UMA PROPOSTA DE METODOLOGIA PARA INDEXAÇÃO AUTOMÁTICA UTILIZANDO SINTAGMAS NOMINAIS

A PROPOSAL OF METHODOLOGY FOR AUTOMATIC INDEXATION USING NOUN PHRASES

Renato Rocha Souza
Doutor em Ciência da Informação,
Professor do Departamento de Organização e Tratamento da Informação, ECI - UFMG
rsouza@eci.ufmg.br

Resumo
Com o aparente esgotamento das estratégias atuais de representação e indexação de documentos, faz-se necessário investigar novas abordagens para sistemas de recuperação de informações. Dentre estas abordagens, há uma vertente que busca levar em conta a semântica intrínseca aos documentos textuais, e uma das formas de fazê-lo é através da utilização de sintagmas nominais como descritores, ao invés de palavras-chave. Uma metodologia para atingir tal propósito, desenvolvida no escopo de uma tese de doutorado, é apresentada neste artigo.

Palavras-chave: Indexação automática. Sistemas de recuperação de informações. Sintagmas nominais.

1 INTRODUÇÃO

Para lidar com os constantes e ininterruptos ciclos de criação e demanda de informação, há muito vêm sendo criados sistemas de recuperação de informações que utilizam diversas tecnologias mecânicas e digitais de computação, para gerenciar grandes acervos de documentos. Podemos citar, dentre eles, a Internet, as intranets empresariais com seus portais corporativos, e as bibliotecas digitais. Muitas pesquisas vêm tentando contribuir para enfrentar alguns dos muitos desafios que surgem quando lidamos com massivas quantidades de dados, como nos grandes acervos de documentos digitais, notadamente quando estes precisam ser regularmente organizados e pesquisados, recuperando em tempo hábil informação relevante para algum objetivo específico.

Neste contexto, este artigo apresenta uma metodologia proposta como resultado de uma pesquisa desenvolvida no âmbito do curso de doutorado do autor, no Programa de Pós Graduação em Ciência da Informação da Universidade Federal de Minas Gerais, para

atribuir automaticamente descritores a documentos textuais digitalizados num processo de indexação automática.

Com o aparente esgotamento das estratégias tradicionais de busca em sistemas de recuperação de informações, entendemos que a melhoria da eficácia do serviço ao usuário dos sistemas depende dos resultados em diversas linhas de pesquisa, em todo o espectro da cadeia de processos de tratamento da informação. Temos como hipótese de trabalho que as principais frentes de atuação são as seguintes:

I. A exploração das informações semânticas e semióticas intrínsecas aos dados, de forma a expandir a compreensão das unidades e padrões de significado em textos, imagens e outras mídias;
II. O desenvolvimento de novas possibilidades de marcação semântica dos dados utilizando-se metalinguagens, criando espécies de índices acoplados aos próprios documentos com termos amplamente consensuais e não ambíguos, para que estes possam ser mais facilmente manipulados e identificados por computadores e outros dispositivos e, como conseqüência, pelos usuários;
III. O desenvolvimento de estratégias de apresentação da informação recuperada nas buscas, sob formas altamente significativas, ou contextuais - como em algumas interfaces gráficas – de forma que as relações entre os conceitos, e em conseqüência, os contextos, sejam evidentes; e também por estratégias que busquem estimular os vários órgãos sensoriais ao mesmo tempo – como nas ferramentas multimídias – para que a absorção das informações pelos usuários seja maior. Através destas interfaces e estratégias, as informações podem ser apresentadas de forma a possuírem conexões visuais aos seus contextos de origem, permitindo ao usuário refinar os resultados através da definição das conexões pertinentes e a exclusão das conexões geradas pelo ruído informacional;
IV. A construção e manutenção de perfis personalizados de utilização, de forma que o SRI "aprenda" com a forma de trabalho do usuário e possa utilizar estas informações específicas para melhorar a estratégia de busca do SRI.

Uma abordagem completa para a organização e a recuperação de informações, visando a melhoria dos Sistemas de Recuperação atuais, deve unir estas estratégias e soluções, buscando:

a) A indexação dos documentos utilizando representações mais significativas, de modo a aumentar e melhorar os pontos de acesso e a relevância das informações recuperadas;
b) Prover uma forma adequada de apresentar as informações recuperadas aos usuários, de maneira que sejam intuitivas e facilmente compreensíveis;
c) Utilizar no processo de indexação padrões universais de registros de metadados para que os sistemas sejam interoperáveis entre si;
d) Adaptar-se continuamente ao usuário, sendo preferível que possa aprender com a forma com que trabalha, de modo que as buscas sejam continuamente refinadas através de um trabalho de personalização.

Existem hoje diversas tentativas, mais ou menos coordenadas, de se abordar estas ações fundamentais, mas uma real integração demandaria a pesquisa em diferentes áreas do conhecimento e campos de pesquisa, como a ciência da informação, a lingüística, a ciência da computação, a sociologia, a antropologia, a comunicação, a psicologia cognitiva, entre outras.

De maneira isolada, há pesquisas em cada uma destas vertentes, mas é pouco explorada a utilização da semântica embutida nos próprios documentos, ou seja, das potencialidades intra-textuais da linguagem natural, para automatizar e melhorar as tarefas de indexação, organização e recuperação de informações. Pesquisas nesta área incluem o uso de estruturas profundas da linguagem natural, como os sintagmas verbais e nominais, para indexação e recuperação (KURAMOTO, 1996 e 1999; MOREIRO et al, 2003).

A pesquisa apresentada neste artigo explora o potencial de uso dos sintagmas nominais como descritores de documentos em processos de indexação. Partiu-se inicialmente da hipótese de que os sintagmas nominais, pelo maior grau de informação semântica embutida, podem vir a se tornar mais eficazes do que as palavras-chave usualmente extraídas e utilizadas como descritores em outros processos automatizados de representação de documentos, tais como os observados nos mecanismos de busca da Internet, ou em sistemas de leitura das palavras-chave fornecidas pelo autor dos documentos.

Alguns trabalhos anteriores se apresentam como marcos a partir dos quais se pretende avançar; dentre eles, a pesquisa sobre a viabilidade do uso dos sintagmas nominais para sistemas de recuperação de informações de KURAMOTO (1996 e 1999), e as ferramentas para marcação sintática do português e automatização da extração de sintagmas nominais desenvolvidas no âmbito dos projetos da Southern Denmark University (BICK, 2000), de VIEIRA (2000) e do PROJETO DIRPI (2001). A partir destes resultados e ferramentas, pretende-se propor uma metodologia de escolha automática de descritores para documentos que utilize os sintagmas nominais em vez de palavras-chave para documentos textuais digitalizados em língua portuguesa. Na seção seguinte, apresentamos alguns conceitos fundamentais ao entendimento da metodologia.

2 SINTAGMAS NOMINAIS E SISTEMAS DE RECUPERAÇÃO DE INFORMAÇÕES

Entendemos por sintagmas certos grupos de palavras que fazem parte de seqüências maiores na estrutura de um texto, mas que mostram um grau de coesão entre eles (PERINI, 1995). Os constituintes ou sintagmas podem ou não ser facilmente identificáveis, sendo que por vezes é necessário recorrer a outros recursos para que seja feita a "demarcação" sintática. Perini acredita que a intuição "subjetiva, mas nem por isso duvidosa" que nos permite separar a oração em seus constituintes imediatos pode ser caracterizada através de critérios puramente formais (1985, pp. 42-43), mas há quem defenda que a identificação dos constituintes é somente completa através de uma abordagem cognitiva e amplamente contextual (LIBERATO, 1997), que só é esperada na análise do discurso e na pragmática; ou através de outros modelos gramaticais, como a análise transformacional (RUWET, 1975, pp.155-212 e 223-279). Para a análise semântica, há também o problema das situações anafóricas, que ocorrem quando uma estrutura de uma oração se apresenta reduzida porque ocorre na vizinhança de outra estrutura oracional de certa forma paralela, dependendo desta para sua total compreensão (PERINI, 1986, p. 57).

De acordo com MIORELLI (2001), os sintagmas nominais podem ser entendidos – e tratados – de forma sintática, privilegiando a forma; ou semântica, buscando os significados maiores; cada uma com suas especificidades e implicações. A abordagem

semântico-pragmática, utilizada por LIBERATO (1997), não prescinde de um "interpretador de contextos", natural na cognição humana, mas dificilmente implementado em heurísticas de inteligência artificial. A forma sintática, como analisada por PERINI (1986, 1995 e 1996) está mais relacionada à estrutura das orações em si, e é mais facilmente tratada computacionalmente. Assim como no trabalho de MIORELLI (2001), esta é a abordagem que será utilizada no âmbito deste projeto, da mesma forma que, provavelmente, em quaisquer abordagens, e com quaisquer ferramentas, que busquem a automatização de extração dos sintagmas nominais.

Sistemas de recuperação de informações usualmente adotam termos índices para indexação de documentos, sendo que estes termos índice são usualmente palavras-chave. Há uma idéia fundamental embutida de que, a semântica dos documentos e das necessidades de informação do usuário pode ser expressas através destes conjuntos de palavras, o que é, claramente, uma grande simplificação do problema, porque grande parte da semântica do documento ou da requisição do usuário é perdida quando se substitui o texto completo por um conjunto de palavras (BAEZA-YATES & RIBEIRO-NETO, 1999, p.19).

Há, na literatura, registros de algumas tentativas de otimizar a organização dos documentos em SRIs através de um processamento aprofundado da linguagem natural dos documentos. Dentre elas, a identificação de "grupamentos de substantivos" (noun groups), ao invés de palavras-chave, se afigura uma boa estratégia para seleção de termos de indexação, uma vez que os substantivos costumam carregar a maior parte da semântica de um documento, ao invés de artigos, verbos, adjetivos, advérbios e conectivos. Esta proposta estabelece uma visão conceitual do documento (ZIVIANI, in BAEZA-YATES & RIBEIRO-NETO, 1999, pp.169-170). Os grupamentos de substantivos são conjuntos de nomes nos quais a distância sintática no texto (medida pelo número de palavras entre dois substantivos) não excede um limite predefinido. Uma metodologia que extrapola esta proposta é a identificação dos sintagmas nominais e o seu uso como descritores, como proposto neste projeto.

SALTON & MCGILL (1983, pp. 90-94) discutem algumas abordagens teóricas para o uso de métodos lingüísticos na recuperação de informações; entre elas, a análise da estrutura sintática (*parsing*) dos documentos de forma a identificar as estruturas

sintagmáticas. Estes autores, entretanto, apontam as dificuldades intrínsecas ao processo de análise semântica através da análise sintática e exemplificam casos em que é impossível o reconhecimento não ambíguo de relações semânticas através dos componentes da sentença, sugerindo que um modelo baseado em gramáticas transformacionais poderia trazer melhores resultados. Neste ponto, parecem então concordar com LIBERATO (1997), que entende que a análise completa das estruturas semânticas só é possível através da análise cognitiva dos contextos. Ao indicar a maior eficácia relativa dos algoritmos de geração de frases, baseadas em freqüência de palavras, talvez aponte uma alternativa para a melhoria do algoritmo proposto neste trabalho. Outra alternativa apontada é a interferência humana no processo de desambiguação através de uma interface, o que seria pouco desejável num processo que pretende ser automático.

Um importante caminho de pesquisa que visa resolver os problemas de desambiguação semântica através da análise dos contextos é resolução de correferência, ou resolução anafórica (VIEIRA, 1998 e 2000; SANT'ANNA, 2000; ROSSI et al, 2001; GASPERIN et al, 2003). Uma cadeia de correferência é uma seqüência de expressões em um discurso que se referem a uma mesma entidade, objeto ou evento. Essas cadeias são úteis para a representação semântica de um modelo de domínio, e podem melhorar a qualidade dos resultados em diversas aplicações de processamento de linguagem natural, como recuperação e extração de informações, geração automática de resumos, traduções automáticas, entre outros (ROSSI et al, 2001). O processo de resolução de correferências envolve a identificação e extração dos sintagmas nominais.

LE GUERN e BOUCHÉ (apud KURAMOTO, 1999) apontam o sintagma nominal como a menor unidade de informação contida em um texto. O grupo de pesquisas SYDO, ao qual pertencem estes pesquisadores, tem como fundamento teórico à utilização de sintagmas nominais como descritores (Ibidem, 1996). Ao trabalhar em parceria com este grupo, KURAMOTO (1999), em sua tese de doutorado, desenvolveu uma pesquisa fundamental para a consideração da utilização de sintagmas nominais como descritores. Já em um trabalho anterior, KURAMOTO (1996) vislumbrou a maquete proposta na tese e já apontava o potencial natural de organização dos sintagmas nominais que, se explorado convenientemente, poderia propiciar aos usuários maior facilidade no uso de um SRI e resultados mais precisos em resposta a um processo de busca de informação.

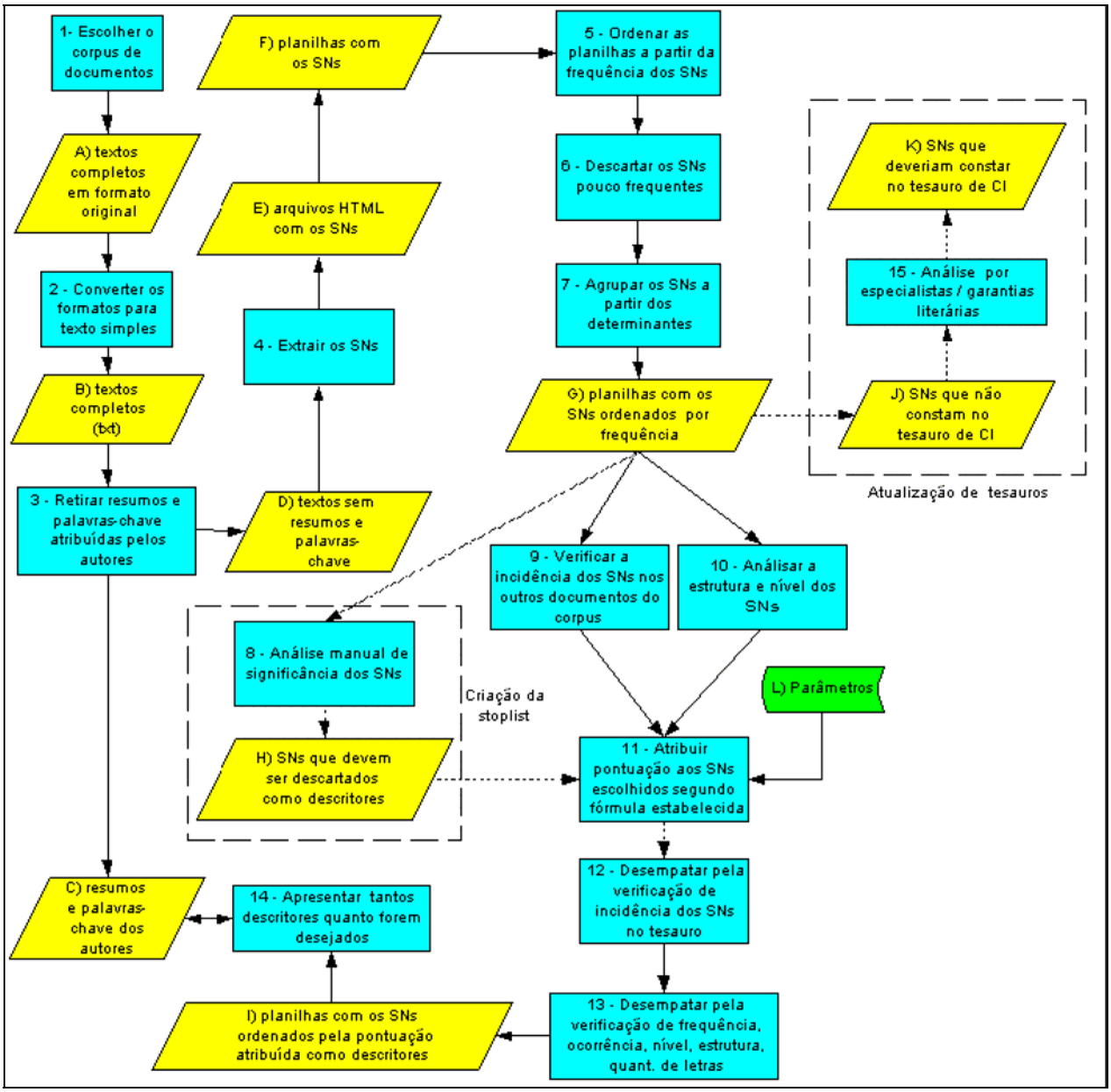
O sistema desenvolvido por Kuramoto pode ser considerado uma inspiração para o presente trabalho, na medida em que, em ambos, busca-se uma alternativa para uma melhor indexação utilizando-se sintagmas nominais. Entretanto, em sua maquete, "a extração dos sintagmas nominais foi realizada de forma manual, 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". (1996, p. 6). Ao menos um sistema deste tipo, entretanto, se encontra hoje disponível, e foi disponibilizado para o propósito deste trabalho (GASPERIN et al, 2003). Uma outra diferença fundamental é o objetivo. Se no projeto de Kuramoto buscava-se apresentar uma maquete de um SRI baseado em sintagmas nominais, o objetivo deste trabalho é propor uma metodologia de auxílio à indexação automática, utilizando uma metodologia aplicada sobre os sintagmas nominais extraídos automaticamente a partir de textos digitalizados em língua portuguesa. Diferenças a parte, o fundo filosófico é bastante comum.

3 A METODOLOGIA PROPOSTA

O objetivo da pesquisa era desenvolver uma metodologia para escolha automática de descritores para documentos textuais digitalizados, em língua portuguesa, utilizando como descritores as estruturas gramaticais conhecidas como sintagmas nominais. O principal pressuposto motivador é a crença de que a utilização de sintagmas nominais como descritores em um processo de indexação automática apresenta vantagens em relação ao uso de palavras-chave, devido ao fato destes possuírem, em comparação, maior densidade informacional, e serem mais bem relacionados ao contexto semântico do documento.

Para a consecução da extração automática de sintagmas nominais foram utilizadas ferramentas de *software*, que serão apresentadas na próxima seção.

Na figura a seguir, podemos notar um detalhamento da metodologia proposta para a indexação automática, com os passos delineados em seguida. A área de conhecimento dos documentos selecionados foi a ciência da Informação, sendo também da CI o tesauro utilizado na metodologia (SOUZA, 2005):



Representação esquemática da metodologia proposta de indexação automática

1. Escolher um *corpus* significativo de documentos reconhecidamente inseridos dentro de uma área de conhecimento;
2. Converter os formatos de arquivo para texto simples;
3. Retirar os resumos e as palavras-chave atribuídas pelos autores;
4. Extrair os sintagmas nominais do corpo do texto;

5. Ordenar os SNs nas planilhas através da verificação da freqüência de ocorrência dos sintagmas nominais nos documentos;
6. Descartar os SNs que apresentam freqüências de ocorrência inferiores a um patamar preestabelecido;
7. Agrupar os SNs remanescentes a partir dos determinantes em suas formas "canônicas", e reordená-los;
8. Analisar manualmente os SNs pré-escolhidos e decidir sobre a sua relevância como descritores, para fins de construção de uma *stoplist* e verificar se algum SN escolhido consta em uma *stoplist*, dinamicamente construída, para, se for o caso, descartá-lo (em 11);
9. Verificar a incidência dos SNs nos outros documentos do *corpus*;
10. Analisar a estrutura e o nível dos SNs;
11. Atribuir pontuação e ranquear os SNs remanescentes de acordo com fórmula estabelecida (explicitada a seguir), levando em conta a freqüência de ocorrências no texto e a freqüência de saturação definida; a quantidade de textos do *corpus* em que ocorrem, a estrutura sintática e o nível do SN. Estes critérios de relevância são regidos por parâmetros (representados na figura em L) a serem sintonizados com a sucessiva aplicação da metodologia, e serão discutidos adiante;
12. Caso ocorram "empates" nos valores da pontuação dos SNs, considerar a ocorrência no tesauro da CI como fator de desempate;
13. Caso ainda ocorram "empates" nos valores da pontuação dos SNs, considerar os seguintes critérios de desempate:
    a. Maior valor absoluto da freqüência de ocorrência;
    b. Menor valor absoluto da ocorrência no *corpus*;
    c. Maiores nível e estrutura do SN;
    d. Maior quantidade de letras do SN;
14. Apresentar tantos descritores quanto forem desejáveis, a partir da lista ranqueada de candidatos a descritores.

Alguns passos opcionais fazem parte de metodologias complementares, e são descritos em SOUZA (2005). Os parâmetros customizáveis propostos, mencionados no

item 11 dos passos descritos acima, possuem a característica de poderem ser alterados dinamicamente, de acordo com a performance dos dados de um *corpus* testado. No entanto, o dimensionamento minucioso destes parâmetros e de suas inter-relações, de modo a oferecer à metodologia uma performance ótima, é um desdobramento da pesquisa que ainda está em curso. Por ora, iremos assumir alguns conjuntos de valores para os quais as observações preliminares conferiram boa performance. Foi proposta uma fórmula para atribuir a pontuação aos SNs extraídos, para efeitos de ranqueamento, como apresentado a seguir:

Pontuação (SN) = [(k1* *freqüência* (Xar)) − (k2* *ocorrência (Ytot))* + *(*k3**CSN)*]

Sendo que:

Pontuação(SN): valor atribuído ao SN de acordo com os critérios apresentados. Quanto maior for este valor, maior será a relevância esperada deste SN como descritor;

freqüência(Xar) = freqüência do SN no artigo, com valor possivelmente limitado à **X** de modo a corrigir distorções;

ocorrência(Ytot) = número de artigos em que o SN ocorre com freqüência maior que **Y**;

X, Y, k1, **k2** e **k3** = constantes a serem ajustadas de acordo com os testes, de modo a conseguir a performance ótima;

CSN = categoria do SN, que assume um valor segundo a estrutura sintática e nível do SN, de acordo com a tabela a seguir:

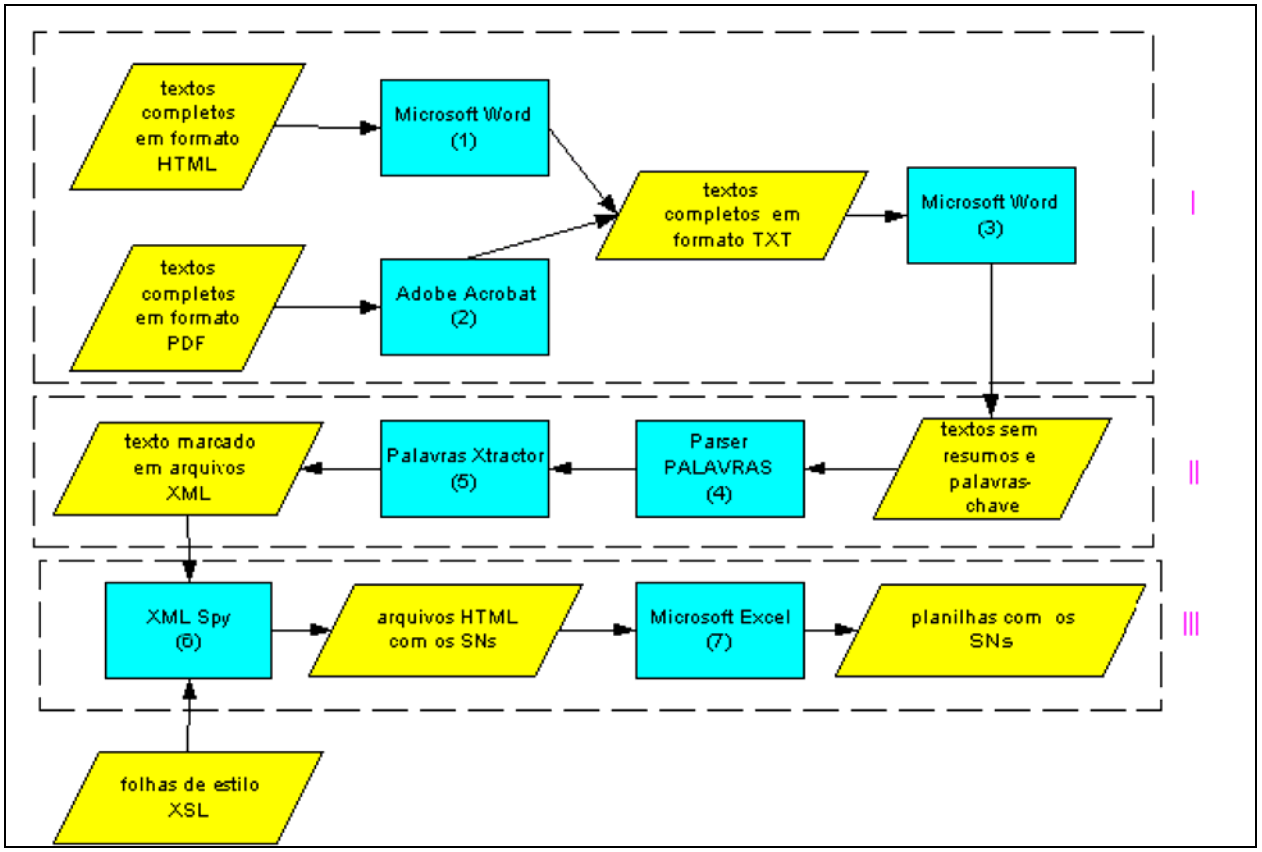
CSN	Estrutura e Nível do SN	Valor associado
1a	Nível 1, estrutura (D + N)	0,25
1b	Nível 1, qualquer estrutura exceto (D + N)	0,75
2	Nível 2, qualquer estrutura	1,0
3	Nível 3, qualquer estrutura	0,75
4	Nível 4, qualquer estrutura	0,5
5	Nível 5 ou superior, qualquer estrutura	0,25

Valor atribuído ao SN de acordo com sua estrutura sintática e nível

A caracterização dos graus de relevância dos SNs como descritores e a conseqüente validação da metodologia será estabelecida através da comparação destes descritores com as palavras-chave e resumos atribuídos pelos autores dos documentos originais.

4 FERRAMENTAS UTILIZADAS

O trabalho de análise, necessário à consecução da metodologia, acima descrita pressupõe um enorme esforço computacional, ao longo do processo. Para que seja possível a análise dos descritores, os SNs tiveram que ser extraídos automaticamente e de forma bastante veloz, mas este processo é composto por várias etapas. A figura a seguir explicita os relacionamentos entre os processos e as ferramentas de software utilizadas:



Ferramentas utilizadas na metodologia

As ferramentas foram utilizadas na seguinte seqüência:

I. Os textos dos corpora foram escolhidos pelo autor e transformados em formato de texto simples, sem caracteres especiais, utilizando as ferramentas ADOBE ACROBAT e MICROSOFT WORD;

II. Em seguida, os textos tratados foram submetidos sucessivamente ao processamento do analisador sintático (parser) "PALAVRAS", da Southern University of Denmark, e ao software "Palavras Xtractor", desenvolvido em conjunto pela Universidade do Vale do Rio dos Sinos (Unisinos) de São Leopoldo, RS, e a Universidade de Évora, em Portugal, tendo como resultado, os documentos sintaticamente marcados em arquivos XML;

III. Após a identificação sintática das palavras dos textos, foi utilizado o software XML SPY para aplicação da transformação XSL nos arquivos XML com uma folha de estilos específica (como explicado na seção 3.1.2), para extração de arquivos HTML com os SNs, e estes SNs foram tratados estatisticamente utilizando o software MICROSOFT Excel.

Não é possível, no escopo deste artigo, apresentar o funcionamento das ferramentas em detalhe. Este detalhamento pode ser verificado em SOUZA (2005).

5 RESULTADOS

Nesta seção serão apresentados os resultados da aplicação da metodologia delineada anteriormente a um *corpus* de 60 documentos textuais digitalizados, dividido nos seguintes conjuntos:

a) O primeiro com 30 textos, sendo que 29 provenientes da Revista *DataGramaZero*, e 1 proveniente da Revista *Ciência da Informação*;
b) O segundo com 30 textos, todos provenientes da Revista *Ciência da Informação*.

Os textos provenientes da revista *Ciência da Informação* apresentam um tamanho ligeiramente maior. A aplicação e análise de forma isolada da metodologia consolidada permitiram vislumbrar as diferenças decorrentes do tamanho dos documentos.

Os valores de parâmetros constantes, como apresentado na metodologia, foram escolhidos de forma arbitrária, e devem ser modificados e testados de forma exaustiva, em

subseqüentes pesquisas, visando refinar paulatinamente a metodologia. Estes valores e parâmetros são apresentados a seguir:

a) O número de descritores escolhidos para cada documento foi calculado tendo como base 1% dos SNs únicos identificados no documento, e levando em conta e os limites inferior de 8 e superior de 15 descritores por documento. Este valor é limitado apenas por uma conveniência metodológica, não havendo limitações reais para a escolha do número de descritores, excetuando o total de SNs extraídos;

b) Seguindo a fórmula introduzida na seção 6.1, os valores escolhidos para as constantes **X, Y, k1**, **k2** e **k3**, nas duas aplicações da metodologia ao *corpus* final são os apresentados na tabela a seguir:

Constantes	Conceituação	Conjunto de valores na primeira aplicação	Conjunto de valores na segunda aplicação
X	Valor máximo a ser considerado para a freqüência do SN no documento, para fins de pontuação.	10	7
Y	Limite inferior de freqüência do SN para o qual k2 se aplica.	3	3
k1	Ponderação da freqüência do SN no documento no cálculo da pontuação.	1	1
k2	Ponderação (negativa) da freqüência do SN no *corpus* de documentos no cálculo da pontuação.	10	15
k3	Ponderação da estrutura do SN no cálculo da pontuação.	10	15

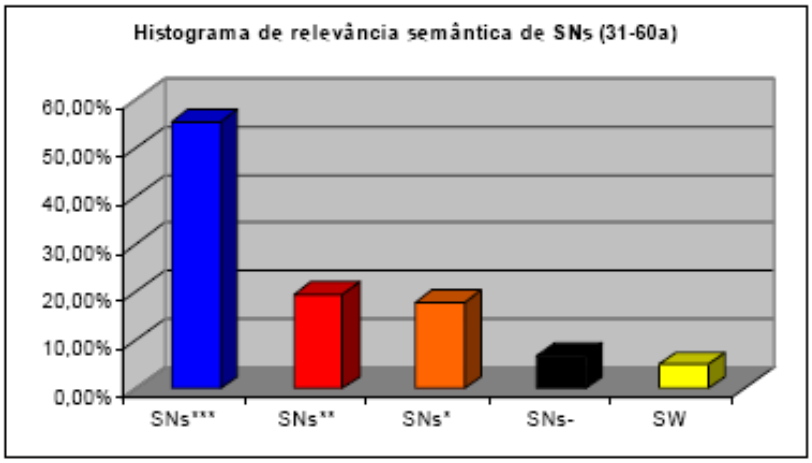
Conjunto de valores utilizados nas aplicações da metodologia

A tabela a seguir apresenta, para os dois conjuntos de parâmetros de aplicação da metodologia apresentados acima, e para os dois subconjuntos de documentos do *corpus*; as médias e os valores percentuais relativos de freqüência de SNs extremamente relevantes como descritores (SNs***), razoavelmente relevantes como descritores (SNs**), moderadamente relevantes como descritores (SNs*) e não relevantes como descritores (SNs-); além da média e o valor percentual dos "*stopwords"* (SW) em relação ao total dos SNs que foram eliminados.

		Textos de 1 a 30 do *corpus*			Textos de 31 a 60 do *corpus*		
I	Primeira aplicação da metodologia	SNs***	138	47,75%	SNs***	179	55,59%
		SNs**	66	22,84%	SNs**	63	19,57%
		SNs*	58	20,07%	SNs*	58	18,01%
		SNs–	27	9,34%	SNs–	22	6,83%
		SW	19	6,17%	SW	17	5,01%
II	Segunda aplicação da metodologia	SNs***	137	47,40%	SNs***	173	52,58%
		SNs**	64	22,15%	SNs**	64	19,45%
		SNs*	56	19,38%	SNs*	64	19,45%
		SNs–	32	11,07%	SNs–	28	8,51%
		SW	5	1,70%	SW	7	2,08%

Resultados da aplicação da metodologia de extração de descritores

Os resultados, na ótica do autor, superaram em muito a expectativa. Podemos perceber que nos piores casos, obtivemos os valores de 47% e 22,15%, para os SNs que constituem descritores de qualidade. Isto representa um total de quase 70% de bons descritores (extremamente relevantes + razoavelmente relevantes), como podemos observar no gráfico a seguir, que representa a primeira aplicação da metodologia, aos textos de 31 a 60:



Exemplo de histograma de resultados da aplicação da metodologia

Neste gráfico, pode-se perceber a quantidade expressiva de bons descritores obtidos pela aplicação da metodologia à parte do *corpus* utilizado, o que confirma o sucesso da metodologia apresentada.

6 CONCLUSÕES

A motivação da pesquisa surgiu da constatação freqüente da impossibilidade de organização manual de grandes acervos de documentos que são continuamente produzidos, como acontece em muitos contextos digitais. Nestes contextos, observamos amiúde processos de indexação automática que buscam descrever os documentos através da análise de freqüência das palavras que neles ocorrem. O objetivo central do trabalho era propor um processo de indexação mais eficaz, que analisasse as palavras e expressões dentro de seus contextos lingüísticos. Para tal, apresentou-se uma metodologia de indexação automática, viabilizando um processo de atribuição de descritores a documentos digitalizados. Estes descritores foram escolhidos através da extração de SNs e análise de fatores como a freqüência de ocorrência destes SNs nos textos dos documentos, no conjunto dos documentos; a estrutura dos SNs; o nível dos SNs e a ocorrência destes em um tesauro de um campo de conhecimento específico. A consideração destes fatores de forma conjunta permite a criação de um ranking de candidatos a descritores, a partir dos SNs extraídos.

A metodologia foi aplicada a um *corpus* de 60 documentos, com dois conjuntos de valores de parâmetros escolhidos, dentre um universo virtualmente ilimitado de possibilidades. Os testes exaustivos com outros conjuntos foram deixados como sugestões para trabalhos futuros.

Os resultados, considerados eminentemente positivos, contrariam experiências anteriores declaradamente malsucedidas, que buscavam a extração de descritores baseando-se em estruturas sintáticas das orações [(EARL, 1970; PAICE, 1981; Fum et. al., 1982) apud LANCASTER, 1993, pp. 250-251]. A bem da verdade, a inexistência até uma década de ferramentas que permitissem a extração automática de SNs é um fator preponderante a ser levado em conta neste sentido.

A teoria desenvolvida por KURAMOTO (1999, 2003), e seu modelo proposto de SRI já apontava alguns caminhos possíveis, embora estes ainda estejam em um estágio

inicial de exploração. A pesquisa desenvolvida em sua tese de doutorado apresenta um modelo de recuperação de informações, baseado em sintagmas nominais, buscando a participação do usuário na definição dos contextos lingüísticos. Infelizmente, não encontramos na literatura científica nacional nenhum indício de continuação destas pesquisas.

Espera-se que a metodologia apresentada – ou qualquer metodologia que derive desta – seja utilizada em situações onde é necessária a atribuição automática de descritores aos documentos, no escopo de funcionamento de SRIs. Usualmente, esta situação acontece quando os documentos são agregados ao sistema em uma taxa que não permite a apreciação manual.

REFERÊNCIAS

BAEZA-YATES, R.; RIBEIRO-NETO, B. **Modern Information Retrieval**. New York: ACM Press, 1999. 511p.

BICK, Eckhard. **Parsers and its applications**. (s/d) Disponível na Internet: http://www.hum.au.dk/lingvist/lineb/home_uk.htm. Consultado em 07/2003

____. **Automatic parsing of Portuguese**. In: Proceedings do II Encontro para o Processamento Computacional do Português Escrito e Falado, SBIA, 1996, Curitiba. Disponível na Internet: http://beta.visl.sdu.dk/~eckhard/postscript/curitiba.ps. Consultado em 07/2003

____. **The VISL System**: research and applicative aspects of IT-based learning. In: Proceedings of NoDaLiDa, 2001,Uppsala. Disponível na Internet: http://stp.ling.uu.se/nodalida01/pdf/bick.pdf. Consultado em 07/2003

GASPERIN, Caroline Varaschin; GOULART, Rodrigo Rafael Vilarreal; VIEIRA, Renata. **Uma Ferramenta para Resolução Automática de Correferência**. In: Anais do XXIII Congresso da Sociedade Brasileira de Computação, VI Encontro Nacional de Inteligência Artificial, v. VII. Campinas, 2003.

GASPERIN, Caroline Varaschin *et alii.* **Extracting XML chunks from Portuguese corpora**. In: Proceedings of the Workshop on Traitement automatique des langues minoritaires. 2003. Batz-sur-Mer.

KURAMOTO, Hélio. Uma abordagem alternativa para o tratamento e a recuperação de informação textual: os sintagmas nominais. **Ciência da Informação**, Brasília, v. 25, n. 2, 1996. Disponível na Internet: http://www.ibict.br/cionline/250296/25029605.pdf. Consultado em 07/2003.

____. **Proposition d'un Système de Recherche d'Information Assistée par Ordinateur Avec application à la langue portugaise**. 1999. Tese (Doutorado em Ciências da Informação e da Comunicação) – Université Lumière - Lyon 2, Paris, França.

LANCASTER, F. W. **Indexação e Resumos**: teoria e prática. Brasília, Briquet de Lemos,1993.

LIBERATO, Yara G. **A Estrutura do Sintagma Nominal em Português**: uma abordagem Cognitiva. 1997. 203 f. Tese (Doutorado em Letras) – Faculdade de Letras, Universidade Federal de Minas Gerais, Belo Horizonte.
MIORELLI, S. T. **Extração do Sintagma Nominal em sentenças em Português**. 2001. 98 f. Dissertação (Mestrado em Ciência da Computação) – Faculdade de Informática, Pontifícia Universidade Católica do Rio Grande do Sul, Porto Alegre.
MOREIRO, José; MARZAL, Miguel Ángel; BELTRÁN, Pilar. **Desarrollo de un Método para la Creación de Mapas Conceptuales.** V ENANCIB, Belo Horizonte, 2003. Anais. 1 CD.
PERINI, Mário A. **A Gramática Gerativa**: introdução ao estudo da sintaxe portuguesa. 2a edição. Belo Horizonte: Vigília, 1985. 254 p.
____. **Gramática descritiva do português**. 2a ed. São Paulo: Editora Ática, 1995. 380p.
PERINI, Mário A. *et alii*. O SN em português: A hipótese mórfica. **Revista de Estudos de Linguagem** - UFMG, Belo Horizonte, jul./dez. 1996. p. 43-56.
PROJETO DIRPI: Desenvolvimento e Integração de Recursos para Pesquisa de Informação. Cooperação Científica e Técnica Luso-Brasileira. ICCTI/GRICES-CAPES, Universidade de Évora, Universidade Nova de Lisboa, Unisinos, PUC-RS. Julho de 2001.
ROSSI, Daniela *et alii.* Resolução automática de Correferência em textos da língua portuguesa. **REIC** Revista de Iniciação Científica da SBC, http://www.sbc.org.br/reic/, v. 1, n. 2, 2001.
RUWET, Nicolas. **Introdução à Gramática Gerativa**. São Paulo: Perspectiva, Editora da Universidade de São Paulo, 1975. 357 p.
SALTON, Gerard; MCGILL, Michael J. **Introduction to modern information retrieval**. New York : Mcgraw-Hill Book Company, 1983. 448 p.
SANT'ANNA, V. **Cálculo de referências anafóricas pronominais demonstrativas na língua portuguesa escrita.** 100 f. 2000. Dissertação (Mestrado em Informática) – Instituto de Informática da PUC-RS – Porto Alegre.
SOUZA, Renato Rocha. **Uma proposta de metodologia para escolha automática de descritores utilizando sintagmas nominais**. 2005. 214 f. Tese (Doutorado em Ciência da Informação) – Escola de Ciência de Informação, Universidade Federal de Minas Gerais, Belo Horizonte.
SPARCK JONES, K.; WILLETT, P. (orgs.). **Readings in Information Retrieval**. San Francisco: Morgan Kaufmann, 1997. 589p.
VIEIRA, R. A review of the Linguistic literature on definite descriptions. **Acta Semiotica et Lingvistica**, v. 7, p. 219-258. 1998.
VIEIRA, R. et al. **Extração de Sintagmas Nominais para o Processamento de Co-referência**. Anais do V Encontro para o processamento computacional da Língua Portuguesa escrita e falada PROPOR, 19-22 Novembro 2000. Atibaia SP.
VISL. About VISL. Disponível na Internet: http://visl.hum.sdu.dk/visl/about/index.html. Consultado em 05/2003.

ABSTRACT
It can be noticed that the indexing and representation strategies nowadays seems to be near the exhaustion, and it is worth to investigate new approaches to the indexing and information retrieving systems. Among these, a branch tries to consider the intrinsic semantics of the textual documents using noun phrases as descriptors instead of single

keywords. We present in this article a methodology that was developed in the scope of a doctorate research.

KEYWORDS: Automatic indexing. Information retrieval. Noun phrases.

Originais recebidos em 13/12/2005.