DataGramaZero - Revista de Ciência da Informação - v.3 n.1 fev/02 *ARTIGO 03*

Sintagmas Nominais: uma Nova Proposta para a Recuperação de Informação

Nominal Groups: a New Purpose to Information Retrieval

por Hélio Kuramoto

Resumo: O uso das palavras como meio de acesso à informação pelos sistemas automatizados de recuperação de informação tem sido a base da maioria dos modelos de recuperação de informação. Apesar de alguns deles terem alcançado relativo sucesso na melhoria da precisão de resultados de uma busca, a meta principal da recuperação de informação, que é a obtenção de todos os documentos pertinentes a uma consulta, não foi atingida. O presente artigo analisa essa questão, discutindo e mostrando a inadequação do uso das palavras nesses modelos, propondo em seu lugar, um outro tipo de unidade de informação: os sintagmas nominais.

Palavras-chave: Indexação Automática; Sintagmas Nominais; Recuperação de Informação; Interface de Busca; Modelo Vetorial; Modelo Booleano, Análise de Referências

Abstract: The use of words as a means of accessing information has been the basis for most of information retrieval models implemented by many information retrieval automatic systems. Despite the success of some of these models in improving the accuracy in search results, the main goal of the information retrieval, which is to find all documents relevant to a user's query, has not been satisfactorily achieved. This paper analyzes this question, discussing and demonstrating the inadequacy of using words in this models and propose in its place another type of information unit: the nominal groups.

Keywords: Automatic Indexing; Nominal Groups; Information Retrieval; Search Interface, Vector Model, Boolean Model, Link Analysis

Introdução

Um dos grandes desafios encontrados na recuperação de informação é como atender às necessidades de informação do usuário de forma rápida e precisa. Várias pesquisas foram e continuam sendo realizadas com o propósito de aumentar a precisão dos resultados de forma que o usuário possa encontrar todos os documentos que atendem às suas necessidades de informação. A grande maioria, quiçá todos os modelos desenvolvidos, elaborados no último século, tem como base a palavra. A partir dessa base, inúmeros métodos de classificação (ranking) foram elaborados e implementados, alguns dos quais foram bem sucedidos, tendo proporcionado melhorias significativas na precisão dos resultados nos procedimentos de recuperação de informação. Entretanto, apesar desse sucesso, esses métodos não são capazes de recuperar todos os documentos relevantes a uma consulta do usuário. O fato é que com o surgimento das novas tecnologias da informação e da comunicação, o volume de informação na rede Internet, disponibilizado aos usuários cresceu rapidamente e de forma totalmente desorganizada. Essa desorganização tem dificultado a recuperação precisa da informação. Exemplos dessa dificuldade são os diversos mecanismos de busca que ao serem demandados por uma determinada informação, nos oferecem centenas e mesmo, não raramente, milhares de referências como resposta. Dessas centenas ou milhares de referências, normalmente, apenas aquelas constantes da primeira página, ou as 10 primeiras são relevantes e nem sempre atendem às necessidades de informação do usuário. Percebe-se, portanto, que os métodos elaborados e implementados não são completamente suficientes para recuperar as informações de forma precisa. A questão que se coloca é: os métodos de classificação ou ranking são ineficazes? Ou seria a unidade básica de acesso à informação, a palavra, que não tem o status de descritor e, portanto, não pode representar adequadamente os documentos?

Essa questão será discutida no presente artigo, apresentando inicialmente alguns dos modelos utilizados na indexação e recuperação de informação. Em seguida farei uma crítica ao uso das palavras como forma de acesso à informação, apresentando as inconveniências de sua utilização, propondo finalmente uma alternativa que poderá facilitar o acesso preciso à informação e trazer inovações na área de recuperação de informação.

Modelos de recuperação de informação

Os métodos utilizados na recuperação de informação têm como base o uso da palavra [1], que representa a unidade básica de acesso à informação. A partir dessa unidade foram desenvolvidos vários modelos com o objetivo de facilitar o acesso à informação e melhorar a precisão do resultado de uma busca ou consulta.

Um dos modelos mais utilizados, na recuperação de informação, é o booleano, o qual baseia-se na teoria dos conjuntos e na álgebra de Boole. As consultas são elaboradas através de expressões de busca combinando termos de indexação e operadores booleanos (*and, or* e/ou *not*). Segundo Ricardo Baeza-Yates & Berthier Ribeiro-Neto (1999), "a grande vantagem desse modelo é a clareza do seu formalismo e a sua simplicidade". No entanto, os resultados obtidos utilizando esse modelo são dotados de pouca precisão.

Um outro modelo, também muito utilizado e conhecido, é o Espaço Vetorial ou simplesmente modelo Vetorial [BAEZA-YATES & BERTHIER, 1999; SALTON & MCGILL, 1999]. Enquanto o Booleano baseia-se na comparação exata [2] entre os termos de uma consulta e aqueles presentes nos documentos, o modelo Vetorial baseia-se na comparação parcial entre a representação dos documentos e a da consulta do usuário. Isto é possível graças à atribuição de pesos aos termos de indexação presentes na consulta e aqueles presentes nos documentos. Essa ponderação, atribuída a cada termo de indexação, permite calcular o grau de similaridade entre um documento e uma consulta. Portanto, a cada documento que atende a uma consulta é atribuído um grau de similaridade, possibilitando, ao sistema de recuperação de informação, apresentar os resultados de uma consulta de forma ordenada, normalmente, em ordem decrescente dos respectivos graus de similaridade. Trata-se de uma maneira de evidenciar a relevância de cada documento em relação a uma dada consulta. Esses pesos são calculados a partir da freqüência de ocorrência do referido termo no documento.

A diferença entre o modelo Booleano e o Vetorial está na forma de estabelecer o resultado de uma busca. No modelo Booleano os pesos são binários (0 e 1), enquanto que no modelo Vetorial, mais do que simplesmente identificar a presença do termo tanto na consulta quanto no documento, ele atribui um peso cujo valor pode variar entre 0 e 1. O valor 0 (zero) indica que o documento não tem qualquer relevância para uma determinada consulta e o valor 1 (um) indica a total relevância do documento em relação à referida consulta. O fato de o peso não ser binário, justifica a denominação do processo de comparação como sendo parcial.

Além desses, outros modelos [BAEZA-YATES & BERTHIER, 1999; HENZINGER, 2000; KLEINBERG, 1998; PÔSSAS et al.,2001; RAMOS, 1999; SALTON & MCGILL, 1983] foram desenvolvidos buscando melhorar a precisão e performance dos sistemas de recuperação de informação. A elaboração desses modelos foi possível graças a inspiração em teorias e cálculos emprestados de disciplinas como probabilidade, estatística, inteligência artificial, bibliometria etc.

Dentre esses modelos, é interessante mencionar aquele que utiliza a análise de referências (*link analysis*) [3, 4] o qual vem sendo utilizado com sucesso por alguns mecanismos de busca (search engines). O Google melhorou consideravelmente a precisão dos seus resultados utilizando esse tipo de análise. A análise de referências assume que é mais difícil manipular as referências que se faz a uma determinada página web. Segundo Eric Ward (2001), a introdução desse tipo de análise, pelos mecanismos de busca, na avaliação de relevância de páginas web para uma consulta é justificada por se tratar de um método útil e não manipulável. Trata-se de uma solução particularizada para as páginas web, tendo em vista que as referidas páginas fazem referência a outras páginas através de hyperlinks, elemento que facilita a navegação na *web*. Os documentos tradicionalmente disseminados em bases de dados textuais ou referenciais não são ou não eram dotados dessa facilidade. No entanto, com a evolução da tecnologia e com a publicação eletrônica, certamente os documentos passarão a incorporar essa facilidade, os *hyperlinks*. A existência dessa facilidade motivou a introdução da "*link analysis*" nos procedimentos de determinação de relevância das

páginas web com relação a uma dada consulta. Deve-se ressaltar, que não se trata de um método ou técnica inéditos, as ferramentas utilizadas são as mesmas oriundas da bibliometria utilizadas para análise de citação.

Segundo Monika Henzinger (2000):

> *"... a análise de referências assume, de forma simplificada, que:*
> ** Uma referência (link) de uma página A para uma página B é uma recomendação da página B pelo autor da página A;*
> ** Se a página A e a página B estão inter-relacionadas através de uma referência, a probabilidade de elas tratarem de um mesmo assunto é maior que se elas não estivessem inter-relacionadas."*

Ora, essa mesma premissa pode ser adotada aos documentos estruturados como artigos, monografias, teses etc., tendo em vista que esses tipos de documentos fazem também referência a outros documentos, assim como são também referenciados. A questão é que as páginas web estão, em sua grande maioria, armazenadas nos bancos de dados dos mecanismos de busca, ou na pior das hipóteses, acessíveis através da Internet. Enquanto os outros tipos de documentos eram e são tradicionalmente impressos e não se encontram, em sua grande maioria, armazenados e nem acessíveis através da Internet.

Além dos modelos adotados utilizarem a extração e indexação das palavras para a recuperação de informação, eles utilizam técnicas ou métodos de determinação do nível de relevância dos documentos recuperados em resposta a uma consulta do usuário. Trata-se, portanto, de métodos de classificação (*ranking*) dos documentos recuperados segundo os seus níveis de relevância. É bem verdade que alguns desses métodos foram bem sucedidos e realmente melhoraram a precisão nos procedimentos de recuperação de informação. No entanto, é também sabido que ainda existem deficiências nesses procedimentos. A hipótese de independência entre os termos de indexação no modelo Vetorial constitui uma dessas deficiências. A melhoria da precisão na recuperação de informação é ainda objeto de várias pesquisas na área de recuperação de informação.

Novos métodos para determinação de evidências de relevância de documentos na recuperação de informação podem surgir. A questão que se coloca doravante é quanto a validade da utilização da palavra como meio de acesso à informação. A utilização dessa unidade seria um fator limitante na melhoria da precisão dos resultados nos procedimentos de recuperação de informação? Os métodos elaborados para determinar a relevância dos documentos recuperados poderiam ser aplicados em modelos que utilizem unidades mais complexas que uma simples palavra no acesso à informação? Essas são algumas das questões que discutiremos a seguir.

Palavra : descritor ou símbolo sem referência?

No último século, a grande maioria dos modelos de recuperação de informação utilizou e utiliza a palavra como forma de acesso à informação. Seria essa a melhor unidade para utilização num procedimento de recuperação de informação? Para responder a essa questão é necessário analisar as características dessa unidade e seus inconvenientes.

Do ponto de vista da língua, a palavra tem as seguintes propriedades:

1. A palavra pode ter vários significados (polissemia);
Exemplo: Chave (solução de um problema);
Chave (ferramenta para abertura de portas);
Chave (ferramenta para apertar parafusos)

2. Duas palavras podem designar um mesmo significado (sinonímia);
Exemplo: Abóbora – Jerimum
3. Duas ou mais palavras podem combinar-se em ordem diferente designando idéias completamente diversas.
Exemplo: crimes, juvenis, vítimas (Vítimas de crimes juvenis);
crimes, juvenis, vítimas (Vítimas juvenis de crimes)

Essas propriedades interferem negativamente na precisão dos resultados em um procedimento de recuperação de informação, podendo aumentar a taxa de ruídos [3] no caso das propriedades 1 e 3 (polissemia e combinação de palavras) ou incrementar a taxa de silêncio [4], no caso de sinonímia. Em outras palavras o usuário poderá receber como resultado de uma busca documentos que não atendem às suas necessidades de informação ou mesmo não recuperar todos os documentos pertinentes existentes na base.

A terceira propriedade apresentada mostra que os termos JUVENIS, CRIMES e VÍTIMAS podem combinar-se para formar o termo VÍTIMAS DE CRIMES JUVENIS, assim como para formar o termo VÍTIMAS JUVENIS DE CRIMES. Utilizando o modelo booleano, a recuperação de informação através da submissão desses mesmos termos, em uma base de dados textual, obteria tanto documentos que tratam de *vítimas de crimes juvenis* quanto documentos que tratam de *vítimas juvenis de crimes*, incrementando, assim, a taxa de ruídos e diminuindo a taxa de precisão [5] do resultado da busca.

Cabe lembrar que antes do surgimento das novas tecnologias da informação e da comunicação, o processo de elaboração da representação de um documento nas antigas bases de dados bibliográficas era realizado por técnicos de informação especializados em indexação. Esses indexadores trabalhavam com o auxílio de vocabulários controlados, tesauros, outras tabelas ou listas que forneciam os descritores adequados à elaboração da representação de cada documento. Esses descritores são termos portadores de informação e que fazem referência a um objeto ou fato do mundo real. Com o alto incremento do registro de informação, em meio magnético, tornou-se inviável a utilização de indexadores para a construção da representação de documentos em uma base de dados. A solução foi criar mecanismos de indexação automática, os quais se baseiam na extração de palavras. Dessa forma, as palavras acabaram por substituir, ainda que inadequadamente, os descritores. Tratam-se de dois elementos distintos, dado que uma palavra nem sempre tem o status de descritor.

Segundo Michel Le Guern (1991) "*a passagem da indexação elaborada por especialistas, indexadores, para a indexação automática não modifica a natureza dos descritores, mas ele obriga a não mais se contentar com uma abordagem intuitiva e empírica.*" [6]. A reflexão apresentada mostra que a indexação automática, mais do que simplesmente viabilizar economicamente a indexação dos documentos de uma base, não poderia ter trocado os descritores por uma outra unidade desprovida de informação. Além disso, a extração de informações de um conjunto de documentos deveria se sistemática e não intuitiva e empírica como é realizado o processo de indexação por parte dos técnicos de informação especializados nesse processo.

A recuperação de informação tratada nesse artigo diz respeito a busca de informação em bases de dados contendo documentos textuais. Esses documentos foram escritos pelo homem utilizando a sua linguagem, a linguagem natural. A redação desse documento é o resultado de um processo de escrita onde o autor exprime as suas idéias através da combinação de um conjunto de palavras. Essas palavras ao se inserirem no texto do documento, assumem um significado específico. As palavras enquanto unidades de um dicionário possuem um conjunto de predicados, não fazem referência a um objeto ou fato do mundo real. Ou seja, não contêm qualquer substância. Elas adquirem essa substância no momento em que se inserem no universo do discurso. Existe, portanto, todo um trabalho intelectual do autor em combinar as palavras dando-lhes significados específicos.

Ora, o processo de indexação, segundo os modelos revistos aqui, faz o percurso inverso daquele

seguido pelos autores no momento da redação de um documento. O processo de indexação extrai cada palavra do texto de um documento e o insere numa lista de palavras ordenadas, de forma a facilitar a recuperação de informação. Esse processo destrói, portanto, o trabalho intelectual do autor do referido documento. As palavras ao serem extraídas de um documento, de um texto, deixam de ter aquele valor específico atribuído e concebido pelo autor quando da sua redação. Essas palavras voltam a ter designações genéricas, ou seja, voltam a ter um conjunto de predicados, sem qualquer referência a um objeto ou fato da realidade extralingüística do autor.

Um exemplo dessa constatação pode ser vislumbrado na frase "*A chave do problema é identificar o valor de Y...*". Nessa frase, a palavra **chave** tem um valor semântico específico, ou seja, ela é utilizada em lugar da palavra "solução". O dicionário Aurélio indica que **chave** é um substantivo feminino podendo ter os seguintes significados:

> *"1) Artefato de metal que movimenta a lingüeta das fechaduras; 2) Instrumento com que se apertam, desapertam, montam ou desmontam vários aparelhos; 3) Peça móvel para fechar orifícios de instrumentos de sopro; 4) Peça com que se dá corda em relógios; 5) Cavilha que atravessa a parte superior do fuso do lagar, prendendo-lhe o peso pelo veio; 6) Lugar que fecha ou domina território e pode ser ponto estratégico contra inimigos; 7) Princípio ou fecho de composição poética ou de outro qualquer trabalho literário; 8) O que prepara, facilita, explica ou inicia: chave do problema; 9) Elemento importantíssimo, decisivo..."*

Verificamos acima, segundo o que consta no dicionário Aurélio, algumas das definições atribuídas à palavra **chave**. Apresentamos apenas 9 das 24 definições constantes daquele dicionário. Portanto, a palavra "chave", de forma isolada, é um elemento genérico que pode vir a ser utilizado em diversos contextos podendo assumir, em um discurso ou texto, uma daquelas definições apresentadas.

Diante do exposto, entendo que o processo de indexação deveria extrair dos documentos informações que possam facilitar a sua recuperação e não símbolos sem referência como o são as palavras.

Michel Le Guern concorda com essa posição ao comentar o que extrair num processo de indexação:

> *"A necessidade de referenciar que é relacionada à natureza de um descritor induz a ver ou extrair uma unidade do discurso e não uma unidade da língua."*.[7]

Esse pesquisador distingue, portanto, a natureza de um descritor daquela de uma palavra, afirmando que o descritor deveria ser uma unidade do discurso e não uma unidade da língua. Portanto, o elemento que deve ser extraído de um documento, para representa-lo deveria possuir o mesmo status de um descritor, de onde se conclui que o uso das palavras para a representação de um documento é inadequado. No lugar das palavras, o sistema de indexação automática deveria reconhecer e extrair unidades do discurso, elementos com o mesmo status dos descritores

Entendo, assim, que a palavra não constitui a melhor unidade a ser utilizada no acesso à informação.

Proposta de um modelo de indexação alternativo

A indexação automática, segundo Michel Le Guern (1991), consiste, da mesma forma que a indexação realizada pelos indexadores, em selecionar, em cada documento, os elementos que permitirão ao usuário recuperá-lo, suprindo, assim, a sua necessidade de informação.

Os elementos, a que se refere Michel Le Guern, são os descritores. Independente da maneira como esses descritores foram extraídos dos ou atribuídos aos documentos, eles devem fazer referência a objetos ou fatos do mundo real. Conforme foi discutido na seção anterior, esses descritores não devem ser símbolos sem referência como o são as palavras. Portanto, esses elementos devem constituir-se de unidades extraídas do discurso. Essa unidade deve ser a menor unidade do discurso que possa servir de base a uma relação referencial autônoma. Essa unidade denomina-se sintagma nominal [9]. O sintagma nominal é a menor parte do discurso portadora de informação. Ao contrário das palavras, os sintagmas nominais não são símbolos sem referências. Não resta dúvida de que os sintagmas nominais possuem uma estrutura sintática, assim como, não resta dúvida, também, que os sintagmas nominais são portadores de uma estrutura lógico-semântica.

Conforme apresentado, os sintagmas nominais possuem uma estrutura sintática, podendo ter diversas configurações em termos sintáticos. Não cabe, neste artigo, aprofundar sobre a composição sintática dos sintagmas nominais. Elaborei, no âmbito de minha tese, um modelo para reconhecimento e extração automática de sintagmas nominais. Tal modelo exige alguns conhecimentos de lingüística, da língua portuguesa e da ciência da computação. Concentraremos, pois, apenas naquilo que concerne o uso dos sintagmas nominais na recuperação de informação. Assim, para simplificar, apresento a seguir um exemplo de sintagma nominal:

I. O estudo da economia da informação

Trata-se de um sintagma nominal complexo, pois, dois outros sintagmas nominais encontram-se embutido nele:

II. A economia da informação
III. A informação

É fácil observar que os sintagmas nominais são compostos de grupos nominais constituídos de uma organização hierárquica em árvore. Diferentemente das palavras, o sintagma nominal quando extraído do texto mantém o significado, o seu conceito. Assim, a questão que se coloca é como utilizá-los em um processo de indexação e recuperação de informação.

A utilização dos sintagmas nominais na recuperação de informação oferece duas alternativas possíveis de implementação em termos de indexação automática e de interfaces de busca.

Uma primeira alternativa seria implementar uma indexação automática nos moldes daquela tradicional baseada em palavras, apenas substituindo os índices contendo as palavras isoladas por índices contendo sintagmas nominais. Essa alternativa inclui também a possibilidade de utilizar os modelos de classificação ou ranking como o modelo Vetorial aplicado aos sintagmas nominais como unidade básica de acesso à informação. A solução que ora se propõe resolve a questão da co-ocorrência de palavras nesse modelo, dado que a organização sintática dos sintagmas nominais a inclui naturalmente. A hipótese de autonomia dos termos persistiria para os sintagmas nominais. Essa hipótese complica quando se trata de co-ocorrência de palavras, dado que as palavras são símbolos sem referência. Para o caso dos sintagmas nominais, entretanto, tal preocupação não tem sentido, dado que os sintagmas nominais comporta uma estrutura lógico-semântica.

Uma segunda alternativa seria o aproveitamento da organização hierárquica em árvore dos sintagmas nominais. O aproveitamento dessa organização não apenas cria um novo conceito em termos de indexação, como também introduz inovação em termos de uma interface de busca.

Considerando o exemplo de sintagma nominal apresentado, ele pode ser organizado segundo os sintagmas que se encontram nele embutidos. Trata-se, portanto, de um sintagma nominal de terceiro

nível, dado que ele contém dois outros sintagmas encadeados em seu interior. A enumeração do nível dos sintagmas nominais poderá ser feita atribuindo-se ao sintagma mais simples (“a informação”) o nível 1, ao sintagma que o contém (“a economia da informação”) seria atribuído o nível 2 e ao sintagma que contém os dois outros (“o estudo da economia da informação”) seria enumerado como sendo o de nível 3, conforme segue:

O estudo da economia da informação (nível 3)
A economia da informação (nível 2)
A informação (nível 1)

Esse tipo de organização permite a navegação na estrutura hierárquica em árvore dos sintagmas nominais. Uma interface de busca, baseada nessa organização, poderia, por exemplo, funcionar da seguinte forma:

1. A interface de busca aguarda que o usuário forneça o centro forneça um termo ou palavra que represente o centro [8] do sintagma nominal de primeiro nível, por exemplo: **informação**.

2. A partir desse termo a interface de busca recupera todos os sintagmas nominais de primeiro nível que tem “informação” como seu centro. No caso, seriam apresentados diversos sintagmas de primeiro nível que tem “informação” como centro do sintagma nominal, inclusive o sintagma “a informação”. A partir desse nível, o usuário seleciona o sintagma que possa vir atender a sua necessidade de informação. Nesse caso, ele escolheria “a informação” e solicita ao sistema que apresente os sintagmas de segundo nível que possua o sintagma nominal “a informação” em sua estrutura.

3. Em seguida a interface apresenta todos os sintagmas nominais do segundo nível que possua o sintagma “a informação” em sua estrutura. Essa navegação continua até o momento em que o usuário encontre o sintagma nominal que mais atenda a sua necessidade de informação. Nesse caso ele seleciona o referido sintagma e solicita que a interface apresente todos os documentos de onde ele foi extraído.

Essa forma de navegar intensifica a interação entre o usuário e o computador. Ou seja, é o próprio usuário que encontra as informações de que necessita mediante o apoio do computador. Observa-se que essa navegação traduz-se em um processo de refinamento de uma busca. À medida que se percorre a árvore de sintagmas nominais, do mais baixo nível para o mais alto nível, o usuário está executando um processo de refinamento de sua demanda de informação. De outra forma, a navegação retroativa, ou seja, de um sintagma de alto nível para um de nível mais baixo possibilita ao usuário a reformulação de sua consulta.

Trata-se, portanto de uma interface inovadora, totalmente diferente daquelas conhecidas comercialmente. Essas interfaces funcionam na base da pergunta e resposta, ou seja, o usuário faz uma consulta e o computador imediatamente traz os documentos recuperados segundo a consulta submetida. Indiretamente, pode-se dizer que as interfaces tradicionais, além de pouco interativas pressupõem que o computador seja capaz de resolver sozinho, mediante apenas a submissão de uma expressão de busca, a necessidade de informação do usuário. A questão é que nem sempre o usuário é capaz de explicitar a sua necessidade de informação numa única expressão de busca.

A interface proposta neste artigo inova a relação usuário – computador, dá ao usuário a oportunidade de orientar e reorientar o computador, bem como a sua consulta, mediante a sua interação com a máquina.

Um outro ponto que difere das interfaces tradicionais é que na minha proposta o usuário navega primeiramente no conjunto de sintagmas nominais até encontrar o termo que melhor se adeqüe à sua necessidade de informação e somente após encontrar esse termo é que ele acessa aos documentos de onde esse termo foi extraído. Esse processo interativo proporciona ao usuário uma aprendizagem sobre o conteúdo da base de dados, ou seja, maior conhecimento da base de dados que ele está consultando. Nos sistemas tradicionais essa forma de aprendizagem é dificultada pelo fato do sistema dar o resultado direto informando os documentos recuperados e também pelo fato dele não oferecer ou apresentar a lista de descritores, dado que ele trabalha com palavras. Assim, o máximo que os sistemas tradicionais poderiam mostrar seria a lista de palavras utilizadas na indexação.

Considerações Finais

A proposta apresentada neste artigo foi objeto de estudos e pesquisa em minha tese de doutoramento. Essa pesquisa compreendeu o desenvolvimento de um protótipo de interface de busca utilizando os sintagmas nominais como forma de acesso à informação. Para testar esse protótipo examinei e extraí cerca de 8800 sintagmas nominais de uma amostra de 15 (quinze) artigos selecionados aleatoriamente da revista Ciência da Informação. A extração dos sintagmas nominais foi realizada por mim utilizando uma abordagem lógico-semântica, tendo em vista que até o presente momento não existe nenhum software capaz de reconhecer, extrair e indexar os sintagmas nominais. Assim, todo o trabalho de reconhecimento e extração de sintagmas nominais foi realizado de forma não automatizada. Essa experiência foi importante para a elaboração de um modelo para o reconhecimento, extração e indexação de sintagmas nominais. Os resultados obtidos com a implementação do protótipo comprovaram a viabilidade técnica de implementar uma interface de busca capaz de navegar em uma estrutura hierárquica em árvore de sintagmas nominais. Após a experimentação do protótipo, realizei estudos com o propósito de estabelecer um modelo para o reconhecimento, extração e indexação dos sintagmas nominais conforme a segunda alternativa de indexação apresentada neste artigo. Esse modelo compõe-se de duas gramáticas, uma descrevendo as unidades léxicas e outra estabelecendo um conjunto de regras sintáticas descrevendo a estrutura sintática dos sintagmas nominais. Com a elaboração da tese concluí todo o embasamento teórico e conceitual necessário à implementação de um Sistema de Recuperação de Informação Assistido por Computador. Considerei tal sistema como sendo **assistido por computador** pelo fato de o computador desempenhar o papel de apoio, de suporte ao usuário, dado que é o usuário quem orienta e obtém às informações de que necessita. O computador apenas segue as orientações do usuário na navegação pela árvore de sintagmas nominais.

As experiências levadas a cabo durante a elaboração de minha tese de doutorado indicaram a viabilidade técnica do uso dos sintagmas nominais, apesar da quantidade reduzida de documentos inseridos na amostra. Os resultados obtidos indicam a necessidade de dar continuidade à pesquisa através de uma avaliação utilizando amostras de dados contendo volume maior de documentos. O tratamento e organização de um volume maior de documentos permitirão conhecer melhor o comportamento dos sintagmas nominais. Além disso, é necessário que essa avaliação seja realizada com a presença de usuários. Somente os usuários poderão avaliar a eficácia dos sintagmas nominais. No entanto, para construir amostras mais consistentes e volumosas, em termos de documentos, é imprescindível contar com ferramentas que possam reconhecer, extrair e indexar os sintagmas nominais. O processo de reconhecimento, extração e indexação não automatizada, além de ser inviável economicamente em se tratando de grandes volumes de documentos, pode prejudicar a uniformidade no processo de reconhecimento, extração e indexação dos sintagmas nominais. A implementação dessas ferramentas tem a vantagem de assegurar a referida uniformidade, tendo em vista que essas ferramentas seguem uma sistemática que é aplicada de forma uniforme em todos os documentos pelo computador. Contrapondo esse fato, posso testemunhar a minha experiência nesse processo de reconhecimento e extração de sintagmas nominais, pois tive a oportunidade de reconhecer e extrair cerca de 8800 deles. Quando é o homem a realizar tal tarefa, ele é influenciado pela subjetividade, pela abordagem lógico-semântica que lhe é peculiar. Esse intervenção subjetiva

impede ou dificulta a manutenção da uniformidade no processo de reconhecimento e extração de sintagmas nominais. O computador não sofre essa influência e tem a vantagem de aplicar sempre as mesmas regras nos mesmos contextos.

Trata-se de uma etapa importante para a continuidade da pesquisa sobre o uso dos sintagmas nominais. A inexistência dessas ferramentas impede uma avaliação mais consistente envolvendo amostras de dados com maior volume de documentos.

Além da implementação de ferramentas de reconhecimento, extração e indexação de sintagmas nominais, deve-se dar continuidade a essa iniciativa implementando interfaces de busca considerando a organização dos sintagmas nominais. Nesse contexto, a efetiva experimentação e avaliação dos sintagmas nominais extraídos de grandes amostras de dados depende da implementação de interface apropriada às características de navegação dos sintagmas nominais. Com uma amostra contendo 15 documentos, verificamos que a quantidade de sintagmas nominais de primeiro nível tendo como centro do sintagma a palavra "informação" a quantidade desses sintagmas chegou a 122. A varredura desses sintagmas através da tela de um computador não é trivial. Trata-se de uma situação similar à dos mapas de conhecimento. Por isso a necessidade de se conceber uma interface que possa manipular grandes quantidades de termos como os sintagmas nominais organizados hierarquicamente. Dentre os últimos desenvolvimentos na área de interfaces de navegação, destaca-se uma em especial, aquela baseada nas equações hiperbólicas, desenvolvida inicialmente pela Xerox [6]. Assim, para que a avaliação junto ao usuário possa ter sucesso, urge que uma interface com as características apresentadas seja implementada e testada. Uma interface com base nessas equações certamente dará ao usuário maior conforto e facilidade de exploração e navegação na árvore de sintagmas nominais.

Vislumbra-se, após as reflexões e experiências realizadas com os sintagmas nominais face às fragilidades que o uso das palavras tem proporcionado à recuperação de informação, a explosão do uso dos sintagmas nominais. Assim, se no século passado o uso da palavra dominou as pesquisas sobre recuperação da informação e outras áreas, espera-se que esse século que se inicia seja dedicado à exploração dos sintagmas nominais.

Notas

[1] Esse termo também conhecido como palavra-chave ou termo de indexação.

[2] No modelo Booleano, a análise de similaridade entre a consulta e um determinado documento atribui peso 1 quando um documento atende à consulta e 0 quando os termos demandados pela consulta não se encontram no referido documento. Portanto, o simples fato de um documento atender à demanda de consulta é suficiente para considerá-lo relevante.

[3] Taxa de ruídos é definida como sendo a relação entre a quantidade de documentos recuperados não pertinentes e a quantidade total de documentos.

[4] Taxa de silencio é definida como sendo a relação entre a quantidade de documentos pertinentes, não recuperados, e a quantidade total de documentos pertinentes na base.

[5] Taxa de precisão é definida como sendo a relação entre a quantidade de documentos pertinentes recuperados e a quantidade total de documentos recuperados.

[6] "Le passage de l'indexation manuelle à l'indexation automatique ne modifie pás la nature des descripteurs, mas il oblige à ne plus se contenter d'une approche intuitive et empirique."

[7] "La nécessité de référer qui est liée à la nature du descripteur engage à y voir une unité de discours, non une unité de la langue.".

[8] Considera-se centro do sintagma nominal de primeiro nível o substantivo principal, núcleo do

sintagma nominal de primeiro nível. Por exemplo: No sintagma nominal de primeiro nível A INFORMAÇÃO CIENTÌFICA, o centro desse sintagma nominal é INFORMAÇÃO.

[9] Michel Le Guern é o autor intelectual e responsável pela idéia de utilizar os sintagmas nominais no lugar das palavras. Todo o arcabouço conceitual foi elaborado por esse lingüista no âmbito do grupo de pesquisa SYDO – SYstème Documentaire. Várias teses sobre o tema foram desenvolvidas e defendidas, assim como artigos escritos por membros do grupo SYDO (SYstème DOcumentaire). Alguns desses trabalhos constam na seção de referências ao final desse documento [BOUCHÉ, LAINÉ & METZGER, 1990; KURAMOTO, 1999; LE GUERN, 1991; LE GUERN, 1994; LAROUK, 1993; METGER, 1988]. A grande maioria dos trabalhos desenvolvidos versam sobre a questão do reconhecimento e extração de sintagmas nominais. A única exceção é a minha tese onde eu pude trabalhar tanto na questão do reconhecimento, extração e indexação de sintagmas nominais, formulando um modelo apropriado para resolver essa questão, quanto na prototipagem de uma interface piloto e experimentação dessa abordagem. Durante a minha pesquisa eu pude contar com a competente orientação desse professor. Os estudos concernentes aos sintagmas nominais não se concentraram apenas na França, tive a oportunidade de encontrar recentemente uma iniciativa de estudos sobre esse tema no departamento de lingüística da UFMG. Trata-se da Profa. Yara Goulart Liberato cuja tese encontra-se também referenciada nesse documento [LIBERATO, 1997].

Referências Bibliográficas

BAEZA-YATES, Ricardo; RIBEIRO-NETO, Berthier. *Modern Information Retrieval*. New York: Addison-Wesley, 1999 .

BOUCHÉ, Richard.; LAINÉ, S.; METZGER, J.-P. Extraction des connaissances à partir d'une collection de documents. In: TOOLS OF KNOWLEDGE ORGANIZATION AND THE HUMAN INTERFACE. *Proceedings*... [s.n.t.] : Darmstadt - Alemanha, 1990.

HENZINGER, Monika. Link Analysis in Web Information Retrieval. *Bulletin of the IEEEComputer Society Technical Commitee on Data Engineering*. v. 23, n. 3, p. 3-8, september 2000. Disponível em: http://research.microsoft.com/research/db/debull/A00sept/issue.htm.

KLEINBERG, J.M. Authorithative sources in a hyperlinked environment. In.: *ANNUAL ACM-SIAM SYMPOSIUM ON DISCRETE ALGORITHMS*, 9., 1998. Proceedings... [s.l.:s.n.] 1998. p. 668-677. Disponível em: http://citeseer.nj.nec.com/1115.html.

KURAMOTO, Hélio. *Proposition d'un système de recherche d'information assistée par ordinateur: avec application au portugais*. 1999. Thèse (Doctorat en Sciences de l'information et de la communication) - Université Lumière–Lyon 2, Lyon, França.

LAMPING John; RAO Ramana; PIROLLI Peter. A focus+context technique based on hyperbolic geometry for visualizing large hierarchies. In: CONFERENCE ON HUMAN FACTORS IN COMPUTING SYSTEMS, 1995, Denver, Colorado. *Proceedings*... [s.n.t.]. Disponível em: http://www.acm.org/sigchi/chi95/proceedings/papers/jl_bdy.htm.

LE GUERN, Michel. Un analyseur morpho-syntaxique pour l'indexation automatique. *Le Français Moderne*. v. 59, n. 1, p. 22-35, juin 1991.

LE GUERN, Michel. Traitement automatique et variation linguistique : la syntaxe des titres. In: *OPÉRATEURS et Constructions Syntaxiques* : évolutions des marques et des distributions du Xvème siècle. Paris : Presses de l'Ecole Normale Supérieure, 1994. p. 75-81.

LE GUERN, Michel. Parties du discours et catégories morphologiques en analyse automatique.In: *LES CLASSES de Mots*. Lyon : Presses Universitaires de Lyon, 1994. p. 207-215.

LAROUK, Omar. *Extraction de connaissances à partir de documents textuels* : traitement automatique de la coordination (connecteurs et ponctuation). 1993. 290f. Thèse (Doctorat d'État em Sciences) - Université Claude Bernard-Lyon 1, Lyon, França.

METZGER, Jean-Paul. *Syntagmes Nominaux et Information Textuelle* : reconnaissance automatique et représentation. 1988. 324f. Thèse (Doctorat d'Etat en Sciences) - Université Claude Bernard-Lyon 1 – Lyon, França.

LIBERATO, Yara G. *A estrutura do SN em português: uma abordagem cognitiva*. Tese (doutorado em Lingüística). 1997. UFMG, Departamento de Lingüística, Belo Horizonte.

PÔSSAS, Bruno et al. Modelagem vetorial estendida por regras de associação. *SIMPÓSIO BRASILEIRO DE BANCOS DE DADOS*, 16., 2001, Rio de Janeiro. Anais... Rio de Janeiro : [s.n.] 2001.

RAMOS, Mônica Gomes. *Uso da teoria de função de crença em sistemas de recuperação da informação*. 1999. Dissertação (Mestrado em Ciência da Informação) – CID, UnB, Brasília.

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

WARD, Eric. How Search Engines Use Link Analysis. In: Search Engine Strategies 2001 Conference, 2001, Dallas, TX. Proceedings... [s.n.t]. Disponível em: http://searchenginewatch.com/searchday/01/sd1219-links.html.

Sobre o autor / About the author:

Hélio Kuramoto
kuramoto@eb.ufmg.br
Docteur em Sciences de l'Information et de la Communication
Pesquisador Visitante da Escola de Ciência da Informação da UFMG
Tecnologista Sênior do Instituto Brasileiro de Informação em Ciência e Tecnologia
Belo Horizonte, MG - Brasil