguagem de comandos" que exigem do usuário o domínio de um grande número de conhecimentos de carácter heterogêneo porque alguns são de ordem computacional, outros concernem a estruturação de dados, e outros enfim relativos ao vocabulário da área de atuação. Esses sistemas são profissionais e exigem, portanto, como toda ferramenta profissional, uma certa formação. Enumerarei em seguida alguns elementos cujo domínio condiciona a utilização desses sistemas:

- Um certo número de comandos para se colocar uma sessão em modo de interrogação, para formular uma expressão de busca, para se visualizar os resultados, para imprimí-los, etc.;
- A indispensável lógica booleana numa expressão de busca multicritérios. Interseção, união e exclusão são operadores indispensáveis para se formular, de maneira mais precisa, uma expressão de busca;
- Os operadores de truncagem, de comparação e de vizinhança;
- A estrutura conceituai da base de dados, os nomes dos campos a pesquisar e as convenções de escrita em cada um desses campos;
- Os termos de indexação (descritores), os léxicos, os tesauros, etc.

Com o crescimento repentino do número de usuários de bases de dados, nos diversos setores de atividades, foi colocado em prova se as IRI são realmente amigáveis. Para se ter acesso à informação, o usuário deve se familiarizar com as IRI e dominar linguagens de busca. Essa dificuldade de familiarização limita o crescimento do número desses usuários (Show *et al).*

O grande problema das interfaces de SRI, naquilo que diz respeito às linguagens de busca, é que elas são, de uma maneira geral, de difícil aprendizagem, especialmente, para os usuários não especializados. Isso é devido ao fato de que elas possuem regras rígidas de sintaxe e utilizam-se de expressões matemáticas complicadas - no caso do SQL -(Wang, 1994).

Além do problema das interfaces dos SRI não serem amigáveis, os resultados fornecidos por esses sistemas, no domínio da informação documentária e textual, não são precisos. Freqüentemente expressa sob a forma de uma combinação booleana de palavras-chave, uma busca de informação textual pode se exprimir também com a ajuda de operadores de proximidade, ou mesmo incluir pesos (determinados automaticamente ou por definição por parte dos usuários) sobre as palavras-chaves. Entretanto, esse tipo de pesquisa pode não ser eficaz, principalmente pelas seguintes razões :

- As palavras do texto podem ter significados ou sentidos diferentes segundo a área do conhecimento. Exemplo: *gota* pode indicar uma pequena quantidade de água que toma uma forma arredondada ou pode significar, no campo da medicina, inflamação dolorosa nas articulações;
- As mesmas palavras podem ser utilizadas em diferentes frases, em ordem ou ligações distintas, e exprimir conceitos totalmente diferentes. Exemplo: análise estatística da informação e análise da informação estatística. Esse exemplo mostra as mesmas palavras, empregadas em ordem diferente em duas frases, e que designam conceitos totalmente diferentes;
- Palavras completamente diferentes podem ser utilizadas para exprimir um mesmo conceito. Exemplo: *Terremoto* e *sismos.*

As técnicas convencionais utilizadas na recuperação de informação, tais como a indexação baseada em palavras-chave ou simplesmente palavras, e o uso de expressões booleanas não podem resolver os problemas apresentados (Smeaton, 1991).

A avaliação de uma expressão na recuperação da informação pode ser vista como uma comparação entre os documentos e a expressão de busca (ou entre a representação dos seus conteúdos). No modelo clássico (Salton & McGill, 1983) essa comparação é feita diretamente. Ou seja, uma comparação baseada na utilização comum das palavras-chave entre o documento e a expressão de busca. O ponto crítico desse modelo é que se um documento semanticamente pertinente não é representado pelas mesmas palavras-chave da expressão de busca, então o documento será julgado, pelo sistema, como não pertinente. Exemplificando, num SRI onde a expressão de busca é feita através de expressões booleanas, um documento indexado pelas palavras *document retrieval* não será encontrado se utilizarmos na expressão de busca o termo *information retrieval.*

Recentemente, algumas aplicações semelhantes aos SRI apareceram na rede Internet. São os denominados Motores de Busca ou *Search Engines,* em inglês. Essas ferramentas são simples de utilizar e fazem acesso rápido aos índices: sem o incômodo de ter que aprender o uso de uma linguagem orientada a comandos para consultas a bases de dados, como se fazia nos anos 80 (Lardy). No entanto, nos deparamos ainda com os mesmos problemas: a) a diversidade de sintaxes para a construção de uma expressão booleana, dado que cada motor de busca possui o seu próprio modo de análise; b) respostas a demandas de informação com uma considerável taxa de ruídos[1]. Para facilitar a tarefa aos usuários, será necessário desenvolver mecanismos

[1] Denomina-se taxa de ruídos a proporção de documentos recuperados não-pertinentes em relação ao total de documentos recuperados.

mais possantes de filtragem de informações.

Assim como os SRI tradicionais, os Motores de Busca fazem, também, a indexação automática baseada nas palavras como descritores. São, usualmente, baseadas no modelo desenvolvido por Gerald Salton. Por outro lado, a falta de precisão nos resultados apresentados por essas ferramentas é muito mais chocante porque as informações contidas nessas bases de dados são mais genéricas, pluridisciplinares e volumosas.

3 PROPOSTA DE UM SISTEMA DE RECUPERAÇÃO DE INFORMAÇÃO

Proporemos, neste documento, um novo Sistema de Recuperação de Informação, tendo como base alguns princípios. Tais princípios são inspirados nas 6 proposições de Italo Calvino para o próximo milênio, de sua obra *Leçons américaines : aide-mémoire pour le prochain millénaire.* São eles: a leveza, a rapidez, a exatidão, a visibilidade, a multiplicidade e a consistência. Apesar desses princípios terem sido formulados para o campo da literatura, parece-nos que eles fornecem bons indícios para a construção de um SRI mais preciso, amigável e leve. Além disso, pode-se adicionar um sétimo princípio, a interatividade. Essa proposta será descrita à medida que se discute cada princípio.

3.1 A Leveza

De acordo com o que se expôs na seção precedente, verifica-se a existência de problemas de utilização dos SRI. Dentre os mais importantes destacamos: o fato de suas interfaces não serem amigáveis e a fraca precisão dos resultados de uma busca de informação.

A falta da qualidade de ser amigável é devida, principalmente, às dificuldades de interação entre o usuário e a interface do SRI, em conseqüência, geralmente, do uso de uma linguagem de comandos que, por sua vez, possui uma sintaxe rígida e às vezes complexa. De fato, o usuário é obrigado a possuir um grande número de conhecimento sobre o sistema, o computador e os comandos. Em função disso, ele é normalmente obrigados a ler volumosos manuais ou mesmo a fazer cursos de acesso a bases de dados. Apesar de já existirem interfaces com facilidades gráficas, no mercado, sem os problemas de uma linguagem orientada a comandos, resta ainda a dificuldade de uso das expressões booleanas e de seus operadores.

Por outro lado, nos resultados fornecidos pelos usuários, em resposta à uma solicitação de informação, encontra-se, freqüentemente uma considerável taxa de

ruídos. Um exemplo são os casos dos resultados de uma simples consulta, utilizando- se um motor de busca qualquer na Internet, onde se encontra, não raramente, qualquer coisa como 15.000, 50.000 referências e mesmo valores superiores a esses. O que torna difícil a tarefa de encontrar a informação que o usuário necessita. Como encontrar as informações de que o usuário necessita num conjunto tão extenso de referências?

Segundo Jin & Fine, “quando um usuário está saturado de informações, a adição de novas informações não o ajuda na tomada de decisão ou na solução de um problema”. Além da dificuldade física de manipular um número de referências tão elevado, perde-se também a motivação de encontrar a informação.

Esses dois problemas discutidos acima dão aos SRI um caráter pesado. Esse carácter não é físico. Não é o SRI, em si, que é pesado. O peso é simbolizado pelas dificuldades encontradas pelos usuários quando esses tentam utilizar um SRI. Encontrar as referências que preenchem as suas necessidades de informação em 50.000 referências, não se trata de uma tarefa leve; ler com atenção manuais de utilização de um banco de dados, composto geralmente por algo em torno de 500 páginas ou mais, também não se trata de uma tarefa leve.

O princípio da leveza, conforme Italo Calvino (1988), “é relacionado à precisão e à determinação, e não ao vago e ao aleatório”. O princípio da leveza é fortemente relacionado ao princípio da exatidão, dado que a leveza depende da precisão. Esse princípio se relaciona não ao software SRI em si, mas àquilo que ele passa aos usuários, no procedimento de interação. Assim, a leveza deve ser traduzida através de resultados mais precisos, de uma interface mais fácil de ser utilizada e mais intuitiva sem, necessariamente, empregar uma linguagem muito complicada. As NTIC oferecem uma ampla gama de ferramentas que podem tomar as interfaces mais amigáveis e mais intuitivas; aliás é isso que se começa a ver nos momentos atuais.

Em relação aos resultados de uma busca, a concepção de um novo SRI deve procurar um novo enfoque de indexação de maneira a diminuir os problemas relacionados com os procedimentos tradicionais de indexação automática (indexação baseada nas palavras), conforme discussão na seção precedente.

Quase todos os SRI existentes atualmente fazem uma indexação baseada, seja na simples extração de palavras encontradas nos textos ou documentos de uma base de dados, seja utilizando-se da mesma extração de palavras acompanhadas da aplicação de um modelo probabilístico/estatístico, através do qual é determinado o grau de pertinência de um descritor. Naquilo que concerne às palavras, observou-se os problemas relacionados à sua utilização numa expressão de busca. Os modelos estatísticos se utilizam da freqüência de ocorrência das palavras nos documentos para

indicar o nível e a ordem de significância destas. Um dos nove postulados de impotência, estabelecidos por Swanson (1988) diz “a estatística de ocorrência de palavras não pode representar o significado da palavra e nem substituí-la. Esse valor, entretanto, pode ser utilizado com eventual sucesso, para sinalizar, ou descobrir potenciais domínios fecundos de textos onde o homem pode procurar o significado ou pertinência”. Quer dizer, a freqüência de aparecimento de uma palavra num texto ou num documento não determina necessariamente sua maior ou menor pertinência em relação ao tema de uma busca. Trata-se de uma das fraquezas do enfoque tradicional. Conforme Swanson, não serão os cálculos estatísticos que vão eliminar as taxas de ruídos encontradas numa resposta a uma solicitação de busca de informação do usuário.

Portanto, é preciso encontrar um enfoque para a indexação de documentos textuais contidos numa base de dados, para que se possa conseguir resultados menos pesados e mais precisos.

Entre as pesquisas sobre a indexação automática, o enfoque desenvolvido no âmbito do grupo SYDO[2] — a utilização dos sintagmas nominais como meio de acesso à informação — pode dar a leveza, esperada e procurada, aos SRI. Pelo menos, esse enfoque pode trazer soluções aos problemas encontrados freqüentemente com os descritores, enquanto palavras-chave ou simples palavras.

Em função dos problemas discutidos sobre a utilização das palavras como descritores—ambígüidades —, os trabalhos do grupo SYDO analisam o problema da indexação sobre um outro aspecto. A indexação propicia a representação do conteúdo de um documento. Portanto, os descritores deveriam representar esse conteúdo, eles deveriam fazer referência aos objetos da realidade extra-linguística do autor do documento indexado, ou seja, do objeto do qual se fala no documento. Em contrapartida, as palavras, enquanto unidades do léxico, não fazem nenhuma referência a nenhum objeto, possuindo apenas predicados ou propriedades. Segundo Michel Le Guern (1982), conforme citação abaixo, um descritor deveria ser um signo com referência; segundo ele as palavras da língua não estão em relação imediata com as coisas. Elas têm um significado, mas não fazem nenhuma referência a um objeto do mundo real.

[2] SYDO - Système Documentaire, grupo de pesquisa — formado no período de 1981 à 1990—constituído pelos quatro centros de pesquisas seguintes:

- Laboratoire d'Informatique Documentaire (LID) de l’Université Claude Bernard Lyon I, (Prof. R. Bouché)
- Centre de Recherches Linguistiques et Sémiologiques (CRLS) de l'Université Lumière Lyon II, (Prof. M. LE GUERN);
- Département linguistique de l’Université de Fribourg (Suisse), (Prof. A. Berrendonner);
- CRISS (Centre de Recherche en Informatique et Sciences Socrates) de l’Université de Grenoble, (Prof. J. Rouault).

> "Não constitui finalidade do descritor a sua visualização através da abstração do valor referencial de suas ocorrências no acervo de documentos. As palavras da língua, enquanto palavras da língua, possuem apenas atributos sem qualquer substância, até que elas façam parte do discurso. Quanto ao descritor, ele representa uma entidade segundo a filosofia de Aristóteles. Assim, o descritor não pode ser considerado, a exemplo das palavras da língua, como um símbolo sem referência".

Segundo Bouché (1990) *et al*, a hipótese primeira do modelo concebido pelo grupo SYDO era que "as partes do discurso construídas em torno do nome ou substantivo (quer dizer o sintagma nominal) são aquelas portadoras de referências aos objetos do universo do discurso e portanto, aquelas que devem ser identificadas".

A definição de um sintagma nominal e de como ele é constituído não serão tratados aqui, dado não ser esse o tema central deste artigo. No entanto, sua definição e alguns conceitos a seu respeito podem ser encontrados no artigo de Michel Le Guern (1991), publicado na revista *Le Français Moderne.*

O modelo proposto pelo grupo SYDO parece muito mais consistente que aquele baseado apenas no uso das palavras como descritores. Aliás, os sintagmas nominais possuem uma organização natural que leva à construção de uma interface diferente daquelas orientadas a uma linguagem de comandos. Ela não exige o uso de operadores booleanos, nem de uma linguagem de comandos. Em consequência, os usuários não têm necessidade de fazer cursos para aprender a utilizá-la. Nada impede, no entanto, de se utilizar os sintagmas nominais como descritores dentro de uma interface tradicional. Diante dessa argumentação, será adotado o uso dos sintagmas nominais como meio de acesso à informação na presente proposta, que tem como precursor o professor Michel Le Guern.

3.2 A rapidez

O princípio da rapidez, como o próprio nome indica, diz respeito à velocidade das coisas, à velocidade de execução de uma tarefa, à velocidade de realização de um procedimento. Num mundo competitivo a rapidez faz a diferença. Trata-se de um princípio muito importante. No entanto, esse princípio não pode ser visto de forma isolada. Existe uma espécie de interdependência entre os princípios. Por exemplo, de nada adianta um SRI ser rápido, se ele não fornece resultados precisos. A falta de precisão obrigará o usuário a refazer a sua expressão de busca. Isto determinará, com

certeza, maior perda de tempo. Esse princípio desempenha um papel importante, também, no procedimento de indexação. O volume de informação vem crescendo progressivamente; assim, é preciso que a atualização de uma base de dados seja feita de uma forma rápida, caso contrário, arrisca-se perder a atualidade das informações e tomar os usuários insatisfeitos. Os procedimentos de indexação devem ser bem estabelecidos e as estruturas de dados devem ser otimizadas, de forma a permitir aos usuários um acesso mais rápido às informações. Evidentemente que tudo isso depende também dos equipamentos de processamento de dados, da velocidade dos discos rígidos *(Winchester),* da velocidade das ferramentas de gerência de arquivos, da velocidade de formatação e apresentação dos dados no monitor. Ou seja, a rapidez é relacionada a um conjunto de fatores, os quais não são completamente controlados pelo grupo de desenvolvimento de softwares. Entretanto, hoje os equipamentos oferecem uma boa velocidade de acesso aos dados, e os utilitários de suporte à gerência de arquivos e de formatação do monitor são bastante eficazes e rápidos. Cabe aos construtores de SRI, portanto, estruturá-los e desenvolvê-los, de forma otimizada, buscando-se a rapidez de acesso à informação, bem como a concepção de interfaces mais ágeis , amigáveis e intuitivas, para abreviar os procedimentos de busca de informação.

Segundo Pierre Lévy, a prosperidade dos povos, das nações, das regiões dependem de uma gerência ótima do conhecimento. Por gerência ótima do conhecimento, entende-se a rapidez de aquisição da informação como um fator importante no processo de geração e difusão do conhecimento. Dessa maneira, pode- se vislumbrar que a prosperidade seja também dependente da velocidade de aquisição de informações. Por outro lado, o usuário é alguém que sempre quer a resposta dentro da maior brevidade possível. É uma exigência de sua atividade técnica. É um dos princípios mais importante para um SRI, porque se trata também de uma questão de mercado: vende-se bem aquilo que é rápido e que tem boa performance.

3.3 A exatidão

A exatidão constitui a precisão das coisas. Esse princípio é fortemente relacionado ao princípio da leveza; eles são interdependentes. Sem a precisão, não existe a leveza e vice-versa. O SRI deve, portanto, fornecer resultados precisos, ou pelo menos oferecer ao usuário meios de intervir no processo de busca de informação, de maneira a obter maior precisão nas respostas finais de um procedimento de recuperação de informação. Nesse aspecto, o enfoque adotado nesta proposta se adapta muito bem. Os sintagmas nominais diminuem os problemas de ambigüidades encontrados nos sistemas tradicionais

de recuperação de informação. Isso acontece porque os sintagmas nominais fazem referência aos objetos da realidade extralinguística dos autores dos documentos. É evidente que se pode ainda encontrar algum tipo de ambigüidade se a base de dados for pluridisciplinar. No entanto, esse é um problema fácil de se resolver, bastando, separar os documentos em bases de dados diferentes, especializadas por áreas do conhecimento.

Os sintagmas nominais têm uma organização natural, possuem uma relação de encadeamento entre si. Será mostrado, a seguir, como os sintagmas nominais se organizam e como podem ser utilizados numa interface de recuperação de informação. Seja o exemplo da figura 1, a frase “A representação do conteúdo do documento”.

Exemplo : A representação do conteúdo do documento

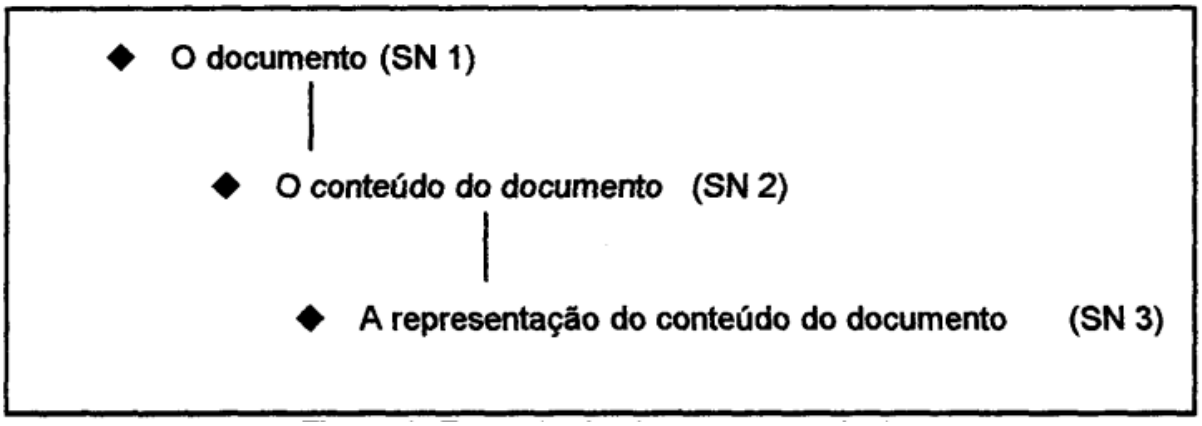

Figura 1. Exemplo de sintagmas nominais

Na figura 1, tem-se três sintagmas nominais, encadeados em três níveis diferentes. Trata-se de uma estrutura em forma de árvore. A extração de todos os sintagmas nominais de um conjunto de documentos de uma base de dados produzirá uma floresta de sintagmas nominais. Esse tipo de organização é apropriado ao desenvolvimento de uma interface navigacional.

A título de ilustração, será mostrado na figura 2 um sub-conjunto de árvores de uma floresta de sintagmas nominais, ou seja, uma subfloresta composta de algumas pequenas árvores.

Na figura 2, os números que aparecem entre parênteses fazem referência à quantidade de documentos de onde se extraiu o respectivo sintagma nominal. O sintagma nominal ***a informação*** foi extraído de 165 documentos. O mesmo faz parte de dois outros sintagmas de segundo nível: a *análise da informação* (extraído de 100 documentos) e *o armazenamento da informação* (extraído de 65 documentos). Observa-se, nessa figura, que o sintagma nominal “A *análise da informação”* está presente em 100

documentos, dos quais, 50 documentos apresentam-no desta forma. Em 29 documentos ele faz parte do sintagma nominal de terceiro nível—"O *estudo da análise da informação".* E em 21 outros documentos ele *faz* parte de um outro sintagma, também de terceiro nível — "O *procedimento de análise da informação".* Nesse exemplo percebe-se claramente o procedimento de refinamento. Observa-se que à medida que se desce a árvore de sintagmas nominais, o usuário faz naturalmente o refinamento da sua busca de informação. Esse processo permite ao usuário não somente o refinamento de sua busca, mas também a reformulação de sua demanda de informação, na medida em que ele tenha oportunidade de voltar para o nível anterior de sintagmas nominais apresentados.

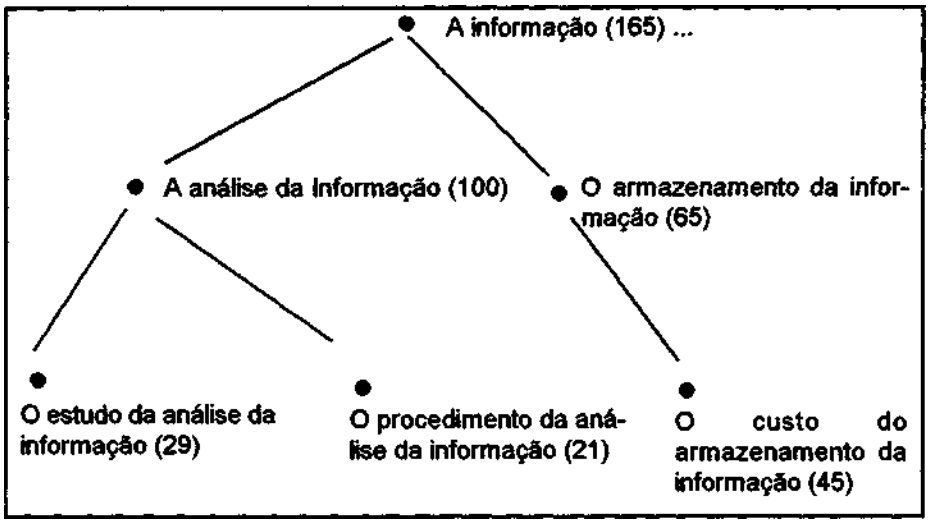

Figura 2. Exemplo de uma pequena floresta de sintagmas nominais.

A idéia geral dessa proposta é que o usuário forneça à interface uma palavra[3] que represente a sua necessidade de informação. A partir dessa palavra, a interface procura e mostra, no monitor, todos os sintagmas nominais de primeiro nível tendo essa palavra como seu centro. Certamente que nesse momento, ter-se-á uma considerável taxa de ruídos, mas cabe ao usuário escolher o sintagma nominal que melhor expresse a sua necessidade de informação e prossiga o procedimento de busca. Em seguida, ele pode solicitar à interface que mostre os documentos de onde o sintagma escolhido foi extraído, ou então, solicitar um refinamento, pedindo ao computador que busque todos

[3] A palavra fornecida será tomada como sendo o centro do sintagma nominal de primeiro nível; ela faz parte da categoria gramatical dos substantivos.

os sintagmas nominais de segundo nível a partir do sintagma nominal de primeiro nível escolhido por ele. A interface repete o mesmo procedimento até que o último nível de sintagmas nominais tenha sido atingido ou até que o usuário tenha encontrado o sintagma que satisfaz às suas necessidades de informação. .

O enfoque proposto orienta, de forma interativa, o usuário na busca da informação, sem se utilizar de uma linguagem de comandos ou de operadores booleanos. Esse procedimento permite aos usuários fazer correções e ajustes, melhorando suas estratégias de busca. Nesse contexto, nem o computador, nem o SRI, processam ou interpretam uma expressão de busca qualquer. São os próprios usuários que conduzem a busca de informação, o que permite maior precisão nas respostas a uma solicitação de busca de informação.

3.4 A visibilidade

Segundo o dicionário Le Robert Micro, “visibilidade é o caráter daquilo que é perceptível pela vista, que é sensível ao olho humano”, enquanto que o Novo Dicionário Aurélio dá como definição, “qualidade de visível”. Às vezes tem-se a impressão que esse princípio, a visibilidade, se trata de uma coisa supérflua, sem importância, e até mesmo que ele seja um princípio ligado unicamente à beleza. Na realidade, a visibilidade é um princípio tão importante quanto os outros. Para se ilustrar essa importância, veja o seguinte exemplo: um usuário tem que escolher entre um sistema que não usa as facilidades gráficas, onde ele é obrigado a digitar tudo, e um sistema cujas informações e comandos se encontram já presentes na tela do monitor com todas as facilidades gráficas. Neste último, o usuário apenas escolhe a opção desejada ou mesmo digita apenas aquilo que for estritamente necessário. É quase certo que ele vá escolher a segunda opção de interface. É a lei do menor esforço. Com certeza, este vai escolher a interface mais agradável visualmente e a mais fácil de se usar, a menos que esta interface seja menos eficaz em termos de performance em relação à precisão das respostas de uma busca de informação. Nota-se aí a importância desse princípio.

Esse princípio reforça a idéia da necessidade de se utilizar vários modos de expressão numa interface de recuperação de informação. É necessário usar não somente o *mouse* e o teclado, mas também as facüdades gráficas e, quando necessário os monitores termo-sensitivos. Todas essas ferramentas poderão dar ao SRI maior visibilidade e torná-lo mais amigável É necessário que os novos SRI sejam mais intuitivos. Por isso, acredita-se que as ferramentas introduzidas pelas NTIC sejam, cada vez mais utilizadas, beneficiando os usuários, tendo em vista que proporcionam

melhor visibilidade e intuição, facilitando ao usuário a interação com os vários sistemas.

O conceito de visibilidade não fica apenas na estética, mas diz respeito também à visualização do conteúdo de uma base de dados. O fato de conhecer o conteúdo de uma base de dados dá ao usuário maiores chances de ganhar tempo e obter as suas respostas dentro de um limite de tempo menor. É importante observar que os tradicionais sistemas de recuperação de informação, aqueles orientados a uma linguagem de comandos, têm uma visibilidade muito pobre, à exceção daqueles que têm a facilidade de mostrar os descritores. Constata-se essa pobreza de visibilidade, dado que o usuário somente vê a quantidade de referências recuperadas, ou apenas as respostas a uma expressão de busca; caso queira ver as referências, ele tem que usar um novo comando para mostrá-las. Em geral não existe uma função que mostre os índices antes de se fazer a busca. Esse tipo de facilidade tem sido apresentada nos sistemas desenvolvidos recentemente, normalmente aqueles que vêm junto com as bases de dados em CD-ROM. Ou seja, *o* fato de não existir uma facilidade de apresentação dos índices previamente, torna difícil ao usuário conhecer a base de dados em que ele está fazendo a pesquisa. Essa dificuldade acaba por provocar uma certa perda de tempo. Em contrapartida, o enfoque proposto neste artigo, permite ao usuário trabalhar com maior visibilidade do que aquela existente nos SRI tradicionais. À medida que ele navega na árvore de sintagmas nominais, passando de um nível a outro, ele vai aprendendo sobre o teor da base de dados. A estrutura, em si, induz o desenvolvimento de uma interface mais visível.

3.5 A multiplicidade

O que é a multiplicidade? É o caráter daquilo que é múltiplo. A necessidade desse princípio se explica pelo fato das coisas não serem únicas, elas são múltiplas. Assim, é preciso tomar cuidado com os diversos aspectos de cada problema, de cada procedimento. Para que um sistema tenha uma boa aceitação e ofereça respostas precisas aos usuário, é necessário que ele ofereça alternativas de escolha, é necessário que ele tenha multiplicidade nos seus mais variados sentidos. Por exemplo, se se quer que as informações tenham maior disponibilidade, ou que uma comunidade maior tenha acesso às informações, é necessário que utilizem não apenas um canal de comunicação ou um tipo de mídia, mas que estejam em vários canais para que possam atingir um grupo maior de usuários. Assim, esse sistema deve estar disponível não apenas na rede local de uma empresa, mas na Internet, em CD-ROM e em outros meios de difusão de informação. O fato de estar na Internet significa que o sistema

será utilizado não somente por um grupo local de pessoas, mas por pessoas de várias partes do mundo, dado que a Internet é uma rede mundial. Assim, é preciso que o sistema ofereça alternativas em termos de mensagens em vários idiomas, ou mesmo que as informações possam estar em mais de um idioma. O inglês é atualmente a língua oficial a nível internacional, e mais de 70% das informações na Internet estão em língua inglesa. Mas, de uma forma geral, os países querem preservar as suas culturas. Nesse sentido, é certo que eles (os países) vão colocar as informações tanto em inglês, quanto em sua língua oficial. Em resumo, certamente que um SRI plurilingüe seja desejável. É também uma questão de mercado.

Sob um outro ponto de vista, existe uma grande diversidade de usuários, muitos inexperientes (noviços), outros especializados (aqueles que já têm o hábito de fazer consultas ou recuperar informações em sistemas automatizados). Naquilo que concerne o domínio do conhecimento, existe também uma grande diversidade de interesse em informação em todos os domínios possíveis e imagináveis. Nota-se aí, um outro tipo de multiplicidade. Quanto aos domínios do conhecimento, é preciso que o SRI seja capaz de fazer buscas nos diversos domínios do conhecimento. Nesse sentido, a proposta ora apresentada se preocupa com essa questão, na medida em que se propõe que o desenvolvimento do software seja modular e adote uma arquitetura aberta, de forma a aceitar módulos de tratamento de informação adaptados a cada domínio do conhecimento.

Quanto ao problema de usuários noviços e usuários experientes, é necessário que o SRI ofereça alternativas, ou seja, que o SRI possua uma interface apropriada aos noviços e outra, aos experientes. A interface proposta, prevê esse tipo de facilidade, oferecendo, inicialmente, o acesso à informação a partir do centro de sintagma[4] de primeiro nível, o que deve facilitar as coisas para os usuários noviços. E para os usuários experientes, a idéia é oferecer o acesso aos sintagmas nominais de mais alto nível, como por exemplo de nível três ou quatro. Esta solução poupa o usuário de ter que passar pelos níveis um e dois.

3.6 A consistência

O significado da palavra consistência, do ponto de vista físico, é o grau de solidez de uma substância. Do ponto de vista figurado, abstrato é também a estabilidade. Entende- se, também, como sinônimo de coerência das coisas. A consistência é a base de um

[4] Centro de sintagma nominal, é o substantivo em tomo do qual se constitui um sintagma nominal. Exemplo: 1) em “a informação científica” o centro de sintagma é "informação", trata-se de um centro de sintagma nominal de primeiro nível; 2) em “o estudo da análise da informação científica”, “estudo” é de um centro de sintagma nominal de terceiro nível.

bom sistema de recuperação de informação. É preciso que um SRI seja coerente. A confiança num SRI depende de sua consistência. Para compreender melhor esse princípio, à luz da construção de um SRI, é necessário ver seus procedimentos e seus componentes. Em relação aos componentes, à nível da interface, é importante observar que as mensagens de ajuda *(Help)* sejam coerentes, sem ambiguidades. E, obviamente, que toda a interação entre o usuário e a interface seja realizada dentro de uma ordem lógica e direta, de maneira a não deixá-lo em dúvida ou indeciso. Com relação ao procedimento de indexação — extração de descritores e construção do índice — o resultado desse procedimento deveria ser um produto consistente, sem ambigüidades. No entanto, não é isso que se verifica nos sistemas tradicionais de recuperação de informação. Uma das razões que induziram à formulação dessa proposta, advém dos problemas encontrados nos procedimentos de recuperação de informação dos tradicionais SRI, onde os descritores, enquanto uma lista de palavras isoladas, provocam uma certa inconsistência. Para isso, deve-se adotar um enfoque que permita maior consistência ao conjunto de descritores resultantes do processo de indexação automática. O enfoque proposto aqui é mais consistente que aquele que utiliza as palavras ou palavras-chave como descritores, porque os sintagmas nominais são menos susceptíveis às ambigüidades encontradas no procedimento tradicional de indexação. Existe, no entanto, um outro tipo de ambigüidade que pode acompanhar os sintagmas nominais, aquela relativa à área do conhecimento. Essa pode ser eliminada através da separação dos documentos em bases de dados especializadas por domínios do conhecimento. Em função desse aspecto, os SRI que adotarem a presente proposta terão que ter módulos de tratamento de informação adaptados a cada domínio do conhecimento.

3.7 A interatividade

Trata-se do último princípio mas nem por isso o menos importante. A recuperação de informação não é uma tarefa exclusiva do computador ou de um SRI, mas certamente será melhor resolvida se existir maior interação entre o usuário e o SRI. Segundo o dicionário *Le Petit Robert,* a interação é a reação recíproca, ou seja, a interdependência. A recuperação de informação é uma tarefa dependente tanto do usuário quanto do SRI. É o usuário que possui dentro de sua cabeça a(s) sua(s) necessidade(s) de informação. Em contrapartida, é o SRI ou o computador que possui e controla as informações armazenadas em uma base de dados, bem como seus procedimentos de acesso à informação. Analisando os SRI tradicionais, percebe-se que eles funcionam como um sistema de pergunta e resposta. Quer dizer, o usuário coloca uma questão através de uma

expressão de busca e o SRI responde a esta questão, fornecendo a quantidade de referências encontradas, sem oferecer nenhuma interatividade entre o momento da solicitação do usuário e o do fornecimento do resultado. Por outro lado, colocar ou exprimir as necessidades de informação de um usuário, numa única expressão de busca, sem conhecer a base de dados, objeto da consulta, é uma tarefa difícil. Faz-se necessário conhecer as características da base de dados onde se deseja fazer a busca. Dentre as características importantes de serem conhecidas por um usuário citamos: a linguagem de indexação; o(s) domínio(s) do conhecimento dos documentos existentes na base de dados, a maneira como esta foi indexada. É bem verdade que atualmente existem SRIs que permitem ou oferecem a possibilidade de se reformular uma expressão de busca, apresentando os termos com a sua pertinência para o refinamento da busca. Ou então, solicitando um documento típico para que ele (o SRI) possa procurar documentos similares. A falta de interatividade é uma característica dos tradicionais sistemas de recuperação de informação, dado que, à época em que foram desenvolvidos, não existiam as facilidades de softwares e periféricos existentes hoje. A proposta que ora se formula neste artigo é baseada fortemente na interação entre o usuário e a interface de busca. Na realidade, é o usuário quem determina ou quem encontra aqüo que quer. A interface apenas procura e apresenta aquilo que o usuário solicita. Assim, é o usuário quem orienta a busca de informação. Ao contrário dos SRI tradicionais, a presente proposta não faz nenhum esforço de interpretação da expressão de busca do usuário.

4 CONCLUSÃO

O tema central deste artigo é a proposição de um novo sistema de recuperação de informação, um novo SRI com um enfoque totalmente diferente daquilo que se vê hoje em dia no mercado. Não se trata somente de uma nova interface, mas de uma nova técnica para o tratamento da informação, de um SRI completo baseado em sete princípios propostos como forma de orientar a construção desse sistema. Um sistema mais leve, mais preciso, mais rápido, mais visível, mais sólido e interativo. Considerando as limitações dos computadores e os problemas existentes no uso dos SRIs tradicionais, esta proposta poderá resolver os problemas colocados na introdução deste documento. O procedimento de indexação proposto consiste na extração dos sintagmas nominais e na sua indexação, como descritor, segundo uma estrutura em árvore. Em relação à interface de recuperação de informação, o enfoque proposto se baseia num procedimento de navegação na estrutura de sintagmas nominais. Como resultado, concebeu-se um SRI que oferece aos usuários a oportunidade de construir sua expressão de busca, de forma indireta, através da navegação na estrutura dos sintagmas nominais até o momento em que seja

encontrado o sintagma nominal que melhor satisfaça à necessidade de informação do usuário. Quando o usuário chega a esse termo, ele solicita à interface ver os documentos de onde o sintagma escolhido foi extraído. Esse procedimento difere dos SRIs tradicionais, onde a solicitação de informação é feita diretamente sem qualquer interação prévia. Esse tipo de procedimento dos SRIs tradicionais, não oferece aos usuários a oportunidade de bem formular as suas expressões de busca, dado que eles não sabem o que existe na base. Não existe um diálogo, nesses SRIs, capaz de dar ao usuário a noção do que existe na base de dados. Os SRIs, deveriam ser mais interativos, permitindo cada vez mais a participação do usuário no procedimento de recuperação de informação. A solução proposta poderá ajudar a tomar a tarefa de recuperação de informação mais interessante e satisfatória. De uma certa forma, propõe-se a construção de um Sistema de Recuperação de Informação Assistida por Computadores, SRIAC.

Segundo Pierre Lévy, nós estamos num novo espaço, o espaço do saber, onde se deve valorizar e sustentar as qualidades e competências do homem. Não se pode mais automatizar tudo, os homens são os únicos elementos que não são automatizáveis. Nesse contexto, a proposta apresentada se enquadra muito bem, porque nos apercebemos das dificuldades encontradas no uso dos SRI tradicionais e modificamos completamente o enfoque tradicional de tratamento e recuperação de informação. No enfoque tradicional, o usuário apenas faz a solicitação de informação através da expressão de busca, e então os únicos a trabalhar são o SRI e o computador. Não existe a mínima intervenção humana no processo de recuperação de informação tradicional. Talvez essa seja uma das razões da falta de precisão nos resultados desses sistemas. Contrapondo os sistemas tradicionais, propõe-se aqui colocar os usuários no procedimento de busca de informação onde eles participem ativamente de forma interativa com o SRI e o computador. É o usuário quem deve decidir a informação que melhor satisfaça à sua necessidade de informação.

5 REFERÊNCIAS BIBLIOGRÁFICAS

1 BOUCHÉ R., LAINÉ, Sylvie, METZGER, J.-P. Extraction de connaissances à partir d'une collection de documents. In.: *Tools of knowledge organization and the human interface,* Congrès organisé par l'ISKO (International Society for Knowledge Organization), Darmnstadt (D.), 14-17 de agosto de 1990.

2 CALVINO, Italo. *Leçons Américaines: aide-mémoire pour le prochain millénaire.* Gallimard. 1988. 197 p.

3 CORET, Annie, MENON, Bruno, SCHIBLER, Danielle, TERRASSE, Christophe. Un système d'indexation structurée à l'INIST. Bilan d'une étude préalable. *Documentaliste - Sciences de l'Information,* v. 31, n. 3. p. 148-158, 1994.

4 HARTER, Stephen P *Online Information Retrieval: Concepts, Principles and Techniques.*

Orlando : Academic Press, 1986. 259 p. (Library and Information Science).
5 LE GUERN, Michel. Les descripteurs d'un système documentaire: essai de définition. In. : Bès, G.C., Fauchère, P.M., Lagueunière, F. *Actes du Colloque Traitement automatique des langues naturelles et systèmes documentaires.* Condenser, supplement I.Université Clermont Ferrand, 1982. p. 163-169.
6 LE GUERN, Michel. Un analyseur morpho-syntaxique pour l'indexation automatique. *Le Français Moderne,* t. 59, n. 1, p. 22-35, Juin, 1991.
7 LEVY, Pierre. *L'Intelligence collective: Pour une anthropologie du cyberspace.* Paris : La découverte, 1997. 246 p.
8 NIE, Jian-Yun. Towards a Probabilistic Modal Logic for semantic-based Information Retrieval. *15th Annual International SIGIR.* 1992. p. 140-151.
9 POLITY Yotta. Evaluation des modes dé recherche en langage naturel. *Documentaliste - Sciences de l'Information,* v. 31, n. 3, p. 136-142, 1994.
10 SALTON, Gerard, MCGILL, Michael J. *Introduction to modern information retrieval.* New York : Mcgraw-Hill Book Company, 1983. 448 p. (Computer Science).
11 SHOW Guan Yeong, KONG, Hinny et LIN, Kenneth Wente. Intelligent user interface to SQL-based database system. *Engineering Application Artificial Intelligence.* v. 6, n. 4, p. 307-316, 1993.
12 SMEATON Alan F. Prospects for intelligent, languaged-based information retrieval. *Online Review,* v. 15, n. 6, p. 373-382, 1991.
13 SWANSON, Don R. Historical Note: Information Retrieval and the Future of an Illusion. *Journal of the American Society for Information Science, v. 38,* n. 2, p.92-98, 1988.
14 WANG, Fangju. Towards a natural language user interface: an approach of fuzzy query *International Journal of Geographical Information Systems, v. 8,* n. 2, p. 143-162,1994.

Informational Retrieval System Assisted by Computer: a proposal

Problems related to the use of System of Information Retrieval based on Italo Calvino's principles as proposed in his book Leçons Américaines: aide-mémoire pour le prochain millénaire

Key words: Information Retrieval Systems. Automatic indexing, information retrieval interface.

Hélio Kuramoto

Engenheiro Eletricista, formado pela Universidade de Brasília, especialista em informática, atualmente funcionário do IBICT/CNPq, licenciado para estudos de doutoramento em ciências da informação e da comunicação na Université Lumière - Lyon 2. Membro do CERSI - Centre d'études et recherche en sciences de reformation et de te communication da ENSSIB - École Nationale Supérieure en Sciences de l'Information et des Bibliothèques.

46, Boulevard du 11 Novembre 1918, Apto 2001
69100 - Villeurbanne - France

Tell.+33 4 78 8903 48

E-mail: kuramotirn@imaginet.fr