Avaliação de critérios para seleção de sintagmas nominais com valor para a recuperação da informação[1]

Evaluation of selection criteria for noun phrases with relevance for information retrieval

Gustavo Diniz do NASCIMENTO[2] 0000-0002-5130-4149

Renato Fernandes CORREA[3] 0000-0002-9880-8678

Resumo

O presente estudo avalia critérios para seleção de sintagmas nominais mais representativos do conteúdo de documentos jurídicos em língua portuguesa. A metodologia da pesquisa consistiu em revisão de literatura brasileira e experimento. No experimento foram aplicados dez critérios de seleção aos sintagmas nominais extraídos de um conjunto de resumos de teses e dissertações. Os critérios foram avaliados quanto à eficácia na seleção de sintagmas nominais relevantes para a recuperação da informação. Por meio do experimento, foram identificados como mais eficazes os critérios de eliminação de sintagmas nominais considerados *stopwords* ou contendo pronomes no núcleo, e os critérios de seleção por posição de ocorrência, nível do sintagma nominal, inverso da frequência nos documentos e frequência de ocorrência em um documento.

Palavras-chave: Indexação automática. Informação jurídica. Representação da informação. Seleção de sintagmas nominais. Sintagmas nominais.

Abstract

This study assesses the criteria for selecting the most representative noun phrases from documents written in Portuguese in the field of law. The research methods were literature review and an experiment. In the experiment, ten selection criteria were applied to noun phrases extracted from a set of abstracts of theses and dissertations. The effectiveness of the criteria was assessed regarding the selection of noun phrases relevant for information retrieval. Through the experiment, the most effective criteria identified were removal of noun *phrases with stopwords value or noun phrases containing pronouns, the selection criteria of noun phrases based on position of occurrence, level of the noun phrase, inverse document frequency, and document occurrence frequency.*

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

[1] Artigo elaborado a partir da dissertação de mestrado de G.D. NASCIMENTO, intitulada "Dos sintagmas nominais aos descritores documentais: estudo de caso na indexação de teses e dissertações da área de direito". Universidade Federal de Pernambuco, 2015.

[2] Universidade Federal de Campina Grande, Biblioteca Central. Campina Grande, PB, Brasil.

[3] Universidade Federal de Pernambuco, Centro de Artes e Comunicação, Departamento de Ciência da Informação. Av. da Arquitetura, s/n., CAC, *Campus* Universitário, 50740-550, Recife, PE, Brasil. Correspondência para/*Correspondence to*: R.F. CORREA. *E-mail*: <renato.correa@ufpe.br>.

Apoio: Fundação de Amparo à Ciência e Tecnologia de Pernambuco (Processo nº APQ-1540-6.07/12).

Recebido em 23 de junho de 2016, versão final reapresentada em 28 de junho de 2017 e aprovado em 28 de julho de 2017.

Como citar este artigo/*How to cite this article*

Nascimento, G. D.; Correa, R. F. Avaliação de critérios para seleção de sintagmas nominais com valor para a recuperação da informação. *Transinformação*, v. 30, n. 2, p. 179-192, 2018. http://dx.doi.org/10.1590/2318-08892018000200004

Introdução

A indexação de documentos é uma das principais atividades desempenhadas com o intuito de organizar a informação produzida, para que esta seja recuperada em outro momento por um indivíduo que dela necessite.

O ambiente virtual, por suas características peculiares, disponibiliza uma maior quantidade de informações, o que acarreta uma necessidade maior de procedimentos automatizados de organização da informação nesse contexto específico. Com o grande volume de informação disponibilizado nos meios virtuais, a indexação automática se mostra como uma alternativa à indexação intelectual, pois permite abarcar toda a informação digital que vem sendo produzida, como também reduz alguns inconvenientes encontrados na indexação intelectual, como por exemplo, a morosidade.

Vieira (1988, p.48) define a indexação automática como "uma operação que identifica, através de programas de computador, palavras ou expressões significativas dos documentos para descrever de forma condensada o seu conteúdo". Portanto, a indexação automática é desempenhada pelo computador, que é programado e preparado por meio de regras para extrair de textos digitais os termos que possam funcionar como pontos de acesso ao documento.

A indexação automática se desenvolveu inicialmente com base na análise da frequência das palavras encontradas nos textos (Borges; Maculan; Lima, 2008). No entanto, a utilização de palavras isoladas como recurso de acesso à informação vem apresentando limitações devido a fenômenos inerentes às línguas naturais, como, por exemplo, a polissemia e a sinonímia.

Nesse contexto, dentre as metodologias de indexação automática que levam em consideração a semântica intrínseca aos documentos, tem-se a baseada na extração de sintagmas nominais relevantes, em vez da utilização de "palavras isoladas" como termos de indexação.

Apontada por Kuramoto (2002) como a menor parte do discurso portadora de informação, o sintagma nominal se constitui de uma palavra sozinha ou de um conjunto de palavras que possui significado e sentido próprio. Trata-se de uma sequência de palavras que determina e modifica uma palavra-núcleo com função de nome ou substantivo. A título de exemplificação, têm-se as expressões "administração pública brasileira", "a Constituição brasileira de 1988" como sintagmas nominais que fazem parte do *corpus* deste trabalho. Nesse sentido, é perceptível que as palavras "pública" e "brasileira" constituem com o nome (núcleo) "administração" uma unidade com sentido específico.

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 do mesmo. Ou seja, a extração automática de sintagmas nominais de um texto não resultará, necessariamente, em sintagmas nominais relevantes. Isso ocorre porque, apesar de os sintagmas nominais possuírem uma grande carga de semântica intrínseca em suas estruturas, nem todos serão representativos do conteúdo informacional desse documento.

Nesse contexto, surge o problema de pesquisa do presente artigo: a necessidade de não apenas extrair os sintagmas nominais, mas selecionar, por meio de critérios específicos, aqueles que sejam representativos dos conteúdos dos documentos e tenham relevância na recuperação da informação de determinado domínio no contexto das bibliotecas digitais e sistemas de recuperação de informação.

Embora a seleção de sintagmas nominais já tenha sido explorada na literatura científica para textos em língua portuguesa, o presente trabalho se diferencia por traçar um panorama mais amplo dos critérios utilizados para realização dessa tarefa, por propor um método para avaliação dos critérios, bem como por avaliar os critérios encontrados em experimento para um domínio específico.

Trans**Informação**, Campinas, 30(2):179-192, maio/ago., 2018 http://dx.doi.org/10.1590/2318-08892018000200004

Portanto, o objetivo geral deste trabalho é investigar os critérios de seleção de sintagmas nominais relevantes para a recuperação da informação para fins de indexação automática, bem como avaliar os critérios presentes na literatura para seleção dos sintagmas nominais mais representativos de documentos jurídicos em língua portuguesa.

Dessa forma, este artigo apresenta os resultados alcançados com a aplicação de dez critérios de seleção de sintagmas nominais em 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 (BDTD-UFPE).

Indexação Automática por sintagmas nominais

A partir da década de 1990, pesquisas voltadas à indexação automática buscaram não só levar em conta a semântica do texto, mas também minimizar alguns problemas identificados na indexação automática baseada em palavras isoladas (Correa; Lapa, 2013).

O pioneiro na ideia de se utilizarem os sintagmas nominais como descritores em lugar das palavras isoladas foi Le Guern (1991). Esse autor é responsável pelo desenvolvimento conceitual acerca daquele recurso como unidade portadora de significado para a indexação e recuperação de informação. O autor faz uma distinção relevante entre descritor e palavra, ao salientar que o "descritor" utilizado para a recuperação de informação deve ser uma unidade do discurso – e não uma unidade da língua, que é um signo isolado e sem significado. Assim, os descritores devem fazer referência à realidade extralinguística do autor, ao passo que a palavra, como unidade da língua, constitui um conjunto de propriedades sem referência à realidade extralinguística.

Complementando esse entendimento, Kuramoto (1995) aponta que as palavras passam a ter valor referencial a partir do momento em que se encontram dentro de um universo do discurso. O referido autor pode ser considerado um dos precursores no estudo da indexação automática por sintagmas nominais para textos escritos em língua portuguesa. Kuramoto (2002) conceitua um sintagma nominal como a menor unidade do discurso portadora de informação, podendo ser tanto uma palavra isolada como um conjunto de palavras que possui semântica e sintaxe, formando uma unidade 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).

Para que ocorra a indexação automática por meio de sintagmas nominais, são necessárias algumas ferramentas ou *softwares* desenvolvidos para tal atividade (Silva; Correa, 2015). As etapas básicas da indexação automática por meio de sintagmas nominais são a extração e a seleção dos mesmos. A extração dos sintagmas nominais corresponde à etapa em que eles são identificados e extraídos do texto. Já a seleção de sintagmas nominais consiste em escolher dentre os sintagmas nominais extraídos aqueles que possam funcionar como expressões representativas do conteúdo dos documentos e tenham relevância na recuperação da informação.

Vários autores se debruçaram em desenvolver métodos de extração e seleção de sintagmas nominais de forma automática para textos em português.

Kuramoto (1995) desenvolveu um protótipo de sistema de recuperação de informação baseado na navegação nas estruturas internas dos sintagmas nominais. Em seguida, foram produzidos outros trabalhos relacionados com a extração e utilização de sintagmas nominais na indexação automática, como, por exemplo, os de Miorelli (2001), Kuramoto (2002), Othero (2004), Maia (2008), Maia e Souza (2010), Morellato (2010), Correa *et al*. (2011), Silva (2014), e Silva e Correa (2015).

Seleção de sintagmas nominais: critérios de seleção

Apesar de desenvolvidos em ambientes distintos e com diferentes objetivos, outros trabalhos contribuíram para os estudos da seleção de sintagmas nominais em textos em língua portuguesa, como os de Souza (2005, 2006), Maia (2008), Maia e Souza (2010), Correa *et al*. (2011), Lopes (2012), Souza e Raghavan (2006, 2014), e Martins (2014).

Em relação à seleção de sintagmas nominais, Correa *et al*. (2011) ressaltam que alguns dos sintagmas nominais extraídos pelos *softwares* não apresentam relevância para o usuário no momento de busca, ou seja, embora sejam sintagmas nominais, não constituem termos de indexação e não correspondem à necessidade de informação do usuário, como também não são representativos do conteúdo daqueles documentos. Tal fato mostra que a extração de sintagmas nominais deve ser acompanhada de estratégias de ordenação conforme sua relevância, ou seja, é necessário proceder a uma seleção ou refinamento na atribuição dos sintagmas nominais aos documentos.

Autores como Correa *et al*. (2011) sugerem que na seleção dos sintagmas nominais sejam levados em conta critérios como frequência e posicionamento, semelhante às propostas existentes para as palavras isoladas.

Assim, por meio de uma pesquisa bibliográfica, identificou-se um conjunto de dez critérios utilizados por pesquisadores na seleção de sintagmas nominais para textos em língua portuguesa. Esses critérios foram aplicados em contextos distintos, visando aplicações distintas (recuperação de documentos, classificação de documentos, criação de ontologias etc.), porém com o mesmo intuito de contribuir para a seleção de sintagmas nominais que funcionem como expressões com valor para a recuperação de informação. Os critérios avaliados na presente pesquisa são:

1) Descarte dos sintagmas nominais que contêm numerais: eliminam-se sintagmas nominais que possuem numerais escritos por meio de algarismos (caracteres numéricos) (Lopes, 2012).

2) Descarte dos sintagmas nominais que possuem como núcleo um pronome: eliminam-se sintagmas nominais que contêm pronomes no início, por se constituírem em expressões anafóricas (que remetem a termos mencionados no texto) (Lopes, 2012).

3) Descarte dos sintagmas nominais iniciados por advérbios: eliminam-se os sintagmas nominais que se iniciam com advérbio, acreditando-se que esses não se referem explicitamente a um termo, mas apenas fazem referência a termos já mencionados (Lopes, 2012).

4) Descarte dos sintagmas nominais em lista de *stopwords* de sintagmas nominais não relevantes: eliminam-se sintagmas comuns em textos científicos que não denotam o assunto tratado pelo documento (Souza, 2005).

5) Detecção de sintagmas nominais com múltiplos adjetivos: selecionam-se os sintagmas nominais que possuem sintagmas nominais implícitos em suas estruturas por meio da existência de múltiplos adjetivos qualificando um substantivo (Lopes, 2012).

6) Detecção de sintagmas nominais com múltiplos adjetivos ligados por conjunção: selecionam-se os sintagmas nominais que possuem sintagmas nominais implícitos em suas estruturas por meio de adjetivos ligados por conjunção que qualificam um substantivo (Lopes, 2012).

7) Estrutura e nível dos sintagmas nominais: são selecionados sintagmas nominais de determinado nível de estrutura sintática. Os sintagmas nominais, ao serem extraídos de dentro de outros sintagmas nominais, vão sendo classificados em níveis (Kuramoto, 1995; Souza, 2005). 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 de nível 1, e assim sucessivamente.

8) Posição do sintagma nominal no documento: selecionam-se os sintagmas nominais que ocorrem em uma das quatro regiões sequenciais em que foram divididos os documentos, com base no percentual de sintagmas nominais extraídos. As regiões denominam-se respectivamente primeiro quartil, segundo quartil, terceiro quartil e

 http://dx.doi.org/10.1590/2318-08892018000200004

quarto quartil, cada uma com aproximadamente 25% dos sintagmas nominais extraídos do documento (Corrêa *et al*., 2011).

9) Frequência de ocorrência dos sintagmas nominais no documento: selecionam-se sintagmas nominais de acordo com a frequência de ocorrência do sintagma em um determinado documento, ou seja, o número de vezes em que ocorre no texto (Souza, 2005).

10) Frequência dos sintagmas nominais no conjunto de documentos: selecionam-se sintagmas nominais de acordo com sua ocorrência nos vários documentos da coleção, ou seja, a frequência nos documentos da coleção (Souza, 2005).

Procedimentos metodológicos

No que diz respeito aos objetivos, o presente estudo se caracteriza como uma pesquisa exploratória. Já em relação aos procedimentos utilizados para coleta dos dados o mesmo se configura como pesquisa bibliográfica, visto que, por meio da revisão da literatura científica brasileira, são utilizados materiais já publicados para levantamento dos critérios de seleção de sintagmas por meio de um experimento pelo qual são obtidos dados empíricos, contribuindo para o conhecimento mais aprofundado acerca da avaliação dos critérios de seleção de sintagmas nominais.

Nesse experimento procurou-se verificar se os critérios são eficazes na seleção desintagmas nominais. Um critério é considerado eficaz na medida em que permite selecionar sintagmas nominais relevantes e eliminar os não relevantes. A categorização dos sintagmas nominais em relevantes ou não relevantes foi feita com base na semelhança – total ou parcial –, com as palavras-chave atribuídas na indexação manual.

A indexação manual, por sua vez, levou em conta a indexação do *corpus* desta pesquisa (no caso, os títulos e os resumos de trinta dissertações e teses indexadas na BDTD-UFPE), realizada por um grupo de quatro bibliotecários pertencentes a bibliotecas especializadas na área jurídica. Essa indexação foi feita de forma livre e sem uso de linguagem documentária. O único parâmetro solicitado foi utilizar de preferência uma quantidade de cinco palavras-chave.

Levando em conta os termos de indexação atribuídos pelos indexadores e pelos autores de cada trabalho, e com base em critérios (como frequência de atribuição do termo, aparecimento do termo no documento e, em caso de empate, a preferência por termos dos autores), foi selecionado um conjunto de cinco termos de indexação ou palavras-chave para cada documento, sendo estes últimos descritores considerados como palavras-chave provenientes da indexação manual.

Para a categorização de cada sintagma nominal extraído como relevante ou não para recuperação da informação em cada documento, fez-se uso da comparação de cada um com as palavras-chave provenientes da indexação manual, identificando, assim, os que eram semelhantes às palavras-chave ou as continham.

Depois de feita essa categorização, procedeu-se à aplicação dos critérios de seleção e avaliação da eficácia de cada critério na seleção de sintagmas nominais. A avaliação dos critérios levou em consideração a capacidade de cada um em selecionar os sintagmas nominais tidos como relevantes e não escolher aqueles considerados como não relevantes para recuperação da informação.

Elaboraram-se cálculos de revocação e precisão dos sintagmas nominais relevantes antes e depois da aplicação de cada critério, para que assim pudessem ser levantados julgamentos de cada um como eficaz ou não na seleção de sintagmas nominais.

Embora tenham sido originalmente propostas na área de recuperação de informação para avaliar a eficácia de sistemas de recuperação de informação quanto à relevância dos documentos retornados para uma busca, a precisão

e revocação são utilizadas neste trabalho como medidas de qualidade na indexação automática. Elas mensuram a eficácia da seleção de sintagmas nominais relevantes pelo sistema de indexação automática quanto aos aspectos de unicidade e completude, respectivamente. Neste trabalho, a utilização das métricas de precisão e revocação na avaliação da qualidade da indexação automática está de acordo com as recomendações metodológicas das competições internacionais de sistemas (Kim *et al*., 2013). A parte empírica da pesquisa constitui-se de nove etapas:

1) Escolha dos documentos: seleção de trinta resumos de teses e dissertações da área de Direito.

2) Coleta dos documentos: os documentos consistem dos 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 os autores de cada documento, visando determinar as palavras-chave ou descritores para cada um deles.

5) Submissão dos documentos coletados ao *software* "Palavras"[4], para identificação dos sintagmas nominais.

6) Extração manual dos sintagmas nominais identificados pelo *software* "Palavras".

7) Definição dos sintagmas nominais relevantes: a categorização dos sintagmas nominais em sintagmas relevantes para recuperação da informação foi feita com base nas palavras-chave atribuídas pela indexação manual. Os sintagmas nominais que eram semelhantes a elas ou que as continham em suas estruturas eram categorizados como sintagmas nominais relevantes, e os que não se encaixavam nessa característica eram marcados como sintagmas nominais não relevantes.

8) Aplicação dos critérios de seleção encontrados na literatura aos sintagmas nominais extraídos: avaliação da eficácia de cada critério para a seleção de sintagmas nominais relevantes.

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.

Visando uma melhor análise e apresentação dos resultados experimentais, os critérios analisados foram enquadrados em cinco categorias: Critérios de descarte/eliminação; Critérios de detecção; Critério de estrutura e nível do sintagma nominal; Critério de posição do sintagma nominal; e Critérios de frequência de ocorrência dos sintagmas nominais.

Para cada critério foram computadas métricas de eficácia na seleção de sintagmas nominais relevantes em cada documento, e, depois, considerados os resultados por documento, sendo computadas as seguintes métricas para toda a coleção:

– Quantitativo de sintagmas nominais relevantes: valor absoluto de sintagmas nominais relevantes entre os sintagmas nominais selecionados e os não selecionados.

– Percentual de sintagmas nominais relevantes: o percentual de sintagmas nominais relevantes entre os sintagmas nominais selecionados ou não selecionados.

– Taxa de revocação: reflete o percentual de sintagmas nominais relevantes selecionados do total de sintagmas nominais relevantes.

– Taxa de precisão: reflete o percentual de sintagmas nominais relevantes selecionados do total de sintagmas nominais selecionados.

[4] O "Palavras", de autoria de Bick (2000), é um *parser* sintático que trabalha com a gramática e o léxico da Língua Portuguesa, fazendo análise de texto. O "Palavras" foi desenvolvido pelo Instituto de Linguagem e Comunicação (ISK) da University of Southern Denmark (SDU).

É importante ressaltar que o modo como foram calculadas as métricas para julgamento da eficácia de cada critério variou de acordo com a natureza de cada um, sendo distinto para os critérios de eliminação, os de detecção, os de frequência de ocorrência dos sintagmas nominais, o de nível do sintagma e o de posição do sintagma no documento.

Quanto ao cálculo das métricas de eficácia, os critérios de eliminação diferem dos demais critérios. Para a maioria dos critérios, os sintagmas nominais selecionados são aqueles em que o critério se aplica. Já para os critérios de eliminação, os sintagmas nominais selecionados são aqueles em que os critérios não se aplicam.

No tocante ao critério do nível dos sintagmas nominais, separou-se cada Sintagma Nominal (SN) em seis categorias (Souza, 2005), as quais foram: nível 1a, nível 1b, nível 2, nível 3, nível 4 e nível 5 ou mais. O "SN de nível 1a" é constituído por um determinante e um substantivo; o "SN de nível 1b" é constituído por qualquer estrutura, exceto a de determinante mais substantivo; o "SN de nível 2" é constituído por dois substantivos; o "SN de nível 3" é formado por três substantivos; o "SN de nível 4" é constituído por quatro substantivos; e os "sintagmas nominais de nível 5 ou mais" são constituídos por sintagmas nominais com 5 substantivos ou mais. As métricas de eficácia foram calculadas para cada nível individualmente.

Em relação ao critério de posição, os sintagmas nominais extraídos de cada documento foram organizados na ordem em que apareciam no texto. Procedeu-se à divisão desses sintagmas nominais em quatro partes do resumo com aproximadamente o mesmo número de sintagmas nominais. Tais regiões, cada uma com aproximadamente 25% dos sintagmas nominais extraídos do documento, foram denominadas: primeiro quartil, segundo quartil, terceiro quartil e quarto quartil. Calcularam-se as métricas para cada uma das quatro partes de cada documento, e depois para toda a coleção.

Quanto ao critério de frequência de ocorrência no documento, procedeu-se à contabilização de quantas vezes cada sintagma nominal ocorria no texto (título e resumo). Analisaram-se as frequências absoluta e normalizada juntas para o cálculo das métricas, para cada documento individualmente, e depois para toda a coleção. Foi utilizada como ponto de corte a frequência absoluta no documento maior que um, pois se percebeu, por meio de experimento prévio, que a frequência normalizada parecia não se mostrar eficaz no *corpus* trabalhado. Depois de aplicado esse ponto de corte, calcularam-se as métricas para sintagmas que se encontravam acima desse ponto.

No que se refere à frequência de ocorrência dos sintagmas no conjunto de documentos, ou seja, na coleção, primeiramente foram computados a frequência nos documentos e o Inverso da Frequência nos Documentos (IDF, *Inverse Document Frequency*), para todos os sintagmas extraídos. Os cálculos das métricas foram feitos tomando como ponto de corte os sintagmas nominais que apareciam em um ou em mais de um documento e que possuíam o IDF maior ou igual a um, ou seja, se um determinado sintagma nominal aparecia uma vez no documento e possuía IDF acima de um, esse sintagma nominal seria selecionado; caso contrário, não.

Resultados

Nesta seção são expostos os resultados alcançados com a aplicação de cada critério, bem como as taxas de revocação e precisão alcançadas com a aplicação dos critérios ao conjunto de sintagmas nominais extraído dos trinta documentos que compuseram o *corpus* desta pesquisa.

Para análise e avaliação de cada critério, e especificamente para a avaliação das taxas de revocação e precisão, fez-se necessário, antes, que elas fossem definidas sem a aplicação de nenhum critério para que pudesse haver comparações e, consequentemente, julgamento quanto à eficácia dos critérios na separação de sintagmas nominais relevantes e não relevantes. As medidas de revocação e precisão neste trabalho são medidas de qualidade na indexação automática, e mensuram a seleção de sintagmas nominais relevantes (contendo as palavras-chave) pelo sistema de indexação automática quanto à unicidade e completude.

As taxas de revocação e de precisão foram inicialmente calculadas sem a aplicação de nenhum critério de seleção de sintagmas nominais, visando estabelecer um parâmetro de comparação com os percentuais alcançados quando da aplicação de cada critério. Assim, as taxas de revocação e precisão foram definidas tomando como base o total de 423 sintagmas nominais relevantes (ou seja, que coincidiam com as palavras-chave atribuídas pela indexação manual) e 1.358 sintagmas nominais não relevantes, resultando, assim, numa precisão de 23,75% e revocação de 100,00% de sintagmas nominais relevantes para recuperação da informação, quando da não aplicação de nenhum critério, ou seja, na seleção de todos os sintagmas nominais extraídos.

Descarte de sintagmas nominais

Esses critérios eliminam os sintagmas nominais geralmente considerados pouco informativos, quais sejam: aqueles que contêm numerais, que possuem um pronome como núcleo ou que se iniciam com advérbio. Para o critério de eliminação de sintagmas nominais *stopwords*, sua lista foi composta durante a análise dos sintagmas nominais dos documentos, constituindo-se de 86 sintagmas nominais vazios de significado.

Na Tabela 1 é apresentado o resumo das aplicações desses critérios, ressaltando a quantidade de sintagmas nominais relevantes e não relevantes selecionados por cada critério. Com base nos dados demonstrados na Tabela 1, em relação ao critério de eliminação de sintagmas nominais com numerais, verifica-se uma boa taxa de revocação (96,40%), mas com baixa precisão (23,30%), abaixo do valor obtido quando da não aplicação de nenhum critério, que é de 23,75%. Como pode ser visto na mesma Tabela, o percentual de sintagmas nominais relevantes eliminados por esse critério foi de 40,50%, ou seja, de todos os sintagmas nominais eliminados por esse critério, quase metade eram relevantes.

Em suma, pode-se concluir que o critério de eliminação de sintagmas nominais contendo numerais não se mostrou eficaz na seleção de sintagmas nominais. Foram eliminados sintagmas nominais relevantes, 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".

Tabela 1. Quantitativo de sintagmas nominais eliminados pelos critérios de descarte.

Critérios de Descarte	Sintagmas Nominais		Percentual de Relevantes
	Relevantes	Não relevantes	%
Descarte (contêm numerais)			
Selecionados (não atendem ao critério)	408	1336	23,30
Não Selecionados (atendem ao critério)	15	22	40,50
Descarte (pronomes como núcleo)			
Selecionados (não atendem ao critério)	423	1355	23,70
Não Selecionados (atendem ao critério)	0	3	0,00
Descarte (iniciam com advérbio)			
Selecionados (não atendem ao critério)	420	1358	23,60
Não Selecionados (atendem ao critério)	3	0	100,00
Descarte (constituem stopwords)			
Selecionados (não atendem ao critério)	423	1272	25,00
Não Selecionados (atendem ao critério)	0	86	0,00

Fonte: Elaborado pelos autores (2015).

http://dx.doi.org/10.1590/2318-08892018000200004

Já em relação ao critério de eliminação de sintagma nominal contendo pronome como núcleo, verifica-se, ainda com base na Tabela 1, uma excelente taxa de revocação (100,00%) e uma taxa de precisão (23,79%), semelhante aos valores encontrados quando da não aplicação de nenhum critério, que são de 100,00% e 23,75%, respectivamente. Assim, é notório que esse critério se mostrou eficaz, pois eliminou somente sintagmas nominais irrelevantes, apesar de serem apenas três. Assim, eliminaram-se sintagmas como "aquela idealizada", "aqueles previstos" e "eles equiparadas", que realmente não são relevantes, confirmando a eficácia desse critério.

Em relação ao critério de eliminação de sintagmas nominais iniciados com advérbio, ele foi aplicado em apenas três sintagmas, os quais eram relevantes, como: "apenas três modos de entidades" e "a mais flagrante violação da isonomia tributária". Ou seja, de todos os sintagmas nominais eliminados por esse critério, 100,00% eram relevantes, apesar de iniciados com advérbios, não devendo ser eliminados. A taxa de revocação foi de 99,20% e a de precisão foi 23,60%, ou seja, abaixo dos valores obtidos sem a aplicação de critérios, que foram de 100,00% e 23,75%, respectivamente. Portanto, o critério em questão não se mostrou eficaz, mesmo tendo uma baixa frequência de aplicação.

No tocante ao critério de eliminação de sintagmas nominais *stopwords*, verifica-se, por meio do Tabela 1, uma taxa de revocação de 100,00% e precisão de 25,00%, pouco superior à taxa alcançada com a não aplicação de nenhum critério (23,75%). Esses valores confirmam o que a literatura diz acerca do uso de lista de *stopwords* em sistemas de indexação automática: a utilização de listas de *stopwords* só contribui para a seleção de termos de indexação, configurando-se um critério eficaz para a seleção de sintagmas nominais. Exemplos de sintagmas nominais categorizados como *stopwords* são: "esta dissertação", "este trabalho" etc.

Detecção de sintagmas nominais com múltiplos adjetivos

Nessa categoria enquadram-se o critério de detecção de sintagmas nominais com mais de um adjetivo qualificando um substantivo e o critério de detecção de sintagmas nominais com múltiplos adjetivos ligados por conjunção qualificando um substantivo.

Na Tabela 2, são expostos os quantitativos de sintagmas nominais relevantes e não relevantes que atenderam a esses dois critérios, seguidos dos comentários acerca do seu desempenho.

Em relação ao critério de detecção de múltiplos adjetivos, pode-se perceber que as taxas de revocação e precisão foram baixas. A taxa de revocação ficou em torno de 28,80%, ou seja, de todos os sintagmas nominais relevantes apenas 28,80% possuíam outros sintagmas nominais implícitos em suas estruturas. A precisão também ficou baixa, em torno de 24,70% – apenas 1,00% superior à taxa quando da não aplicação de nenhum critério,

Tabela 2. Quantitativo de sintagmas nominais selecionados pelos critérios de detecção.

Critérios de Detecção	Sintagmas Nominais		Percentual de Relevantes
	Relevantes	Não relevantes	%
Múltiplos Adjetivos			
Não Selecionados (não atendem ao critério)	301	986	23,30
Selecionados (atendem ao critério)	122	372	24,70
Múltiplos Adjetivos Ligados por Conjunção			
Não Selecionados (não atendem ao critério)	419	1340	23,80
Selecionados (atendem ao critério)	4	18	18,10

Fonte: Elaborado pelos autores (2015).

que foi de 23,75%. Além disso, o percentual de sintagmas nominais relevantes que atendem e não atendem ao critério são próximos, fazendo inferir que esse critério não ajuda a separar sintagmas nominais relevantes dos não relevantes. Assim, vê-se esse critério como não eficaz, não só pelo fato de ter alcançado baixas taxas de revocação e precisão, mas também pelo fato de apresentar um comportamento neutro no tocante à separação de sintagmas nominais relevantes e não relevantes.

Já em relação ao critério de detecção de sintagmas nominais com múltiplos adjetivos ligados por conjunção, percebe-se que o percentual de sintagmas nominais relevantes que atenderam ao critério foi muito baixo (em torno de 0,9%), bem como também o percentual dos sintagmas nominais não relevantes que atenderam a esse critério. A precisão também se mostrou baixa, alcançando apenas 18,1%, ou seja, de todos os sintagmas nominais que atenderam ao critério, apenas quatro (18,1%) eram relevantes. Portanto, esse critério não se mostrou eficaz em conseguir separar sintagmas nominais relevantes dos não relevantes.

Estrutura e nível dos Sintagmas Nominais

Os dados mostrados na Tabela 3 conduziram para a escolha de um ponto de corte para a análise do critério de modo geral, permitindo o levantamento da taxa de revocação e precisão para toda a coleção e, posteriormente, o julgamento da eficácia do critério para a seleção de sintagmas nominais relevantes. Levando em conta que a maior parte dos sintagmas nominais relevantes se encontram categorizados nos níveis 1b, 2 e 3, bem como o fato de que a precisão assume valores mais altos e cresce conforme se selecionam os sintagmas nominais de nível maior ou igual a dois, foi escolhido o ponto de corte de nível do sintagma nominal maior ou igual a dois, sendo encontrada uma taxa de revocação de 74,20% e precisão de 41,50% para esse critério em todo o *corpus*. 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 esse critério se mostrou eficaz, pois conseguiu separar os sintagmas nominais relevantes dos não relevantes, como já ressaltado por Souza (2006).

Posição do sintagma nominal no documento

Os sintagmas nominais extraídos de cada documento foram organizados na ordem em que apareciam no texto, procedendo-se à categorização das ocorrências dos sintagmas nominais em quatro regiões sequenciais de posição no texto, denominadas primeiro quartil, segundo quartil, terceiro quartil e quarto quartil (Tabela 4).

Os dados demonstram melhores valores de revocação e precisão para o primeiro quartil, seguido do quarto. Assumindo como o ponto de corte os sintagmas nominais ocorridos no primeiro quartil (título e primeiros

Tabela 3. Taxas de revocação e precisão alcançadas com a seleção de cada nível de sintagma nominal.

Nível	Sintagmas Nominais		Revocação	Precisão
	Relevantes	Não Relevantes	%	
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: Elaborado pelos autores (2015).

 http://dx.doi.org/10.1590/2318-08892018000200004

parágrafos do documento) tem-se uma revocação de 40,1% e precisão de 38,6%. Assim, verificou-se que os primeiros sintagmas extraídos, retirados do título e do primeiro trecho do resumo de cada documento, constituíam-se bons candidatos a sintagmas nominais relevantes.

Esse critério mostra 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 SN relevantes. Portanto, esse critério alcançou um bom desempenho, mostrando-se eficaz para selecionar sintagmas nominais relevantes.

Frequência de ocorrência do sintagma nominal

Essa categoria de seleção de sintagmas nominais engloba dois critérios: seleção por frequência de ocorrência dos sintagmas nominais no documento; e seleção de sintagmas nominais pela frequência dos sintagmas no conjunto de documentos (Tabela 5).

Em relação ao critério de frequência de ocorrência dos sintagmas no documento, verifica-se que a frequência de ocorrência absoluta acima de 1 demonstra melhores resultados em comparação com a frequência de ocorrência

Tabela 4. Taxas de precisão e revocação para as quatro partes dos documentos.

Posição dos sintagmas nominais	Sintagmas Nominais		Precisão	Revocação
	Relevantes	Não Relevantes	%	
Primeiro quartil (0 a 25%)	170	270	38,6	40,1
Segundo quartil (26 a 50%)	77	366	17,3	18,2
Terceiro quartil (51 a 75%)	78	366	17,5	18,4
Quarto quartil (76 a 100%)	98	356	21,5	23,1

Fonte: Elaborado pelos autores (2015).

Tabela 5. Taxas de precisão e revocação para as faixas de frequência de ocorrência no documento e frequência nos documentos da coleção.

	Sintagmas Nominais		Precisão	Revocação	
	Relevantes	Não Relevantes	%		
Frequência de ocorrência					
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	
Frequência nos documentos					Valor de IDF
Ocorre somente em 1	394	1218	24,4	93,1	1,477
Ocorre somente em 2	23	73	23,9	5,4	1,176
Ocorre somente em 3	5	22	18,5	1,1	1,000
Ocorre somente em 4	0	8	0,0	0,0	0,875
Ocorre em 5 ou mais	1	37	2,6	0,2	[0,435 0,778]

Fonte: Elaborado pelos autores (2015).

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, obtendo-se 11,10% de revocação e 40,10% de precisão. Essa baixa revocação se deu pelo fato de que grande parte dos sintagmas nominais relevantes possuía frequência de ocorrência de apenas uma vez. Entretanto, o valor de precisão se mostrou bem melhor que o obtido quando da não aplicação de nenhum critério (23,75%). Apesar de perceber um comportamento razoável desse critério neste experimento, acredita-se que o seu desempenho seria bem melhor em um *corpus* com textos completos. Todavia, o critério se mostrou eficaz em separar sintagmas nominais relevantes e não relevantes.

Já em relação ao critério de frequência dos sintagmas nominais nos documentos, verifica-se que a maior parte dos sintagmas nominais relevantes aparece em apenas um documento dos trinta que compõem o *corpus*. Analisando a quarta coluna da Tabela 5, pode-se perceber que, à medida que os sintagmas nominais aparecem em mais documentos, o percentual dos relevantes diminui, demonstrando assim o que foi visto na literatura acerca da frequência dos sintagmas nominais nos documentos, ou seja, sintagmas nominais que aparecem em muitos documentos tendem a possuir pouco poder discriminatório e ser irrelevantes para recuperação da informação. Verifica-se também que os percentuais de sintagmas nominais relevantes apresentam melhores taxas para as faixas de frequência menores ou iguais a três. Levando em consideração o Inverso da Frequência nos Documentos (IDF), percebe-se que o IDF acima de um é o que apresenta um melhor rendimento em termos de selecionar sintagmas nominais relevantes.

Levando em consideração a seleção dos sintagmas nominais que tiveram frequência na coleção maior ou igual a um e IDF maior que um, alcançou-se 98,58% de revocação e 24,40% de precisão. Assim, tem-se que esse critério, como percebido na literatura, é eficaz para a seleção de sintagmas nominais relevantes.

Considerações finais

Fazendo uma síntese e uma reflexão acerca dos critérios de seleção aplicados no experimento desta pesquisa, levando em conta a eficácia em selecionar os sintagmas nominais relevantes e evitar os não relevantes, perceberam-se comportamentos distintos dos critérios analisados.

No que se refere aos critérios de descarte de sintagmas nominais, foi possível perceber que o critério de eliminação daqueles com numerais não se mostrou eficaz, uma vez que o uso de números (dígitos numéricos) na área jurídica é recorrente na designação de leis, o que explica sua presença nos sintagmas nominais relevantes da área. Em contraposição, o critério de eliminação de sintagma nominal contendo pronome como núcleo e o critério de eliminação de sintagmas nominais *stopwords* se mostraram eficazes, pois eliminaram somente sintagmas nominais irrelevantes. Em relação ao critério de eliminação de sintagmas nominais iniciados com advérbio, esse não se mostrou eficaz, por eliminar sintagmas relevantes.

No tocante aos critérios de detecção de múltiplos adjetivos e adjetivos ligados por conjunção, pode-se perceber que esses critérios não ajudaram a separar sintagmas nominais relevantes dos não relevantes, não se mostrando eficazes na seleção de sintagmas nominais.

Em relação ao critério de nível do sintagma nominal, esse se mostrou eficaz na seleção de sintagmas nominais, corroborando as pesquisas realizadas por Souza (2006).

O critério de posição do sintagma nominal se mostrou eficaz em selecionar sintagmas nominais. Percebeu-se que a posição do sintagma dentro do documento tem relação direta com o seu potencial discriminatório, ou seja, sintagmas nominais que se encontram em partes mais importantes dos documentos tendem a ser relevantes.

Em relação ao critério de frequência de ocorrência dos sintagmas no documento, esse se mostrou eficaz na seleção de sintagmas nominais, conforme já ressaltado na literatura científica, demonstrando assim que sintagmas nominais que se repetem em um documento tendem a ser mais representativos e relevantes do que outros que nele aparecem com menor frequência.

Trans**Informação**, Campinas, 30(2):179-192, maio/ago., 2018 http://dx.doi.org/10.1590/2318-08892018000200004

O critério de frequência dos sintagmas nominais nos documentos se mostrou eficaz na seleção dos sintagmas nominais considerados relevantes, ressaltando que o critério de IDF obteve melhor desempenho quando comparado com o de frequência de ocorrência no documento.

Portanto, foram identificados como mais eficazes e promissores na seleção de sintagmas nominais os critérios de eliminação de sintagmas nominais considerados *stopwords* ou contendo pronomes no núcleo, e os critérios de seleção por posição de ocorrência, nível do sintagma nominal, inverso da frequência nos documentos e frequência de ocorrência no documento.

Todas as considerações acerca dos critérios mencionados anteriormente são feitas para o contexto em que esta pesquisa foi desenvolvida, ou seja, no domínio específico do Direito e com materiais informacionais específicos dentro desse domínio, como informações doutrinárias e resumos de dissertações e teses.

Nesse contexto restrito de aplicação, este trabalho confirma a eficácia de alguns critérios de seleção de sintagmas nominais, ressaltando também a pertinência de aplicação desses mesmos critérios em *corpus* mais diversificado, tendo em vista a verificação de seu comportamento e sua consolidação como critérios eficazes na seleção de sintagmas nominais relevantes para teses e dissertações do domínio jurídico.

Com base nos dados obtidos com esta pesquisa, verificam-se pertinentes contribuições, especialmente no que diz respeito ao levantamento de critérios de seleção de sintagmas nominais aplicados na literatura científica. Além disso, o experimento realizado conduziu a uma avaliação detalhada dos critérios levantados, os quais foram aplicados em domínio específico, mas cujo método de avaliação pode ser aplicado em documentos de outros domínios e a outros critérios de seleção.

Adicionalmente, os resultados deste trabalho contribuem diretamente para o desenvolvimento de uma metodologia específica para seleção de sintagmas nominais para teses e dissertações do domínio jurídico, uma vez que se perceberam quais critérios foram eficazes para a seleção de sintagmas nominais relevantes. Por outro lado, ressalta-se a pertinência de se analisarem os critérios em *corpus* maiores e mais diversificados, tendo em vista que a pequena amostra aqui analisada – trinta documentos de um mesmo programa de pós-graduação –, pode ter se constituído numa limitação para os resultados alcançados.

Em relação a trabalhos futuros, acredita-se que podem ser desenvolvidos estudos nas seguintes vertentes: aplicação dos critérios de seleção em um *corpus* com mais documentos e mais diversificado quanto às fontes, buscando validar a generalidade do comportamento de cada critério; alterações no método de avaliação, especialmente na definição das palavras-chave como representativas dos documentos, ou seja, buscando realizar a indexação com mais consistência, fazendo uso de uma maior padronização da indexação manual, visto que todas as etapas posteriores do experimento são influenciadas por essa primeira fase; aplicação dos critérios de seleção estudados nessa pesquisa em *corpus* de outros domínios científicos, com vistas a verificar a generalidade das conclusões sobre a eficácia de alguns critérios; e a proposição da combinação dos critérios de seleção de sintagmas nominais num método de ordenamento, de forma a obter os sintagmas nominais de cada texto em ordem decrescente de importância, bem como a avaliação de tais métodos de ordenamento.

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 artigo.

Colaboradores

Todos os autores contribuíram na concepção e desenho do estudo, análise de dados e redação final.

Referências

Bick, E. *The Parsing System "Palavras"*: Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework. 2000. 505 f. Thesis (Doctoral degree in Linguistic) – University Of Arthus, Arthus, 2000.

Borges, G. S. B.; Maculan, B. C. M.; Lima, G. A. B. Indexação automática e semântica: estudo da análise do conteúdo de teses e dissertações. *Informação e Sociedade*: *Estudos*, v. 18, n. 2, p. 181-193, 2008.

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

Correa, R. F.; Lapa, R. C. Panorama de estudos sobre indexação automática no âmbito da Ciência da Informação no Brasil (1973-2012). *Ciência da Informação*, v. 42, n. 2, p. 255-273, 2013.

Kim, S. N. *et al*. SemEval-2010 Task 5: Automatic keyphrase extraction from scientific articles. *Lang Resources and Evaluation*, v. 47, n. 3, p.723-742, 2013. http://dx.doi:10.1007/s10579-012-9210-3

Kuramoto, H. 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, H. Sintagmas nominais: uma nova proposta para a recuperação de informação. *DataGramaZero*: *Revista de Ciência da Informação*, v. 3, n. 1, 2002. Disponível em: <http://brapci.inf.br/index.php/article/download/7479>. Acesso em: 23 mar. 2018.

Le Guern, M. Un analyseur morpho-syntaxique pour l'indexation automatique. *Le Français Moderne*, Tome LIX, n. 1, p. 22-35, 1991.

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

Maia, L. C. G. *Uso de sintagmas nominais na classificação automática de documentos eletrônicos*. 2008. 158 f. Tese (Doutorado em Ciência da Informação) – Universidade Federal de Minas Gerais, Belo Horizonte, 2008.

Maia, L. C. G.; Souza, R. R. Uso de sintagmas nominais na classificação automática de documentos eletrônicos. *Perspectivas em Ciência da Informação*, v. 15, n. 1, p. 154-172, 2010.

Martins, A. L. *O uso do sintagma nominal na recuperação de documentos*: proposta de um mecanismo automático para classificação temática de textos digitais. 2014. 192 f. Tese (Doutorado em Ciência da Informação) – Universidade Federal de Minas Gerais, Belo Horizonte, 2014.

Miorelli, S. T. *Extração do sintagma nominal em sentenças em português*. 2001. 98 f. Dissertação (Mestrado em Ciência da Computação) – Pontifícia Universidade Católica do Rio Grande do Sul, Porto Alegre, 2001.

Morellato, L. V. *Metodologia computacional para identificação de sintagmas nominais na língua portuguesa*. 2010. 112 f. Dissertação (Mestrado em Ciência da Computação) – Universidade Federal do Espírito Santo, Vitória, 2010.

Othero, G. A. *Grammar Play*: um parser sintático em Prolog para a língua portuguesa. 2004. 265 f. Dissertação (Mestrado em Letras) – Pontifícia Universidade Católica do Rio Grande do Sul, Porto Alegre, 2004.

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

Silva, T. J.; Correa, R. F. Ferramentas para indexação automática: uma análise comparativa entre o OGMA, Parser PALAVRAS, LX-Parser e a extração manual de sintagmas nominais. In: Encontro Nacional de Pesquisa em Pós-Graduação em Ciência da Informação, 16., 2015, João Pessoa. *Anais*... João Pessoa: PPGCI/UFPB, 2015. p. 1-20.

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

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, p. 42-59, 2006. Número Especial. http://dx.doi.org/10.5007/1518-2924.2006v11nesp1p42

Souza, R. R.; Raghavan, K. S. A methodology for noun 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 e Tecnologia (ITEC)*, v. 1, n. 1. p. 5-16, 2014.

Vieira, S. B. Indexação automática e manual: revisão de literatura. *Ciência da Informação*, v. 17, n. 1, p. 43-57, 1988.