XVII Encontro Nacional de Pesquisa em Pós-Graduação em Ciência da Informação – **(XVII ENANCIB)**

GT 2 - Organização e Representação do Conhecimento

SINTAGMAS NOMINAIS COM VALOR DE DESCRITORES: critérios para seleção

NOUN PHRASES WITH VALUE OF DESCRIPTORS: CRITERIA FOR SELECTION

Gustavo Diniz Nascimento[1] e Renato Fernandes Corrêa[2]

Modalidade: Comunicação Oral

Resumo: Através de pesquisa bibliográfica e experimento, investiga critérios para seleção de sintagmas nominais com valor de descritores a partir das metodologias de seleção de sintagmas nominais para fins de indexação automática de documentos em língua portuguesa. Via experimento envolvendo análise de um corpus de documentos, avalia critérios para seleção de sintagmas nominais mais representativos de documentos da área jurídica escritos em língua portuguesa. No experimento foram aplicados dez critérios de seleção identificados na literatura aos sintagmas nominais extraídos de um conjunto de resumos de teses e dissertações da área jurídica. Os sintagmas nominais selecionados foram comparados com os descritores atribuídos através da indexação intelectual com o intuito de mensurar a utilidade ou não desses critérios no que diz respeito à seleção de sintagmas nominais tidos como descritores. Cada critério obteve um nível de eficácia diferente, sendo a maioria considerada útil para a seleção de sintagmas nominais.

Palavras-chave: Sintagmas Nominais. Indexação Automática. Representação da Informação. Seleção de Sintagmas Nominais. Informação Jurídica.

Abstract: *This work, through bibliographic research and an experiment, investigates the criteria for selection of noun phrases with descriptor value for automatic indexing of*

[1]Graduado em Biblioteconomia. Especialista em Biblioteconomia. Mestre em Ciência da Informação pela Universidade Federal de Pernambuco e Bibliotecário/Documentalista da Universidade Federal de Campina Grande-PB.

[2]Doutor em Ciência da Computação. Docente do Programa de Pós-Graduação em Ciência da Informação da UFPE.

documents written in Portuguese. Based in an experiment involving analyze of corpus, it evaluates identified criteria in selection of the most representative noun phrases of legal documents written in Portuguese. In the experiment were applied ten selection criteria to noun phrases extracted from a set of abstracts of theses and dissertations in the legal area. We compare the selected noun phrases with descriptors assigned by indexers in order to measure the usefulness or not of these criteria in the selection of noun phrases with value of descriptor. Each criteria had distinct behavior, where the majority was useful for the selection of noun phrases.

Keywords: *Noun phrases. Automatic indexing. Information representation. Noun phrases selection. Legal Information.*

1 INTRODUÇÃO

A indexação de documentos é uma das principais atividades desempenhadas com o intuito de organizar a informação que vem sendo produzida e que em outro momento deverá ser utilizada por um indivíduo que dela necessita.

O ambiente virtual, por suas características peculiares, disponibiliza uma maior quantidade de informações, o que acarreta em uma necessidade maior de procedimentos de organização dessas informações. Com o grande volume de informação disponibilizado nos meios virtuais, a indexação automática se mostra como uma alternativa à indexação intelectual.

A indexação automática se desenvolveu inicialmente baseando-se nas palavras que se encontravam nos textos, no entanto, a utilização desse recurso, ou seja, a palavra "isolada" do texto, como recurso de acesso a informação veio se mostrando inviável devido a fenômenos inerentes à língua portuguesa, como, por exemplo, a polissemia, a sinonímia. Nesse contexto, dentre as metodologias de indexação automática que vem se aprimorando, identificam-se algumas que levam em consideração a semântica intrínseca aos documentos, como, por exemplo, a utilização de sintagmas nominais como descritores representativos de conteúdos de documentos ao invés da utilização das "palavras isoladas". Segundo Kuramoto (2002. p. 6) "O sintagma nominal é a menor parte do discurso portadora de informação", em outras palavras, é o conjunto de palavras que possuem significado, possuem um sentido.

Apesar do grande potencial que os sintagmas nominais têm no que se refere à representatividade informacional, é preciso que se tenha em mente que nem todo sintagma nominal que se encontra em um texto tem potencial para representar o conteúdo temático desse texto. Ou seja, a extração automática de sintagmas nominais de um texto não resultará,

necessariamente, em descritores documentais[3]. Isto ocorre porque apesar dos sintagmas nominais possuírem uma grande carga de semântica intrínseca em suas estruturas, nem todos serão representativos do conteúdo informacional desse documento.

Nesse contexto, surge a necessidade de não apenas extrair os sintagmas nominais, mas selecionar os que realmente possam funcionar como descritores documentais dos documentos em bibliotecas digitais e sistemas de recuperação de informação.

Este trabalho tem como objetivo geral investigar as metodologias de seleção de sintagmas nominais com valor de descritores para fins de indexação automática de textos escritos em português e, com isso, avaliar os critérios presentes na literatura para selecionar os sintagmas nominais mais representativos de um determinado documento em língua portuguesa.

Dessa forma, este artigo apresenta os resultados alcançados com revisão de literatura sobre a temática e com a avaliação de dez critérios de seleção de sintagmas nominais aplicados a um conjunto de trinta documentos compostos por títulos e resumos de dissertações e teses da área jurídica indexados na Biblioteca Digital de Teses e Dissertações da Universidade Federal de Pernambuco.

2 INDEXAÇÃO AUTOMÁTICA E CRITÉRIOS DE SELEÇÃO DE SINTAGMAS NOMINAIS

No que se refere à conceituação da indexação automática, Vieira (1988, p. 48), de forma simples e sucinta, define-a como "uma operação que identifica, através de programas de computador, palavras ou expressões significativas dos documentos para descrever de forma condensada o seu conteúdo".

A indexação automática desenvolveu-se, inicialmente, baseando-se na frequência de ocorrência das palavras que ocorriam no documento, onde as palavras que tinham uma frequência alta de ocorrência tendiam a serem candidatas a descritores do documento a que pertenciam (BORGES; MACULAN; LIMA, 2008).

A utilização das "palavras isoladas" na indexação automática foi, aos poucos, mostrando-se ineficiente, isso, causado por fenômenos da língua natural, como a sinonímia, a polissemia, e a combinação de palavras em ordens distintas, causando assim, diversos

[3] Descritores documentais são considerados no escopo deste trabalho, como termos de indexação relativos a um documento, convencionalmente, palavras ou conjunto de palavras que representem conceitos relacionados ao(s) assunto(s) principal(is) desse documento.

significados. Com isso, foram sendo desenvolvidas pesquisas voltadas para a indexação automática, buscando sempre outras metodologias que levassem em conta a semântica do texto e que também suprisse alguns problemas identificados na indexação baseada apenas nas palavras isoladas. É nesse contexto que surgem pesquisas em indexação automática que fazem uso dos sintagmas nominais.

O pioneiro na ideia de se utilizar os sintagmas nominais como descritores ao invés das palavras isoladas foi Michel Le Guern (1991). Esse autor é responsável pelo desenvolvimento conceitual acerca desse recurso como unidade portadora de significado para a indexação e recuperação de informação. O referido autor faz uma distinção relevante entre descritor e palavra, uma vez que o descritor utilizado para a recuperação de informação deveria ser uma unidade do discurso e não uma unidade da língua (signo isolado sem significado).

Os descritores deveriam fazer referência à realidade extralinguística do autor. A palavra como uma unidade da língua constitui um conjunto de propriedades, no entanto sem referência a realidade extralinguística. Complementando esse entendimento, Kuramoto (1995) diz que as palavras passam a ter valor referencial a partir do momento que as mesmas se encontram dentro de um universo do discurso.

No âmbito da Recuperação de Informação (RI), os sintagmas nominais podem ser utilizados como termos de indexação e de busca em Sistemas de Recuperação de informação (KURAMOTO 1995; 2002).

Vários autores se debruçaram em desenvolver métodos e instrumentos de extração de sintagmas nominais de forma automática. Já outros se voltaram mais para a questão da seleção dos sintagmas. Kuramoto (1995) pode ser considerado um dos precursores nesses estudos para a língua portuguesa.

Para que ocorra a indexação automática por meio de sintagmas nominais, são necessárias algumas ferramentas ou softwares desenvolvidos para tal atividade (SILVA; CORRÊA, 2015). Contudo as etapas básicas da indexação automática por meio de sintagmas nominais são: extração dos sintagmas nominais e seleção dos sintagmas nominais, chegando assim aos sintagmas que possam funcionar como descritores documentais dos documentos.

Apesar de serem desenvolvidos em ambientes distintos e com diferentes objetivos, alguns trabalhos que contribuem para os estudos da seleção de sintagmas nominais em língua portuguesa são os de: Souza (2005; 2006), Maia (2008), Maia e Souza (2010), Corrêa et al. (2011), Lopes (2012), Souza e Raghavan (2006, 2014), e Martins (2014).

Em relação à seleção de sintagmas nominais, Corrêa et al. (2011) afirmam que alguns dos sintagmas nominais extraídos pelas ferramentas não apresentam relevância para o usuário

no momento de busca, ou seja, embora sejam sintagmas nominais, não constituem descritores e não correspondem a necessidade de informação do usuário como também não são representativos dos assuntos daqueles documentos. Tal fato mostra que a extração de sintagmas nominais deve ser acompanhada de estratégias de seleção ou ordenação por relevância dos sintagmas nominais. Os autores sugerem que seja levado em conta critérios como frequência e posicionamento, semelhante às propostas existentes para as palavras isoladas.

Martins (2014) desenvolveu estudo voltado para o uso dos sintagmas nominais na recuperação de documentos, mais precisamente, voltado para a classificação automática de documentos digitais. Em relação à seleção de sintagmas nominais, percebem-se algumas contribuições da pesquisa de Martins (2014) para este trabalho. No que se refere à frequência de ocorrência de sintagmas nominais, e sabendo que essa métrica é comumente utilizada para a seleção de sintagmas nominas que possam funcionar como descritores,

Apesar de não levar em consideração a frequência de ocorrência dos sintagmas nominais, nem ponderação e nem pontuação dos sintagmas nominais, Martins (2014) contabiliza a frequência de ocorrência dos sintagmas nominais extraídos, com vistas a demonstrar a importância dessa métrica para atividades de indexação, recuperação e, também, classificação de documentos. Nessa contabilização, o referido autor se utiliza dos sintagmas nominais que apareceram pelo menos sete vezes no documento. Esse limiar foi escolhido após a observação de que o número de repetições para o sintagma que mais aparece no documento, em relação ao segundo, tem uma queda abrupta. O mesmo acontece até o sétimo sintagma que mais aparece. A partir desse ponto, existe uma tendência em variar apenas minimamente o aparecimento do próximo sintagma, em relação aos anteriores (que apareceram mais vezes).

Nos trabalhos de Souza (2005; 2006), Maia (2008), Maia e Souza (2010) e Souza e Raghavan (2006, 2014) os seguintes critérios fizeram parte das diferentes metodologias para selecionar sintagmas nominais: frequência de ocorrência; estrutura e nível; descarte de sintagmas nominais não significativos; inverso da frequência no conjunto de documentos.

O trabalho de Lopes (2012) apresenta critérios para extração de termos a partir de sintagmas nominais extraídos automaticamente, sendo tais critérios equivalentes aos seguintes para seleção de sintagmas nominais: descarte dos sintagmas nominais que contêm numerais; descarte dos sintagmas nominais que possuem como núcleo um pronome; descarte dos sintagmas nominais que iniciam com advérbios; detecção de sintagmas nominais implícitos através da remoção sucessiva de adjetivos; e detecção de sintagmas nominais implícitos quando um substantivo é qualificado por mais de um adjetivo ligado por conjunção.

Assim, por meio de uma pesquisa bibliográfica, identificou-se um conjunto de dez critérios exibido no Quadro 1 que utilizados por pesquisadores na seleção de sintagmas nominais. Esses critérios foram aplicados em contextos distintos, com objetivos distintos (recuperação de documentos, classificação de documentos, criação de ontologias, etc.), porém com o intuito comum de contribuir para refinar a seleção de sintagmas nominais.

Quadro 1 – Critérios de seleção de sintagmas nominais.

Critérios
Frequência de ocorrência dos sintagmas nominais no documento
Inverso da frequência dos sintagmas nominais no conjunto de documentos (IDF)
Estruturas e níveis dos sintagmas nominais
Descarte dos sintagmas nominais sem lista de *stopwords* de sintagmas nominais não descritores
Descarte dos sintagmas nominais que contêm numerais
Descarte dos sintagmas nominais que possuem como núcleo um pronome
Descarte dos sintagmas nominais que iniciam com advérbios
Detecção de sintagmas nominais implícitos através da remoção sucessiva de adjetivos
Detecção de sintagmas nominais implícitos quando um substantivo é qualificado por mais de um adjetivo ligado por conjunção.
Posição do sintagma nominal no documento

Fonte: desenvolvido pelo autor.

Por limitações de espaço, a descrição de cada critério se encontra na seção 4, juntamente com a análise dos resultados alcançados com a aplicação dos mesmos no experimento.

3 MÉTODO

No que diz respeito aos objetivos, o presente estudo se caracteriza como uma pesquisa exploratória (GONSALVES 2007). Já em relação aos procedimentos utilizados para coleta dos dados, o mesmo se configura como uma pesquisa bibliográfica (GIL 2007), visto que, se utiliza de materiais já publicados para a obtenção dos dados, bem como pesquisa empírica (MICHEL, 2009). A pesquisa empírica é realizada neste trabalho por meio de um experimento, por meio do qual são obtidos dados empíricos, contribuindo para o conhecimento mais aprofundado acerca da avaliação dos critérios de seleção úteis para a escolha de sintagmas nominais. A pesquisa empírica é aquela que busca respostas e soluções através de observação e prática dos fenômenos que embasam suas conclusões (MICHEL, 2009). Nessa tipologia de pesquisa, apenas o aporte teórico não é suficiente para melhor compreensão e entendimento do fenômeno estudado.

Esse experimento procurou verificar se os critérios contribuem ou não para a seleção de sintagmas nominais, ao passo em que selecionam sintagmas descritores e eliminam sintagmas não descritores. Para a categorização de cada sintagma nominal em descritor ou não, fez-se uso da comparação de um conjunto de sintagmas nominais com palavras-chave provenientes da indexação manual, identificando, assim, os sintagmas nominais descritores, ou seja, os que são semelhantes às palavras-chave ou as contêm. Depois de feita essa categorização dos sintagmas nominais, procedeu-se com a aplicação dos critérios de seleção, identificando o comportamento de cada critério ao selecionar ou eliminar sintagmas nominais descritores. Elaboraram-se cálculos de revocação e precisão para cada critério para que assim pudessem ser levantados julgamentos de cada um como útil ou não.

O experimento constitui-se de nove etapas, idealizadas a partir dos trabalhos de Souza (2005), Lopes (2012) e Silva (2014). As etapas são descritas brevemente a seguir:

1. *Escolha dos documentos – seleção de trinta resumos de teses e dissertações da área de Direito;*
2. *Coleta dos documentos – os documentos são os Títulos e Resumos de cada dissertação e tese que foram transcritos para arquivos texto;*
3. *Indexação manual dos documentos coletados (corpus) - Indexação realizada por quatro bibliotecários/indexadores.*
4. *Definição dos descritores de cada documento - Análise comparativa da indexação feita pelos quatro bibliotecários e o autor para cada documento visando determinar os descritores documentais;*
5. *Submissão dos documentos coletados ao software PALAVRAS para identificação dos sintagmas nominais;*
6. *Extração manual dos sintagmas nominais identificados pelo software PALAVRAS;*
7. *Definição dos sintagmas nominais descritores – a categorização dos sintagmas nominais em sintagmas nominais descritores e em sintagmas nominais não descritores, essa categorização foi feita com base nas palavras-chave atribuídas pela indexação manual, onde os NS que eram semelhantes a elas ou que as continham em suas estruturas eram categorizados como sintagmas nominais descritores, e os que não se encaixavam nessa característica eram marcados como sintagmas nominais não descritores;*
8. *Aplicação dos critérios de seleção encontrados na literatura aos sintagmas nominais extraídos - verificar a viabilidade de cada critério para a seleção de sintagmas nominais que funcionem como Descritores Documentais. Os critérios encontram-se explicitados no quadro 1 na seção 2.*
9. *Análise e comparação dos valores de revocação e precisão obtidos na aplicação de cada critério em cada documento e no corpus como um todo.*

Uma descrição mais detalhada das etapas que constituíram o experimento deste trabalho pode ser consultada em (NASCIMENTO, 2015).

A avaliação de cada critério levou e consideração a capacidade do mesmo em selecionar os sintagmas nominais que estavam categorizados como sintagmas descritores, bem como também a capacidade de não selecionar os sintagmas não descritores.

Os critérios analisados enquadram-se em 5 (cinco) categorias de critérios, são elas: Critérios de descarte/eliminação; Critérios de detecção/adição; Critério de nível do sintagma; Critério de posição dos sintagmas; e, por fim, Critérios de frequência de ocorrência dos sintagmas.

Assim, as métricas utilizadas para os critérios de descarte/eliminação de sintagmas nominais foram: Quantitativo de sintagmas nominais descritores e sintagmas nominais não descritores que atendem ou não ao critério; a taxa de revocação, refletindo o percentual de sintagmas nominais descritores selecionados do total de sintagmas nominais descritores; e a precisão, refletindo o percentual de sintagmas nominais descritores selecionados do total de sintagmas nominais selecionados.

Os critérios de detecção/adição, por diferirem dos de eliminação, exigiram os cálculos de revocação e precisão diferenciados, pois neste caso os sintagmas nominais retornados são os que o critério é aplicado, sendo assim, critérios de adição. A análise feita deste critério também difere um pouco dos critérios de eliminação. Assim, as métricas utilizadas foram: Quantitativos de sintagmas nominais descritores e não descritores que atenderam ao critério, bem como as taxas de revocação e precisão, essas taxas refletindo, respectivamente, o percentual de sintagmas nominais descritores selecionados do total de sintagmas nominais descritores e o percentual de sintagmas nominais descritores selecionados do total de sintagmas nominais selecionados.

No tocante ao critério de nível dos sintagmas nominais, levantaram-se dados quantitativos referentes aos sintagmas nominais descritores e não descritores e seus respectivos níveis. Separaram-se os sintagmas nominais em seis categorias, as quais são: nível 1a, nível 1b, nível 2, nível 3, nível 4 e nível 5 ou mais (+). A revocação foi alcançada dividindo-se o total de sintagmas nominais descritores de cada nível individualmente pelo total de sintagmas nominais descritores de toda a coleção (soma de todos os sintagmas nominais descritores de cada documento). Já a precisão foi alcançada, dividindo-se o total de sintagmas nominais descritores de cada nível pelo total de sintagmas nominais descritores e não descritores que cada nível.

Em relação ao critério de posição dos sintagmas nominais, buscou-se verificar se existia uma relação entre relevância e o quesito posição dos sintagmas, ou seja, se os primeiros sintagmas extraídos, os quais vieram do título e do primeiro trecho do resumo de

cada documento se constituíam de sintagmas nominais descritores. Assim, os sintagmas nominais extraídos de cada documento foram organizados na ordem em que apareciam no texto e procedeu-se com a divisão desses sintagmas nominais em quatro partes do resumo. Essas quatro partes correspondiam, em ordem, aos sintagmas nominais extraídos do titulo e do resumo de cada documento. Calcularam-se as taxas de revocação e precisão de cada uma das quatro partes de cada documento, bem como de toda a coleção. A revocação foi obtida por meio da divisão do total de sintagmas nominais descritores de cada parte pelo total de sintagmas nominais descritores de toda a coleção. Já a precisão foi alcançada por meio da divisão entre o total de sintagmas nominais descritores de cada parte pelo total de sintagmas nominais descritores e não descritores de cada parte.

Por fim, em relação critério de frequência de ocorrência, procedeu-se com a contabilização de quantas vezes cada sintagma nominal ocorria no texto/documento. Analisaram-se duas frequências (absoluta e normalizada) juntas para que fossem calculadas as taxas de precisão e revocação de cada documento individualmente, tendo em vista verificar o comportamento de cada documento individual, bem como levantamento de uma média de toda a coleção.

Inicialmente, aplicou-se um ponto de corte onde se selecionou os sintagmas que possuíam frequência no documento maior que um (1) e frequência normalizada maior que 0,02739. O ponto de corte da frequência normalizada variava, tendo em vista que em alguns documentos o menor valor era 0,055556 ou 0,03335. Em suma, o ponto de corte efetivo foi a frequência absoluta, pois percebeu-se que a frequência normalizada parecia não se mostrar útil no corpus trabalhado. Depois de aplicado esse ponto de corte, calculou-se a precisão, dividindo-se a quantidade de sintagmas nominais descritores que se encontravam acima desse ponto de corte pela quantidade total de sintagmas que se encontravam acima do referido ponto de corte. A revocação foi calculada, dividindo-se a quantidade de sintagmas nominais descritores que se encontravam acima do referido ponto de corte pela quantidade total de sintagmas nominais descritores do documento inteiro. Mais adiante, expõem-se as taxas de revocação e precisão, bem como as médias e o desvio padrão.

No que se refere à frequência de ocorrência na coleção, os cálculos foram feitos tomando como ponto de corte os sintagmas nominais que apareciam em um ou em mais de um documento e que possuíam IDF maior ou igual a 1, ou seja, se um determinado sintagma nominal aparecia uma vez no documento e possuíam IDF acima de um, esse sintagma nominal seria selecionado, todavia se esse sintagma nominal, hipoteticamente, aparecesse em mais de um documento e possuísse IDF abaixo de um (1,0), o mesmo não seria selecionado.

4 ANÁLISE DOS RESULTADOS EXPERIMENTAIS

Nesta seção apresenta-se a descrição dos critérios e os resultados alcançados com a aplicação de cada um ao conjunto de sintagmas nominais extraídos dos trinta documentos que compuseram o *corpus* desta pesquisa.

Para análise e avaliação de cada critério, e especificamente avaliação das taxas de revocação e precisão de cada critério fez-se necessário, antes, que se definisse a taxa de revocação e precisão sem a aplicação de nenhum critério para que pudesse haver comparações e, consequentemente, expor julgamento, quanto a utilidade dos critérios na separação dos sintagmas nominais descritores dos não descritores.

Levando em conta todos os sintagmas nominais extraídos dos documentos da coleção, tem-se que o corpus é constituído de 423 sintagmas nominais descritores e 1358 sintagmas nominais não descritores. O que resulta numa precisão de 23,75% e revocação de 100% quando da não aplicação de nenhum critério, ou seja, na seleção de todos os sintagmas nominais extraídos.

4.1 DESCARTE DOS SINTAGMAS NOMINAIS QUE CONTÊM NUMERAIS

Esse critério elimina os sintagmas nominais que possuem numerais escritos por extenso ou por meio de caracteres numéricos. A utilização desse critério baseou-se na suposição de que os sintagmas com numerais, em sua maioria, não seriam tão potenciais e não serviriam como sintagmas nominais descritores. No Quadro 2 é apresentado o resumo da aplicação desse critério, ressaltando a quantidade de sintagmas nominais descritores e não descritores que atendem ou não o critério.

Quadro 2 – Quantitativo de sintagmas nominais descritores ou não eliminados pelo critério: descarte de sintagmas nominais com numerais.

CRITÉRIO DE ELIMINAÇÃO (NUMERAIS)	SINTAGMAS DESCRITORES	SINTAGMAS NÃO DESCRITORES	PERCENTUAL DE SINTAGMAS NOMINAIS DESCRITORES
SINTAGMAS NOMINAIS SELECIONADOS (não atendem ao critério)	408	1336	**23,3%**
SINTAGMAS NOMINAIS NÃO SELECIONADOS (atendem ao critério)	15	22	**40,5%**

Fonte: desenvolvido pelo autor.

Com base nos dados demonstrados no Quadro 2, verifica-se que esse critério alcançou uma boa taxa de revocação de 96,4%, no entanto, a precisão ficou em 23,3%, abaixo da média quando da não aplicação de nenhum critério que é de 23,75%. Como pode ser visto na quarta coluna do Quadro 17, o percentual de sintagmas nominais descritores que foram eliminados pelo referido critério foi de 40,5%, ou seja, de todos os sintagmas nominais eliminados por este critério, quase metade deles eram descritores, demonstrando que o critério não se mostrou viável.

Em suma, pode-se concluir que o critério não obteve um comportamento tão satisfatório para a seleção de sintagmas nominais, tendo em vista os dados demonstrados anteriormente, bem como também, sintagmas que eram descritores e foram eliminados como, por exemplo, "a audiência pública introduzida no direito brasileiro pelas leis nº 9.868/99 e 9.882/99" e "a constituição brasileira de 1988".

4.2 DESCARTE DE SINTAGMAS NOMINAIS QUE POSSUEM PRONOME COMO NÚCLEO

Esse critério elimina os sintagmas nominais que possuem um pronome como núcleo. Utilizou-se esse critério acreditando que o mesmo seria viável, tendo em vista que os sintagmas com pronomes no núcleo não possuem sentido autocontido, pois se referem a outros termos mencionados anteriormente, ou seja, construções anafóricas. No Quadro 3 são expostos os quantitativos de sintagmas nominais descritores e não descritores que atenderam a esse critério.

Quadro 3 – Quantitativo de sintagmas nominais descritores ou não eliminados pelo critério: descarte de sintagmas nominais com pronomes.

CRITÉRIO DE ELIMINAÇÃO (PRONOMES)	SINTAGMAS NOMINAIS DESCRITORES	SINTAGMAS NOMINAIS NÃO DESCRITORES	PERCENTUAL DE SINTAGMAS NOMINAIS DESCRITORES
SINTAGMAS NOMINAIS SELECIONADOS (não atendem ao critério)	423	1355	**23,79%**
SINTAGMAS NOMINAIS NÃO SELECIONADOS (atendem ao critério)	0	3	**0%**

Fonte: desenvolvido pelo autor.

Por meio do Quadro 3, pode-se ver uma excelente taxa de revocação de 100% e uma taxa de precisão de 23,79%, semelhante à média encontrada quando da não aplicação de nenhum critério que é de 23,75%. Assim, é notório que esse critério se comportou de forma

positiva, ao passo que eliminou somente sintagmas nominais irrelevantes, apesar de serem somente três. Assim, eliminaram-se sintagmas como: “aquela idealizada”, “aqueles previstos” e “eles equiparadas”, o que demonstra a eliminação de sintagmas nominais que realmente não servem como descritores documentais.

4.3 DESCARTE DE SINTAGMAS NOMINAIS QUE INICIAM COM ADVÉRBIO

Esse critério elimina os sintagmas nominais que iniciam com advérbio, acreditando que alguns sintagmas nominais que iniciam com advérbios não se referem explicitamente a um termo, mas apenas fazem referência a termos já mencionados. Mais adiante, no Quadro 4, são expostos os quantitativos de sintagmas nominais descritores e não descritores que atenderam a esse critério.

Quadro 4 – Quantitativo de sintagmas nominais descritores ou não eliminados pelo critério: descarte de sintagmas nominais que iniciam com advérbio.

CRITÉRIO DE ELIMINAÇÃO (ADVÉRBIO)	SINTAGMAS DESCRITORES	SINTAGMAS NÃO DESCRITORES	PERCENTUAL DE SINTAGMAS NOMINAIS DESCRITORES
SINTAGMAS NOMINAIS SELECIONADOS (não atendem ao critério)	420	1358	**23,6%**
SINTAGMAS NOMINAIS NÃO SELECIONADOS (atendem ao critério)	3	0	**100%**

Fonte: desenvolvido pelo autor.

Pode-se perceber que o critério de eliminação foi aplicado em apenas três sintagmas, os quais eram sintagmas nominais descritores, ou seja, de todos os sintagmas nominais eliminados pelo referido critério, 100% era de sintagmas descritores, o que vem a demonstrar que o critério eliminou sintagmas nominais que não deveriam ser eliminados, pois apesar de iniciarem com advérbios, eles eram descritores. A taxa de revocação foi de 99,2% e de precisão foi 23,6%. A taxa de precisão abaixo da média obtida sem a aplicação de critério que foi de 23,75% confirma que o critério em questão se comportou de forma negativa. Conclui-se que, mesmo tendo uma baixa frequência de aplicação, esse critério se mostrou falho ao eliminar sintagmas nominais descritores, como: “apenas três modos de entidades” e “a mais flagrante violação da isonomia tributária”.

4.4 DESCARTE DE SINTAGMAS NOMINAIS CATEGORIZADOS COMO S*TOPWORDS*

A utilização desse critério se baseou no fato que alguns sintagmas nominais, apesar de terem estrutura de sintagmas, não possuem significado específico, pois são usuais, gerais, termos que são utilizados convencionalmente em textos científicos como, por exemplo, “este trabalho”, “esta pesquisa” e assim por diante. Esta lista de stopwords foi composta durante a análise dos sintagmas nominais dos documentos. O quadro 5 mostra o quantitativo de sintagmas nominais eliminados pelo critério, bem como o total de sintagmas nominais stopwords, que são 86.

Quadro 5 – Quantitativo de sintagmas nominais descritores ou não eliminados pelo critério: descarte de sintagmas nominais Stopwords.

CRITÉRIO DE ELIMINAÇÃO (STOP WORDS)	SINTAGMAS NOMINAIS DESCRITORES	SINTAGMAS NOMINAIS NÃO DESCRITORES	PERCENTUAL DE SINTAGMAS NOMINAIS DESCRITORES
SINTAGMAS NOMINAIS SELECIONADOS (não atendem ao critério)	423	1272	**25%**
SINTAGMAS NOMINAIS NÃO SELECIONADOS (atendem ao critério)	0	86	**0%**

Fonte: desenvolvido pelo autor.

Por meio dos dados explicitados no Quadro 5, encontra-se a taxa de revocação de 100% e a de precisão de 25%, pouco superior à taxa alcançada com a não aplicação de nenhum critério (23,75%). Estes valores confirmam por sua vez o que a literatura diz acerca do uso de lista de *stopwords* em sistemas de indexação automática, ou seja, a utilização de listas de *stopwords* só contribui para a seleção de termos de indexação. Exemplos de termos categorizados como *stopwords*: “esta dissertação”, “este trabalho” etc.

4.5 DETECÇÃO DE SINTAGMAS NOMINAIS IMPLÍCITOS POR MEIO DA REMOÇÃO SUCESSIVA DE ADJETIVOS

Esse critério de adição seleciona os sintagmas nominais que possuem sintagmas nominais implícitos em suas estruturas por meio da existência de múltiplos adjetivos em torno de um único nome. Esse critério foi utilizado acreditando que os sintagmas constituídos por mais predicados, no caso, adjetivos, são mais potenciais no que se refere ao sentido. Mais adiante, no Quadro 6, são expostos os quantitativos de sintagmas nominais descritores e não descritores que atenderam a esse critério.

Quadro 6 – Quantitativo de sintagmas nominais descritores ou não selecionados pelo critério: detecção de sintagmas nominais por meio da remoção adjetivos múltiplos.

MÚLTIPLOS ADJETIVOS	SINTAGMAS NOMINAIS DESCRITORES	SINTAGMAS NOMINAIS NÃO DESCRITORES	PERCENTUAL DE SINTAGMAS NOMINAISDESCRITORES
SINTAGMAS NOMINAIS NÃO SELECIONADOS (não atendem ao critério)	301	986	**23,3%**
SINTAGMAS NOMINAIS SELECIONADOS (atendem ao critério)	122	372	**24,7%**

Fonte: desenvolvido pelo autor.

Com base no quadro anterior, pode-se perceber que as taxas de revocação e precisão foram baixas, onde a revocação ficou em torno de 28,8%, ou seja, de todos os sintagmas nominais descritores apenas 28,8% eram sintagmas nominais descritores que possuíam outros sintagmas nominais implícitos em suas estruturas. A precisão, da mesma forma que a revocação, também ficou baixa, em torno de 24,7%, apesar de superior à média da não aplicação deste critério que foi de 23,75%.

Verificou-se que tanto a revocação quanto a precisão foram baixas, demonstrando que o percentual de sintagmas nominais descritores que atendem e não atendem ao critério são próximos, fazendo inferir que esse critério não ajuda a separar sintagmas nominais descritores de sintagmas nominais não descritores. Assim, vê-se esse critério como ineficiente, não só pelo fato de ter alcançado baixas taxas de revocação e precisão, e sim pelo fato de apresentar um comportamento neutro no tocante a separação de sintagmas nominais descritores de sintagmas nominais não descritores.

4.6 DETECÇÃO DE SINTAGMAS NOMINAIS IMPLÍCITOS POR MEIO DE CONJUNÇÃO ENTRE ADJETIVOS

Esse critério seleciona os sintagmas nominais que possuem sintagmas nominais implícitos em suas estruturas por meio de adjetivos ligados por conjunção. A suposição de uso desse critério foi pelo fato de acreditar que os sintagmas nominais constituídos por mais adjetivos são mais potenciais, do mesmo modo que o critério anterior. Mais adiante, no Quadro 7, são expostos os quantitativos de sintagmas nominais descritores e não descritores que atenderam a esse critério.

Quadro 7 – Quantitativo de sintagmas nominais descritores ou não selecionados pelo critério: detecção de sintagmas nominais por meio De múltiplos adjetivos ligados por conjunção.

MÚLTIPLOS ADJETIVOS LIGADOS POR CONJUNÇÃO	SINTAGMAS NOMINAIS DESCRITORES	SINTAGMAS NOMINAIS NÃO DESCRITORES	PERCENTUAL DE SINTAGMAS NOMINAIS

			DESCRITORES
SINTAGMAS NOMINAIS NÃO SELECIONADOS (não atendem ao critério)	419	1340	**23,8%**
SINTAGMAS NOMINAIS SELECIONADOS (atendem ao critério)	4	18	**18,1%**

Fonte: desenvolvido pelo autor.

O Quadro anterior demonstra que o percentual de sintagmas nominais descritores que atenderam ao critério foi muito baixo, bem como também os sintagmas nominais não descritores que atenderam a esse critério. A maior parte dos sintagmas nominais descritores não atende ao critério, onde a taxa de revocação foi muito baixa, ficando em torno de 0,9%. A precisão também se mostrou baixa, ao passo que alcançou apenas 18,1%, ou seja, de todos os sintagmas nominais que atenderam ao critério, apenas quatro (18,1%) eram sintagmas nominais descritores, demonstrando assim que esse critério não se mostrou tão efetivo em conseguir separar sintagmas nominais descritores de sintagmas nominais não descritores.

Assim, verifica-se que esse critério não se comportou de forma positiva, pois selecionou uma quantidade bastante baixa de sintagmas nominais descritores.

4.7 ESTRUTURA E NÍVEL DOS SINTAGMAS NOMINAIS

Os sintagmas nominais, ao passo que vão sendo extraídos de outros sintagmas nominais, vão sendo classificados em níveis (KURAMOTO, 1995). Sendo assim, um sintagma nominal mais simples é definido como de nível 1. Já o sintagma nominal de nível 2 será aquele que contém o sintagma nominal de nível 1, e assim sucessivamente.

Para elaboração dos cálculos levou-se em consideração os sintagmas nominais de nível 1a, 1b, 2, 3, 4 e 5 ou mais, sendo o "sintagma nominal de nível 1a" constituído por um determinante seguido de nome ou substantivo, o "sintagma nominal de nível 1b" constituído por qualquer estrutura envolvendo um substantivo exceto a determinante seguido de nome, o "sintagma nominal de nível 2" constituído por dois nomes, o "sintagma nominal de nível 3" formado por três nomes, o "sintagma nominal de nível 4" constituído por quatro nomes e os "sintagma nominais de nível 5 ou mais" constituídos por sintagmas nominais com 5 nomes ou mais. Esses dados resultantes da aplicação do critério encontram-se no Quadro 8.

Quadro 8 – Taxas de revocação e precisão alcançadas com a análise de cada nível de sintagma nominal.

NÍVEL	SINTAGMAS NOMINAIS	SINTAGMAS NOMINAIS NÃO	REVOCAÇÃO	PRECISÃO

	DESCRITORES	DESCRITORES		
1a	27	400	6,3%	6,3%
1b	82	362	19,3%	18,4%
2	135	378	31,9%	26,3%
3	90	136	21,2%	39,8%
4	41	51	9,6%	44,5%
5+	48	31	11,3%	60,7%

Fonte: desenvolvido pelo autor.

Os dados demonstrados no Quadro 8 conduziram para a escolha de um ponto de corte para a análise do critério nível de modo geral, permitindo o levantamento da taxa de revocação e precisão para toda a coleção, e posteriormente o julgamento da utilidade do critério de nível para a seleção de sintagmas nominais descritores. Tendo sido escolhido o ponto de corte de nível do sintagma nominal maior ou igual a dois, verificou-se uma taxa de revocação de 74,2% e precisão de 41,5%. Apesar de uma revocação mais baixa, foi obtida uma precisão bem acima da obtida sem aplicação de nenhum critério que é de 23,75%. Portanto este critério se mostrou útil, ao passo que conseguiu separar os sintagmas nominais descritores dos sintagmas nominais descritores.

4.8 POSIÇÃO DO SINTAGMA NOMINAL NO DOCUMENTO

Este critério consistiu em selecionar os sintagmas nominais que ocorriam em uma das quatro regiões sequenciais em que foram divididos os documentos com base no percentual de sintagmas nominais extraídos, denominadas respectivamente de primeiro quartil, segundo quartil, terceiro quartil e quarto quartil. Supôs-se que a posição dos sintagmas influencia diretamente no potencial informativo de cada sintagma, ou seja, acredita-se que os sintagmas que estão no início de um resumo de um trabalho acadêmico são mais potenciais informacionalmente do que os que estão no meio do resumo, os quais geralmente se referem aos procedimentos metodológicos da pesquisa. Os sintagmas nominais extraídos de cada documento foram organizados na ordem em que apareciam no texto e procedeu-se com a categorização das ocorrências dos sintagmas nominais nestas quatro regiões, cada uma com aproximadamente 25% dos sintagmas nominais extraídos do documento.

O Quadro 9 demonstra o total de sintagmas nominais descritores e não descritores, bem a precisão e revocação em cada uma das partes dos documentos.

Quadro 9 – Taxas de precisão e revocação para as quatro partes dos documentos.

POSIÇÃO	SINTAGMAS NOMINAS DESCRITORES	SINTAGMAS NOMINAIS NÃO DESCRITORES	PRECISÃO	REVOCAÇÃO
primeiro quartil (0-25% dos sintagmas nominais)	170	270	38,6%	40,1%
segundo quartil (26 a 50% dos sintagmas nominais)	77	366	17,3%	18,2%
terceiro quartil (51 a 75% dos sintagmas nominais)	78	366	17,5%	18,4%
quarto quartil (76 a 100% dos sintagmas nominais)	98	356	21,5%	23,1%

Fonte: desenvolvido pelo autor.

Os dados demonstram melhores revocação e precisão para o primeiro e último quartil, sendo melhor no primeiro quartil. Assumindo o ponto de corte como sendo os sintagmas nominais ocorrendo no primeiro quartil (título e primeiros parágrafos do documento) tem se uma revocação de 40,1% e precisão de 38,6%.

Esse critério mostrou que a posição do sintagma nominal tem relação com o seu potencial discriminatório, ou seja, sintagmas nominais que se encontram em partes mais relevantes de documentos, tendem a ser melhores candidatos a descritores documentais. Além disso, esse critério alcançou um bom desempenho no quesito precisão em selecionar sintagmas nominais descritores.

4.9 FREQUÊNCIA DE OCORRÊNCIA DO SINTAGMA NOMINAL NO DOCUMENTO

Esse critério verifica a frequência absoluta de um determinado sintagma nominal no documento, supondo que os sintagmas que se repetem com frequência em um determinado documento são mais importantes do que outros que aparecem poucas vezes. Essa suposição já foi constatada em outras pesquisas, como também já foi ressaltado na literatura da área. Mais adiante são expostas as frequências de ocorrências dos sintagmas nominais, bem como a precisão e revocação para cada faixa de frequência.

Quadro 10 – Taxas de precisão e revocação para as faixas de frequência de ocorrência no documento.

FREQUÊNCIA DE OCORRÊNCIA DOS SINTAGMAS NOMINAIS	SINTAGMAS NOMINAIS DESCRITORES	SINTAGMAS NOMINAIS NÃO DESCRITORES	PRECISÃO	REVOCAÇÃO
Ocorre somente 1 vez	376	1288	22,5%	88,8%
Ocorre somente 2 vezes	26	59	30,5%	6,1%
Ocorre somente 3 vezes	13	7	65,0%	3,0%
Ocorre somente 4 vezes	5	1	83,3%	1,1%

Ocorre 5 vezes ou mais	3	3	50,0%	0,7%

Fonte: desenvolvido pelo autor.

Levando em consideração mais o valor da precisão, verifica-se que a frequência de ocorrência absoluta acima de 1 demonstra melhores resultados em comparação com a frequência de ocorrência de apenas uma única vez. Visando uma melhor precisão, elaborou-se o cálculo de revocação e precisão utilizando o ponto de corte de frequência absoluta maior que um e obteve-se: 11,1% de revocação e 40,1% de precisão. Essa baixíssima revocação se deu pelo fato de que grande parte dos sintagmas nominais descritores encontrava-se dentro da faixa de frequência de ocorrência de apenas uma vez, entretanto o valor de precisão se mostrou bem melhor que a obtida quando da não aplicação de nenhum critério (23,75%). Apesar de perceber um comportamento "tímido" desse critério neste experimento, acredita-se que o seu desempenho seria bem melhor em um corpus com textos maiores, como, por exemplo, artigos e dissertações completas. Todavia, o critério, se mostrou útil em separar sintagmas nominais descritores e sintagmas nominais não descritores.

4.10 INVERSO DA FREQUÊNCIA DOS SINTAGMAS NOMINAIS NO CONJUNTO DE DOCUMENTOS

A seguir, encontra-se o Quadro 11, no qual são expostos dados quantitativos referentes à quantidade de sintagmas nominais descritores e não descritores que aparecem em apenas um documento da coleção, em dois documentos da coleção, três, quatro e em cinco ou mais documentos. Esse critério pauta-se na ideia de que a ocorrência de um determinado SN em demasia em vários documentos pode demonstrar que esse SN é comum e possui pouco poder discriminatório.

Quadro 11 – Quantitativo de sintagmas nominais descritores ou não que apareceram em 1, 2, 3, 4 ou 5 ou mais documentos da coleção.

FREQUENCIA NOS DOCUMENTOS	SINTAGMAS NOMINAIS DESCRITORES	SINTAGMAS NOMINAIS NÃO DESCRITORES	PRECISÃO	REVOCAÇÃO	VALOR DE IDF
Ocorre somente em 1 documento	394	1218	24,40%	93,1%	1,477121
Ocorre somente em 2 documentos	23	73	23,90%	5,4%	1,176091
Ocorre somente em 3 documentos	5	22	18,50%	1,1%	1
Ocorre somente em 4 documentos	0	8	0,00%	0,0%	0,875061
Ocorre em 5 ou mais documentos	1	37	2,60%	0,2%	[0,43472857 0,778155125]

Fonte: desenvolvido pelo autor.

Com base no Quadro 11, verifica-se que a maior parte dos sintagmas nominais descritores aparece em apenas um documento dos trinta que compõe o corpus. Analisando a quarta coluna do quadro acima, pode-se perceber que a medida que os sintagmas nominais aparecem em mais documentos, o percentual de sintagmas nominais descritores diminui, demonstrando assim o que foi visto na literatura acerca da frequência demasiada de sintagmas nominais, ou seja, sintagmas nominais que aparecem com frequência em muitos documentos, tendem a possuir pouco poder discriminatório, não servindo como descritores documentais. Verifica-se também que os percentuais de sintagmas nominais descritores apresentam melhores taxas para as faixas de frequência menores ou iguais a três. Levando em consideração o IDF, percebe-se que o IDF igual ou acima de um é o que tem um melhor rendimento em termos de selecionar sintagmas nominais descritores.

Levando em consideração os sintagmas nominais que tiveram frequência na coleção maior ou igual a um e IDF maior que um, alcançou-se 98,58% de revocação e 24,4% de precisão. Assim, conclui-se que esse critério, como percebido na literatura, contribui diretamente para a seleção de sintagmas nominais descritores.

5 CONSIDERAÇÕES FINAIS

Fazendo uma síntese e uma reflexão acerca dos critérios de seleção aplicados no experimento desta pesquisa, quanto à utilidade em selecionar sintagmas nominais descritores e evitar sintagmas nominais não descritores, percebeu-se que dos dez critérios analisados, seis se mostraram úteis.

Dos quatro critérios de eliminação, dois se mostraram úteis alcançando bons desempenhos. O critério de eliminação de *stopword* e o de eliminação de sintagmas nominais com pronomes mostraram-se de utilidade, já os de eliminação de descarte de sintagmas nominais com numerais e descarte de sintagmas nominais que iniciavam com advérbio não alcançaram resultados satisfatórios.

No tocante aos critérios de adição (múltiplos adjetivos e adjetivos ligados por conjunção), os dois utilizados neste trabalho não serviram e consequentemente não contribuíram para a seleção de sintagmas nominais, se mostrando assim inviáveis neste contexto.

No que se refere ao nível do sintagma nominal, pôde-se perceber que realmente o nível dos sintagmas nominais tem relação com o seu potencial semântico, como já ressaltado por Souza (2006).

Outro critério que se mostrou viável, ainda que pouco estudado até o momento, foi o de posição do sintagma nominal, o qual evidenciou que a posição do sintagma nominal tem relação com o seu potencial discriminatório. Este critério se mostrou útil nesta pesquisa.

Percebeu-se também que os dois critérios de ordenação se mostraram úteis para a seleção de sintagmas nominais descritores, ressaltando que o critério de inverso da frequência nos documentos (*IDF*) obteve melhor desempenho quando comparado com o de frequência de ocorrência no documento.

Todas as considerações acerca dos critérios mencionados anteriormente são feitas para o contexto onde essa pesquisa foi desenvolvida, ou seja, em um domínio específico (Direito) e com materiais informacionais específicos dentro desse domínio (informações doutrinárias – Dissertações e Teses). Entretanto, este trabalho confirma a utilidade de alguns critérios de seleção de sintagmas nominais já experimentados em outros domínios, sugerindo sua experimentação e análise para outros domínios em que se deseje aplicar a indexação automática por sintagmas nominais.

Acredita-se, com base nas considerações feitas acerca da utilidade de cada critério de forma individual, que o julgamento dos critérios analisados nesta pesquisa contribui diretamente para o desenvolvimento de uma metodologia de avaliação e seleção de sintagmas nominais para este domínio específico, pois se pôde perceber quais critérios foram uteis e quais não contribuíram para a seleção de sintagmas nominais relevantes e a eliminação de sintagmas nominais irrelevantes

Em relação a trabalhos futuros, acredita-se que podem ser desenvolvidos estudos nas seguintes vertentes: experimentação em corpus de texto completo, identificando o comportamento de cada critério em textos maiores e não somente nos títulos e resumos dos documentos; investigar o impacto da definição das palavras-chave descritoras nos resultados obtidos e experimentação de maior padronização na etapa da indexação manual; investigar uma metodologia de seleção envolvendo conjuntamente os critérios considerados úteis; e aplicação dos critérios de seleção estudados nessa pesquisa em corpus de outros domínios científicos, com vistas a verificar a generalidade das conclusões sobre a utilidade dos critérios.

Agradecimentos

À Fundação de Amparo à Ciência e Tecnologia de Pernambuco (FACEPE) pelo fomento ao projeto intitulado "Mapeador Temático de Teses e Dissertações" (processo número APQ-1540-6.07/12), tornando possível a elaboração deste trabalho.

REFERÊNCIAS

BORGES, Graciane Silva Bruzinga; MACULAN, Benildes Coura Moreira dos.; LIMA, Gercina Ângela Borém de. Indexação Automática e Semântica: estudo da análise do conteúdo de teses e dissertações. **Informação & Sociedade**: estudos, João Pessoa-PB, v. 18, n.2, p. 181-193, mai./ago. 2008.

CORRÊA, R. F. et. al. Indexação e recuperação de teses e dissertações por meio de sintagmas nominais. **AtoZ,** Curitiba, v. 1, n. 1, p. 11-22, jan./jun. 2011.

GIL, Antonio Carlos. **Métodos e Técnicas de Pesquisa Social**. 4.ed. São Paulo: Atlas, 2002.

GONSALVES, Elisa Pereira. **Conversas sobre iniciação à pesquisa científica**. 4.ed. revisada e ampliada. Campinas, SP: Alínea, 2007.

KURAMOTO, Hélio. Uma abordagem alternativa para o tratamento e a recuperação de informação textual: os sintagmas nominais. **Ciência da Informação**, v. 25, n. 2, p. 1- 18, 1995.

KURAMOTO, Hélio. Sintagmas Nominais: uma nova proposta para a recuperação de informação. **DataGramaZero –** revista de ciência da informação. Rio de Janeiro, v. 3, n. 1, fev. 2002. Não paginado.

LE GUERN, Michel. **Unanalyseurmorpho-syntaxique pour I'indexation automatique**. Le Français: Moderne, juin, 1991.

LOPES, Lucelene. **Extração automática de conceitos a partir de textos em língua portuguesa**. 2012, 156 f. Tese (Doutorado em Ciência da Computação). Faculdade de Informática, Pontifícia Universidade Católica do Rio Grande do Sul, Porto Alegre, 2012.

MAIA, Luiz Cláudio Gomes. **Uso de sintagmas nominais na classificação automática de documentos eletrônicos**. 2008, 158 f. Tese (Doutorado em Ciência da Informação). – Escola de Ciência da Informação, Universidade Federal de Minas Gerais Minas Gerais, 2008.

MARTINS, Agnaldo Lopes. **O uso do sintagma nominal na recuperação de documentos [manuscrito]**: proposta de um mecanismo automático para classificação temática de textos digitais. 2014, 192 f. Tese (Doutorado em Ciência da Informação) – Escola de Ciência da Informação, Universidade Federal de Minas Gerais Minas Gerais, 2014.

MICHEL, Maria Helena. **Metodologia e Pesquisa Científica em Ciências Sociais**. São Paulo: Atlas, 2009.

NASCIMENTO, Gustavo Diniz do. **Dos Sintagmas Nominais aos Descritores Documentais:** estudo de caso na indexação de teses e dissertações da área de direito. 2015, 198 f. Dissertação (Mestrado) – Mestrado em Ciência da Informação, Universidade Federal de Pernambuco, Recife-PE, 2015.

SOUZA, Renato Rocha. **Uma proposta de metodologia para a escolha automática de descritores utilizando sintagmas nominais.** 2005. 197 f. Tese (Doutorado) – Curso de Doutorado em Ciência da Informação, Escola de Ciência da Informação, Universidade Federal de Minas Gerais – UFMG, Belo Horizonte, 2005.

_____. Uma proposta de metodologia para indexação automática utilizando sintagmas nominais. **Encontros Bibli:** Revista Eletrônica de Biblioteconomia e Ciência da Informação**.** Florianópolis, v. 11, n. esp., p. 42-59, 1º sem. 2006.

SOUZA, R. R.; RAGHAVAN, K. S. A methodology for noum phrase-based automatic indexing. **Knowledge Organization,** v. 33, n. 1, p. 45-56, 2006.

SOUZA, R. R.; RAGHAVAN, K. S. Extraction of keywords from texts: an exploratory study using Noun Phrases. **Informação & Tecnologia (ITEC)**. Marília/ João Pessoa. v. 1, n. 1. p. 5-16, jan./jun., 2014.

SILVA, Tiago José da. **Indexação automática por meio da extração e seleção de sintagmas nominais em textos em língua portuguesa**. 2014, 144 f. Dissertação(Mestrado) – Mestrado em Ciência da Informação, Universidade Federal de Pernambuco, Recife-PE, 2014.

SILVA, T. J. da; CORRÊA, R. F. FERRAMENTAS PARA INDEXAÇÃO AUTOMÁTICA: uma análise comparativa entre o OGMA, Parser PALAVRAS, LX-Parser e a extração manual de sintagmas nominais. **Anais do XVI Encontro Nacional de Pesquisa em Pós-Graduação em Ciência da Informação(ENANCIB)**, João Pessoa-PB, 2015.

VIEIRA, Simone Bastos. Indexação automática e manual: revisão de literatura. **Cin. Inf**. Brasília, v.17, n. 1, p. 43-57, jan./jun. 1988.