Indexação e recuperação de teses e dissertações por meio de sintagmas nominais

Renato Fernandes Corrêa¹, Darliane Goes de Miranda², Camila Oliveira de Almeida Lima³, Tiago José da Silva⁴

Resumo

Introdução: Aborda a utilização dos sintagmas nominais no processo indexação automática das teses e dissertações depositadas na Biblioteca Digital de Teses e Dissertações da UFPE (BDTD-UFPE), considerando a hipótese de que os sintagmas nominais consistiriam numa melhor unidade de conhecimento para a indexação e recuperação de informação que as palavras isoladas, permitindo aumentar a satisfação da necessidade de informação do usuário durante a busca pela informação. Discute-se sobre o estado da arte dos sintagmas nominais e de sua extração automática, bem como sua aplicação na indexação automática e recuperação de informação. **Método:** Analisa, com base em ferramenta para análise de texto (OGMA), a aplicabilidade da extração de sintagmas nominais na indexação automática e recuperação de informação de teses e dissertações no contexto da BDTD-UFPE. Com base em resumos da área de Direito, Computação e Nutrição, definiram-se os valores totais para cada uma das variáveis observadas, o que permitiu avaliar a extração de sintagmas nominais através dos percentuais de precisão de sintagmas nominais relevantes; da taxa de erro ao extrair cadeias de caracteres que não constituem sintagmas nominais, e; do percentual de sintagmas nominais extraídos não relevantes. **Resultados:** O processo de extração de sintagmas nominais através do OGMA teve diferentes desempenhos para cada programa de pós-graduação, sendo obtido melhor desempenho (melhor índice de precisão) para resumos de Direito, seguidos dos de Computação e Nutrição. Esta diferença de desempenho pode em parte ser explicada pela diferente natureza dos termos técnicos presentes nos resumos. **Conclusões:** Embora existam limitações nas ferramentas disponíveis, a aplicação de métodos automatizados de extração e indexação por sintagmas nominais mostra-se promissora, pois os sintagmas nominais se configuram como melhores descritores e pontos de acesso aos documentos, eliminando os problemas causados pela sinonímia e a polissemia das palavras isoladas.

Palavras-chave

Sintagmas nominais; Recuperação de informação; Indexação automática; Teses e dissertações.

Introdução

Atualmente, vivencia-se uma transição das atividades relacionadas com a manipulação, a edição, o armazenamento, a distribuição das informações e a sua recuperação da forma impressa para a forma digital.

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.).

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, pesquisados, recuperando em tempo hábil informação relevante para algum objetivo específico.

Um dos grandes desafios encontrados na recuperação de informação é como atender às necessidades de informação do usuário de forma rápida e eficaz. Várias pesquisas foram e continuam sendo realizadas com o propósito de aumentar a precisão e revocação dos resultados de forma que o usuário possa encontrar todos os documentos que atendem às suas necessidades de informação.

O rápido desenvolvimento de novas tecnologias da informação criou um ambiente propício para o experimento de novos métodos de recuperação e de indexação da informação. A extração automática de sintagmas nominais faz parte da tentativa de indexar os documentos automaticamente.

O objetivo deste trabalho é discutir como os sintagmas nominais podem ser utilizados na indexação automática de teses e dissertações e prover uma melhor recuperação de informação na Biblioteca Digital de Teses e Dissertações da UFPE (BDTD-UFPE). Através de estudos de casos, analisa-se a extração automática de sintagmas nominais e a utilização dos mesmos como descritores e pontos de acesso a teses e dissertações de três programas de pós-graduação da UFPE.

Metodologia

Para a realização deste artigo foram inicialmente definidas as formas como essa pesquisa seria estruturada e fundamentada, em termos de revisão de literatura, estudo de uma ferramenta de extração de sintagmas nominais e planejamento e realização dos estudos de caso.

Primeiramente, para coletar informações a cerca do objeto de estudo, foram realizadas buscas por palavra-chave em *sites* como o Google Acadêmico, SciELO e BRAPCI. Os dois últimos escolhidos por oferecerem um acervo de materiais científicos confiáveis e em maior quantidade e o primeiro por ser um *site* de busca de referência mundial.

As buscas utilizadas para tal pesquisa nos três *sites* foram: sintagmas nominais, “sintagmas nominais”, sintagma + nominal, recuperação + informação + sintagma + nominal, extração + sintagma + nominal, sintagma + nominal + indexação.

Após a coleta de artigos científicos, o material foi organizado de acordo com a relevância para este artigo, uma vez que o foco é discutir a forma como os sintagmas nominais podem ser utilizados em sistemas de recuperação de informação. Após o estudo destes artigos, foi elaborada a revisão de literatura, descrita na próxima seção, que busca abordar os seguintes pontos: esclarecer termos como sintagmas nominais, sistemas de recuperação de informação, extração de sintagmas nominais e descrever a estrutura e uso de *softwares* de extração de sintagmas nominais.

Após o levantamento do estado da arte dos sintagmas nominais e sua extração, foram planejados e realizados os estudos de caso. Os estudos de caso consistiram da extração de sintagmas nominais, através da ferramenta OGMA¹, dos resumos de teses e dissertações de três programas de pós-graduação da UFPE: ciência da computação, direito e nutrição. Foram selecionados trinta resumos, procurando abranger as três grandes áreas do conhecimento (humanas, exatas e saúde), sendo desta forma escolhidos: dez resumos de Direito, dez de Ciências da Computação e dez de Nutrição. Os resultados dos estudos de caso são apresentados na seção Resultados.

Os resumos foram extraídos do sistema TEDE da BDTD-UFPE a partir dos metadados no formato MTD-BR das primeiras teses e dissertações inseridas no sistema para cada programa, ou seja, aquelas com menor valor do campo identificador. Definiu-se como resumo das teses e dissertações os valores dos seguintes seis campos presentes nos metadados das mesmas na seguinte ordem: título, titulação, resumo, assunto (contendo as palavras-chave), nome do programa, grande

¹ Ferramenta para análise de texto, cálculo da similaridade entre documentos e extração de sintagmas nominais (<http://www.luiзмаia.com.br/ogma/>)

área do programa. Foi incluído o caractere ponto como separador de campos, bem como múltiplos valores de um campo como no caso das palavras-chave.

O seguinte roteiro foi utilizado para extrair os sintagmas nominais através do OGMA:

- a) abrir o texto de cada resumo na ferramenta;
- b) etiquetar os termos do resumo aberto;
- c) extrair os Sintagmas Nominais Pontuados do resumo etiquetado.

A partir deste roteiro foi possível verificar o resultado dos processos de etiquetagem dos termos e da extração dos sintagmas nominais pelo OGMA.

A avaliação do processo de extração automática de sintagmas nominais dos resumos de cada programa de pós-graduação foi realizada através do cálculo e análise dos percentuais de precisão em extrair sintagmas nominais relevantes como descritores, a taxa de erro ao extrair cadeias de caracteres que não constituem sintagmas nominais e o percentual de sintagmas nominais extraídos não relevantes como descritores.

Para cálculo dos percentuais foi necessária a classificação dos possíveis sintagmas nominais extraídos como verdadeiros sintagmas nominais ou falsos sintagmas nominais, sendo esta classificação realizada com base nas definições de sintagmas nominais descritas na revisão de literatura. Bem como a classificação dos verdadeiros sintagmas nominais extraídos em relevantes ou não como descritores para o resumo de onde foram extraídos, sendo esta última realizada com base na análise de assunto do resumo e as palavras-chaves constantes no resumo da respectiva tese ou dissertação.

Estado da arte

Na presente seção apresenta-se o referencial teórico que norteou a pesquisa e deu

embasamento para a realização e conclusão deste trabalho.

Serão abordados nas subseções seguintes os conceitos de recuperação de informação e análise de texto, definição dos sintagmas nominais e a utilização dos mesmos pelos sistemas de recuperação de informação. Por fim, será apresentada uma ferramenta para extração de sintagmas nominais, descrevendo como se dá o processo de extração automática de sintagmas nominais.

Recuperação de informação

A recuperação de informação (RI) é uma área que lida com o armazenamento de documentos e a recuperação automática de informação associada aos mesmos (BAEZA-YATES & RIBEIRO-NETO, 1999). Um objeto informacional é geralmente constituído de texto, tais como documentos diversos, páginas web e livros, embora possa conter outros tipos de conteúdo, tais como imagens, áudios, gráficos e figuras. A representação e organização desses objetos devem permitir às pessoas o acesso à informação relevante a partir da expressão de uma necessidade de informação, por exemplo, por meio de uma consulta. A essência da RI consiste na busca de documentos relevantes a uma dada consulta que expressa a necessidade de informação do usuário.

De acordo com Baeza-Yates & Ribeiro-Neto (1999) os sistemas de recuperação de informação (SRIs) são “sistemas que lidam com as tarefas de representação, armazenamento, organização e acesso aos itens de informação” e devem apresentar o conteúdo do documento ao usuário de uma maneira que lhe permita uma rápida seleção dos itens que satisfazem total ou parcialmente a sua necessidade de informação, formalizada através de uma expressão de busca.

Um sistema de recuperação de informação normalmente implementa uma ferramenta de busca passiva de informações que computa a relevância potencial dos documentos encontrados a partir da análise de similaridade,

podendo ser definido como um conjunto de dados padronizados, armazenados em meio eletrônico, utilizados para identificar informação e fornecer sua localização.

Atualmente os SRIs trabalham com diversos modelos para trazer a informação até o usuário. Os modelos utilizados na recuperação de informação têm como base o uso da palavra como unidade básica de acesso à informação. Vários modelos foram desenvolvidos com o objetivo de facilitar o acesso à informação, porém os modelos mais utilizados pelos SRIs são o modelo Espaço Vetorial e o modelo Booleano.

Para que um SRI possa responder as demandas dos usuários com tempos de resposta aceitáveis, primeiramente é preciso que os documentos que estão na base de dados sejam submetidos a um tratamento. Esse procedimento permite a extração de descritores e sua estruturação com o intuito de garantir um rápido acesso às informações (SOUZA et al, 2007). Um descritor consiste de uma palavra ou expressão que identifica, geralmente para fins de indexação, determinado conceito ou tema.

De acordo com Souza (2006), o processo de indexação produzindo uma lista de descritores visa à representação dos conteúdos dos documentos com o objetivo de extrair as informações contidas nesses documentos organizando-as para permitir a recuperação destes últimos. Porém, na maioria dos SRIs atuais, os descritores não passam de palavras isoladas extraídas dos documentos que constituem o acervo.

Os primeiros estudos sobre avaliação de sistemas da informação relacionavam o nível de satisfação do usuário com o resultado apresentado pelo sistema de recuperação de informação, sendo quantificado pela porcentagem de resultados relevantes retornados por sua pesquisa. À medida que novos métodos, instrumentos e ferramentas com foco na relevância do resultado retornado dos SRIs foram utilizados, surgiu também a necessidade de identificar qual método - ou quais - seria capaz de retornar o maior número

de documentos relevantes para o usuário como resultado de uma consulta.

Análise de texto

Segundo Maia (2008), a análise de texto corresponde a uma área que envolve outras subáreas como, por exemplo, a mineração de texto e a área de processamento de linguagem natural (PLN). A PLN também é uma subárea da inteligência artificial e da linguística que estuda os problemas da geração e tratamento automático de línguas naturais.

A mineração de textos refere-se ao processo de extração de informação útil (que possa gerar conhecimento) em documentos de textos não-estruturados. É basicamente um conjunto de métodos usados para organizar e descobrir informações em bases de textos. Com base no conhecimento a ser extraído, a mineração de textos define técnicas de extração de padrões ou tendências de grandes volumes de textos em linguagem natural, normalmente, para objetivos específicos.

Já o processamento de linguagem natural, de acordo com Maia (2008), trata-se de um conjunto de métodos formais para analisar textos e gerar frases escritas em um idioma humano. Um dos objetivos finais da PNL é fornecer aos computadores a capacidade de entender e compor textos, ressaltando que “entender” um texto significa reconhecer o contexto, fazer análise sintática, semântica, léxica e morfológica, criar resumos, extrair informação, interpretar os sentidos e até aprender conceitos com os textos processados.

É neste contexto de análise textual que são apresentados os sintagmas nominais, pois com o aparente esgotamento das estratégias atuais de indexação e representação de documentos, faz-se necessário investigar novas abordagens para sistemas de recuperação de informação. 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 isoladas.

Sintagmas nominais

Segundo Dubois-Charlier (1977), chama-se sintagma uma sequência de palavras que constituem uma unidade. Um sintagma é uma associação de elementos compostos em um conjunto, organizados em um todo, funcionando conjuntamente. Sintagma significa, por definição, organização e relações de dependência e de ordem à volta de um elemento essencial.

O sintagma nominal “é a menor parte do discurso portadora de informação” (KURAMOTO, 1995), onde os signos linguísticos ligam-se uns aos outros formando grupos ao redor de substantivos. O sintagma nominal quando extraído do texto mantém o significado, o seu conceito. Por exemplo, na frase “O estudo dos sistemas de informação”, poderíamos extrair três sintagmas nominais, seriam eles:

- a) o estudo dos sistemas de informação;
- b) os sistemas de informação;
- c) informação.

De acordo com Miorelli (2001), os sintagmas nominais podem ser entendidos e tratados de forma sintática (privilegiando a forma) ou semântica, buscando os significados mais amplos, cada uma com suas especificidades e implicações. Segundo Perini (1998), os sintagmas nominais possuem duas estruturas: uma à esquerda do núcleo do sintagma que podem ser compostas por determinantes, possessivos, quantificadores e outras classes de palavras e, a estrutura à direita do núcleo que é composta por modificadores, que por sua vez podem ser classes abertas ou outros sintagmas.

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. Os sintagmas nominais geralmente são extraídos do texto

e analisados a fim de facilitar o processo de indexação automática.

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. Com isso, muitos pesquisadores viram a possibilidade de utilizá-los em um processo de indexação e recuperação de informação.

Segundo Kuramoto (2002), 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. 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.

A grande maioria dos modelos de recuperação de informação utilizou e utiliza a palavra como forma de acesso à informação (KURAMOTO, 2002). Entretanto, a palavra possui algumas propriedades que geram alguns inconvenientes, como por exemplo, a sinonímia (concordância dos diversos nomes dados a uma mesma coisa) e a polissemia (ligada ao fato de uma palavra ter muitas significações). Isto gera problemas de ambiguidade não tratados pelos atuais sistemas de recuperação de informação, que podem retornar documentos que contêm palavras com significados diferentes da solicitada pelo usuário.

A utilização dos sintagmas nominais em substituição às palavras isoladas eliminaria estes inconvenientes.

Para que os sintagmas nominais possam ser utilizados nos sistemas de recuperação de informação faz-se necessário realizar o processo de extração dos sintagmas nominais.

A utilização dos sintagmas nominais no processo de recuperação da informação

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

Os sistemas de recuperação de informação usualmente adotam termos índices para indexação de documentos, sendo que estes termos índice são usualmente palavras isoladas. Há uma ideia fundamental embutida de que a semântica dos documentos e as das necessidades de informação do usuário pode ser expressa através destes conjuntos de palavras. Porém, isto 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).

O processo de recuperação de informação baseado em sintagmas nominais, segundo Kuramoto (2002), aponta um potencial natural de organização que, se explorado convenientemente, poderia propiciar aos usuários maior facilidade no uso de um SRI e o retorno de resultados mais precisos em resposta a um processo de busca de informação. Ainda segundo o mesmo autor, a organização baseada em sintagmas nominais permite a navegação na estrutura hierárquica em árvore dos sintagmas nominais.

Considerando o exemplo de sintagma nominal citado anteriormente, 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 (“informação”) o nível 1, ao sintagma que o contém (“os sistemas de informação”) seria atribuído o nível 2 e ao sintagma que contém os dois outros (“o estudo dos sistemas de informação”) seria enumerado como sendo o de nível 3, conforme abaixo:

- a) o estudo dos sistemas de informação (nível 3);
- b) os sistemas de informação (nível 2);
- c) informação (nível 1).

Uma interface de busca, baseada nessa organização, poderia funcionar da seguinte forma:

- a interface de busca aguarda que o usuário forneça um termo ou palavra que represente o centro do sintagma nominal de primeiro nível, por exemplo: informação;
- 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;
- 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.

extração de sintagmas nominais em textos escritos em português foi o *software* OGMA.

Ferramentas para extração de sintagmas nominais

Dentro do referencial teórico não foram encontrados muitos trabalhos, nem ferramentas que tinham como foco a extração de sintagmas nominais. A única ferramenta encontrada para

O OGMA é uma ferramenta para análise de texto, cálculo da similaridade entre documentos e extração de sintagmas nominais. O aplicativo foi desenvolvido por Maia (2008) na ferramenta *Visual Studio.NET* em linguagem C#. O OGMA realiza também a identificação da classe do sintagma nominal, bem como o cálculo da

Quadro 1 – Regras de extração de Sintagmas Nominas do software OGMA

AR ← AD	de ← AR	AV ← AV ad	re ← SU	NS ← MD NS
AR ← AI	de ← PD	MD ← AV MD	de ← PP	NS ← NS pr NS
AJ ← VP	de ← PI	MD ← MD co MD	re ← NP	NS ← NS pr de NS
NU ← NR	qu ← AJ	NS ← NS MD	NS ← re	NS ← NS co NS
NU ← NC	qu ← NU	co ← CO	MD ← qu	NS ← NS co de NS
CO ← VG	qu ← PS	pr ← PR	SN ← NS	NS ← AV NS
CO ← CJ	ad ← AV		AV ← ad	SN ← de SN

Fonte: Maia (2008).

pontuação do mesmo como descritor de forma automática.

Para realizar a extração de sintagmas nominais o OGMA faz uso de um léxico da língua portuguesa construído a partir do vocabulário utilizado pelo dicionário BR.ISPELL e uma lista de 475 palavras irrelevantes criada tendo como base a gramática de Tufano (1990).

O léxico é utilizado para etiquetar cada palavra do texto com as possíveis classes gramaticais correspondentes.

Para resolver problemas de ambiguidade, o OGMA forma uma lista com todas as combinações de etiquetas encontradas para palavras de uma frase e submete cada combinação às regras para extração dos sintagmas nominais. Por exemplo, na frase: “O mato estava grande”, a ferramenta etiquetaria o texto da seguinte forma:

“O/AD mato/VBSU estava/VB grande/AJ”. Posteriormente, o OGMA submete às regras de extração as duas versões da frase etiquetada com diferentes combinações de etiquetas: “O/AD mato/VB estava/VB grande/AJ”; e “O/AD mato/SU estava/VB grande/AJ”.

Para extrair os sintagmas nominais, o OGMA faz uso do conjunto de regras presentes no Quadro 1, aplicando regra por regra na ordem de leitura até obter um sintagma nominal cujo símbolo é SN. Estas regras atuam sobre as etiquetas (que representam as classes gramaticais) atribuídas às palavras, visando marcar o início e fim do sintagma em cada sentença do texto. A correspondência entre as etiquetas presentes nas regras e as respectivas classes gramaticais podem ser obtidas em MAIA (2008).

Os Sintagmas Nominais encontrados entram em uma lista geral de Sintagmas Nominais da frase, e os duplicados são eliminados. Este tratamento possibilitou resolver o problema da ambiguidade de forma bem eficiente.

Resultados

Com o objetivo de avaliar o processo de extração de sintagmas nominais, utilizou-se a ferramenta OGMA nos resumos de teses e dissertações de três programas de pós-graduação da UFPE: Ciência da Computação, Direito e Nutrição.

Após a extração de sintagmas nominais de cada resumo através do OGMA, contabilizaram-se o número de possíveis sintagmas nominais

Tabela 1 – Resultado da extração de sintagmas nominais nos resumos de Teses e Dissertações em Direito/UFPE

Resumo	Possíveis sintagmas extraídos	Não constituem sintagmas	Constituem sintagmas não relevantes	Constituem sintagmas relevantes
D1	46	17	10	19
D2	39	12	4	23
D3	54	18	4	32
D4	55	20	7	28
D5	44	13	6	25
D6	41	9	7	25
D7	55	16	10	29
D8	38	14	9	15
D9	29	5	9	15
D10	63	23	6	34
Total	464	147	72	245

Fonte: os autores.

Tabela 2 – Avaliação da extração de sintagmas nominais nos resumo de Teses e Dissertações dos programas de Ciência da Computação, Direito e Nutrição da UFPE

Programa	Precisão	Taxa de Erro	Percentual de não relevantes
Computação	0,47	0,33	0,20
Direito	0,53	0,32	0,16
Nutrição	0,47	0,33	0,20
Média	0,49	0,32	0,19
Desvio	0,03	0,01	0,03

Fonte: os autores.

extraídos, quantos destes não constituem sintagmas nominais; quantos destes se constituem em sintagmas, mas não são relevantes como descritores para o respectivo resumo, e; quanto destes se constituem em sintagmas nominais relevantes como descritores para o respectivo resumo. Estes dados foram organizados em tabelas para cada programa de pós-graduação, como a mostrada na Tabela 1 para o programa de pós-graduação em direito.

A partir das três tabelas construídas (uma para cada programa de pós-graduação estudado) definiram-se os valores totais para cada uma das variáveis observadas, o que permitiu avaliar a extração de sintagmas nominais através dos percentuais de precisão em extrair sintagmas nominais relevantes; da taxa de erro ao extrair cadeias de caracteres que não constituem

sintagmas nominais, e; o percentual de sintagmas nominais extraídos não relevantes. Estes percentuais estão organizados na Tabela 2.

Na Tabela 2, pode-se observar que a taxa de erro do OGMA na identificação de sintagmas nominais foi em média de 42%, indicando que 42% dos possíveis sintagmas nominais extraídos pelo OGMA não são de fato sintagmas nominais. Os demais 58% dos possíveis sintagmas nominais extraídos são de fato sintagmas nominais. Entretanto, em média, a precisão na extração de sintagmas nominais relevantes dos resumos foi de 44%, significando que menos da metade dos possíveis sintagmas nominais extraídos pelo OGMA são de fato sintagmas nominais relevantes como descritores para os resumos, sendo 14% sintagmas nominais irrelevantes como descritores.

Gráfico 1: Desempenho na extração automática de sintagmas nominais

Fonte: os autores.

A partir da Tabela 2, foi construído o Gráfico 1, onde se pode observar com mais clareza que o processo de extração de sintagmas nominais através do OGMA teve diferentes desempenhos para cada programa de pós-graduação, sendo obtido melhor desempenho (melhor índice de precisão) para resumos de Direito, depois

Computação e Nutrição. Esta diferença de desempenho pode em parte ser explicada pela diferente natureza dos termos técnicos presentes nos resumos destes três programas: enquanto a terminologia da área do Direito faz uso vocábulos mais comuns (provavelmente mais presentes no léxico do OGMA), em Computação há o

Quadro 2 – Sintagmas nominais relevantes extraídos do resumo D6 –
do programa de pós-graduação de Direito da UFPE

A tutela [dos direitos da personalidade](#) por meio da [aplicabilidade direta do princípio da dignidade da pessoa humana](#) nas relações de direito privado.

[Mestre em Direito.](#)

O presente trabalho faz [um estudo do sistema de proteção dos direitos da personalidade no ordenamento jurídico brasileiro](#) com vistas a demonstrar [a possibilidade de aplicação direta do princípio constitucional da dignidade da pessoa humana nas relações de direito privado](#). Esta possibilidade de apresenta a partir da constatação da mudança paradigmática da codificação experimentada com a promulgação do Código Civil de 2002 que adotou como uma das suas orientações metodológicas a técnica [das cláusulas gerais](#). Por meio da efetividade das chamadas [cláusulas gerais](#), [o sistema normativo de direito privado](#) deixa de ser um sistema fechado e passa a estabelecer conectividade com todo o ordenamento. [A personalidade humana](#) não se restringe a um requisito formal do sujeito de direito. Ela é mais do que isso, é um valor juridicamente tutelado. A tutela da personalidade confere unidade ao ordenamento jurídico, e torna evidente [a superação da tradicional dicotomia entre o direito público e o direito privado](#). [A aplicação direta do princípio da dignidade da pessoa humana nas relações de direito privado](#) reflete e superação dessa dicotomia e estabelece [um verdadeiro sistema de proteção da pessoa humana no ordenamento jurídico brasileiro](#).

[Cláusulas gerais](#). [Dignidade da pessoa humana](#). [Personalidade](#). [DIREITO](#).

[Pós-Graduação em Direito.](#)

Ciências Humanas, Letras e Artes.

Fonte: os autores.

uso frequente de estrangeirismos e siglas, e em Nutrição há uso frequente de siglas, unidades de medidas e nomes científicos de animais, plantas e substâncias.

Entende-se que a precisão na extração de sintagmas nominais relevantes em torno de 50%, gera uma sobrecarga sobre profissional da informação atuando como indexador, que teria que avaliar os possíveis sintagmas nominais extraídos e praticamente descartar metade dos mesmos.

Entretanto, avaliando a qualidade dos sintagmas nominais relevantes extraídos como descritores dos resumos das teses e dissertações, percebe-se que este trabalho de indexação semi-automática, mesmo que enfadonho, é recompensador e que a aplicação dos sintagmas nominais na indexação de teses e dissertações é eficaz. Sustentando esta afirmativa, apresenta-se a taxa de revocação das palavras-chaves das teses e dissertações nesta lista de sintagmas nominais relevantes extraídos que é de 0,88 para Direito, 0,84 para Computação e 0,83 para Nutrição. Em outras palavras, cerca de 80% das palavras-chaves atribuídas às teses e dissertações são extraídas (recuperadas) como sintagmas nominais. Além disso, apresentam-se grifados no Quadro 2, para fins de exemplificação, os sintagmas nominais relevantes extraídos do resumo D6 do programa de pós-graduação de Direito, neste resumo a extração de sintagmas nominais obteve um dos melhores desempenhos. Observa-se, para o Quadro 2, que foi usado um código de cores para realçar quais sintagmas nominais estão relacionados a qual palavra-chave.

Analisando o Quadro 2, percebe-se que todas as palavras-chaves foram detectadas como sintagmas nominais (antepenúltima linha do quadro), e que os sintagmas nominais são mais específicos que as palavras-chaves que neles estão contidos. Este é o caso, por exemplo, da palavra-chave Personalidade, que fica menos ambígua e contextualizada através dos sintagmas: a personalidade da pessoa humana; os direitos da personalidade; a personalidade no ordenamento jurídico brasileiro; um estudo do sistema de

proteção dos direitos da personalidade no ordenamento jurídico brasileiro. O conceito ordenamento jurídico, que também se constitui como um importante descritor, também foi detectado como parte de sintagmas nominais, e complementar bem a lista de palavras-chaves.

- alguns dos sintagmas nominais não representam relevância para o usuário no momento da busca, ou seja, embora sejam sintagmas, não constituem descritores e não corresponderiam à necessidade de informação do usuário no momento da busca por aquele documento. Tal fato mostra que a extração de sintagmas deve ser acompanhada de estratégias de ordenação por relevância dos sintagmas, levando em contra critérios de frequência e posicionamento, semelhantemente às propostas existentes para palavras isoladas. O OGMA gera uma pontuação de cada sintagma nominal que indicaria os mais relevantes como descritores, entretanto nem sempre a lista de sintagmas extraídos ordenada em ordem decrescente de pontuação aponta no topo os sintagmas mais relevantes;
- nem todos os sintagmas extraídos pelo OGMA são de fato sintagmas nominais, os motivos para a ferramenta retornar tais falsos positivos pode ser atribuída a um problema na etiquetagem do texto e uma possível falha nas regras de extração estabelecidas pelo OGMA;
- nem todos os sintagmas nominais são extraídos pelo OGMA como, por exemplo, os nomes próprios. Isto se deve a uma falha nas regras de extração estabelecidas pelo OGMA, que não contempla a possibilidade de dois ou mais substantivos adjacentes formarem um sintagma nominal;
- os sintagmas nominais extraídos classificados como relevantes constituem bons descritores para os resumos, constituindo em bons pontos de acesso as teses e dissertações no processo de recuperação de informação, de semântica muito mais precisa que

o conjunto de palavras isoladas que os compõem.

Baseando-se nestas observações, conclui-se que os processos de extração de sintagmas nominais e ordenação por relevância precisam ser mais profundamente investigados visando melhora na extração de sintagmas nominais relevantes como descritores.

Conclusão

Um dos principais objetivos dos profissionais e cientistas da informação é garantir o acesso rápido e preciso à informação solicitada pelo usuário. A busca por novos métodos, novos caminhos para satisfazer essa necessidade são implementados e testados com o intuito de chegar mais próximo do ideal, ou seja, retornar o maior número de documentos relevantes à necessidade de informação do usuário.

A aplicação de métodos automatizados de extração e indexação por sintagmas nominais, embora se reconheçam as limitações nas ferramentas disponíveis, mostra-se bastante promissora. Os sintagmas nominais se apresentam como uma alternativa ao uso de palavras isoladas nos sistemas de recuperação de informação, pois se configuram como melhores descritores e pontos de acesso aos documentos, eliminando os problemas causados pela sinonímia e a polissemia das palavras isoladas.

A literatura é carente de ferramentas para extração de sintagmas nominais em textos escritos em português, sendo o OGMA a única ferramenta disponível que se teve conhecimento durante a execução do estudo.

É possível perceber que os métodos para extração de sintagmas nominais implementados no OGMA precisam ser aperfeiçoados a fim de diminuir a taxa de falsos sintagmas nominais extraídos. Além disso, a extração dos sintagmas nominais não garante por si só a seleção de bons descritores, sendo necessário que a ferramenta de extração de sintagmas possa fazer a análise

dos textos e pontuar os sintagmas de acordo com a pontencialidade de serem bons descritores, este método de ordenação também precisa ser aprimorado no OGMA.

Assim, a extração automática de sintagmas nominais e utilização dos mesmos na indexação e recuperação de informação se constitui um amplo espaço de pesquisa e desenvolvimento de novos trabalhos. Para trabalhos futuros na temática, pretende-se avaliar a revocação na extração de sintagmas nominais; investigar com maior profundidade o processo de etiquetagem dos termos, extração e pontuação de sintagmas nominais com a finalidade de propor novos métodos; bem como investigar e aplicar outros métodos e ferramentas da área de processamento de linguagem natural nestes processos.

Referências

- BAEZA-YATES, R.; RIBEIRO-NETO, B. **Modern information retrieval**. New York: ACM Press, 1999.
- KURAMOTO, H. Sintagmas nominais: uma nova proposta para a recuperação de informação. **DataGramaZero**: revista de Ciência da Informação, v. 3, n. 1, 2002.
- _____. Uma abordagem alternativa para o tratamento e a recuperação de informação textual: os sintagmas nominais. **Ciência da Informação**, Brasília, v. 25, n. 2, 1995.
- LE GUERN, M. Un analyseur morpho-syntaxique pour l'indexation automatique. **Le Français Moderne**, v. 59, n. 1, p. 22-35, juin 1991.
- MAIA, L. C. G. **Uso de sintagmas nominais na classificação automática de documentos eletrônicos**. 2008. Tese (Doutorado em Ciência da Informação) – Universidade Federal de Minas Gerais – UFMG. Belo Horizonte, 2008.
- 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.
- PERINI, M. A. **Gramática descritiva do Português**. 3 ed. São Paulo: Ática, 1998.

SOUZA, R. R.; ALVARENGA NETO, R. C. D. de;
MENDES, K. C. I. Mapeamento semântico através da
análise de ocorrência de descritores sobre gestão do
conhecimento. **Transinformação**, v. 19, n. 1, p. 19-30,
2007.

SOUZA, R. R. Uma proposta de metodologia para
indexação automática utilizando sintagmas nominais.
Encontros Bibli: revista eletrônica de Biblioteconomia
e Ciência da Informação, v. 11, n. esp., p. 42-59, 2006.

TUFANO, D. **Estudos de língua e literatura**. 4. ed. São
Paulo, Moderna, 1990.

Indexing and information retrieval of theses and dissertations through noun phrases

Abstract

Introduction: Discusses the use of noun phrases in the automatic indexing process of theses and dissertations deposited in the UFPE Digital Library of Theses and Dissertations (BDTD-UFPE), on the assumption that noun phrases consist of a better knowledge unit for indexing and information retrieval that individual words, allowing an adequate response to the users information need when searching for information. It presents the state of the art of noun phrases and their automatic extraction process, as well as its applicability in automatic indexing and information retrieval. Method: Based on text analysis tool (OGMA), analyses the applicability of the extraction of noun phrases in automatic indexing and information retrieval of thesis and dissertations in the context of BDTD-UFPE. Applied to abstracts from Law, Computer and Nutrition thesis and dissertations, the variables could be observed, allowing the research team assess the extraction of noun phrases using: the percentage of accuracy of relevant noun phrases; the error rate extract strings that are not noun phrases, and; the percentage of non relevant noun phrases extracted. Results: The process of extracting noun phrases by OGMA showed different performances for each graduate program, with better performance (better accuracy rate) for abstracts from Law Thesis and Dissertations, followed by Computer and Nutrition ones. This performance difference can be partly explained by the different nature of technical terms presented in the abstracts. Conclusions: It concludes that although there are limitations in the available tools, the application of automated methods of extraction and indexing by noun phrases appears to be quite promising, since the noun phrases are configured as best descriptors and access to documents, eliminating the problems caused by synonymy and polysemy of isolated words.

Keywords

Noun phrase; Information retrieval; Automatic indexing;
Theses and dissertations.

Agradecimentos

Os autores gostariam de agradecer a Fundação de Amparo à Pesquisa do Estado de Pernambuco (FACEPE) pelo fomento dado ao projeto "Mapeador de teses e dissertações da UFPE (MTD-UFPE)" cujos equipamentos foram utilizados nesta pesquisa e ao laboratório Liber da UFPE pela estrutura física disponibilizada para realização deste trabalho.

Recebido em 14 de fevereiro de 2011

Aceito em 15 de fevereiro de 2011

¹ Bacharel em Gestão da Informação (UFPE), Doutor em Ciência da Computação (UFPE). Professor adjunto - Universidade Federal de Pernambuco (UFPE)/Departamento de Ciência da Informação.
renato.correa@ufpe.br

² Bacharel em Gestão da Informação (UFPE). Estudante de graduação - Universidade Federal de Pernambuco (UFPE)/ Departamento de Ciência da Informação.
darlygoes@gmail.com

³ Bacharel em Gestão da Informação (UFPE). Estudante de graduação - Universidade Federal de Pernambuco (UFPE)/ Departamento de Ciência da Informação.
camila.oalima@gmail.com

⁴ Bacharel em Gestão da Informação (UFPE). Estudante de graduação - Universidade Federal de Pernambuco (UFPE)/ Departamento de Ciência da Informação.
tiago.stallin@gmail.com

Como citar este artigo:

CORRÊA, R. F.; MIRANDA, D. G. de; LIMA, C. O. de A.; SILVA, T. J. da. Indexação e recuperação de teses e dissertações por meio de sintagmas nominais. **AtoZ**, Curitiba, v. 1, n. 1, p. 11-22, jan./jun. 2011. Disponível em: <<http://www.atoz.ufpr.br>>. Acesso em: