Artigo Original

Encontros Bibli

SELEÇÃO DE SINTAGMAS NOMINAIS NA INDEXAÇÃO AUTOMÁTICA

Noun phase selection in automatic indexing

Gustavo Diniz do **NASCIMENTO**
Biblioteca Central, Universidade Federal de Campina Grande, Campina Grande, Brasil
dinizufcg@hotmail.com
https://orcid.org/0000-0002-5130-4149

Renato Fernandes **CORREA**
Departamento de Ciência da Informação, Universidade Federal de Pernambuco, Recife, Brasil
renato.correa@ufpe.br
https://orcid.org/0000-0002-9880-8678

Mais informações da obra no final do artigo

RESUMO

Objetivo: Sintetizar e classificar critérios de seleção de sintagmas nominais utilizados em metodologias de indexação automática por sintagmas nominais para textos escritos em língua portuguesa.
Métodos: A metodologia da pesquisa tem natureza exploratória, de cunho bibliográfico, e tem como método procedimental a análise de conteúdo. As metodologias de seleção de sintagmas nominais baseiam-se em critérios como: frequência absoluta de ocorrência, frequência normalizada de ocorrência, frequência inversa nos documentos, não ocorrência em lista de sintagmas nominais pouco significativos, na estrutura gramatical e no nível dos sintagmas nominais.
Resultados: Quanto ao escopo dos critérios, predominam em número os baseados em características do sintagma nominal (estrutura gramatical, nível, conteúdo lexical) e quanto a adoção predominam os baseados no conteúdo do documento e no conteúdo do corpus.
Conclusões: A principal contribuição deste estudo consiste do panorama dos critérios de seleção de sintagmas nominais relevantes para textos em português.

PALAVRAS-CHAVE: Indexação automática. Sintagmas nominais. Seleção de sintagmas nominais. Língua portuguesa. Recuperação da informação.

ABSTRACT

Objective: this study aims to synthetize and classify the noun phrases selection criteria present in methods for automatic indexing by noun phrases of texts written in Portuguese.
Methods: The research methodology has an exploratory nature and bibliographic character, and has the content analysis as procedural method. The bases of the noun phrases selection methodologies are criteria as absolute frequency of occurrence, normalized frequency of occurrence, inverse document frequency, non-occurrence in list of stopwords, and the grammatical structure and level of noun phrases.
Conclusions: As for the criteria scope, predominates in quantity those based on the noun phrases characteristics (grammatical structure, level, lexical content), in adoption predominates those based on the document content and the corpus content.
Results: The main contribution of this work is the panoramic overview of the noun phrases selection criteria for texts written in the Portuguese idiom.

KEYWORDS: Automatic indexing. Noun phrases. Noun phrase selection. Portuguese language. Information retrieval.

1 INTRODUÇÃO

Este estudo tem como propósito investigar métodos de indexação automática por sintagmas nominais quanto aos critérios utilizados para a seleção de sintagmas nominais relevantes como descritores documentais.

No presente trabalho, a indexação pode ser entendida como a forma de caracterizar os conceitos que estão presentes na escrita de certo tipo de documento de modo a permitir a sua recuperação posterior. O propósito da indexação é representar tematicamente os conteúdos dos documentos através de termos organizados em índices. Tais termos podem ser denominados: descritores, palavras-chave ou metadados de assunto.

Araújo e Oliveira (2011, p. 41), em relação à indexação, assinalam que essa "é uma das principais atividades desenvolvidas numa Biblioteca ou Unidade de Informação".

No entanto, a indexação manual, realizada pelo profissional especializado, não consegue abarcar todo o volume de informação que vem sendo produzido em meio digital.

Como alternativa para dar conta do tratamento temático das informações produzidas no ambiente digital, ganha destaque a indexação automática. A indexação automática, além de ser rapidamente realizada, permitindo abarcar a massa documental em ambiente digital, minimiza o problema da subjetividade encontrado na indexação manual, uma vez que sendo desempenhada pelo ser humano está sujeita a influências externas que podem afetar o resultado do processo.

Vieira (1988, p. 48), de forma simples e sucinta, conceitua a indexação automática como sendo "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 baseia-se comumente na identificação de palavras significativas que ocorrem no texto dos documentos. Várias tentativas vêm sendo desenvolvidas com o intuito de se utilizar como pontos de acesso aos documentos outras unidades em Sistemas de Recuperação de Informação (SRIs). Nesse contexto, uma alternativa que vem se mostrando promissora é a utilização dos sintagmas nominais presentes nos textos como pontos de acesso à informação.

A indexação automática por sintagmas nominais tem como ponto norteador a extração de elementos textuais descritores de assuntos contidos nos documentos, elementos que façam referência a objetos e fatos do mundo real, ou seja, os sintagmas nominais.

Michel Le Guern (1991) traz contribuições pertinentes no que diz respeito à utilização dos sintagmas nominais como descritores documentais. Encontra-se em Brito (1992) um estudo que pode ser considerado um dos precursores, juntamente com Le Guern (1991), no que se refere ao uso dos sintagmas nominais em sistemas de indexação automática.

Kuramoto (1995) também possui contribuições nos primeiros estudos voltados para a utilização de sintagmas nominais em sistemas de indexação e recuperação de documentos em língua portuguesa.

Contudo, não é somente a extração dos sintagmas nominais que resultará na obtenção de descritores documentais para fins de indexação e recuperação de documentos. É preciso atentar-se para a seleção dos sintagmas nominais mais relevantes.

Nesse contexto, tem-se como objetivo geral sintetizar e classificar os critérios de seleção de sintagmas nominais utilizados em pesquisas sobre indexação automática por sintagmas nominais de documentos em língua portuguesa. Como objetivos específicos, têm-se:

- identificar pesquisas sobre indexação automática por sintagmas nominais para textos em português;
- levantar nas pesquisas identificadas critérios para a seleção de sintagmas nominais para fins de indexação automática;
- sintetizar e classificar os critérios de seleção de sintagmas nominais utilizados para selecionar os sintagmas nominais com valor de descritor dos documentos.

As seções que se seguem estão assim estruturadas: na seção 2, contextualiza-se a indexação automática, define-se o processo de indexação automática por sintagmas nominais, discutem-se os trabalhos que tratam do tema para textos em português e define-se a seleção de sintagmas nominais como uma etapa deste processo; na seção 3 é apresentada a metodologia da pesquisa; a seção 4 apresenta e discute os resultados da síntese e categorização dos critérios de seleção de sintagmas nominais; e a seção 5 contempla as considerações finais.

2 INDEXAÇÃO AUTOMÁTICA

A indexação automática muitas vezes é rotulada como indexação assistida por computador e indexação semiautomática. No entanto, essa variedade de expressões designa conceitos distintos, uma vez que a indexação assistida por computador refere-se a programas que auxiliam na representação temática de documentos realizada pelo homem, já a indexação semiautomática refere-se àquela que ocorre em sistemas que indexam automaticamente os documentos e posteriormente os termos são validados pelo profissional, e a indexação automática é realizada completamente pelo computador através de software desenvolvido para desempenhar tal atividade (CORRÊA; LAPA, 2013).

O surgimento da indexação automática está relacionado com o uso de programas computacionais para geração de índices pré-coordenados. Nesse contexto, verificam-se o desenvolvimento do *Selective Listing Combination* (SLIC), o *PREserved Context Indexing System* (PRECIS), entre outros. É a partir do sistema *Key Word in Context* (KWIC) que a indexação se baseia nas palavras significativas dos títulos dos documentos (BORGES; MACULAN; LIMA, 2008).

Vieira (1988) conclui que a indexação automática produz resultados na recuperação da informação no mínimo equivalentes aos obtidos pela indexação manual, quando são utilizadas as palavras isoladas como descritores documentais.

Apesar de ter sido a palavra isolada a primeira unidade base para a indexação automática, essa foi se mostrando aos poucos ineficiente para fins de representação e recuperação de informação, devido aos fenômenos linguísticos como a sinonímia e a polissemia que se encontram nas línguas naturais. Nesse contexto, as pesquisas em indexação automática buscam cada vez mais desenvolver metodologias que se preocupam com a sintaxe e com a semântica imbuídas nos textos, resultando no surgimento da indexação automática por sintagmas nominais.

2.1 Indexação Automática por Sintagmas Nominais

Pode-se considerar Michel Le Guern (1991) como responsável pelo desenvolvimento conceitual acerca dos sintagmas nominais como unidade portadora de significado. Esse autor faz uma distinção relevante entre descritor e palavra. Para ele, o descritor utilizado para a recuperação da informação deveria ser uma unidade do discurso

como os sintagmas nominais e não uma unidade da língua como as palavras (signo isolado sem significado definido). Complementando esse entendimento, Kuramoto (1995) afirma que as palavras passam a ter valor referencial a partir do momento que as mesmas se encontram dentro de um universo do discurso.

Outro estudo que se deteve ao uso dos sintagmas nominais como instrumentos de indexação e que pode ser considerado um dos primeiros estudos nessa vertente para a língua portuguesa foi o de Brito (1992). Tal autor apresentou uma visão diferente sobre a análise e descrição linguística, fundada sobre uma descrição mais rica dos fenômenos linguísticos e que está na origem das reflexões sobre o tratamento automático da informação com base nos sintagmas nominais.

Segundo Perini (2010).

> [...] o sintagma nominal tem **potencial referencial** ao contrário de outros sintagmas da língua. O potencial referencial do sintagma nominal é sua propriedade semântica básica, e condiciona o modo como ele é construído internamente. Não é possível fazer referência a uma entidade do mundo usando a língua a não ser com um sintagma nominal. (PERINI, 2010, p. 252, grifo nosso).

Kuramoto (2002, p. 6), no tocante a conceituação do sintagma nominal, diz que "um sintagma nominal é a menor unidade do discurso portadora de informação". Um sintagma nominal pode ser tanto uma palavra isolada como também um conjunto de palavras que possuem sintaxe e semântica.

Seguem alguns exemplos de sintagmas nominais e suas várias possibilidades de formação: "A Ciência da Informação", onde se tem um determinante (A), um nome (Ciência) e um modificador (da Informação); "Um certo Bibliotecário", onde se tem um determinante (Um), seguido de outro determinante (certo), seguido de um nome (Bibliotecário).

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.

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 de sintagmas nominais.

Kuramoto (1995; 2002) pode ser considerado um dos precursores nos estudos sobre extração automática de sintagmas nominais para textos em língua portuguesa.

Para que ocorra a indexação automática por sintagmas nominais, são necessárias ferramentas ou softwares que permitam a realização desta atividade. A seguir verificam-se, por meio do Quadro 1, as ferramentas necessárias à execução da indexação automática por sintagmas nominais: Etiquetadores; Identificadores de sintagmas nominais; Extratores de sintagmas nominais; e Selecionadores de sintagmas nominais.

Nesse contexto, a indexação automática inicia-se com a retirada das palavras que compõem o texto e em seguida a etiquetagem dos vocábulos extraídos com classes gramaticais, ou seja, essa primeira etapa consiste na categorização das palavras em substantivo, adjetivo, advérbio etc.

Feito isso, inicia-se a execução da segunda ferramenta que é a identificação dos SNs por meio das regras de formação de sintagmas nominais, baseando-se nas classes gramaticais das palavras atribuídas na etapa anterior. Essa segunda etapa confronta as palavras categorizadas com o conjunto de regras de formação de SNs. Os identificadores de SNs apresentam os SNS de forma destacada no próprio texto ou em uma representação arbórea da oração.

Após a identificação dos SNs, inicia-se a extração dos sintagmas identificados anteriormente. Os extratores de sintagmas nominais apresentam os SNS extraídos do texto em uma lista ou em arquivo separado.

Quadro 1 – Ferramentas essenciais para a indexação automática por sintagmas nominais

Ferramentas	**Funções desempenhadas**
Etiquetadores 1 ª Ferramenta: Etiquetagem das palavras do texto.	Os etiquetadores (*Taggers*) têm como função identificar e rotular as palavras que compõem um texto em determinadas classes gramaticais. Essa ferramenta categoriza as palavras e as rotulam com etiquetas correspondentes às classes gramaticais.
Identificadores de sintagmas nominais 2 ª Ferramenta: Identificação dos sintagmas nominais.	Os identificadores de sintagmas nominais têm como propósito analisar um determinado texto observando as sequências de léxicos e respectivas etiquetas gramaticais a fim de aplicar as regras gramaticais de formação dos sintagmas nominais (gramática sintagmática). Com base nesse cotejamento de "léxico etiquetado+ regras", são identificadas as sequências que se constituem sintagmas nominais.
Extratores de sintagmas nominais 3 ª Ferramenta: Extração de sintagmas nominais.	Os extratores realizam, além da atividade de identificação dos sintagmas nominais, a extração dos sintagmas nominais, mostrando os sintagmas nominais fora do texto, ou seja, tendo como saída uma lista de sintagmas nominais.
Selecionadores de sintagmas nominais 4 ª Ferramenta: Seleção dos sintagmas nominais .	A seleção se faz necessária, pois, muitas vezes, os sintagmas nominais extraídos pela máquina não são descritores do conteúdo de determinado documento. Essa ferramenta escolhe, com base em determinados critérios, sintagmas nominais com valor de descritor dentre os sintagmas nominais extraídos.

Fonte**:** desenvolvido pelos autores.

Por fim, a quarta ferramenta, os selecionadores de sintagmas nominais, seleciona os SNs extraídos com base em determinados critérios. Essa seleção se faz necessária

para ordenação dos SNs por relevância, tendo em vista que certos sintagmas nominais não são sintagmas representativos dos assuntos do documento. Assim, a seleção irá escolher os sintagmas que possam ser descritores do conteúdo do documento.

Com base no que foi visto até o momento, pode-se estruturar o processo de indexação por meio de sintagmas nominais em três etapas gerais, as quais são expostas no Quadro 2. Na primeira etapa é feita a identificação dos SNs, por meio da etiquetagem das palavras isoladas e em seguida do confronto dessas palavras com o conjunto de regras de formação de SNs. Após essa primeira etapa é efetuada a extração dos SNs, essa segunda etapa consiste em mostrar os SNs de forma destacada no texto ou como um arquivo separado. Finalizando, é efetuada a seleção dos SNs extraídos, identificando os que são mais descritivos do conteúdo do documento do qual foram extraídos.

Quadro 2 – Etapas do processo de indexação automática por sintagmas nominais.

Processo de indexação automática por sintagmas nominais	
1ª Etapa	Identificação dos sintagmas nominais através das subetapas de “etiquetagem” e de “cotejamento dos léxicos etiquetados com as regras de formação dos sintagmas nominais”
2ª Etapa	Extração dos sintagmas nominais do texto, mostrando-os em listas ordenadas ou ordenáveis.
3ª Etapa	Seleção dos sintagmas nominais, com base em critérios que os classifique como descritores documentais

Fonte: desenvolvido pelos autores

Fazendo analogia com a indexação humana, do mesmo modo que o indexador identifica e seleciona os termos mais representativos de um determinado documento, a máquina também deve ser capaz de identificar, extrair e selecionar os sintagmas nominais mais apropriados para a descrição do conteúdo de um documento.

É importante salientar que existem ferramentas que desempenham uma ou todas as três etapas da indexação automática por sintagmas nominais. Como, por exemplo, o software OGMA de Maia (2008), que identifica, extrai e seleciona os sintagmas nominais. Outra ferramenta frequentemente utilizada para identificar os sintagmas nominais em textos em português é o parser PALAVRAS (SILVA; CORRÊA, 2015).

2.2 Seleção de Sintagmas Nominais

No contexto da indexação automática por sintagmas nominais, a seleção de sintagmas com valor de descritor documental é uma tarefa importante a ser automatizada, e se constitui o foco deste estudo.

Corrêa et al. (2011) afirmam que alguns dos sintagmas nominais extraídos pelos sistemas não apresentam relevância para o usuário no momento de busca, ou seja, embora sejam sintagmas nominais, não constituem descritores. Tais sintagmas nominais irrelevantes não correspondem à necessidade de informação do usuário bem como não são representativos do assunto daqueles documentos. Tal fato mostra que a extração de sintagmas nominais deve ser acompanhada de estratégias de ordenação por relevância dos sintagmas nominais. Os autores sugerem que seja levado em conta critérios como frequência e posicionamento, semelhante aos métodos de indexação automática para as palavras isoladas.

Complementando o que os autores acima mencionaram, Lopes (2012, p. 33) afirma que: “Um aspecto importante para a recuperação de informações textuais é o passo posterior à extração de termos, que consiste em escolher dentre os termos extraídos aqueles que são portadores de valor conceitual, e não apenas terminológico”. Em (LOPES, 2012) foi realizada a extração de conceitos relevantes a partir de sintagmas nominais extraídos de corpus de documentos de determinados domínios, com vistas à criação de recursos linguísticos e ontologias. Os critérios apontados pela autora na seleção dos sintagmas nominais contendo conceitos podem ser adaptados para a seleção de sintagmas nominais mais relevantes para cada documento.

Souza (2005) pode ser considerado pioneiro nos estudos de critérios de seleção de sintagmas nominais em textos em português. Na sua tese é proposta uma metodologia consolidada para seleção de sintagmas nominais relevantes em artigos científicos escritos em português da área de Ciência da Informação. A metodologia proposta é baseada nos critérios de frequência no documento, inverso da frequência no corpus, e numa classificação baseada no nível e estrutura gramatical do sintagma nominal.

Os trabalhos de Souza (2006), Souza, Alvarenga Neto e Mendes (2007), Maia (2008), Maia e Souza (2010), Souza e Raghavan (2006, 2014), e Martins (2014) se constituem em desdobramentos da pesquisa de Souza (2005) quanto à seleção de sintagmas nominais.

São apresentadas na seção 4 a síntese e classificação dos critérios de seleção de sintagmas nominais adotados nos trabalhos que investigaram a indexação automática por sintagmas nominais para textos em português. A seguir, pontua-se sobre os métodos e ferramentas que guiaram a presente pesquisa.

3 MÉTODO

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 a mesma se configura como uma pesquisa bibliográfica, visto que, se utiliza de materiais já publicados para a obtenção dos dados.

O método procedimental de síntese e classificação dos critérios de seleção dos sintagmas nominais tem como base a análise de conteúdo (BARDIN, 2011) dos trabalhos científicos coletados.

A análise de conteúdo constitui uma metodologia de pesquisa usada para descrever e interpretar o conteúdo de documentos e textos. Esta forma de análise, conduzindo a descrições sistemáticas, quantitativas ou qualitativas, contribui para a interpretação das mensagens, atingindo uma compreensão de seus significados num nível que vai além de uma leitura superficial. Para Bardin (2011) a análise de conteúdos e constitui de várias técnicas onde se busca descrever o conteúdo emitido no processo de comunicação, seja ele por meio de falas ou de textos. Desta forma, a técnica é composta por procedimentos sistemáticos que proporcionam o levantamento de indicadores (quantitativos ou não) permitindo a realização de inferência de conhecimentos.

A análise de conteúdo foi utilizada na categorização e filtragem dos trabalhos que realizavam a indexação automática por sintagmas nominais, depois para filtrar os que aplicavam a seleção de sintagmas nominais, e por último para identificar, categorizar e sintetizar os critérios de seleção aplicados por cada trabalho.

Foram coletados e analisados: artigos científicos, dissertações e teses. Os artigos científicos foram recuperados nas seguintes bases de dados: Base de dados referencial de artigos de periódicos em Ciência da Informação - BRAPCI[1], Scientific Eletronic Library Online - SciELO[2] e Google Acadêmico[3]. As dissertações e teses foram recuperadas por

[1]Universidade Federal do Paraná – UFPR. Base de dados referenciais de artigos de periódicos em ciência da informação – BRAPCI. Disponível em: http://www.brapci.ufpr.br. Acesso em 07 jan. 2019.
[2]SciELO. Disponível em: http://www.scielo.org . Acesso em: 07 jan. 2019.

meio de Biblioteca Digital de Teses e Dissertações do Instituto Brasileiro de Informação em Ciência e Tecnologia – BDTD/IBICT[4]. A expressão de busca submetida nas bases foi: “sintagmas nominais” AND “indexação”. Tal expressão de busca foi realizada em todos os campos sem delimitação de período temporal de publicação.

4 ANÁLISE DOS RESULTADOS

No Quadro 3 são sistematizadas as pesquisas que contribuíram com os estudos de indexação automática por sintagmas nominais, fazendo uso de critérios de seleção de sintagmas nominais.

Quanto ao foco, algumas pesquisas voltam-se para diferentes tarefas, como a classificação automática de documentos, a indexação automática, a extração de conceitos, e a criação de ontologias. Todavia essas tarefas estão relacionadas com a indexação automática por meio de sintagmas nominais.

Quanto aos critérios utilizados na seleção de sintagmas nominais, percebe-se a adoção mais frequente de determinados critérios em detrimento de outros. A maioria dos trabalhos é de coautoria de Souza e decorrem da pesquisa inicial da tese de doutorado do mesmo autor. Nas pesquisas realizadas por Souza e coautores, são adotados os seguintes critérios em ordem de frequência: os critérios de frequência de ocorrência no documento, classificação do sintagma nominal quanto ao nível e estrutura, inverso da frequência de ocorrência no corpus, e eliminação de sintagmas nominais presentes em lista de expressões pouco significativas.

Em contraste, no trabalho de Lopes (2012), com a finalidade de extração de conceitos de corpus de documentos, são utilizados critérios de seleção relacionados à estrutura gramatical dos sintagmas nominais. São adotados também critérios de frequência de ocorrência e inverso da frequência de ocorrência em corpus, esses dois critérios constituem o índice tf-dcf¸ criado pela própria autora.

[3]Google Acadêmico. Disponível em: http://scholar.google.com.br . Acesso em: 07 jan. 2019.
[4]Biblioteca Digital de Teses e Dissertações / Instituto Brasileiro de Informação em Ciência e Tecnologia. Disponível em: http://bdtd.ibict.br . Acesso em: 07 jan. 2019.

Quadro 3 – Síntese dos trabalhos que utilizaram critérios para a seleção de sintagmas nominais.

Autores	**Foco da pesquisa**	**Critérios utilizados na seleção de sintagmas nominais**
Souza (2005, 2006), Souza e Raghavan (2006, 2014)	Propõe uma metodologia para indexação de documentos digitalizados de textos completos por meio da identificação, extração e seleção de sintagmas nominais. Apresenta método para seleção de sintagmas nominais a partir de textos.	• Frequência de ocorrência dos sintagmas nominais no texto do documento (absoluta ou normalizada). • Inverso da frequência de ocorrência dos sintagmas nominais no conjunto de documentos (*corpus).* • Classificação CNP quanto ao nível e estrutura do sintagma nominal. • Eliminar sintagmas nominais em *stoplist* de sintagmas nominais não descritores.
Souza, Alvarenga Neto e Mendes (2007)	Proposta metodológica de construção de um mapa terminológico-conceitual. O objeto de estudo desse artigo foi à área "Gestão do Conhecimento".	• Frequência de ocorrência dos sintagmas nominais no texto do documento. • Inverso da frequência de ocorrência dos sintagmas nominais no conjunto de documentos (*corpus).* • Estruturas e níveis dos sintagmas nominais.
Maia (2008), Maia e Souza (2010)	Uso dos sintagmas nominais pontuados como características na classificação por similaridade e criação de aglomerados de documentos eletrônicos.	• Frequência de ocorrência dos sintagmas nominais no texto do documento • Classificação CNP quanto ao nível e estrutura do sintagma nominal. • Eliminar sintagmas nominais em *stoplist* de sintagmas nominais não descritores Obs.: o Software OGMA proposto por Maia (2008) realiza a identificação da classe gramatical do Sintagma, bem como o cálculo de pontuação do mesmo como descritor baseando-se no método proposto por Souza (2005).
Lopes (2012)	Extração de conceitos relevantes a partir de sintagmas nominais, para um determinado domínio, com vistas à criação de recursos linguísticos e ontologias	• Frequência de ocorrência de sintagmas nominais no texto do documento. • Inverso da frequência de ocorrência de sintagmas nominais em outros documentos de domínios diferentes (domínio contrastante). • Heurísticas de descarte: descarte dos sintagmas nominais que contêm numerais; descarte dos sintagmas nominais que contenham símbolos; descarte dos sintagmas nominais que possuem como núcleo um pronome; e descarte dos sintagmas nominais que iniciam com advérbios. • Heurísticas de inclusão: detecção de sintagmas nominais contidos em sintagmas nominais maiores através da remoção sucessiva de adjetivos, detecção de estruturas gramaticais múltiplas com o uso de conjunções (substantivo qualificado por mais de um adjetivo).
Martins (2014)	Apresenta um estudo sobre o uso dos sintagmas nominais como características na classificação automática de documentos quanto ao assunto tratado em cada documento, apoiando-se no Processamento de Linguagem Natural.	• Frequência de ocorrência dos sintagmas nominais no documento, levando em consideração somente os sintagmas nominais que ocorrem pelo menos sete vezes no documento (limiar mínimo de frequência no documento).

Fonte: desenvolvido pelos autores

No Quadro 4 foram organizados sistematicamente os critérios utilizados pelas pesquisas descritas no Quadro 3, classificando-os quanto ao escopo de aplicação de cada critério e apontando os trabalhos que fizeram uso desses critérios.

Quadro 4 – Critérios de seleção quanto ao escopo de aplicação dos critérios e trabalhos que aplicaram.

Critério	**Escopo de aplicação do Critério**	**Trabalhos que fizeram uso do critério de seleção**
Frequência de ocorrência dos sintagmas nominais em um documento	DOCUMENTO	Souza (2005, 2006), Souza, Alvarenga Neto e Mendes (2007), Maia (2008), Maia e Souza (2010), Lopes (2012), Souza e Raghavan (2014) e Martins (2014).
Frequência de ocorrência normalizada do sintagma nominal em um documento	DOCUMENTO	Souza e Raghavan (2014)
Inverso da frequência de ocorrência dos sintagmas nominais em um conjunto de documentos ou *corpus* (IDF)	CORPUS	Souza (2005, 2006), Souza, Alvarenga Neto e Mendes (2007), e Souza e Raghavan (2014).
Frequência de ocorrência de sintagmas nominais em conjunto de documentos de domínio diferente (*corpus* contrastante)	CORPUS CONTRASTANTE	Lopes (2012)
Estruturas e níveis dos sintagmas nominais	SN	Souza (2005, 2006), Souza, Alvarenga Neto e Mendes (2007), Maia (2008), Maia e Souza (2010) e Souza e Raghavan (2006, 2014).
Ocorrência do sintagma nominal em tesauro	SN	Souza (2005, 2006)
Eliminar sintagmas nominais contidos em *stoplist* de sintagmas nominais não descritores	SN	Souza (2005, 2006) Maia (2008), Maia e Souza (2010)
Descarte dos sintagmas nominais que contêm numerais	SN	Lopes (2012)
Descarte dos sintagmas nominais que possuem como núcleo um pronome	SN	Lopes (2012)
Descarte dos sintagmas nominais que iniciam com advérbios	SN	Lopes (2012)
Detecção de sintagmas nominais contidos em sintagmas nominais maiores através da remoção sucessiva de adjetivos	SN	Lopes (2012)
Detecção de sintagmas nominais em estruturas gramaticais múltiplas com o uso de conjunções, quando um substantivo é qualificado por mais de um adjetivo	SN	Lopes (2012)

Fonte**:** desenvolvido pelos autores.

Quanto ao escopo de aplicação de cada critério de seleção de sintagmas nominais com valor de descritor, foram utilizadas as categorias: SN – quando o critério se aplica sobre o sintagma nominal de acordo com o nível, conteúdo léxico ou estrutura gramatical do mesmo; DOCUMENTO – quando o critério se aplica sobre a lista dos sintagmas nominais de um documento; CORPUS – quando o critério se aplica sobre a lista dos

sintagmas nominais de um conjunto de documentos ou corpus; CORPUS CONTRASTANTE – quando o critério se aplica sobre listas de sintagmas nominais, cada lista advinda de corpus de domínio temático distinto.

Verificou-se que a frequência de ocorrência é um critério comumente utilizado para a classificação do sintagma nominal como descritor ou não, sendo considerado fundamental para a seleção de sintagmas nominais. Essa frequência de ocorrência demonstra em alguns casos o potencial informativo de um determinado sintagma nominal, bem como também demonstra o caráter genérico do mesmo ao ocorrer em vários documentos, daí a necessidade do uso também da frequência inversa de documentos como critério de seleção.

Ainda sobre a frequência de ocorrência, vale salientar a pertinência da frequência normalizada e não somente da frequência absoluta. Souza e Raghavan (2014) propõem o uso da frequência normalizada, pois a frequência absoluta de um sintagma nominal pode ser bem maior em um documento que seja mais extenso em comparação com outro menor. A medida normalizada é a divisão da frequência absoluta pelo número total de sintagmas nominais ocorridos no documento, sendo uma medida de importância relativa que independe do tamanho do documento.

Martins (2014), em sua tese, contabiliza a frequência de ocorrência dos sintagmas nominais extraídos, com vistas a demonstrar a importância dessa métrica para atividades de indexação e 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 aparecia no documento, em relação ao segundo, tinha uma queda abrupta. O mesmo acontecia até o sétimo sintagma mais frequente no documento. A partir desse ponto, o autor percebeu que existia uma tendência em diminuir minimamente o aparecimento do próximo sintagma, em relação aos anteriores (que apareciam mais vezes).

No tocante a estrutura dos sintagmas nominais, Souza (2005) demonstrou a importância de se levar em conta, paralelamente à frequência de ocorrência, o nível e a estrutura dos sintagmas nominais. O nível de um sintagma nominal corresponde ao número de substantivos que o compõem. Sintagmas de nível 1, 2 ou mais possuem potencialidades diferentes, como por exemplo, um sintagma nominal complexo pode muitas vezes não ser conciso, objetivo. Este autor propõe uma tabela com diferentes pontuações para os sintagmas nominais de acordo com o nível e estrutura dos mesmos,

fazendo uso de uma classificação dos sintagmas nominais (CSN). Propõe também um método para cálculo da pontuação de cada sintagma nominal em função do CSN, onde sintagmas nominais relevantes recebem maiores valores de pontuação.

Lopes (2012) além de fazer uso de critério baseado em frequência no documento, já utilizado por outros autores, faz uso de critério que penaliza os sintagmas nominais que ocorrem com frequência em documentos de outros domínios, ou seja, corpus contrastante, julgando-os como sintagmas nominais que possuem pouco valor conceitual. Além disso, essa autora em sua pesquisa faz uso de um conjunto de heurísticas com o intuito de eliminar sintagmas que não funcionem como conceitos, e extrair sintagmas que funcionem como conceitos. Tais heurísticas podem ser adaptadas para critério de seleção de sintagmas nominais. Tais critérios levam em conta a estrutura gramatical dos sintagmas nominais, tendo em vista eliminar sintagmas nominais contendo números, iniciados com advérbio ou tendo pronome como núcleo, e detectar sintagmas nominais implícitos dentro de outros sintagmas nominais tratando os múltiplos adjetivos associados aos substantivos.

Embora não tenham utilizado critérios para selecionar sintagmas nominais, e por isto não componham os quadros anteriores, alguns trabalhos encontrados suscitam a aplicação de certos critérios de seleção.

Um desses foi o trabalho de Mesquita, Souza e Porto (2013), no qual indicam a potencialidade do uso do posicionamento dos sintagmas nominais no texto como um critério de seleção. Os autores realizaram um estudo, onde identificaram as características de teses de doutorado de oito áreas de conhecimento. As partes das teses analisadas foram a “introdução”, o “desenvolvimento” e a “conclusão”. Por meio da pesquisa desses autores, pôde-se perceber que as partes constitutivas das teses, de diferentes áreas do conhecimento, apresentam diferentes quantidades de sintagmas nominais e distintos comportamentos, como, por exemplo, as teses da área de Letras, as quais apresentaram maiores porcentagens de sintagmas nominais identificados, tendendo assim a possuir maior densidade de conceitos associados aos sintagmas nominais.

O trabalho descrito anteriormente corrobora com a indagação de Corrêa et al. (2011): “Será que não seria o caso de se levar em consideração o posicionamento dos sintagmas nominais do mesmo modo que se verifica nos sistemas de indexação baseados em palavras isoladas?”. Apesar desse critério ainda não ter sido utilizado na seleção de sintagmas nominais para textos em português, acredita-se que a utilização do elemento “posicionamento” contribuirá para a indexação automática por sintagmas

nominais, do mesmo modo que ocorre na indexação manual, na qual são levadas em consideração as partes mais informativas dos documentos.

Em (CORRÊA; BAZÍLIO 2017) é apontado também um critério que pode ser adaptado para a seleção de sintagmas nominais. Neste trabalho foi traçado um perfil descritivo das sequências ou padrões de etiquetas gramaticais das palavras-chaves presentes no texto de documentos e que foram extraídas ou não como sintagmas nominais pelo software OGMA. Os padrões de etiquetas mais frequentes na estrutura das palavras-chaves podem ser utilizados para selecionar sintagmas nominais que contenha tais padrões como os provavelmente mais relevantes. Porém, este critério ainda não foi avaliado em pesquisas envolvendo a seleção de sintagmas nominais para textos escritos em língua portuguesa.

Quanto ao escopo dos critérios de seleção de sintagmas nominais, percebe-se que a maioria dos critérios se baseia em características inerentes aos sintagmas nominais, seja quanto ao nível do sintagma nominal, a estrutura gramatical do mesmo, ou o conteúdo léxico em termos de expressões presentes em *stoplist*. Levando em conta a frequência de adoção em metodologias de seleção de sintagmas nominas, ganham destaque os critérios baseados no conteúdo do documento e conteúdo do corpus, sendo tal conteúdo representado em termos de frequência de ocorrência de cada sintagma nominal em cada documento e no conjunto de documentos respectivamente.

Borges e Lima (2015) realizaram estudo voltado para a identificação de critérios utilizados na construção de softwares de indexação automática no período de 1950 a 2008. Como resultado da referida pesquisa, são apontados oito principais critérios para a extração de termos relevantes na indexação automática. Os critérios identificados como importantes foram: 1. Formatação de frases-termo (*word phrase formation*) através da união de palavras adjacentes; 2. Frequência absoluta de ocorrência de termos no texto como critério de ordenamento por importância; 3. Identificação de palavras por comparação com uso de dicionário, buscando corrigir erros ortográficos; 4. Identificação de radicais de palavras (*word stemming*) visando a conflação de termos; 5. Lista de palavras proibidas (*stoplist* / *stopwords*) a fim de descartar palavras irrelevantes; 6. Peso numérico (*term weighting*) através da razão entre a frequência no documento e a frequência na coleção; 7. Posição do termo no texto como critério de ordenamento por importância; 8. Vocabulário semântico/Vocabulário de cabeçalhos conceituais/Tesauro como fonte de termos relevantes a serem cotejados no texto dos documentos.

Embora os critérios apontados por Borges e Lima (2015) tenham como escopo majoritário a extração de palavras isoladas como termos de indexação, tais critérios ratificam a importância dos critérios encontrados por este trabalho, uma vez que se verificam semelhanças entre os critérios considerados relevantes por ambos os trabalhos, como: Formatação de frases-termo (semelhante à extração de sintagmas nominais); Frequência absoluta de ocorrência de termo no texto; Lista de palavras proibidas (semelhante à lista de sintagmas nominais não descritores); Peso numérico (que é proporcional à frequência inversa de ocorrência de termos nos documentos); e Posição do termo no texto.

5 CONSIDERAÇÕES FINAIS

As metodologias de seleção de sintagmas nominais em textos em língua portuguesa baseiam-se em critérios como frequência absoluta de ocorrência, na frequência normalizada de ocorrência, na frequência inversa de ocorrência nos documentos, na não ocorrência em lista de sintagmas pouco significativos, na estrutura gramatical e no nível do sintagma nominal.

Apesar da não aplicação do critério de posição do termo no texto nos trabalhos voltados para indexação automática por meio de sintagmas nominais, Corrêa et al. (2011) e Borges e Lima (2015) apontam para a relevância deste critério, ratificando o fato de que determinadas partes de um texto são potencialmente mais relevantes do que outras.

O escopo dos critérios de seleção de sintagmas nominais encontrados é em sua maioria baseados em características do sintagma nominal quanto à estrutura gramatical, nível e conteúdo lexical. Levando em conta a adoção de cada critério nas pesquisas, predominam os baseados no conteúdo do documento e no conteúdo do corpus.

Apesar da limitação quanto ao número reduzido de trabalhos encontrados que realizam a seleção de sintagmas nominais, este artigo se mostra como um trabalho pertinente e fundamental para o desenvolvimento de pesquisas voltadas para a indexação automática por meio de sintagmas nominais, uma vez que sintetiza e categoriza critérios de seleção de sintagmas nominais, etapa crucial para a obtenção de sintagmas nominais relevantes na descrição dos assuntos dos documentos.

Como sugestões de trabalhos futuros, aponta-se a necessidade de que mais estudos se voltem especificamente para a seleção de sintagmas nominais, propondo e avaliando critérios para seleção, analisando detidamente a eficácia de cada um dos

critérios de seleção de sintagmas nominais, bem como também a aplicação combinada de múltiplos critérios de seleção no contexto de indexação automática por sintagmas nominais.

REFERÊNCIAS

ARAUJO, Eliany Alvarenga; OLIVEIRA, Marlene de. A produção de conhecimentos e a origem das bibliotecas. In.: OLIVEIRA, Marlene de. (Organizadora). **Ciência da Informação e Biblioteconomia:** novos conteúdos e espaços de atuação. 2. ed. Belo Horizonte: UFMG, 2011.

BARDIN, Laurence. **Análise de conteúdo**. São Paulo: Edições 70, 2011, 229 p.

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.

BORGES, Graciane Silva Bruzinga; LIMA, Gercina Ângela Borém de. O. Desenvolvimento de softwares de indexação automática: breve avaliação dos principais critérios. **Informação & Tecnologia**, v. 2, n. 2, 2015.

BRITO, Marcilio de. Sistemas de Informação em linguagem natural: em busca de uma indexação automática. **Ciência da Informação**, Brasília, v.21, n.3, p. 223-232, set./dez. 1992.

CORRÊA, Renato Fernandes 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. DOI: http://dx.doi.org/10.5380/atoz.v1i1.41280.

CORRÊA, Renato Fernandes; BAZÍLIO, Luiz Henrique Teixeira. Análise da extração de descritores como sintagmas nominais através do software OGMA. **Encontros Bibli:** revista eletrônica de biblioteconomia e ciência da informação, Florianópolis, v. 22, n. 50, p. 44-58, set. 2017. DOI: https://doi.org/10.5007/1518-2924.2017v22n50p44.

CORRÊA, Renato Fernandes; LAPA, Remi Correia. Panorama de Estudos sobre indexação automática no âmbito da Ciência da Informação no Brasil (1973-2012). **Ciência da Informação**. Brasília, DF, v. 42, n. 2, p. 255-273, maio/ago. 2013.

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.Nãopaginado.

LE GUERN, Michel. **Unanalyseurmorpho-syntaxique pour l'indexationautomatique**. 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.

MAIA, Luiz Cláudio Gomes; SOUZA, Renato Rocha. Uso de sintagmas nominais na classificação automática de documentos eletrônicos. **Perspectivas em Ciência da Informação,** v. 15, n.1, p. 154-172, 2010.

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.

MESQUITA, Luiz Antônio Lopes; SOUZA, Renato Rocha; PORTO, Renata Maria Abrantes Baracho. Características de Teses de oito áreas de conhecimento: uma análise para o desempenho de indexação automática através de sintagmas nominais. In.: ENCONTRO NACIONAL DE PESQUISA EM CIÊNCIA DA INFORMAÇÃO, 14., 2013, Santa Catarina. **Anais...** Florianópolis, SC, 2013.

PERINI, Mário A. **Gramática do português brasileiro**. São Paulo: Parábola editorial, 2010. 336 p.

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.

SOUZA, Renato Rocha. 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. DOI: https://doi.org/10.5007/1518-2924.2006v11nesp1p42.

SOUZA, Renato Rocha; ALVARENGA NETO, Rivadávia Correa Drummond de; MENDES, Kellen Christina Ignácia. Mapeamento semântico através da análise de ocorrência de descritores sobre gestão do conhecimento. **Transinformação,** Campinas, v. 19, n.1, p. 19-30, jan./abr., 2007.DOI: http://dx.doi.org/10.1590/S0103-37862007000100002.

SOUZA, Renato Rocha; RAGHAVAN, Koti S. A methodology for noun phrase-based automatic indexing. **Knowledge Organization,** v. 33, n. 1, p. 45-56, 2006.

SOUZA, Renato Rocha; RAGHAVAN, Koti 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; CORRÊA, Renato Fernandes. Ferramentas para indexação automática: uma análise comparativa entre o OGMA, Parser PALAVRAS, LX-Parser e a extração manual de sintagmas nominais. In.: XVI Encontro Nacional de Pesquisa em Pós-Graduação em Ciência da Informação, 2015, João Pessoa. **Anais do XVI Encontro Nacional de Pesquisa em Pós-Graduação em Ciência da Informação**. João Pessoa: PPGCI/UFPB, 2015. p. 1-20.

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

NOTAS

AGRADECIMENTOS
Não se aplica.

CONTRIBUIÇÃO DE AUTORIA

Concepção e elaboração do manuscrito: G. D. Nascimento, R. F. Correa
Coleta de dados: G. D. Nascimento, R. F. Correa
Análise de dados: G. D. Nascimento, R. F. Correa
Discussão dos resultados: G. D. Nascimento, R. F. Correa
Revisão e aprovação: G. D. Nascimento, R. F. Correa

CONJUNTO DE DADOS DE PESQUISA

Todo o conjunto de dados que dá suporte aos resultados deste estudo foi publicado no próprio artigo.

FINANCIAMENTO
Não se aplica.

CONSENTIMENTO DE USO DE IMAGEM
Não se aplica.

APROVAÇÃO DE COMITÊ DE ÉTICA EM PESQUISA
Não se aplica.

CONFLITO DE INTERESSES
Não se aplica.

PUBLISHER –
Universidade Federal de Santa Catarina. Programa de Pós-graduação em Ciência da Informação. Publicação no Portal de Periódicos UFSC. As ideias expressadas neste artigo são de responsabilidade de seus autores, não representando, necessariamente, a opinião dos editores ou da universidade.

EDITORES –
Enrique Muriel-Torrado, Edgar Bisset Alvarez, Camila Barros.

HISTÓRICO –
Recebido em: 28/06/2018– Aprovado em: 07/02/2019