Ciência da Informação - Vol 25, número 2, 1995 - Artigos

Uma abordagem alternativa para o tratamento e a recuperação de informação textual : os sintagmas nominais

Hélio Kuramoto

Resumo

O presente trabalho propõe uma abordagem alternativa de tratamento e recuperação de informação, onde os sintagmas nominais desempenham o principal papel. Trata-se da construção de um protótipo, de sistema de recuperação de informação capaz de navegar uma estrutura em árvore de sintagmas nominais. Para exploração deste protótipo, foi utilizada uma amostragem de 15 artigos em língua portuguesa, de onde foram extraídos os sintagmas nominais. O processo de extração foi manual, simulando um processo automatizado.

Palavras-chave

Sintagmas nominais; Indexação automática; Sistemas de recuperação de informação.

INTRODUÇÃO

Durante os últimos anos, um volume crescente de informações tem sido registrado em várias bases de dados, nos mais diversos domínios do conhecimento e sob diversas formas (numéricas, textuais, imagens etc.). Considerando que os recursos informacionais estão cada vez mais acessíveis aos usuários finais, incluindo os pessoais, o principal problema é saber como acessar tais recursos de forma fácil e precisa. É neste sentido que se faz necessário utilizar os Sistemas de Recuperação de Informação (SRI)[1].

Constata-se, no entanto, mediante a literatura consultada e observações empíricas, que os usuários têm encontrado dificuldades no uso de tais sistemas, seja porque eles possuem uma interface pouco amigável, seja porque os resultados das consultas às bases de dados nem sempre são precisos.

Os sistemas tradicionais de recuperação de informação possuem interfaces, na sua maioria, orientadas a comandos que possuem regras rígidas de utilização e exigem dos usuários o domínio de um grande número de conhecimentos; alguns destes relacionados à informática, outros relacionados à estruturação dos dados e outros em função do vocabulário da área de especialização. Os SRI são sistemas profissionais e exigem, como toda ferramenta profissional, uma formação adequada para a sua utilização (Polity, 1994, p. 137-138). Dentre os elementos que devem ser necessariamente dominados para a utilização correta de um SRI, destacam-se:

- um certo número de comandos para se colocar o sistema em modo de consulta, para se escrever uma expressão de busca dentro de uma sintaxe correta, para visualizar o resultado de uma consulta etc.;

- a indispensável lógica booleana em uma consulta de múltiplos critérios a qual, para oferecer bons resultados, exige a utilização de operadores como intersecção, união, exclusão, comparação, de proximidade etc.;

[1] Harter define um sistema de recuperação de informação (SRI) como um dispositivo que se interpõe entre os usuários e a coleção de informação. Strzalkowski complementa esta definição, afirmando que tais sistemas têm como função típica a de selecionar documentos de uma base de dados, em resposta a uma questão do usuário, apresentando-os em ordem de pertinência, enquanto Salton e Mcgill conceituam técnicamente um SRI como um sistema que trata da representação, do armazenamento, da organização e do acesso aos ítens de informação.

- a estrutura conceitual da base de dados, os nomes de campos a consultar e as convenções de escrita de cada um destes campos;

- os termos de indexação, os léxicos, os tesauros etc.

Show *et alli* (1993, p. 307) observam que o crescimento repentino do número de usuários[2] nos vários setores de atividades colocou em prova a facilidade de utilização das interfaces de consulta dos sistemas tradicionais de informação. As dificuldades enumeradas acabam por limitar o crescimento do número de usuários.

É bem verdade que existem atualmente alguns SRI, principalmente aqueles distribuídos junto às bases de dados em CD-ROM, que se têm preocupado em amenizar as dificuldades citadas, introduzindo a utilização de menus, de janelas e de outras facilidades gráficas. Contudo, tais facilidades não atendem completamente aos anseios dos usuários.

Por outro lado, as questões relativas à precisão dos resultados de uma busca são ligadas às técnicas de indexação e de recuperação de informação. Uma busca de informação, nestes SRI, é quase sempre expressa sob a forma de uma combinação booleana de palavras-chave (eventualmente usando operadores de proximidade). Entretanto, estes mecanismos podem não ser eficazes, principalmente pelas seguintes razões:

- as palavras do texto podem ter sentidos ou significados diferentes conforme a área do conhecimento. Exemplo: *Humor* pode indicar, em fisiologia, uma quantidade de líquido contido em um corpo organizado ou, na linguagem do dia-a-dia, a capacidade de perceber, de apreciar e de expressar o que é cômico ou divertido;

- as mesmas palavras podem ser utilizadas em diferentes frases exprimindo conceitos diferentes em cada uma destas. Exemplo: as vítimas juvenis dos crimes; as vítimas dos crimes juvenis;

- palavras completamente diferentes podem ser utilizadas para exprimir o mesmo conceito. Exemplo: A expressão **os livros manuscritos** pode ser também expressa como **as obras escritas à mão.**

Segundo Smeaton (1991, p. 373), as abordagens convencionais[3] utilizadas na construção dos atuais sistemas de recuperação de informação são incapazes de resolver tais problemas.

Assim, visando a melhorar os SRI no que se refere aos aspectos apresentados, vêm-se desenvolvendo técnicas de tratamento e recuperação de informação, baseadas no Tratamento Automático da Linguagem Natural (TALN). É neste contexto que se apresenta a experimentação de uma abordagem alternativa de tratamento e de acesso à informação na qual os sintagmas nominais desempenham papel fundamental.

USO DOS SN NA INDEXAÇÃO DE UMA BASE DE DADOS TEXTUAL

Para que um SRI possa responder às demandas dos usuários com tempos de respostas aceitáveis, é preciso que os documentos constantes da base de dados sejam submetidos a um tratamento prévio. Este procedimento permite a extração dos descritores e sua estruturação com vistas a um acesso rápido às informações. Este tratamento preliminar consiste da indexação automática, no caso das bases de dados contendo textos completos. Por outro lado, no tratamento de bases de dados bibliográficos, existem alguns produtores que se utilizam de uma indexação mista (automática e manual), enquantos outros preferem realizar o tratamento usando apenas a

[2] O crescimento repentino do número de usuários de bases de dados se deve principalmente à popularização dos microcomputadores e às facilidades de acesso aos grandes centros distribuidores de informação por intermédio das redes públicas de computadores (Internet).

[3] Considera-se, neste contexto, abordagens convencionais aquelas que adotam a indexação baseada em palavras e as buscas de expressões booleanas.

indexação manual, que é realizada por técnicos especializados. Os produtos deste tratamento são palavras e/ou grupos de palavras chamados de descritores e, muitas vezes, de palavras-chave.

O processo de indexação produzindo uma lista de descritores visa à representação dos conteúdos dos documentos. Ou seja, este processo tem como objetivo extrair as informações contidas nos documentos, organizando-as para permitir a recuperação destes últimos. Assim, os descritores deveriam ser, obrigatoriamente, portadores de informação de maneira a relacionar um objeto da realidade extra-linguística com o documento que traz informações sobre este objeto. Contudo, na maioria dos SRI convencionais, os descritores não passam de uma simples lista de palavras extraídas dos documentos que constituem as bases de dados. Le Guern, em artigo publicado na revista Le Français Moderne (1984, p. 165-166) faz distinção entre palavra e descritor, conforme mostra a citação a seguir :

« Não constitui finalidade do descritor a sua visualização *mediante a 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."* [4]

Desta forma, os descritores deveriam fazer referência à realidade extralingüística do autor. Os descritores, se tidos como uma lista de palavras consideradas isoladamente, são elementos portadores apenas de propriedades. A palavra como uma unidade da língua constitui um conjunto de propriedades, mas sem referência à realidade extralingüística. Não se deve confundir significado com referente. As palavras passam a ter um valor referencial a partir do momento em que elas passam a fazer parte do universo do discurso.

Para exemplificar, considere-se a palavra **livro**. Isoladamente, ela possui um conjunto de predicados (segundo Aurélio Buarque de Hollanda Ferreira, em seu *Novo Dicionário da Língua Portuguesa*, **livro** é, dentre outras definições, uma reunião de folhas ou cadernos, soltos, cosidos ou por qualquer outra forma presos por um dos lados, e enfeixados ou montados em capa flexível ou rígida). Verifica-se que este termo não possui nenhuma relação com o universo do discurso. É a intervenção da lógica intensional, que se caracteriza por tratar de elementos sem referência e sem classe, constituídas apenas de propriedades. Denominam-se esses termos de predicados livres.

Por outro lado, o termo **livro de bolso** faz referência a uma classe de objetos do universo do discurso. Ambas as palavras, livro e bolso, consideradas isoladamente são palavras da língua, são conjuntos de propriedades. O relacionamento destas duas palavras constitui a passagem da lógica intensional para a lógica extensional. É a passagem do genérico para o específico. A lógica extensional trata dos elementos que estão em relação com o mundo, que fazem referência a uma classe de objetos, é a compreensão das coisas. Denominam-se esses termos de predicados complexos ou predicados relacionados. A oposição das lógicas intensional e extensional é a oposição entre a língua e o discurso.

Acrescentando-se um determinante (artigo definido) ao termo, **o livro de bolso**, este passa a fazer referência precisa a uma classe de objetos do qual se trata no universo do discurso. Em outras palavras, procedeu-se a um fechamento lógico. Este fechamento caracteriza o que se denomina de sintagma nominal (SN)[5], que é a menor parte do discurso portadora de informação.

---

[4] « La finalité du descripteur exclut qu'on puisse l'envisager en faisant abstraction de la valeur référentielle de ses occurrences dans le corpus. Les mots de la langue, en tant qu'ils sont mots de la langue, ne signifient que des attributs, et non des substances, tant qu'ils ne sont pas mis en oeuvre dans le discours. Le descripteur, quant à lui signifie une entité au sens de la philosophie d'Aristote. Le descripteur ne peut donc pas être considéré, à l'instar des mots de la langue comme un symbole sans référence. »

[5] Na língua portuguesa nem sempre um sintagma nominal é precedido de um determinante, enquanto, na língua francesa, a ausência de determinantes é uma rara exceção. Isto pode ser explicado pela inexistência na língua portuguesa de artigos partitivos e também pela não utilização de artigos precedendo núcleos de sintagmas nominais abstratos. Essa constatação dificulta a identificação e extração dos SN, razão pela qual realizamos o

Neste caso, a palavra **livro** constitui o centro do SN, ou o núcleo do SN; e **bolso** caracteriza a identificação de uma classe de livros.

Assim, em um SRI, os descritores deveriam ser representados pelos SN, que é o fundamento teórico dos trabalhos conduzidos pela equipe SYDO[6].

No entanto, a simples substituição dos descritores pelos SN em um SRI convencional não resolve completamente os problemas discutidos na Introdução. É importante ressaltar que o SN possui um potencial natural de organização (existência de relações de encadeamento entre um SN de um dado nível com um outro de nível imediatamente superior) que, se explorado convenientemente, propiciaria aos usuários maior facilidade no uso de um SRI e resultados mais precisos em resposta a um processo de busca de informação.

A título de exemplo, considere o seguinte SN:

AS CARACTERÍSTICAS DO MEIO AMBIENTE DO MUNDO DOS NEGÓCIOS

SN 1: os negócios

SN 2: o mundo dos negócios

SN 3: o meio ambiente do mundo dos negócios

SN 4: as características do meio ambiente do mundo dos negócios

O SN, como apresentado, mostra o seu potencial de estruturação por meio de relações de encadeamento. A análise do sintagma nominal, no exemplo, permitiu a extração do SN — **o meio ambiente do mundo dos negócios**. A partir deste SN, pode-se visualisar um outro SN embutido — **o mundo dos negócios** — que, por sua vez, possui um quarto SN — **os negócios** — que representa o nível mais inferior[7]. Percebe-se, neste exemplo, a existência de quatro SN encadeados que, enumerados em ordem crescente (do SN mais simples ao mais complexo), levam à classificação do SN original como sendo de nível 4 (SN 4).

Com base nas características apresentadas, os SN podem ser organizados sob uma estrutura arborescente. Desta maneira, o SRI, atendendo a uma demanda do usuário, feita mediante o fornecimento de um centro de SN de seu interesse, (por exemplo : **negócios**), apresentaria todos os SN 1 relativos a esta demanda, inclusive o SN **os negócios**. Uma vez conhecida a lista de SN 1, o usuário poderá restringir o seu perfil de busca escolhendo um SN 1, por exemplo, **os negócios**, e solicitar os SN 2 relacionados a este SN 1. O SRI apresentaria todos os SN 2, inclusive o SN **o mundo dos negócios**. O processo de refinamento é realizado por meio da

---

estudo « L'omission d'article dans le discours dans la langue portugaise », visando a identificar um procedimento que permita a identificação e extração automática dos SN.

[6] O grupo SYDO (Systèmes Documentaires) foi constituído pelos seguintes centros de pesquisa: Laboratoire d'Informatique Documentaire (LID) da Université Claude Bernard (prof. R. Bouché), Centre de Recherches Linguistiques et Sémiologiques (CRLS) da Université Lumière Lyon II (prof. M. Le Guern), Département de Linguistique da Université de Fribourg (Suisse) (prof. A. Berrendonner), Centre de Recherches en Informatique et Sciences Sociales (CRISS) da Université de Grenoble II (prof. J. Rouault).

[7] Os sintagmas nominais, à medida que são extraídos de um outro SN, são classificados por níveis. Assim, o sintagma mais simples é denominado SN de nível 1. Constitui SN de nível 2 aquele a partir do qual foi extraído o de nível 1 e, assim sucessivamente. Os SN em seus vários níveis serão representados, no contexto deste artigo, na forma seguinte: SN 1 = sintagma nominal de primeiro nível, SN 2 = sintagma nominal de segundo nível etc.

passagem pelos vários níveis de uma estrutura arborescente de SN[8], dado que o SN vai se tornando mais específico à medida que se atingem os níveis mais elevados da estrutura. Ao percorrê-la, o usuário está, na realidade, delimitando, ou melhor, qualificando a sua necessidade de informação. Cabe, portanto, ao usuário identificar o nível em que as suas necessidades de informação serão atendidas.

METODOLOGIA PARA EXPERIMENTAÇÃO DOS SN

Com vistas à experimentação do uso dos SN, desenvolveu-se, no âmbito do curso de DEA[9] da École Nationale Supérieure en Sciences de l'Information et des Bibliothèques (ENSSIB), um protótipo de interface de recuperação de informação utilizando-os como meio de acesso à informação. Este trabalho contou com as seguintes etapas:

1) constituição de uma amostragem de 15 artigos — extraídos da revista *Ciência da Informação* — escritos em língua portuguesa;
2) construção de uma interface de busca utilizando-se de um sistema de gerência de bancos de dados relacionais;
3) estruturação dos SN e exploração da base de dados mediante o protótipo desenvolvido.

A etapa de constituição da amostragem de artigos e de extração dos SN precedeu as outras, com vistas a obter um melhor conhecimento do seu comportamento, pré-requisito para o desenvolvimento do protótipo.

CONSTITUIÇÃO DA AMOSTRAGEM DE ARTIGOS

Para constituição da amostragem de artigos, selecionaram-se documentos tendo como tema central a ciência da informação. A importância da definição de um domínio para a amostragem de artigos está estritamente relacionada com o critério de obtenção de menor ocorrência de ambigüidades entre os SN. Segundo Minsky:

" Na linguagem natural, as ambigüidades não advêm apenas do fato de que as palavras podem ser regrupadas de diversas maneiras, mas ainda do fato de que cada palavra pode ter diferentes sentidos..."[10]

Assim, mesmo um SN simples, constituído apenas de um determinante e de um substantivo –por exemplo os sistemas, os dados, as informações, a rede etc.– é passível de ser ambíguo. Portanto, a construção de uma base de dados textual contendo documentos pertencentes a um só domínio do conhecimento poderá diminuir, ou mesmo evitar, a ocorrência de ambigüidades proporcionando melhor precisão aos resultados de uma busca, o que tornará os SRI mais eficazes.

Os níveis dos SN determinam a altura da sua estrutura arborescente. Nesta proposta, foram escolhidos artigos (documentos) que possuíssem SN contendo até cinco níveis. A adoção deste

---

[8] Constatou-se empiricamente, utilizando a maquete desenvolvida nesta experimentação, que a quantidade de SN de segundo nível em relação a um dado SN de primeiro nível pode ser maior que o total de SN de primeiro nível. Por exemplo : a resposta à demanda do centro de SN informação foi de 122 SN de primeiro nível e a resposta à demanda do SN de primeiro nível a informação foi de 172 SN de segundo nível. Por outro lado, verificou-se que o SN a informação indexava 15 documentos na base, enquanto o SN de segundo nível a análise da informação indexava apenas 1 (um) documento. Confirma-se, neste exemplo, que a passagem de um dado nível a um superior na árvore de SN proporciona maior refinamento no processo de seleção dos documentos.

[9] DEA - Diplôme d'Études Approfondies

[10] Marvin MINSKY. *Semantic Information Processing*. Cambridge, Mass. : M.I.T. Press, 1969, p. 18, citado por Hubert L. Dreyfus no seu livro *Intelligence Artificielle : mythes et limites*

critério é necessária para que o protótipo possa construir uma estrutura arborescente capaz de oferecer ao usuário facilidades de refinamento das buscas.

EXTRAÇÃO DOS SINTAGMAS NOMINAIS

A extração dos sintagmas nominais foi realizada de forma manual, simulando uma extração automática. Este procedimento foi adotado em função da não-existência ainda de um sistema de extração automática de SN em acervos contendo documentos em língua portuguesa.

A semelhança da estrutura sintática das línguas francesa e portuguesa permitiu a utilização de algumas das regras básicas de reescritura dos SN adotadas para a língua francesa.

A tarefa de extração dos SN não foi trivial, pois estes nem sempre se apresentavam de forma clara. A ocorrência, normal, em todo texto em linguagem natural, de anáforas[11] e de elipses[12], dificultou a identificação dos SN. Deve-se observar, contudo, que estas dificuldades serão maiores em um processo automatizado. Estas e outras dificuldades encontradas no procedimento de extração dos SN serão discutidas em seguida:

a) Sintagmas nominais escondidos em frases com fatoração

Entendem-se por frases com fatoração aquelas que contêm uma seqüência de palavras que precedem um outro conjunto de palavras coordenadas pelas conjunções **e / ou**. A dificuldade maior na identificação dos SN em frases desse tipo é devida ao fato de que eventualmente o SN é constituído de toda a frase, ou, às vezes, a mesma poderá dar origem a vários SN. Assim, por exemplo: **o processo de negociação dos setores privado e público.**

Percebe-se claramente, neste exemplo, que o SN de nivel 1 é **os setores privado e público**, porque a palavra **setores** está no plural fazendo referência aos dois adjetivos simultaneamente — **privado e público** — que se encontram no singular.

Em contrapartida, encontraram-se frases sem evidências sintáticas capazes de indicar, como no último exemplo, os verdadeiros SN. Nestes casos, optou-se pela combinação dos termos. Assim, a frase como um todo constituia um dos SN, bem como as resultantes da combinação de cada palavra coordenada com o complemento. Exemplo: **a análise, interpretação, avaliação e comunicação da informação pelos meios convenientes.**

Como resultado da combinação, obtiveram-se os seguintes SN:

- a análise da informação pelos meios convenientes;
- a interpretação da informação pelos meios convenientes;
- a avaliação da informação pelos meios convenientes;
- a comunicação da informação pelos meios convenientes;

Existe igualmente uma outra forma de frases com fatoração na qual as palavras coordenadas aparecem entre parênteses, como um complemento combinatório do termo, ou da frase que precede o parêntese. Exemplo: **profundas transformações (políticas, econômicas, sociais, tecnológicas).**

---

[11] Em lingüística, segundo Ducrot e Todorov (1972, p. 358), um segmento do discurso é dito anafórico quando, para interpretá-lo (inclusive do ponto de vista literário), for necessário se reportar a um outro segmento do mesmo discurso. Utilizaremos, neste artigo, o termo fonte da **anáfora** para designar o segmento do discurso ao qual o segmento anafórico se reporta.

[12] A figura de sintaxe **elipse** é definida por Cunha e Cintra (1991, p. 613) como sendo a omissão de um termo que o contexto ou a situação permitem facilmente suprimir. Enquanto, Ducrot e Todorov (1972, p. 354) definem a **elipse** como sendo a supressão de um dos elementos necessários a uma construção sintática completa.

A combinação dos termos teve como produto os seguintes SN:

- profundas transformações políticas;
- profundas transformações econômicas;
- profundas transformações sociais;
- profundas transformações tecnológicas;

No entanto, é preciso prestar atenção na construção entre parênteses, pois nem sempre esta solução pode ser adotada. Nos casos de construções explicativas ou siglas, o sistema deverá extrair os SN como se fosse em uma outra frase independente.

b) Artigo zero

De forma diferente da língua francesa, na qual são raros os SN com ausência de um determinante, na língua portuguesa é frequente a ausência de determinantes. Razão pela qual algumas regras estabelecidas para a língua francesa não foram utilizadas.

No procedimento de extração dos SN, constatou-se que 28,89% dos SN não eram precedidos de qualquer determinante. Em uma amostra de 6010 SN, 1736 SN não são precedidos por nenhum determinante. Percebe-se que a inexistência de determinantes nos SN em língua portuguesa é relativamente grande.

Uma das diferenças entre as duas línguas é que na língua portuguesa não existe o artigo partitivo. No francês, diz-se **je mange de la viande**, enquanto que em português diz-se **eu como carne**. Esse fato pode ser uma das causas para o elevado índice de ausência de artigos nos SN extraídos de documentos escritos na língua portuguesa. Um outra constatação observada no processo de extração dos SN é que boa parte da ausência é verificada quando o centro do SN é um substantivo abstrato ou um substantivo no plural. Exemplo: a acumulação de **riquezas**, a adoção de **estratégias** etc.

A construção de um sistema automatizado de extração dos SN em acervos contendo documentos em língua portuguesa exigirá o estabelecimento de regras que busquem solucionar tal problema. Um estudo com este objetivo já foi realizado[13].

c) Cálculo das anáforas

Os elementos anafóricos, em português, aparecem freqüentemente mediante partículas como os pronomes.

Encontrou-se certa dificuldade na extração de SN contendo anáforas. Nos casos em que a fonte da anáfora se encontrava próxima, dentro do mesmo parágrafo, foi possível resolvê-lo com relativa facilidade. Entretanto, quando a fonte da anáfora se encontrava nos parágrafos precedentes, a tarefa de extração dos SN tornava-se bastante difícil. Mesmo assim, tentou-se identificar os verdadeiros SN.

No entanto, não foi possível resolver dois casos de anáforas:

- aquelas sem fonte explícita no texto, tais como **nesse sentido** (em que sentido?), **nossa experiência** (de quem? do autor? dos técnicos de informação?) etc. A dificuldade na solução

[13] KURAMOTO, Hélio. *L'omission d'article dans le discours dans la langue portugaise.* Rapport d'Etudes. 1996. 17 p.

desse tipo de anáfora é devida ao fato de que a fonte se encontra na interpretação das idéias contidas no documento;

- o segundo caso de anáfora não resolvido é constituído de termos cujas fontes se encontram na história dos acontecimentos, tais como **esse período pré-industrial**, **esse sistema de comunicação** etc. Nestes casos, tais SN foram extraídos na forma como se encontravam no texto.

Os SN resultantes da solução das anáforas são construções repetitivas como se pode notar pelo seguinte exemplo:

- uma categoria de clientes conscientizados dos **seus** direitos a produtos e serviços de alta qualidade

cuja solução é:

- uma categoria de clientes conscientizados dos direitos dos **clientes** a produtos e serviços de alta qualidade

Uma maneira elegante de resolver tal problema é substituir os elementos anafóricos apenas no momento em que a extração dos SN o envolvesse. Por exemplo:

SN 6 : uma categoria de clientes conscientizados dos seus direitos a produtos e serviços de alta qualidade

SN 5 : clientes conscientizados dos seus direitos a produtos e serviços de alta qualidade

SN 4 : **os direitos dos clientes** a produtos e serviços de alta qualidade

SN 3 : *produtos e serviços de alta qualidade*

SN 2 *: produtos de alta qualidade*

SN 2 *: serviços de alta qualidade*

SN 1 : *alta qualidade*

SN 1 : os clientes

d) Cálculo das elipses

O problema ligado a este tipo de figura de sintaxe depende da capacidade de percepção da falta de alguma palavra no contexto de uma frase. É preciso, para identificá-la, analisar não somente as frases precedentes, mas também as frases seguintes. Exemplo:

- uma visão de longo prazo que assegure não só a sobrevivência (?), como também o crescimento da organização

Qual o complemento do termo **sobrevivência**. **Sobrevivência** de quem? A solução encontrada estava na frase seguinte: **o crescimento da organização**.

O SN completo é:

- uma visão de longo prazo que assegure não só a sobrevivência da organização, como também o crescimento da organização

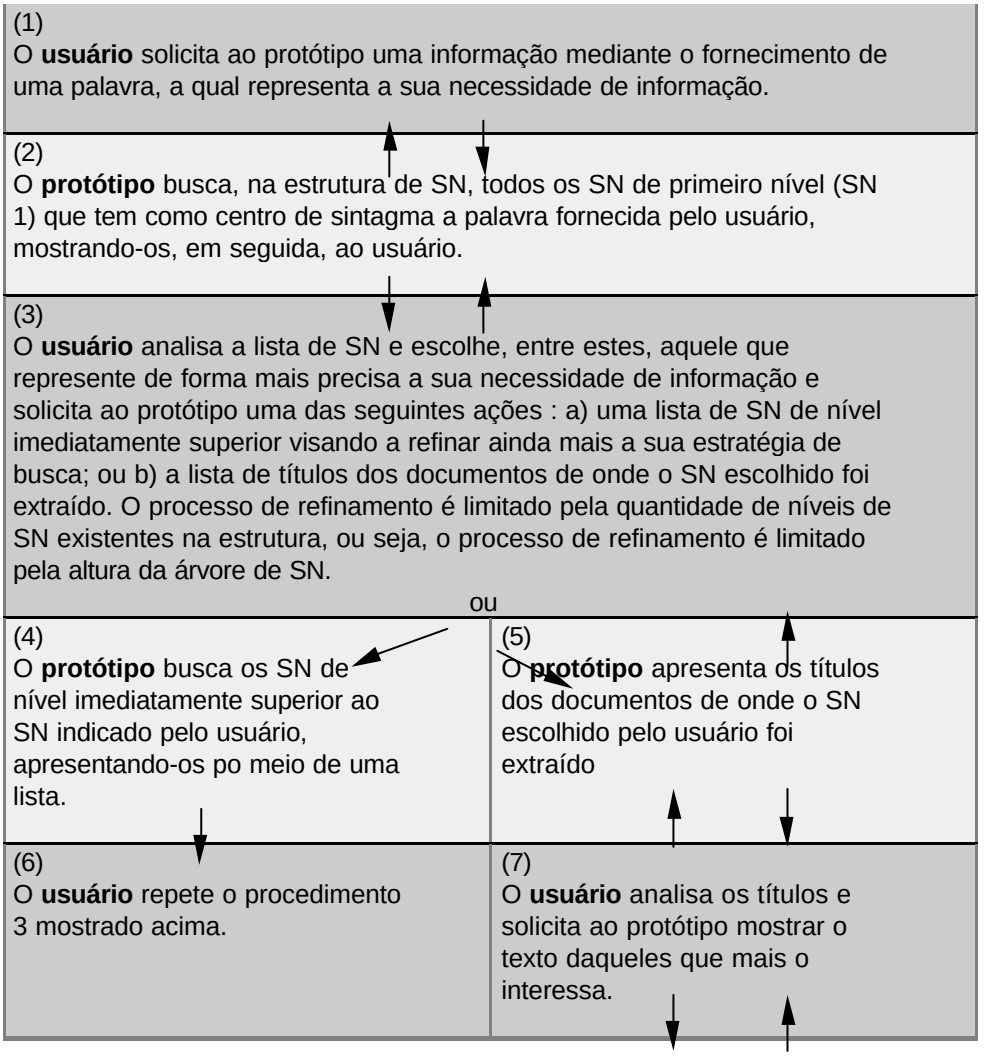
Em um procedimento manual, a solução é sempre muito simples, tendo em vista a lógica e a percepção de quem está extraindo os SN. Contudo, em um procedimento automático, a solução não parece ser tão trivial.

O procedimento de extração dos SN produziu um conjunto de 8.818 SN. Dada a extensão das estatísticas dos sintagmas nominais, suas análises e suas tipologias, estes itens não serão mostrados nesse artigo. No entanto, o leitor poderá encontrar tais elementos na referência bibliográfica (Kuramoto, 1995).

CONSTRUÇÃO DO PROTÓTIPO: DESENHO DA INTERFACE DE BUSCA

O desenho da interface de busca baseou-se no encadeamento hierárquico existente entre os SN, conforme discutido na seção 2 deste artigo. A idéia básica foi construir uma interface capaz de varrer a estrutura arborescente dos SN de maneira interativa com o usuário. A figura 1 descreve de maneira esquemática esta interação usuário-protótipo. A interação começa pelo protótipo, que torna disponível uma tela onde o usuário faz a sua solicitação de informação mediante o fornecimento de uma palavra (centro de SN 1) que represente a sua necessidade de informação. A partir deste ponto, o esquema da figura 1 mostra toda a interação usuário-protótipo. A forma escolhida para a interação entre o usuário e a interface de busca do protótipo foi aquela orientada por menus, em função das facilidades de implementação e de ajustes. No entanto, nada impede o desenvolvimento de um SRI orientado a comandos em linguagem natural. A idéia básica era desenvolver uma interface capaz de criar os menus dinamicamente à medida que os SN fossem recuperados e mostrados na tela. Esta apresentação constitui o próprio menu, a partir do qual o usuário faz a escolha do SN que mais lhe interessa e solicita ao sistema mostrar os SN relacionados de nível imediatamente superior, ou apresentar os documentos dos quais o SN escolhido foi extraído.

FIGURA 1.
Procedimentos de interação usuário Û protótipo



	(8) O **protótipo** mostra o documento de acordo com a escolha do usuário.

Para o desenvolvimento do protótipo, utilizou-se o sistema Microsoft Access 2.0. Trata-se de um sistema gerenciador de bancos de dados relacionais. A escolha deste sistema foi feita considerando-se mais os aspectos de rapidez e facilidade de desenvolvimento/ajustes oferecidas pelo mesmo, do que pela sua velocidade de processamento. Sabia-se, previamente, das dificuldades de manipulação de textos por este tipo de sistema e, como conseqüência, do baixo desempenho em termos de tempo de resposta. Em contrapartida, o Access oferece facilidades de criação de interfaces gráficas totalmente compatíveis com o sistema Windows, portanto, com todas as facilidades de criação de janelas, botões etc.

ORGANIZAÇÃO DOS SINTAGMAS NOMINAIS COMO ESTRUTURA DE BUSCA

Os sistemas de gerência de bases de dados, como o Access, têm a vantagem de já ter embutido, em seu escopo, um conjunto de métodos de acesso necessários à implementação de qualquer sistema. Tendo em vista que o Access é um sistema de gerência de bancos de dados relacionais[14], construiu-se um modelo de dados considerando as características dos SN observadas durante a sua extração. De forma geral, o modelo de dados parte da existência de duas entidades : os artigos (documentos da base de dados) e os sintagmas nominais. As características dos SN podem ser enumerados como seguem:

1) os artigos (documentos da base de dados) são enumerados seqüencialmente a partir de 1;

2) para cada artigo, os parágrafos são enumerados seqüencialmente a partir de 1;

3) a cada artigo corresponde um título de um tamanho aproximado de 256 caracteres;

4) a cada artigo corresponde um texto de tamanho maior que 256 caracteres;

5) um mesmo sintagma nominal pode aparecer em vários artigos;

6) um mesmo sintagma nominal pode aparecer em vários parágrafos de um determinado artigo;

7) os sintagmas nominais podem ser classificados em até cinco níveis, assim definido no escopo do presente trabalho;

8) um mesmo sintagma nominal pode ser classificado em mais de um nível;

9) existe associação entre os sintagmas nominais de um determinado nível com outros de um nível imediatamente superior;

10) um sintagma nominal pode ser associado a vários sintagmas nominais de nível imediatamente superior;

11) existe um centro de sintagma nominal associado a cada sintagma de primeiro nível;

12) os sintagmas nominais de primeiro nível tem como associação de nível inferior os centros de sintagmas;

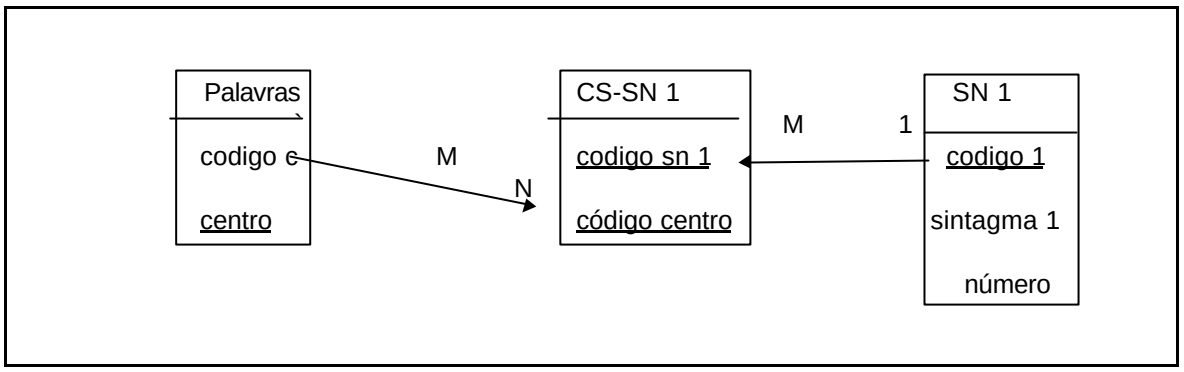
13) os sintagmas nominais são associados a cada artigo e a cada parágrafo de onde foram extraídos.

[14] Os sistemas de gerência de bancos de dados relacionais são baseados na abordagem de modelos de dados relacionais, os quais possuem uma base teórica já bem conhecida dos técnicos de informática e de parte dos técnicos de informação. Assim, não será tratada, neste artigo, a base teórica deste modelo. No entanto, os leitores poderão encontrar, na seção Bibliografias, algumas publicações sobre este tema.

Da análise das características citadas, foram desenvolvidas as várias tabelas[15] e associações necessárias para a construção da estrutura arborescente dos SN. Devido ao fato de a descrição das tabelas ser longa, não cabe, neste artigo, apresentá-la. No entanto, serão mostrados diagramas que ilustram como foi implementada a estrutura arborescente dos SN. Estas estruturas constituem os métodos de acesso às informações mediante a varredura dos SN nos seus vários níveis hierárquicos. A estrutura necessária para acessar os SN 1 a partir da demanda de um usuário, por meio de uma palavra, é mostrada na figura 2.

FIGURA 2

Estrutura de dados para acessar os SN 1 a partir de uma palavra



Na figura 2, tem-se a associação das tabelas Palavras, CS-SN 1 e SN 1. Os nomes dos elementos de dados sublinhados representam as chaves de cada tabela. Na tabela Palavras, estão agrupados todas as palavras (**centro**) que representam os centros de SN 1. Para cada **centro** foi atribuído um código chamado **código c**. Na tabela CS-SN 1, tem-se a associação dos códigos dos centros de SN 1 com os códigos dos SN 1. Esta figura mostra que para cada centro de SN 1 existem vários SN 1. A indicação na seta da associação da tabela Palavras com a tabela CS-SN 1 define que, na tabela Palavras, podem existir M ocorrências de um código de centro de SN 1. O mesmo pode ocorrer na tabela CS-SN 1, em que este código pode verificar-se N vezes. Esta indicação traduz a idéia de que para cada SN 1 pode existir mais de um centro de SN 1. Isto se explica pela existência, no contexto de um SN, de palavras que são tão importantes quanto o centro de sintagma.

Exemplo : **o sistema de informação.** Neste SN 1, o centro de sintagma é **sistema**. Entretanto, a palavra **informação** tem tanta importância quanto o próprio centro de sintagma. Em casos semelhantes a estes, adotou-se o procedimento de também associar estas palavras como sendo centro de sintagma, o qual denominamos também de centro complementar de SN.

A figura mostra também que cada centro de SN 1 pode estar associado a mais de um SN 1. Esta indicação é dada pela seta que associa a tabela SN 1 à tabela CS-SN 1, onde o número 1 significa que, na tabela SN 1, existe uma só ocorrência de um determinado código de SN 1, enquanto, na tabela CS-SN 1, existem M ocorrências deste código.

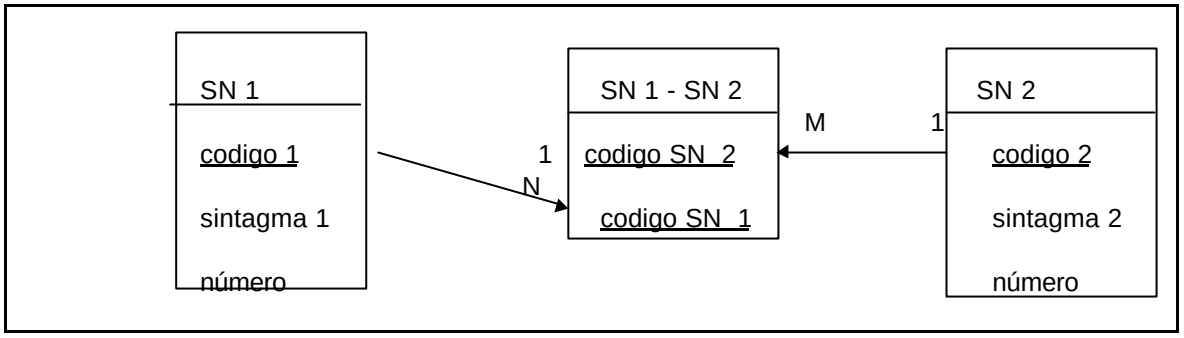
Na tabela SN 1, existe um terceiro elemento de dados chamado **número**, que indica a quan-tidade de artigos de onde um determinado SN 1 foi extraído. Na apresentação de cada SN 1 relacionado

[15] Denomina-se ***tupla*** um conjunto de valores representando cada elemento de dados. Exemplo: <5, a informação, 7, 8> Este conjunto de valores corresponde a uma tupla, onde **5** é o código do registro, **a informação** é um valor atribuído ao elemento de dados sintagma nominal, **7** corresponde ao número do artigo de onde este SN foi extraído e **8** ao número do parágrafo de onde foi extraído o referido SN. Um conjunto de tuplas denomina-se ***relação***. A ***relação*** é representada nas implementações dos sistemas gerenciadores de bancos de dados relacionais por ***tabelas***, na qual as colunas representam os atributos ou elementos de dados e as linhas formam as tuplas. No contexto deste trabalho, será usado preferencialmente o termo ***tabelas*** para representar o conjunto de ***tuplas.***

com um centro de SN 1, escolhido pelo usuário, é apresentado também o número de referências de onde o referido SN foi extraído.

A figura 3 ilustra a estrutura de dados construída para a busca dos SN 2 a partir de um SN 1 selecionada pelo usuário.

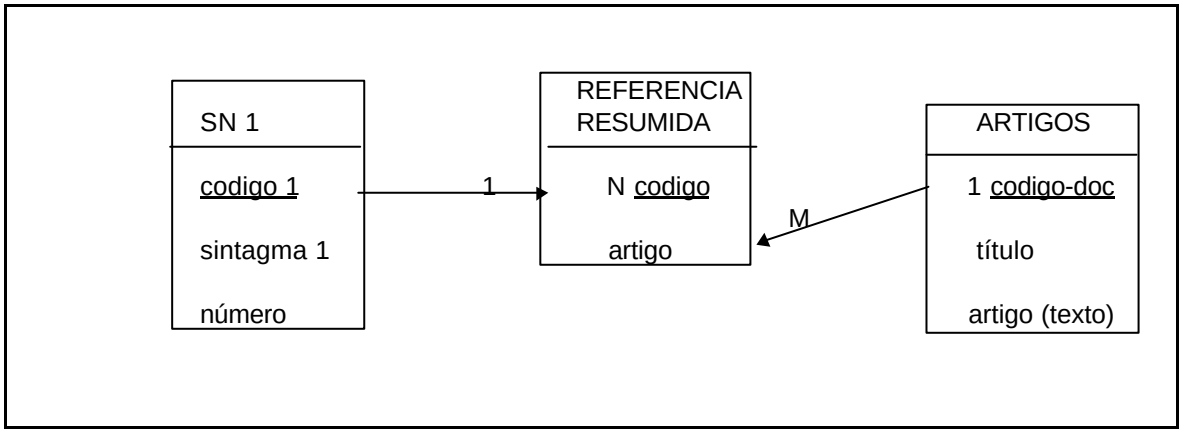
FIGURA 3
Estrutura de dados para acessar os SN 2 a partir de um SN 1



Nesta ilustração, de maneira análoga à da figura 2, construiu-se uma associação de tabelas de forma a facilitar a busca dos SN 2 a partir de um SN 1 escolhido pelo usuário. Percebe-se, analogamente, que um dado SN 1 pode estar associado a vários SN 2 e vice-versa. Isto traduz a idéia de que um SN 2 pode ter embutido mais de um SN 1. Esta estrutura atende às características dos SN listadas no início desta seção.

As estruturas necessárias para varrer os SN de nível 3, 4 e 5 são semelhantes àquelas apresentadas nas figuras 2 e 3. A única diferença com relação a estas duas figuras é o nome de cada elemento de dados que corresponde àqueles relativos aos SN de nível 3, 4 e 5.

FIGURA 4
Estrutura de dados para o acesso aos títulos e textos dos artigos



A figura 4 ilustra uma estrutura de dados construída para acessar o texto do artigo a partir da escolha de um dado SN, no caso um SN 1. Esta estrutura permite ao protótipo atender a uma demanda do usuário no sentido de visualizar todos os títulos e textos dos artigos de onde um SN 1 foi extraído. Outras associações semelhantes a estas da figura 4 foram criadas para o acesso aos documentos a partir de SN de qualquer um dos cinco níveis previstos no protótipo. Para evitar uma descrição repetitiva do modelo de dados, não serão apresentadas as outras associações.

É importante observar que todas as tabelas contendo os SN nos seus vários níveis têm como chave de acesso um código numérico único de SN. Para tanto, construiu-se uma tabela contendo os SN onde estes são identificados por meio de um código numérico. Não existe nenhum impedimento técnico por parte do sistema Access quanto ao uso do próprio texto dos SN como chave de acesso

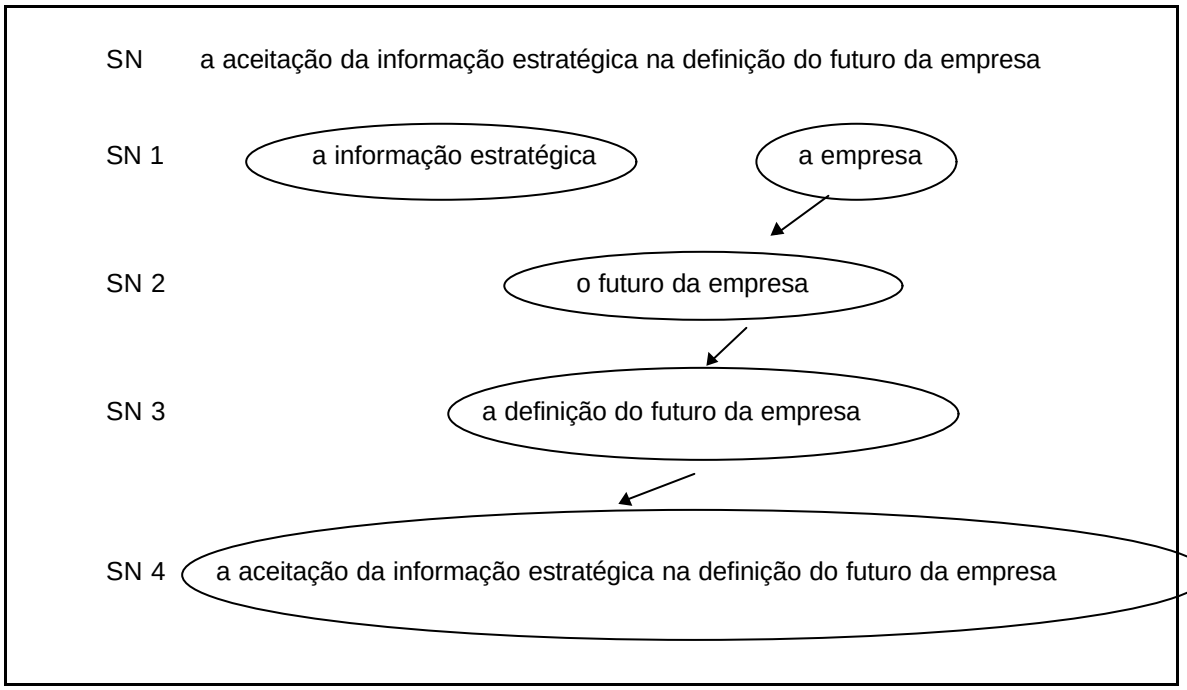
às informações. Deve-se ressaltar que, apesar da lentidão que este tipo de chave de acesso provoca, as estruturas de dados seriam mais simples e fáceis de manusear. Contudo, optou-se pela utilização das chaves numéricas identificando cada SN com o intuito de obter maior velocidade de acesso aos SN e às informações.

EXPLORAÇÃO DO PROTÓTIPO

O protótipo foi desenvolvido em duas fases. Na primeira fase, consideram-se as características observadas na extração dos SN. Nesta fase, além da implementação do protótipo, fez-se a carga de cinco artigos juntamente com a construção da árvore de SN relativos a estes artigos. A construção deste protótipo inicial permitiu a observação de algumas características dos SN não percebidas durante o procedimento de sua extração. Construiu-se um segundo protótipo com base nestas observações. Neste ultimo protótipo, foi inserido todo o acervo de 15 artigos e estruturada a árvore de SN. A primeira observação realizada com a experimentação do primeiro protótipo diz respeito à forma como foram construídas as associações entre os SN. A abordagem inicial utilizada na estruturação dos SN considerava que os níveis dos SN fossem absolutos e hierarquizados, ou seja, um SN de primeiro nível somente se associaria a um SN de segundo nível e este somente poderia se associar a um SN de terceiro nível e assim sucessivamente. Verificou-se, contudo, ao explorar o protótipo inicial, que existem SN portadores de uma característica especial: a existência de SN complementados por dois ou mais SN 1 distintos.
O problema encontrado neste tipo de SN é que ele é classificado considerando-se a quantidade de SN encadeados. Como conseqüência, o nível do SN é determinado pelo número máximo de SN embutidos. A título de exemplo, considere o SN da figura 5.

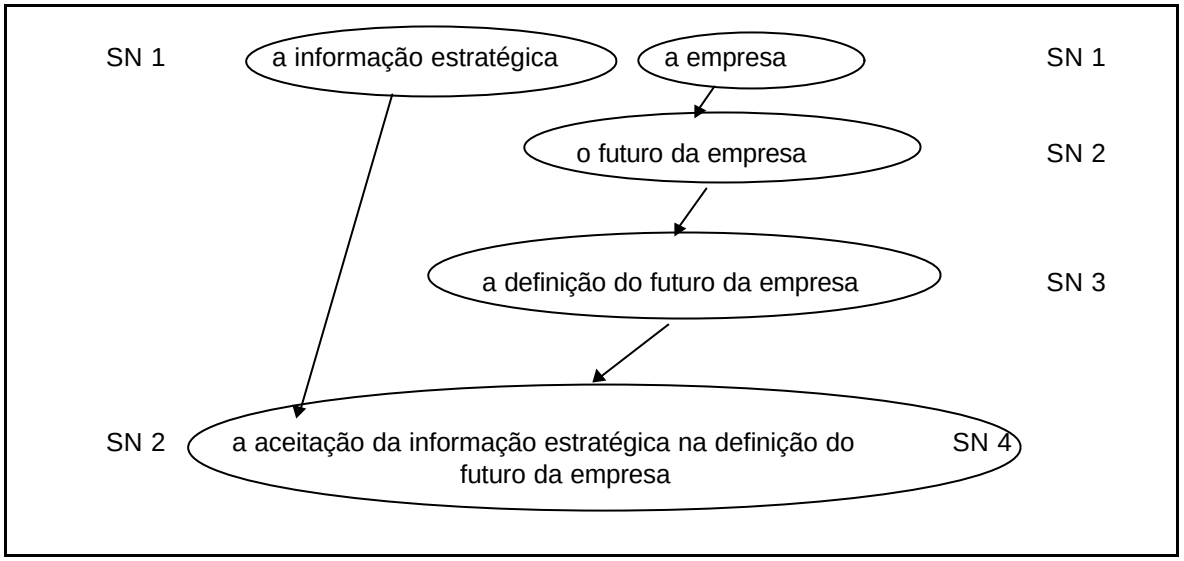
A análise deste SN permite a extração de dois SN de primeiro nível: **a informação estratégica** e **a empresa**. O SN **a empresa** é extraído do SN de segundo nível, **o futuro da empresa**, que, por sua vez, é extraído do SN de terceiro nível, **a definição do futuro da empresa**, que é extaído do SN original, **a aceitação da informação estratégica na definição do futuro da empresa**, de quarto nível. Por outro lado, o SN de primeiro nível **a informação estratégica** é extraído também diretamente do mesmo SN considerado de quarto nível. Percebe-se na figura 5 que a partir do SN **a empresa** se pode chegar até o SN de quarto nível. Entretanto, o mesmo não acontece utilizando o SN **a informação estratégica**.

FIGURA 5
Estrutura em árvore de SN considerando os níveis de forma absoluta



A solução adotada para este caso foi a utilização de uma classificação relativa dos SN, conforme mostra a figura 6.

FIGURA 6
Estrutura em árvore de SN considerando a relatividade dos seus níveis



Desta maneira, o SN original é considerado de quarto nível em relação ao SN **a empresa**. E o mesmo SN é classificado como sendo de segundo nível em relação ao SN **a informação estratégica**. Esta solução não exigiu nenhuma modificação na estrutura de dados do protótipo, mas uma simples modificação nos procedimentos de carga e construção da árvore dos SN. Em resumo, o SN **a aceitação da informação estatégica na definição do futuro da empresa** foi inserido tanto na estrutura contendo os SN 2, quanto na estrutura contendo os SN 4.

Uma segunda observação diz respeito à definição dos centros de SN de primeiro nível. Considerou-se como centro de SN de primeiro nível o núcleo do SN, ou seja, o substantivo. Por exemplo, no SN **o sistema**, o centro deste SN é o substantivo **sistema**. Constatou-se, no entanto, que a determinação do centro de SN nem sempre é tão óbvia como neste exemplo. Nos casos em que o SN é complementado por uma expansão preposicional, sabe-se, *a priori*, que o centro de SN é representado pelo substantivo que antecede a expansão preposicional. Assim, no SN **o sistema de informação** o centro de SN é **sistema**. O termo **informação** apenas identifica o tipo de sistema. Desta maneira, se o usuário solicitar os SN 1 a partir da palavra **informação**, o citado SN não aparecerá na resposta do protótipo, pois **informação** não constitui centro do SN **o sistema de informação**. Percebeu-se, nesse tipo de SN, que existem outros termos que são tão importantes quanto ao próprio centro de SN. Estes termos foram definidos como centros complementares de SN. A não indexação destes Centros Complementares de SN provocaria aumento na taxa de silêncio[16]. A solução para este problema pode ser obtida mediante a escolha entre duas alternativas: **a)** criação de uma tabela específica para os centros complementares de SN; ou **b)** inclusão dos centros complementares na tabela de centros de SN. A solução **a** seria mais interessante do ponto de vista de homogeneidade de estrutura de dados. Entretanto, ela perde em termos de eficiência, tendo em vista a inserção de mais um nível no processo de varredura da árvore de SN e conseqüentemente um tempo de resposta maior nos procedimentos de busca. A solução **b**, apesar das questões de homogeneidade da estrutura de dados, é mais recomendável em função do menor tempo de resposta que ela proporciona nos procedimentos de recuperação da informação.

[16] Define-se como taxa de silêncio o percentual de documentos existentes em uma base de dados, mas não recuperados em uma busca.

Um terceiro aspecto do comportamento dos SN, percebido durante a fase de exploração do protótipo, foi o fraco resultado obtido quando de uma busca envolvendo os SN que tinham como centro a palavra **informação**. Verificava-se que o protótipo encontrava SN como **a informação**, **a informação científica**, **a informação técnica** etc. Contudo, o protótipo não encontrava SN como **as informações**, **as informações científicas**, **as informações industriais**, **as informações organizacionais**, **as informações técnicas** etc. Um problema parecido acontecia também com os centros de SN que se encontravam em outros idiomas, como francês e inglês, no caso **information**.

Adotou-se como solução o uso de uma tabela de equivalência entre os termos flexionados (em número e eventualmente em gênero) e os termos em outros idiomas com os códigos dos respectivos centros de SN. Assim, a varredura dos SN de primeiro nível passa a ser feito mediante esta tabela de equivalência (=Tabela Palavras) e não pela tabela de Centros de SN de primeiro nível. A vantagem de se adotar esta solução é que não houve nenhum ônus em termos do tempo de resposta, nem tampouco em termos da estrutura de dados.

Existem outras soluções, como, por exemplo, o uso de operadores de truncamento. Tal solução tem o inconveniente de introduzir mais um item de conhecimento necessário para o usuário manusear o protótipo, ou seja, é mais um tipo de operação que o usuário tem de dominar. Além disto, as flexões dos termos e as traduções destes para um outro idioma nem sempre possuem a mesma raiz.

CONCLUSÃO

A utilização dos sintagmas nominais como estrutura de acesso à informação contida em uma base de dados textual se apresenta como uma alternativa aos sistemas tradicionais de recuperação de informação. O protótipo desenvolvido com base nesta abordagem tornou o acesso à informação uma tarefa simples e convergente[17] para o usuário. Ela não tem os inconvenientes dos tradicionais sistemas de recuperação de informação onde os usuários são obrigados a dominar uma série de conhecimentos, dentre os quais o aprendizado de um novo idioma que é representado nesses sistemas por uma linguagem orientada a comando. Em contrapartida, o protótipo desenvolvido exige básicamente o uso do *mouse*, sem exigir do usuário o emprego de uma linguagem de busca. Todas as funções se apresentam na tela para o usuário, ou seja, existem apenas duas operações básicas que o usuário deve dominar: a) informar ao protótipo o centro do SN que atende à sua necessidade de informação; b) escolher uma função que se apresenta na tela em forma de ícones[18]. Deve-se ressaltar ainda que o protótipo desenvolvido permite ao usuário um aprendizado maior do conteúdo da base de dados, tendo em vista que o usuário, na realidade, é guiado pelo protótipo na varredura da árvore dos SN. Este passeio por toda a estrutura acaba por permitir a visualização mais fácil do conteúdo de uma base de dados.

Vale ressaltar também o fato de que o protótipo pode ser utilizado para o acesso a acervos de documentos escritos em outros idiomas que possuam estruturas de SN semelhantes àquelas existentes na língua portuguesa e francesa. Assim, o protótipo poderá ser utilizado para implementação de bases de dados contendo documentos em outros idiomas, desde que observada a restrição citada.

Além destes aspectos, esta experimentação baseada nos SN como meio de acesso à informação mostrou ser possível o desenvolvimento de um sistema de recuperação de informação mediante um sistema comercial de gerência de bancos de dados.

---

[17] Entende-se por convergente o fato do protótipo conduzir o usuário a um maior refinamento de uma consulta. Em outras palavras, o protótipo proporciona ao usuário um caminho que leve a satisfação das suas necessidades de informação.

[18] Denominam-se ícones os botões (pequenos quadrados ou retângulos) contendo uma figura mnemônica representando a ação a ser tomada pelo protótipo quando eles são acionados, com auxílio do *mouse*, pelos usuários. Os ícones são utilizados principalmente nas aplicações compatíveis com o Windows.

Apesar dos bons resultados obtidos, este trabalho não está totalmente concluído, pois existem vários problemas a resolver, tanto em nível da extração automática dos SN, quanto em nível da construção da interface de busca.

Quanto à extração automática dos SN, encontram-se as seguintes questões: a) a resolução dos elementos anafóricos; b) a resolução dos problemas de elipses; c) a identificação dos SN sem determinação (artigo zero), fato comum na língua portuguesa.

Outro aspecto a ser considerado no processo de automação da extração dos sintagmas nominais diz respeito ao procedimento de determinação dos centros de SN de primeiro nível, bem como dos centros complementares de SN. A eficácia desta abordagem é diretamente proporcional à boa determinação destes centros de SN, tendo em vista que os mesmos representam o ponto de partida para o acesso à informação. A presente experimentação mostrou ainda que o procedimento de busca de informação a partir do centro de SN de primeiro nível é mais apropriado aos usuários novatos e àqueles que não possuem bem definida a sua necessidade de informação. Para um usuário especialista ciente de sua necessidade de informação, a demanda de informação a partir de centros de SN de níveis mais elevados seria recomendável, pois, assim, ele não teria de passear por toda a árvore de SN para encontrar a informação. Em outras palavras, ele teria um acesso mais rápido à informação.

Quanto aos aspectos relacionados com a interface de recuperação de informação, a principal questão que se coloca é quanto à sua avaliação em termos de precisão nos resultados de uma consulta. A avaliação desta interface somente poderá ser feita com a participação do usuário, dada a necessidade de conhecer as reações dos usuários quanto ao uso dessa abordagem. Os

resultados desta avaliação darão maiores subsídios para o ajuste da interface no sentido de melhorá-la em nível de interação com o usuário e também no sentido de aperfeiçoar essa abordagem.

Apesar da viabilidade demonstrada pela construção de um protótipo utilizando-se de sistemas de gerência de bancos de dados comerciais (SGBD), é preciso considerar as limitações que tais sistemas impõem, quanto em nível do tempo de resposta no acesso à informação, ou seja, quanto aos limites dos tamanhos dos SN. A utilização de tais sistemas é condicionada ao tipo de aplicação que se deseja desenvolver. Para o desenvolvimento de um sistema de recuperação de informação de uso profissional, recomenda-se o desenvolvimento de um sistema completo de tratamento e recuperação de informações utilizando-se de linguagens de programação tais como C++ ou Pascal. O uso destas lingugagens permite maior controle das estruturas de dados, dos procedimentos de acesso à informação e de interação com o usuário, proporcionando melhores tempos de respostas que os SGBD comerciais.

Pode-se dizer, finalmente, que existem hoje condições favoráveis ao desenvolvimento de um sistema completo de tratamento e recuperação de informação, baseado na abordagem proposta por este trabalho. Isto é possível por duas razões:

1) do ponto de vista de tratamento da informação (extração dos SN e indexação): existem vários trabalhos desenvolvidos no âmbito do grupo SYDO para o idioma francês, que demonstram a factibilidade de semelhante desenvolvimento para a língua portuguesa;

2) do ponto de vista de sistema de recuperação de informação, o presente trabalho vem confirmar a viabilidade de se construir sistemas de recuperação de informação tendo os SN como meio de acesso à informação textual.

São estas as motivações para continuar trabalhando, no âmbito de meu programa de tese de doutorado, em um sistema de indexação automática e recuperação de informação textual aplicado ao português, tendo os sintagmas nominais como base.

REFERÊNCIAS BIBLIOGRÁFICAS

BERRENDONNER, Alain. *Grammaire pour une analyseur : aspects morphologiques.* Université des Sciences Sociales de Grenoble. Grenoble, Novembre 1990. 88 p. Collection Les cahiers du Criss.

BOUCHÉ, Richard. « Le Syntagme Nominal, une Nouvelle Approche des Bases de Données Textuelles ». *Meta*. 1989, vol. 34, n°. 3. p. 428-434.

BRITO, Marcílio de. *Réalisation d'un analyseur morpho-syntaxique pour la reconnaissance du syntagme nominal : utilisation des grammaires affixes*. Lyon, 1991. 221 p. Thèse de doctorat d'Etat, Université Claude Bernard, Lyon I.

BRITO, Marcilio de. « Sistemas de Informação em Linguagem Natural: em busca de uma indexação automática ». *Ciência da Informação*. 1992, vol. 21, n°. 3. p. 223-232.

CODD, E. F. « A relational model for large shared data banks ». *CACM*. 1970, vol. 13, nº 6.

CODD, E. F. « Further normalization of the relational model ». *Data Base Systems, Courant computer science symposium 6,* 1971. Rustin R. Editeur, Englewood Cliffs, New Jersey 1972.

CUNHA, Celso et CINTRA, Lindsey. *Nova Gramática do Português Contemporâneo*. Lisboa : Edições João Sá da Costa, 1991.734 p.

DUCROT, Oswald & TODOROV, Tzvetan. *Dictionnaire Encyclopédique des Sciences du Langage.* Paris : Editions du Seuil, 1972. 470 p.

HARTER, Stephen P. *Online Information Retrieval : Concepts, Principles and Techniques*. Orlando : Academic Press. Inc., 1986. 259 p. (Library and Information Science).

KURAMOTO, Hélio. *Maquette d'un système de recherche d'information en utilisant des syntagmes nominaux.* Villeurbanne, 1995. Mémoire du DEA. École Nationale Supérieure des Sciences de l'Information et des Bibliothèques.

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, supplément I, Université Clermont Ferrand, 1982. p. 164-169.

LE GUERN, Michel. « Un analyseur morpho-syntaxique pour l'indexation automatique », *Le Français Moderne*. Juin, 1991, t. LIX, n°. 1, p. 22-35.

METZGER, J-P. *Syntagmes Nominaux et information textuelle : reconnaissance automatique et représentation*. Lyon, 1988. 324 p. Thèse de doctorat d'Etat, Université Claude Bernard, Lyon I.

POLITY, Yolla. « Evaluation des modes de recherche en langage naturel ». *Documentaliste - Sciences de l'Information*. 1994, vol. 31, n°. 3. p. 136-142.

SALTON, Gerard et MCGILL, Michael J. *Introduction to modern information retrieval*. New York : Mcgraw-Hill Book Company, 1983. 448 p. (Computer Science).

SHOW Guan Yeong, KONG, Hinny et LIN, Kenneth Wente. « Intelligent user interface to SQL-based database system ». *Engineering Application Artificial Intelligence*. 1993, vol. 6, n°. 4. p. 307-316.

SMEATON Alan F. « Prospects for intelligent, languaged-based information retrieval ». *Online*

Review. 1991, vol. 15, n°. 6. p. 373-382.

STRZALKOWSKI, Tomek. « Natural language processing in large-scale text retrieval tasks ». *Text REtrieval Conference (TREC-1).* Gaithersburg, 1993. p. 173-187.

VETTER, Max. *Modélisation des données : Approches globale et orientée objets*. Paris : Dunod Informatique, 1992.

An alternative approach to the treatment and retrieval of textual information: nominal syntagmas

Abstracts

An alternative approach for treating and retrieving information is proposed, where nominal syntagmas play the principal role. It is the construction of a prototype of an information retrieval system capable of navigating a tree structure of nominal syntagmas. To examine this prototype a sample was used of 15 articles in the Portuguese language, from which the nominal syntagmas were extracted. The extraction process was done manually simulating an automated process.

Keywords

Nominal syntagmas; Automatic indexing; Information retrieval systems.

Hélio Kuramoto
Funcionário do Instituto Brasileiro de Informação em Ciência e Tecnologia (IBICT)
Atualmente em tese de doutoramento pela Université Lumière Lyon 2, com bolsa do Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq)
Membro associado do Centre d'Études et de Recherches en Sciences de l'Information (CERSI) de l'École Nationale Superiéure des Sciences de l'Information e des Bibliothèques (ENSSIB)
E-mail : kuramoto@enssib.fr