XVIII ENCONTRO NACIONAL DE PESQUISA EM CIÊNCIA DA INFORMAÇÃO – ENANCIB 2017

GT-02 – Organização e Representação do Conhecimento

A REVOCAÇÃO NA INDEXAÇÃO AUTOMÁTICA POR SINTAGMAS NOMINAIS DE ARTIGOS DE PERIÓDICOS EM CIÊNCIA DA INFORMAÇÃO

Victor Galvão Celerino - Universidade Federal de Pernambuco (UFPE)

Renato Fernandes Corrêa - Universidade Federal de Pernambuco (UFPE)

THE RECALL IN AUTOMATIC INDEXING BY NOUN PHRASES OF ARTICLES IN INFORMATION SCIENCE

Modalidade da Apresentação: Comunicação Oral

Resumo: Investiga a utilização dos sintagmas nominais no processo de indexação automática de artigos de periódicos da área de Ciência da Informação. Tem como objetivo verificar a validade da hipótese que a indexação automática por sintagmas nominais do título e resumo permite obter um bom nível de revocação das palavras-chave, o que motiva o uso do título e resumo como entrada para sistemas de indexação automática por sintagmas nominais na construção de bases de dados científicas. A pesquisa é exploratória e experimental (empírica), pautada em estudo de caso. Avalia a revocação das palavras-chave dos autores na indexação automática por sintagmas nominais do título e resumo dos 60 artigos de periódicos do corpus de Souza (2005), utilizando as palavras-chave como padrão de referência de qualidade na indexação. O experimento consiste em: extrair os sintagmas nominais dos documentos compostos por título e resumo através da plataforma de processamento e extração de informação denominada PyPLN; comparar os sintagmas nominais extraídos com as palavras-chave; e mensurar a revocação das palavras-chave. A análise da revocação das palavras-chave na indexação automática por sintagmas nominais, indicou que em 66,6% dos documentos (40 documentos) foi obtido um nível de revocação igual ou superior a 50%, em 26,6% (16 documentos) o nível de revocação ficou entre 14% e 43%, e em 6,6% (4 documentos) o nível de revocação foi 0%, sendo a média de revocação obtida de 56% das palavras-chave por documento. Portanto, conclui-se que a indexação automática por sintagmas nominais do título e resumo dos artigos científicos apresentou bons resultados quanto ao nível de revocação das palavras-chave e que com o tratamento de casos especiais, como a utilização de termos estrangeiros e de caracteres especiais, os resultados podem ser melhorados.

Palavras-Chave: Indexação Automática; Sintagmas Nominais; Artigos De Periódicos Científicos; Ciência Da Informação; Revocação.

Abstract: This work investigates the use of the noun phrases in the process of automatic indexing of articles of periodicals of the area of Information Science. The goal of this work is to verify the validity of the hypothesis that automatic indexing by noun phrases of the title and abstract allows a good

level of recall of the keywords, which motivates the use of the title and abstract as input for automatic indexing systems by noun phrases in the construction of scientific databases. The research is exploratory and experimental (empirical), based on a case study. It evaluates the recall of authors' keywords in the automatic indexing by noun phrases of the title and abstract of the articles of 60 scientific articles of the corpus of Souza (2005), using the keywords as standard reference of indexing quality. The experiment consists of the following: the extraction of the noun phrases from the titles and abstracts of the documents through the platform of processing and extraction of information called PyPLN; the comparison of the extracted noun phrases with the keywords; and measuring of the recall of keywords. The analysis of the recall of keywords by noun phrases shows that in 66,6% of the documents (40 documents) a level of recall equal or greater than 50% was obtained, in 26,6% (16 documents) the level of recall was between 14% and 43%, and in 6,6% (4 documents) the level of recall was 0%. Thus, the average recall obtained was 56% of the keywords by document. Therefore, it concludes that the automatic indexation by noun phrases of the title and abstract of scientific articles obtains good results regarding the level of recall of the keywords and that, with the treatment of special cases, such as the use of foreign terms and special characters, the results may improve.

Keywords: Automatic Indexing; Noun Phrases; Scientific Journal Articles; Information Science; Recall.

1 INTRODUÇÃO

Com o desenvolvimento de novas tecnologias, a publicação e uso da informação em formato digital e o constante aumento no volume de informações, a automatização do tratamento e a organização da informação tem sido cada vez mais demandado. Entre os métodos nesta linha, se destaca a indexação automática.

A indexação automática surgiu devido a necessidade de lidar com a enorme quantidade de informação a ser organizada e disponibilizada, já que a indexação manual não permite em tempo hábil, organizar tamanha quantidade de informação.

A indexação automática começou a ser estudado a partir da década de 1970. Assim como a indexação manual, o processo de indexação automática é bastante complexo e envolve diversos aspectos cognitivos, linguísticos e tecnológicos, formalizados em *softwares* ou programas de computador.

O processo de indexação manual produz termos ou palavras-chave, propostos pelo criador do documento ou por um profissional indexador, para representar o documento. Esses termos ou palavras-chave auxiliam na recuperação do documento através de um Sistema de Recuperação da Informação (SRI).

Entretanto, a maioria dos atuais sistemas de indexação automática, presentes nos SRIs, associam palavras isoladas como ponto de acesso aos documentos, o que gera problemas na representação doas assuntos, visto que as palavras isoladas são insuficientes para descrever os documentos, por não possuírem valor discursivo (BAEZA-YARES; RIBEIRO-NETO, 1999, p.19).

Assim como Baeza-Yares e Ribeiro-Neto (1999), Kuramoto (1995) e Souza (2005) afirmam que a utilização de palavras isoladas para a representação do conteúdo do documento não é adequada, pois causa uma diminuição da realidade extralinguística do autor e não se constituem numa unidade do discurso, mas somente da língua.

Portanto, é preciso que os descritores escolhidos para representarem os documentos contextualizem e representem a informação sem descaracterizá-la. Para isso, Kuramoto (1995) propôs a utilização dos sintagmas nominais como uma alternativa para as palavras isoladas, pois os sintagmas nominais agregam valor semântico à descrição do documento, se constituem em uma unidade do discurso e, portanto, são melhores descritores.

Um sintagma é uma unidade sintática composta por palavras organizadas hierarquicamente em torno do núcleo sintático. Ademais, são classificados de acordo com a função do seu núcleo. No caso dos sintagmas nominais, o núcleo exerce a função de nome (substantivo, pronome substantivo, numeral ou palavra substantivada).

O problema da presente pesquisa é a avaliação da qualidade na indexação automática por sintagmas nominais extraídos dos títulos e resumos dos artigos de periódicos de Ciência da Informação presentes no *corpus* de Souza (2005). A qualidade na indexação automática será mensurada por meio da revocação das palavras-chave dos artigos nos sintagmas nominais extraídos automaticamente.

Diante desse cenário, este artigo tem como objetivo verificar a validade da hipótese de que a indexação automática por sintagmas nominais extraídos do título e resumo possuem bom nível de revocação das palavras-chave.

Existem duas justificavas para o desenvolvimento deste artigo. A primeira é validar se a utilização do título e resumo é suficiente como entrada para sistemas de indexação automática, pois a indexação do texto completo é computacionalmente mais custosa. A segunda é a dificuldade em avaliar a qualidade da indexação automática realizada com a utilização do texto completo.

Para extração dos sintagmas nominais do título e resumo de cada artigo do *corpus* de Souza (2005), será utilizado o PyPLN que consiste de uma plataforma de processamento e extração de informação, desenvolvida pela Escola de Ciências Aplicadas da Fundação Getúlio Vargas (COELHO et al., 2013). Ao final, a revocação das palavras-chave por meio da extração de sintagmas nominais será mensurada.

2 FUNDAMENTAÇÃO TEÓRICA

Nesta seção é apresentado o referencial teórico que norteia a pesquisa e o desenvolvimento deste trabalho.

As subseções seguintes abordarão a temática principal da pesquisa: indexação automática. Sendo descrita em termos de seus conceitos e tipos. Em seguida, será apresentada a utilização dos sintagmas nominais na indexação automática, com o objetivo de explicar o motivo da utilização deles na indexação automática.

2.1 Indexação Automática

Existem diversas definições para indexação automática. Vieira (1988) define a indexação automática como uma tarefa na qual o computador é responsável pela análise dos textos e construção de índices de assuntos, possibilitando a recuperação do documento.

Para Hjørland (2008), a indexação automática é um procedimento feito por algoritmos que funcionam em uma base de dados onde estão presentes as representações dos documentos (textos completos ou parciais, registros bibliográficos, etc.).

Segundo Borges (2009), o processo de indexação automática (também chamado de indexação assistida por computador ou indexação semiautomática) se trata de um modelo de extração com características estatísticas e probabilísticas. Essas características se devem ao fato da utilização de técnicas onde são considerados fatores como a ocorrência de palavras e a repetição de palavras.

Segundo Gil Leiva (1997, p. 53-54), a indexação automatizada se apresenta em três diferentes conceitos:

1. Indexação auxiliada por computador – São utilizados programas de computador que armazenam os termos extraídos pela indexação manual e auxiliam o indexador apresentando-lhe notas e termos relacionados.
2. Indexação semiautomática – Trata-se de uma indexação automática, mas que no final do processo é necessária a validação dos termos por um profissional indexador.
3. Indexação automática – É o processo onde um programa de computador extrai ou atribui termos para a indexação de um documento.

A principal característica da indexação automática é o uso do computador para a sua realização, e isso se deve à necessidade de agilizar a indexação visando acompanhar o ritmo do crescimento informacional. Porém, assim como a indexação manual, a indexação automática também apresenta problemas que, em sua maioria, são causados por erros na

configuração dos sistemas de indexação automática, fazendo com que alguns documentos não consigam ser recuperados (GIL LEIVA; FUJITA, 2012).

De acordo com Lancaster (2004), a indexação automática pode ser aplicada de duas formas: por atribuição ou por extração.

A indexação por extração automática é feita com base na frequência e na posição das palavras ou as expressões presentes no texto do documento, sendo então extraídas para representá-lo (LANCASTER, 2004).

Segundo Borges; Maculan e Lima (2008), a indexação automática por atribuição é um pouco mais complicada em comparação com a indexação automática por extração, porém apresenta uma eficiência maior. Na indexação automática por atribuição é utilizado um instrumento de controle terminológico (vocabulário controlado). A dificuldade apresentada nesse processo de indexação é que para os programas de computador é difícil interpretar o texto do documento e atribuir corretamente os termos do vocabulário controlado. Por exemplo, a frase "Após a queda do muro, a Alemanha ocidental e a Alemanha oriental..." na indexação manual pode ser indexada como "Guerra Fria", mas para o computador esse nível de interpretação é mais difícil (O'CONNOR, 1965 apud LANCASTER, 2004).

Autores como Edmundson (1969), Garvin (1969) e Salton (1973) defendiam que os estudos sobre o processamento de informação e linguística computacional deveriam ser voltados as propriedades estruturais e semânticas da linguagem natural. Portanto, já era evidente que as relações semânticas tinham grande importância, permitindo através da análise semântica da linguagem natural identificar estruturas de formação de conceitos que possibilitam a escolha de termos representativos com significado.

Buscando concretizar o processamento sintático e semântico dos textos na indexação automática, a partir da década de 1990, estudiosos passaram a pesquisar a aplicação dos sintagmas nominais na indexação automática. Dando continuidade as pesquisas nesta linha para textos em português do Brasil, autores como Kuramoto (1995), Souza (2005), Borges (2009), Corrêa et al. (2011), Silva (2014) e Nascimento (2015), destacam a importância da semântica e da sintaxe para a indexação automática, com o objetivo de melhorá-la, e tendo como principal preocupação o significado dos descritores de assunto atribuídos aos documentos.

2.2 Indexação Automática por Sintagmas Nominais

Para compreender melhor o motivo pelo qual pesquisadores passaram a indicar a utilização dos sintagmas nominais na indexação automática, é preciso primeiro entender o que são os sintagmas.

As orações, dentro de um texto, são organizadas de acordo com leis sintagmáticas, e a elas são atribuídas grupos de unidades de significado, chamados de sintagmas. De acordo com Perini (2005), os sintagmas são grupos de unidades que compõem sequências maiores de forma coesa e são subdivisões presentes naturalmente nas orações.

Na semântica, os sintagmas são unidades que possuem significados únicos e coerentes, e são classificados de acordo com as funções que desempenham. Quando possuem a função de predicado, são classificados como sintagma verbal (SV), mas se possuem função de substantivos, são classificados como sintagma nominal (SN) (PERINI, 2005, p. 43-44).

De acordo com Othero (2009), os sintagmas são compostos por uma estrutura sintática e seguem determinadas regras que não permitem a dispersão das palavras. Em uma estrutura sintática, o posicionamento de cada palavra é importante, pois a organização de cada palavra dentro da sentença é que dá sentido e forma à um sintagma.

Os sintagmas definem relações de dependência na oração, estabelecendo ordens de subordinação para os elementos presentes na frase e se dividem nos seguintes tipos, com base na função que exercem:

- Sintagma nominal;
- Sintagma adjetival;
- Sintagma verbal;
- Sintagma preposicional;
- Sintagma adverbial.

Os sintagmas nominais possuem um núcleo (elemento fundamental) que pode ser composto por um nome (substantivo, pronome substantivo, numeral ou palavra substantivada) e é acompanhado de duas estruturas que são definidas como pré ou pós-nucleares, como ilustrado no Quadro 1.

Quadro 1: Elementos dos sintagmas nominais.

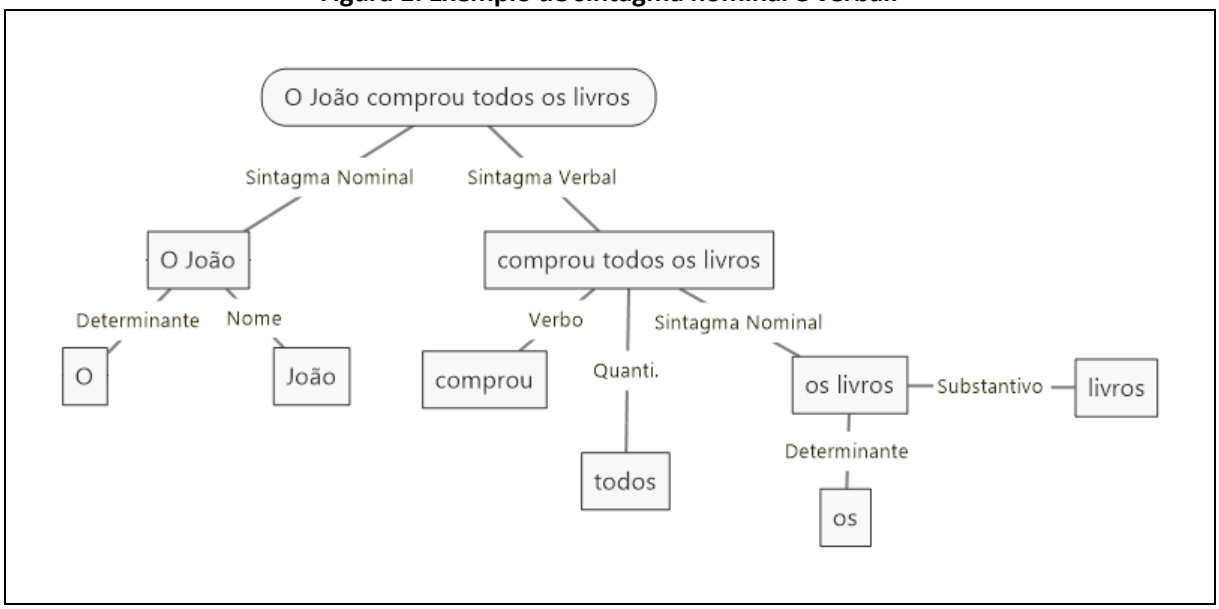
Elementos que compõem os sintagmas nominais		
Elementos pré-nucleares	Núcleo	Elementos pós-nucleares
Predeterminantes, determinantes, quantificadores, possessivos sintéticos, numeral.	Nome (substantivo, pronome substantivo, numeral ou palavra substantivada).	Modificadores (palavra ou conjunto de palavras que qualificam o núcleo, restringem o sentido do núcleo, inclusive outros nomes que podem ser núcleos também).

Fonte: PERINI (2010), p. 259.

A estrutura pré-nuclear pode apresentar os seguintes elementos: predeterminantes, determinantes, quantificadores, possessivos sintéticos e numeral. A estrutura pós-nuclear pode apresentar palavras que qualifiquem o núcleo, os modificadores (PERINI, 2008).

Na Figura 1 temos o exemplo de uma estrutura sintagmática de uma oração que é composta por sintagmas do tipo nominal e verbal. Ao analisar a Figura 1, percebe-se que dentro de uma mesma oração é possível a presença de dois ou mais tipos de sintagmas. Alguns sintagmas podem aparecer dentro de outros sintagmas, como foi o caso do sintagma nominal “os livros”, que estava dentro do sintagma verbal “comprou todos os livros”.

Figura 1: Exemplo de sintagma nominal e verbal.



Fonte: Autoria nossa. – 2017.

Como os sintagmas são compostos por estruturas sintáticas que seguem determinadas regras, consequentemente, os sintagmas nominais também possuem regras quanto à sua formação. Fundamentado por Miorelli (2001) e Santos (2005), Silva (2014, p. 50) elaborou um conjunto de regras quanto à formação dos sintagmas nominais (Quadro 2):

Quadro 2: Regras de formação dos sintagmas nominais.

Regras	Exemplos
Regra geral: DET + MOD + N + MOD	A interdisciplinar Ciência da Informação
Regra 1: DET + N + MOD	A Ciência da Informação
Regra 2: N + MOD	Informação estratégica
Regra 4: DET + N	A informação
Regra 5: N	Informação
Regra 6: DET + N + DET + N + MOD	A filosofia e a ciência juntas
Regra 7: DET + DET + N + MOD	A minha recuperação da informação
Regra 8: MOD + N + MOD	Grande área da informação
Regra 9: DET + DET + N	Uma certa área

Fonte: Silva (2014, p. 50), baseado em Miorelli (2001) e Santos (2005).

Segundo Kuramoto (1995), os sintagmas nominais que possuem outros sintagmas nominais, podem ser classificados em níveis. Os níveis variam de acordo com a quantidade de sintagmas nominais presentes na oração. Com isso, um sintagma nominal que não possui outro sintagma nominal em sua estrutura, é considerado nível 1. Já um sintagma nominal que possui um sintagma nominal de nível 1 em sua estrutura, é considerado nível 2, e assim sucessivamente. Exemplificando, teríamos: “Sistema de Classificação de Documentos” como um sintagma nominal de nível 3; “Classificação de Documentos” como um sintagma nominal de nível 2; e “Documentos” como um sintagma nominal de nível 1.

Mas por que então é indicada a utilização dos sintagmas nominais na indexação automática, ao invés das palavras isoladas? A proposta de utilização dos sintagmas nominais teve início quando se percebeu que a utilização das palavras isoladas não era suficiente para a representação e recuperação da informação.

Diante desse cenário, passou-se a estudar os aspectos semânticos e sintáticos para a indexação automática. O resultado desses estudos foi que a utilização de grupos de substantivos ou grupos nominais (do inglês *noun groups*), apresentavam mais qualidade para a indexação, pois possuem maior valor semântico quando comparados a outros termos morfossintáticos da linguagem.

No contexto das pesquisas sobre a extração de grupos nominais, foi proposta a extração automática e uso dos sintagmas nominais (do inglês *noun phrases*) e, ao usá-los no processo de indexação, notou-se uma melhora nos resultados do processo de representação e recuperação da informação, como conclui, por exemplo, a pesquisa realizada por Kuramoto (1995).

Michel Le Guern (1991) é considerado o pioneiro nas pesquisas sobre o uso dos sintagmas nominais na indexação automática. Ele tinha como proposta a troca das palavras

isoladas por sintagmas nominais como descritores da informação, pois os sintagmas nominais são portadores de significado para a indexação e recuperação da informação. Além disso, ele ressaltava que existe uma diferença entre descritor e palavra, sendo o descritor uma unidade do discurso e a palavra uma unidade da língua, unidade essa que não possui significado para a indexação e recuperação da informação.

Portanto, a diferença da indexação automática com palavras isoladas e da indexação automática por sintagmas nominais é quanto a sua significação. Durante a indexação de um documento, deve-se extrair descritores que facilitem sua recuperação. Logo, as palavras (símbolos sem referências) não são adequadas (KURAMOTO, 2002).

De acordo com Kuramoto (2002, p.6), sintagmas nominais são a menor unidade do discurso portadora de informação e podem se apresentar como palavra isolada ou um conjunto de palavras, ambas com valor semântico e sintático.

Assim como na indexação por palavras isoladas, o processo de indexação automática por sintagmas nominais deve seguir algumas etapas para ser realizado. Nascimento (2015) apresenta um quadro que descreve as etapas do processo de indexação automática por sintagmas nominais (Quadro 3).

Quadro 3: Etapas da indexação automática por sintagmas nominais

Processo de indexação automática por meio de sintagmas nominais	
1ª Etapa	*Identificação dos sintagmas nominais* através das subetapas de "etiquetagem" e de "cotejamento dos léxicos etiquetados com as regras dos sintagmas nominais"
2ª Etapa	*Extração dos sintagmas nominais* do texto, mostrando-os em listas, por exemplo.
3ª Etapa	*Seleção dos sintagmas nominais* com base em critérios que os classifiquem como "Bons Descritores"

Fonte: Nascimento, 2015.

A 1ª etapa é responsável pela etiquetagem e identificação dos sintagmas nominais através de *softwares*. Nessa etapa alguns *softwares* apresentam também o cotejamento dos léxicos etiquetados com as regras dos sintagmas nominais.

Na 2ª etapa é realizada a extração dos sintagmas nominais identificados e é gerada uma lista destes sintagmas nominais.

Por fim, na 3ª etapa, são selecionados os sintagmas nominais considerados "bons descritores" dos assuntos do texto. Esse processo de seleção é feito através da pontuação dos sintagmas nominais com base em critérios que tentam estimar sua relevância para o texto.

Nos parágrafos a seguir desta seção, serão descritos os trabalhos relacionados à presente pesquisa, em ordem cronológica.

Caso o leitor deseje obter uma revisão mais ampla das pesquisas sobre indexação automática no Brasil, sugerimos a leitura do trabalho dos autores Corrêa e Lapa (2013), por apresentarem um panorama dos estudos sobre a indexação automática no âmbito da ciência da informação no Brasil dentro do período de 1973 a 2012.

No trabalho de Souza (2005), o objetivo era propor uma metodologia que viabilizasse o processo indexação automática através da extração de sintagmas nominais e da seleção dos mesmos através da análise de fatores como: a frequência de ocorrência desses sintagmas nominais nos textos dos documentos, no conjunto dos documentos; a estrutura e nível dos sintagmas nominais e a ocorrência desses em tesauro de um campo de conhecimento específico. Souza (2005) compilou o *corpus* que é utilizado nesse artigo, porém o autor fez uso do texto completo para a extração e seleção dos sintagmas nominais, enquanto no presente artigo foram utilizados apenas o título e resumo de cada documento.

A pesquisa de Maia e Souza (2010) tinha como objetivo verificar se ocorria o aprimoramento na classificação de documentos eletrônicos através de técnicas e algoritmos de mineração de texto (análise de texto), quando eram utilizadas como características ou termos de indexação os sintagmas nominais em detrimento do uso das palavras ou termos isolados. Os resultados obtidos por Maia e Souza (2010) apontaram que a utilização de sintagmas nominais na classificação automatizada de documentos gerava índices semelhantes à utilização dos termos isolados sem *stopwords*.

O trabalho desenvolvido por Corrêa et al. (2011) abordou a utilização dos sintagmas nominais no processo indexação automática das teses e dissertações depositadas na Biblioteca Digital de Teses e Dissertações da UFPE (BDTD-UFPE), com a hipótese de que os sintagmas nominais seriam uma melhor unidade de conhecimento para a indexação e recuperação de informação quando comparadas com as palavras isoladas. A conclusão apontada pela pesquisa é que deve ser dada mais atenção para a ordenação por relevância dos sintagmas nominais e que a extração dos sintagmas nominais não é por si só suficiente para a indexação e recuperação de informação em ambientes digitais, visto que alguns sintagmas nominais extraídos não possuem valor representativo do assunto dos documentos.

No artigo desenvolvido por Souza e Raghavan (2014), foi discutida uma nova abordagem quanto a seleção dos sintagmas nominais como descritores dos documentos. O

resultado obtido com a pesquisa foi que a atribuição de pesos aos sintagmas nominais com base apenas na frequência não é satisfatória. Porém, notou-se uma melhora quando utilizada a frequência normalizada e a frequência inversa no cálculo dos pesos, e uma melhora um pouco maior ao utilizá-las em conjunto com a classificação dos sintagmas nominais em níveis.

Martins (2014) realizou uma pesquisa com o objetivo de avaliar o uso de sintagmas nominais como fontes de dados para um sistema automático de classificação de documentos textuais armazenados no formato digital. Para isso, ele desenvolveu uma metodologia dividida em duas etapas: a primeira realizava um teste qualitativo na comparação entre as representações dos documentos do *corpus*; e a segunda etapa utilizou o *software* SVMLight para classificar automaticamente as representações dos documentos. Os resultados da pesquisa indicaram que a utilização do processo de *stemming* foi mais satisfatória, se comparada com a utilização do próprio sintagma no momento de treinar o classificador automático. Martins (2014) afirma que os trabalhos que usaram apenas sintagmas puros apresentaram, em média, 80% de classificação satisfatória, ao passo que com a utilização do *stemming* nos sintagmas nominais esse percentual foi de 100%.

A pesquisa desenvolvida por Silva e Corrêa (2015) avaliou e comparou as seguintes ferramentas de extração automática de sintagmas nominais: Parser PALAVRAS, OGMA e LX-Parser. Utilizando como referência a extração manual de sintagmas nominais, o objetivo dos autores era destacar quais ferramentas poderiam auxiliar na extração automática de sintagmas nominais. Os resultados indicaram que o LX-Parser, em alguns casos, obtinha uma performance superior aos outros. Porém, o número de identificação de falsos sintagmas nominais foi maior no LX-Parser em comparação ao PALAVRAS e o OGMA. A ferramenta PALAVRAS apresentou uma taxa de erro de 6%, e isso, segundo os autores, corrobora com as demais pesquisas que indicam o PALAVRAS como ferramenta com maior potencial para identificação de sintagmas nominais.

Nascimento e Corrêa (2016) avaliaram critérios para a seleção de sintagmas nominais relevantes na representação dos assuntos dos documentos. Nesse estudo foram avaliados dez critérios para a seleção dos sintagmas nominais e foram definidos níveis de eficácia para cada critério. Os resultados obtidos pelos critérios avaliados foram distintos, indicando quais são mais eficazes ou úteis na seleção de sintagmas nominais com valor de descritor de assunto.

Diferentes ferramentas foram utilizadas nas pesquisas relatadas para realizar o processo de indexação automática por sintagmas nominais. Algumas dessas ferramentas

foram o Parser PALAVRAS, o LX-Parser e o OGMA. Essas ferramentas podem ser classificadas como: ferramentas de etiquetagem, ferramentas de identificação de sintagmas nominais, ferramentas de extração de sintagmas nominais, e ferramentas de seleção de sintagmas nominais.

As ferramentas de etiquetagem são responsáveis por identificar e rotular, através de etiquetas, as palavras presentes em um texto com base em suas classes gramaticais para, em seguida, ser feita a identificação dos sintagmas nominais. Todas as ferramentas citadas anteriormente realizam esta tarefa.

As ferramentas de identificação têm como objetivo analisar as sequências de léxicos e respectivas etiquetas gramaticais, através dessa análise, aplicar as regras gramaticais de formação dos sintagmas nominais, a fim de identificar as sequências de palavras que constituem um sintagma nominal. Todas as ferramentas citadas anteriormente também realizam esta tarefa.

As ferramentas de extração de sintagmas nominais são responsáveis tanto pela identificação como também pela extração dos sintagmas nominais presentes no documento. Além disso, ela produz uma lista, separada do documento original, apresentando todos os sintagmas nominais. Dentre as ferramentas citadas, esta tarefa só é realizada pela ferramenta OGMA.

Por fim, as ferramentas de seleção de sintagmas nominais são responsáveis por selecionar os sintagmas nominais que possuem valor descritivo de assunto para o documento, e, consequentemente, descartar os sintagmas nominais com significado vazio. O processo de seleção é realizado com base em critérios que determinam a importância de cada sintagma nominal. Dentre as ferramentas citadas, apenas a OGMA possui um procedimento de pontuação que permite a seleção de sintagmas nominais.

Outra ferramenta capaz de realizar a etiquetagem, a identificação e extração de sintagmas nominais é o PyPLN (COELHO et al., 2013), que foi escolhida para o desenvolvimento dessa pesquisa. O PyPLN foi desenvolvido através de um projeto de pesquisa da Escola de Matemática Aplicada da Fundação Getúlio Vargas, e trata-se de uma Plataforma de Processamento de Linguagem Natural desenvolvida na linguagem de programação Python que utiliza o Parser PALAVRAS como analisador sintático-semântico e permite realizar diversas análises linguísticas em documentos de texto.

3 METODOLOGIA

O presente artigo tem como objetivo validar a hipótese que a indexação automática por sintagmas nominais do título e resumo permite obter um bom nível de revocação das palavras-chave. Sendo verdadeira a hipótese, estaria justificado o uso do título e resumo como entrada para sistemas de indexação automática por sintagmas nominais na construção de bases de dados científicas.

Este artigo tem caráter teórico-prático e foi desenvolvido através das abordagens qualitativa (estudo de caso) e quantitativa (comparação e mensuração da revocação das palavras-chave). Trata-se de uma pesquisa exploratória de cunho experimental e por isso é realizado um estudo de caso com o objetivo de estudar exaustivamente e com profundidade a revocação das palavras-chave dos documentos que compõem o corpus de 60 artigos de periódicos em Ciência da Informação coletados por Souza (2005).

Foram definidas etapas com o objetivo de orientar o desenvolvimento do experimento.

A primeira etapa foi a extração manual do título e resumo de cada um dos 60 documentos do corpus e salvá-los em arquivos de texto. Esses arquivos em formato de texto serão submetidos separadamente ao PyPLN, pois a extração dos sintagmas nominais é feita apenas documento por documento, e não todos ao mesmo tempo. Ainda na primeira etapa, foram coletadas e organizadas, em uma planilha eletrônica, todas as palavras-chave dos documentos do *corpus*, mantendo uma indicação do documento ao qual pertencem, com o objetivo de facilitar o teste de revocação das palavras-chave.

A segunda etapa corresponde a extração dos sintagmas nominais dos arquivos de texto pelo PyPLN. Para realizar a extração, os arquivos no formato texto foram submetidos ao PyPLN. Cada resumo continha em média 126 palavras, sendo o menor resumo com 32 palavras e o maior com 314, e a soma dos resumos tinha 7582 palavras. Cada título continha em média 11 palavras, sendo o menor com 4 e o maior com 21, e a soma de todos os títulos tinha 698 palavras. Após submeter os resumos e títulos ao PyPLN, o processo de extração dos sintagmas nominais levou, em média, 5 minutos por documento.

A terceira etapa consiste na coleta e organização dos sintagmas nominais extraídos pela plataforma PyPLN de cada arquivo. Os sintagmas nominais extraídos pelo PyPLN, apresentam alguns caracteres especiais, como: “_” (sublinha), “*” (asterisco), aspas, parênteses, chaves, sinais de pontuação (? ! , . : ;). Alguns desses caracteres especiais são utilizados para destacar o referente e o quantificador. Por exemplo, no sintagma nominal

extraído: “o *valor de_ _o conhecimento”, o caractere sublinha é utilizado para denotar a separação de uma contração, e o asterisco para denotar o referente do sintagma nominal (núcleo). Além disso, a lista de sintagmas nominais extraídos pelo PyPLN contém os sintagmas nominais na sua forma máxima e os respectivos sintagmas nominais de níveis inferiores embutidos no primeiro, o que levou à necessidade de filtrar somente os sintagmas nominais máximos extraídos, a fim de evitar a contagem duplicada de sintagmas nominais contendo as palavras-chaves.

Por fim, a quarta etapa trata da comparação das palavras-chave com os sintagmas nominais máximos extraídos através do PyPLN e da mensuração da revocação das palavras-chave. Para isso, foi utilizada uma planilha eletrônica, onde eram registrados, para cada documento: o número de palavras-chave encontradas nos sintagmas nominais de nível máximo, o número total de palavras-chave e o número de sintagmas nominais extraídos. Nessa etapa, a revocação das palavras-chave foi mensurada através do cálculo da razão entre o número de palavras-chave recuperadas em sintagmas nominais de nível máximo, dividido pelo total de palavras-chave do documento.

4 ANÁLISE DOS RESULTADOS

Nesta seção é apresentada a análise da revocação das palavras-chave dos documentos a partir da indexação automática por sintagmas nominais extraídos do título e resumo dos mesmos. Os níveis da revocação das palavras-chaves para cada documento do corpus podem ser observados na Tabela 1.

Tabela 1: Resultado da Revocação.

VALOR DA REVOCAÇÃO DAS PALAVRAS-CHAVE POR DOCUMENTO							
Documento	**Revocação**	**Documento**	**Revocação**	**Documento**	**Revocação**	**Documento**	**Revocação**
DOC. 1	66,7%	DOC. 16	20,0%	DOC. 31	60,0%	DOC. 46	83,3%
DOC. 2	100,0%	DOC. 17	0,0%	DOC. 32	66,7%	DOC. 47	33,3%
DOC. 3	60,0%	DOC. 18	25,0%	DOC. 33	80,0%	DOC. 48	66,7%
DOC. 4	60,0%	DOC. 19	40,0%	DOC. 34	66,7%	DOC. 49	0,0%
DOC. 5	60,0%	DOC. 20	25,0%	DOC. 35	60,0%	DOC. 50	50,0%
DOC. 6	42,9%	DOC. 21	20,0%	DOC. 36	50,0%	DOC. 51	80,0%
DOC. 7	20,0%	DOC. 22	55,6%	DOC. 37	50,0%	DOC. 52	80,0%
DOC. 8	100,0%	DOC. 23	66,7%	DOC. 38	100,0%	DOC. 53	60,0%
DOC. 9	80,0%	DOC. 24	66,7%	DOC. 39	60,0%	DOC. 54	14,3%
DOC. 10	40,0%	DOC. 25	50,0%	DOC. 40	100,0%	DOC. 55	50,0%
DOC. 11	25,0%	DOC. 26	75,0%	DOC. 41	100,0%	DOC. 56	20,0%
DOC. 12	50,0%	DOC. 27	33,3%	DOC. 42	80,0%	DOC. 57	71,4%
DOC. 13	33,3%	DOC. 28	42,9%	DOC. 43	100,0%	DOC. 58	71,4%
DOC. 14	33,3%	DOC. 29	100,0%	DOC. 44	66,7%	DOC. 59	0,0%

DOC. 15	0,0%	DOC. 30	66,7%	DOC. 45	80,0%	DOC. 60	100,0%
						MÉDIA	56,0%
						DESVIO	29,0%

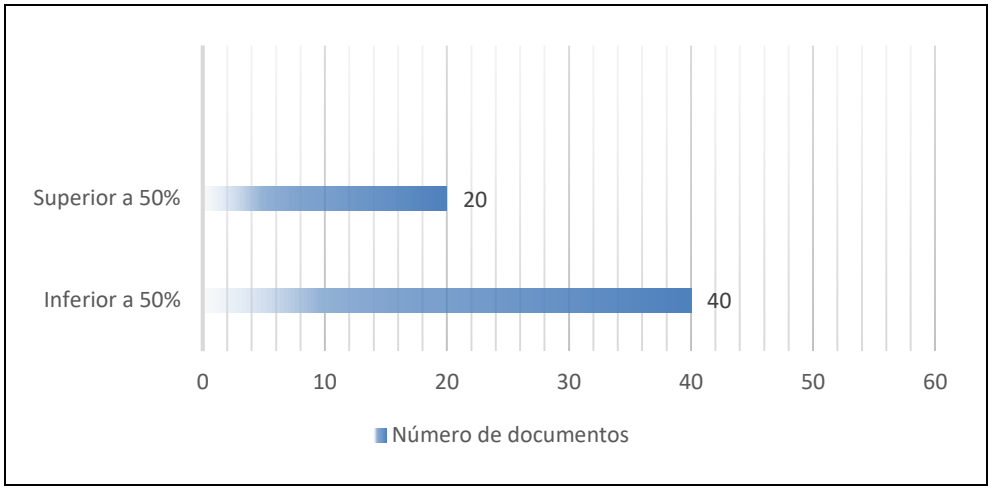
Fonte: Autoria nossa - 2017.

A média da revocação das palavras-chave nos 60 documentos foi de 56%, o que significa que em média mais da metade das palavras-chave dos autores foram recuperadas através dos sintagmas nominais extraídos automaticamente pelo PyPLN. Um resultado bastante expressivo, pois a grande maioria dos documentos apresentou a possibilidade de ser recuperado. Essa média pode ser melhorada se alguns problemas específicos, como a utilização de caracteres especiais e utilização de palavras estrangeiras, fossem tratados no processamento automático dos textos.

Outro resultado interessante foi o desvio padrão de 29%. O desvio padrão significa a medida de variação ou dispersão obtida pela média da revocação, ou seja, ele mede a variabilidade dos valores em torno da média. Quanto mais baixo o desvio padrão, os dados tendem a estar mais próximos da média. Portanto, a maioria (65%) dos valores de revocação das palavras-chaves ficaram entorno da média, variando entre 27% e 85% de revocação.

Como demonstrado no Gráfico 1, dos 60 documentos, 20 deles (33,3%) tiveram um índice de revocação inferior a 50%. Dentre esses, em quatro documentos (6,6%) o valor de revocação foi nulo (0%), pois não foram recuperadas palavras-chave nos sintagmas nominais extraídos pelo PyPLN. Entretanto, em 40 documentos (66,6%), houve uma revocação igual ou superior a 50% das palavras-chave.

Gráfico 1: Quantidade de documentos por nível de revocação.



Fonte: Autoria nossa - 2017.

Quatro documentos não conseguiram ter nenhuma de suas palavras-chave recuperadas nos sintagmas nominais de nível máximo extraídos pelo PyPLN, ou seja, apenas 6,6% dos documentos. Percebeu-se que o motivo da baixa revocação das palavras-chave nesses documentos foi a ausência da palavra-chave no título e no resumo do documento, o que impossibilitava o PyPLN de recuperá-las através da extração de sintagmas nominais.

Outro fator que também contribuiu para que alguns outros documentos tivessem um valor de revocação inferior a 50% foi que, em alguns casos, o formato da palavra-chave se aproximava de algum sintagma nominal de nível máximo extraído pelo PyPLN, mas como não era exatamente igual, não foi contabilizado no cálculo de revocação. Como exemplo, podemos citar o caso do documento 15, no qual a palavra-chave “Ensino e pesquisa” se aproximava do sintagma nominal extraído “Relação Ensino-Pesquisa”.

Apesar do número de documentos com índice de revocação inferior a 50% ter sido expressivo (33,3%), o número de documentos que tiveram uma revocação igual ou superior a 50% foi bastante satisfatório. No total, foram 40 documentos, ou seja, 66,6% dos documentos, sendo que, dos 40 documentos, 8 deles (13%) tiveram uma revocação de 100% das palavras-chave.

Ao analisar os resultados apresentados pelos 40 documentos com revocação superior a 50%, ficou claro que a positividade desses resultados não se deve a quantidade de sintagmas nominais extraídos, mas sim ao fato de que as palavras-chave apresentadas pelos documentos estavam presentes no resumo ou no título.

Nos 8 documentos em que foi obtida uma revocação de 100%, as palavras-chave utilizadas não apresentavam o uso de nenhum caractere especial como o travessão, aspas, etc.

Em alguns casos, o uso do travessão no resumo ou no título do documento impossibilitou a extração do sintagma nominal. Por exemplo, no documento 42, existem duas palavras-chave: Literatura Cinzenta e Literatura Branca. Porém o PyPLN não conseguiu extrair esses dois sintagmas nominais, porque no texto foram utilizadas aspas nas palavras: “Cinzenta” e “Branca”, o que resultou em uma queda na revocação desse documento.

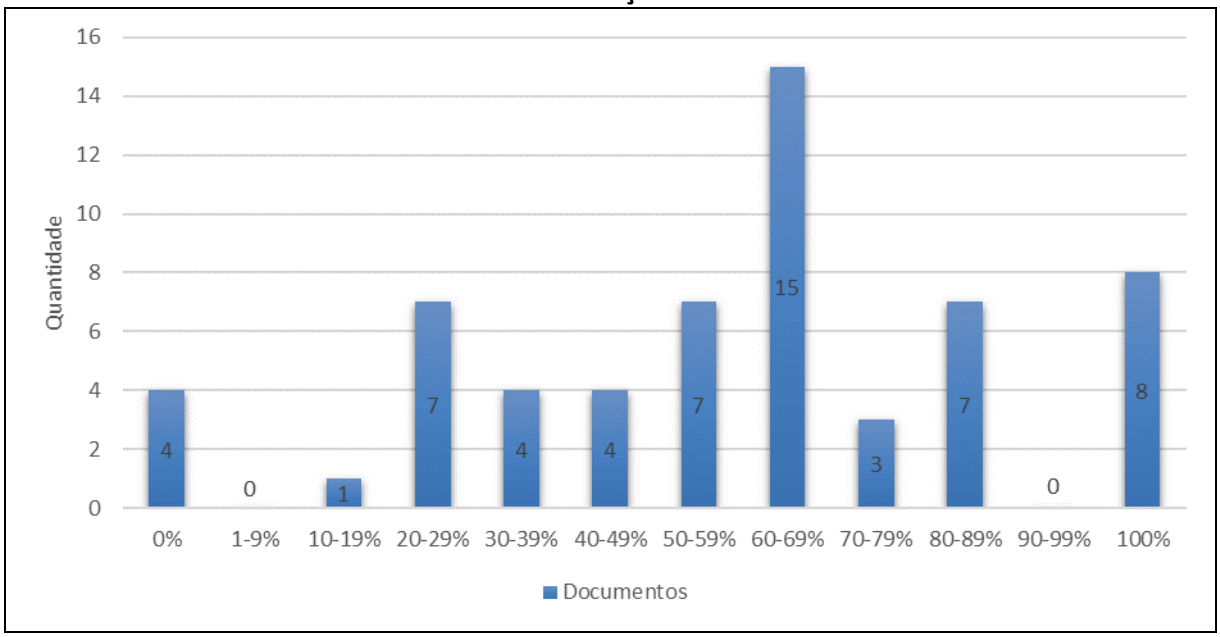
Outro fator que contribuiu para que algumas palavras-chave não conseguissem ser recuperadas foi o uso de alguns termos estrangeiros. Em determinados documentos o PyPLN conseguiu extrair o sintagma nominal corretamente, mas em outros não. Como exemplo,

temos o documento 54, que possui o sintagma nominal “Informacion Literacy”, que o PyPLN extraiu separadamente cada palavra como um sintagma nominal.

Na totalidade dos resultados, a revocação obtida foi bastante satisfatória, pois grande parte dos documentos apresentaram a possibilidade de serem recuperados, mesmo apresentando um nível de revocação um pouco baixo. Ao todo, 93,33% dos documentos (56 dos 60 documentos) poderiam ser recuperados em uma base de dados que utilizasse os sintagmas nominais extraídos do título e do resumo do documento como descritores documentais, como ilustrado no Gráfico 2.

Com relação ao total de palavras-chave recuperadas por nível de revocação, 4 documentos com 0% de revocação não obtiveram nenhuma palavra-chave recuperada do total de 13 palavras-chave; 16 documentos com nível de revocação entre 10% e 49% tiveram 23 palavras-chave recuperadas do total de 78 palavras-chave; 32 documentos com nível de revocação entre 50% e 89% tiveram 98 palavras-chave recuperadas do total de 148 palavras-chave; e 8 documentos com nível de revocação de 100% tiveram um total de 20 palavras-chave recuperadas por meio de sintagmas nominais.

Gráfico 2: Quantidade de documentos por faixas de valores de revocação.



Fonte: Autoria nossa - 2017.

5 CONSIDERAÇÕES FINAIS

Através da análise dos resultados deste trabalho, conclui-se que é viável a utilização da indexação automática por sintagmas nominais do título e do resumo na construção de bases de dados científicas na área de Ciência da Informação. Pois tal indexação permite um nível de revocação médio de mais da metade das palavras-chaves por documento (contidas nos sintagmas nominais extraídos), o que constitui em um bom nível de revocação das palavras-chave dos autores.

A análise da revocação na indexação automática indicou que em 66,6% dos documentos (40 documentos) foi obtido um nível de revocação igual ou superior a 50%, em 26,6% (16 documentos) o nível de revocação ficou entre 14% e 43%, e em 6,6% (4 documentos) o nível de revocação foi 0%, resultando em uma média de revocação de 56% das palavras-chave por documento.

É importante destacar que, em sua maioria, a extração dos sintagmas nominais através do PyPLN foi bastante satisfatória quanto ao nível de revocação das palavras-chave, e este nível pode ser melhorado se forem feitos alguns ajustes no tratamento de termos estrangeiros e na utilização dos caracteres especiais.

Neste trabalho, ficou também evidente que uma limitação da indexação automática por sintagmas nominais do título e resumo dos artigos científicos, está na falta de uniformidade quanto a expressividade e coerência da descrição e representação dos trabalhos através do título, resumo e palavras-chave. Entende-se que a maior adoção pela comunidade científica das normas existentes para elaboração de resumo e atribuição de termos advindos de tesauros especializados, possam minimizar tal limitação no futuro.

Como trabalhos futuros, seria interessante: estudar formas de normalizar os sintagmas nominais para obter descritores mais semelhantes as palavras-chave; investigar métodos para a eliminação de sintagmas nominais vazios de significado; comparar a revocação obtida na indexação automática do título e resumo com a revocação obtida na indexação automática do texto completo por Souza (2005), visando validar a hipótese que a indexação automática com sintagmas nominais do título e resumo tem níveis de revocação próximos da indexação automática com sintagmas nominais do texto completo; avaliar a revocação e a precisão na recuperação das palavras-chave nos sintagmas nominais extraídos do título e resumo através de outros *softwares*, com a finalidade de aperfeiçoar os sistemas de indexação automática por sintagmas nominais.

REFERÊNCIAS

BAEZA-YARES; RIBEIRO NETO, B. **Modern Information Retrieval**. [S.L]: ACM Press, 1999.

BORGES, G. S. B. **Indexação automática de documentos textuais:** critérios essenciais. 2009. Tese (Doutorado em Ciência da Informação) - Escola de Ciência da Informação, Universidade Federal de Minas Gerais, Belo Horizonte, 2009.

COELHO, F. C. et al. PyPLN: a Distributed Platform for Natural Language Processing. [S.l]: Journal of Machine Learning Research. 2013.

CORRÊA, R. F. et al. Indexação e recuperação de teses e dissertações por meio de sintagmas nominais. **AtoZ: Novas Práticas em Informação e Conhecimento**, v. 1, n. 1, p. 11-22, 2011.

CORRÊA, 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, 2015.

EDMUNDSON, H. P. New methods in automatic extracting. **J. ACM**, v. 16, n. 2, p. 264- 285, 1969.

GARVIN, P. L. et al. **Some opinions concerning linguistics and reformation processing.** Washington, D. C.: Center for Applied Linguistics, 1969.

GIL LEIVA, I. **La automatización de la indización, propuesta teórico-metodológica**: aplicación al área de biblioteconomía y documentación. 1997. Tese – Universidad de Murcia, Murcia, Espanha, 1997.

GIL LEIVA, I.; FUJITA, M. S. L. (org). **Política de Indexação**. São Paulo: Cultura Acadêmica; Marília: Oficina Universitária, 2012.

HJØRLAND, B. Automatic Indexing. Lifeboat for Knowledge Organization. [S.L]:[S.N], 2008. Disponível em: <http://www.iva.dk/bh/lifeboat_ko/CONCEPTS/automatic_indexing.htm >. Acesso em: 15 ago. 2016.

KURAMOTO, H. Sintagmas Nominais: uma nova proposta para a recuperação de informação. **DataGramaZero Revista de Ciência da Informação**. v. 3, n. 1, 2002.

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

LANCASTER, F. W. **Indexação e Resumos**: teoria e prática. Tradução de Antonio Agenor Briquet de Lemos. 2. ed. revista e atualizada. Brasília, DF: Briquet de Lemos, 2004.

LE GUERN, M. Un analyseur morpho-syntaxique pour l'indexation automatique. **Le Français Moderne**. v. 59, n. 1, p. 22-35, 1991.

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

MAIA, L. C. U. 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.

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

NASCIMENTO, G. D. **Dos sintagmas nominais aos descritores documentais:** estudo de caso na indexação de teses e dissertações da área de direito. 2015. Dissertação (Mestrado em Ciência da Informação) – Departamento de Ciência da Informação, Universidade Federal de Pernambuco, Recife, 2015.

NASCIMENTO, G. D.; CORRÊA, R. F. Sintagmas nominais com valor de descritores: critérios para seleção. **XVII Encontro Nacional de Pesquisa em Ciência da Informação**, v. 17, 2016.

OTHERO, G. A. **A gramática da frase em português**: algumas reflexões para a formalização da estrutura frasal em português. Porto Alegre: EDIPUCRS, 2009.

PERINI, M. A. **Gramática descritiva do português**. 4.ed. São Paulo: Ática, 2005

SALTON, G. Recent studies in automatic text analysis and document retrieval. **Journal of the Association for Computing Machinery**, v. 20, n. 2, p. 258-27, 1973.

SANTOS, C. N. **Aprendizado de máquina na identificação de sintagmas nominais:** o caso do português brasileiro. 2005. Dissertação (Mestrado em Sistemas e Computação) – Instituto Militar de Engenharia, Rio de Janeiro, 2005.

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

SILVA, T. J.; CORRÊA, R. F. Ferramentas para indexação automática: uma análise comparativa entre o OGMA, parser palavras, lx-parser e a extração manual de sintagmas nominais. **XVII Encontro Nacional de Pesquisa em Ciência da Informação**, v. 16, 2015.

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

SOUZA, R. R; RAGHAVAN, K. S. A extração de palavras-chave a partir de textos: um estudo exploratório utilizando sintagmas. **Informação & Tecnologia**, v. 1, n. 1, p. 5-16, 2014. Disponível em: <http://basessibi.c3sl.ufpr.br/brapci/v/a/15114>. Acesso em: 26 Jan. 2017.