ARTIGO

Recebido em:
02/09/2016

Aceito em:
20/05/2017

Encontros Bibli: revista eletrônica de biblioteconomia e ciência da informação, v. 22, n. 50, p. 44-58, set./dez., 2017. ISSN 1518-2924. DOI: 10.5007/1518-2924.2017v22n50p44

Análise da extração de descritores como sintagmas nominais através do software OGMA

Analysis of extraction of descriptors as noun phrases through the OGMA software

Renato Fernandes CORRÊA (renato.correa@ufpe.br)*
Luiz Henrique Teixeira BAZÍLIO (henriquecatende@hotmail.com)**
* Docente do Programa de Pós-Graduação em Ciência da Informação - UFPE.
** Graduando em Gestão da Informação - UFPE.

Resumo
Analisa a indexação automática por sintagmas nominais de documentos, compostos por título e resumo de 30 teses e dissertações, escritos em português e de três áreas do conhecimento diferentes. O método de pesquisa é categorizado como exploratório, com base em revisão de literatura e experimento computacional. O experimento consistiu na análise da saída do software OGMA quando aplicado ao corpus de documentos e a mensuração do nível de revocação das palavras-chaves. Durante a análise, foram observadas quais palavras-chave indicadas pelos autores estavam nos documentos e depois observou-se quais palavras-chave presentes nos documentos foram extraídas ou não como sintagmas nominais pelo software. Foi traçado um perfil descritivo das sequências ou padrões de etiquetas gramaticais de cada grupo de palavras-chaves presentes – as extraídas e não extraídas como sintagmas nominais. Conclui-se que da totalidade de palavras-chaves informadas pelos autores 68% se encontravam no título ou resumo da tese ou dissertação, dessas 66% foram extraídas como sintagmas nominais, correspondendo ao nível de revocação de palavras-chaves presentes alcançado pelo software OGMA. As palavras-chaves presentes e não extraídas na grande maioria apresentavam substantivos ou adjetivos etiquetados com classe gramatical errada pelo software, e por isso não foram extraídas. As palavras-chaves presentes e extraídas eram na maioria substantivos isolados (30%), substantivos seguidos de adjetivo (28%) e substantivo seguido de preposição e substantivo (19%). O OGMA alcançou um bom nível de revocação das palavras-chaves presentes, e este nível ainda pode ser aumentado em até 34% com ajustes no etiquetador gramatical do software.
Palavras-chaves: Indexação automática; Sintagmas Nominais; Palavras-chaves; Teses e dissertações; software OGMA.

Abstract
This work investigates automatic indexing by noun phrases of documents containing title and abstract of 30 theses and dissertations written in Portuguese and of three different areas of knowledge. The research method is exploratory and based on literature review and an experiment. The experiment consisted of the OGMA software output analysis for the document corpus and the measurement of the level of recall of keywords present in the documents. It shows a descriptive profile of the sequences of grammatical labels for keywords present extracted and not extracted as noun phrases. It is concluded that 68% of the totality of keywords informed by the authors were in the title or abstract of the thesis or dissertations, of these 66% were extracted as noun phrases, which corresponds to the recall level of keywords present reached by OGMA software. Keywords present and not extracted had mainly nouns or adjectives labeled with incorrect grammatical category by the software. Keywords present and extracted were mostly single nouns (30%), noun-adjective pair (28%) and noun-preposition-noun trigram (19%). The OGMA obtained a good level of recall of keywords present, and this level can increases in almost 34% with adjustments in the part-of-speech tagger.
Keywords: Automatic indexing; Noun Phrases; Keywords; Theses and dissertations; OGMA software.

1 INTRODUÇÃO

Nos últimos 50 anos, a busca por documentos digitais tem sido realizada através da menor parte de um texto: a palavra. Métodos baseados na recuperação da informação através de expressões de busca formadas por palavras mostraram-se eficazes em alguns casos, porém ineficazes em outros. A ineficácia ocorre porque um conjunto de palavras por si só não expressa de forma completa um assunto específico, para que se recuperem documentos sobre essas palavras, pois as palavras isoladas são passiveis de interpretações polissêmicas (KURAMOTO, 2002).

As novas propostas de indexação, com grandes volumes de documentos digitais envolvem a indexação automática de documentos. A indexação automática apresenta algumas vantagens em relação à indexação intelectual: os descritores são homogêneos ou consistentes em suas regras de representação; e a rapidez na realização da indexação em grandes quantidades de documentos.

A indexação automática, segundo Robredo (1982), ocorre "na comparação de cada palavra do texto com uma relação de palavras vazias de significado, previamente estabelecidas, que conduz, por eliminação, a considerar as palavras restantes do texto como palavras significativas". Porém tal definição encontra-se não mais condizente com o atual contexto das pesquisas que envolvem indexação automática, numa definição mais atual, Corrêa e Lapa (2013) definem a indexação automática como “um conjunto de operações realizadas pelo computador, de natureza estatística, linguística, ou de programação, destinado a selecionar termos como elementos descritivos de um documento pelo processamento automático de seu conteúdo”.

No entanto, a atribuição de palavras isoladas no processo de indexação automática pode resultar em representações dúbias, devido a características intrínsecas das palavras isoladas, como a polissemia (a palavra com mais de um significado) e a sinonímia (várias palavras significando uma mesma coisa).

O uso de sintagmas nominais no processo de indexação automática tem se mostrado um promissor campo de estudo, devido ao fato de que um sintagma nominal é menos passível dessas características das palavras isoladas, por ser constituído por uma estrutura frasal, ou seja, uma sequência de palavras ligadas a um substantivo ou nome.

O sintagma nominal carrega a ideia do autor através de uma frase que contenha informações passiveis de serem entendidas e de serem recuperadas pelo usuário fora do contexto do documento. O sintagma nominal é construído de forma logico-semântica, ou seja, cada palavra obedece a uma ordem sequencial. Possui como núcleo um substantivo, sendo geralmente antecedido por um determinante e sendo sucedido por modificadores. Dependendo do sintagma nominal, outros sintagmas nominais podem ser encontrados em sua estrutura sintática, dando um aspecto mais completo ao primeiro. Por exemplo, “O estudo da economia da informação” é um sintagma nominal complexo, pois outros sintagmas nominais encontram-se embutidos nele: “o estudo”; “a economia da informação”; “a informação”.

Para Kuramoto (2002), o sintagma nominal pode ser compreendido como “a menor parte do discurso portadora de informação”, isso significa que um sintagma nominal em sua composição serve para descrever de forma fiel as intenções do autor na hora de escrever seu conhecimento em um suporte textual.

A indexação automática por sintagmas nominais de um documento ocorre em duas etapas sequenciais: na extração dos sintagmas nominais (KURAMOTO, 2002) – onde as expressões ou sequências de palavras que constituem sintagmas nominais são identificados e extraídos do texto do documento; e na seleção dos sintagmas nominais (SOUZA, 2006) (SOUZA; RAGHAVAN 2014) – onde os sintagmas nominas extraídos são pontuados quanto a importância como descritor documental,

permitindo o ordenamento e a seleção dos mais prováveis a se constituírem descritores do documento.

O sintagma nominal é extraído automaticamente, através da etiquetagem das palavras com classe gramatical e casamento da sequência de etiquetas das palavras com as regras de formação de sintagmas nominais. Tais regras variam de acordo com o idioma do texto.

Em 2002, já existiam ferramentas que possibilitavam a extração automática de sintagmas nominais para textos em português do Brasil, essas ferramentas computacionais são conhecidas como analisadores sintáticos e buscam identificar as classes gramaticais e os elementos sintáticos e semânticos que compõem cada sentença do texto. Posteriormente, outros softwares chamados extratores de sintagmas nominais foram desenvolvidos (SILVA 2014). Tais softwares realizam uma análise sintática superficial buscando extrair somente os sintagmas nominais do texto. Entretanto, atualmente para o idioma português do Brasil, somente o software OGMA é de caráter gratuito e atualmente acessível na internet (SILVA; CORRÊA 2015).

O estudo em questão aborda a aplicação da indexação automática por sintagmas nominais, em teses e dissertações armazenadas na Biblioteca Digital de Teses e Dissertações da UFPE (BDTD-UFPE) nas áreas de Direito, Computação e Nutrição.

O presente trabalho decorre a complementar o estudo realizado por Corrêa et al (2011), onde foi avaliada a indexação automática por sintagmas nominais dos documentos constituídos por título, resumo e palavras-chave de 30 teses e dissertações do BDTD-UFPE através do software OGMA, sendo analisados os sintagmas nominais extraídos quanto à corretude e a relevância com base no julgamento dos autores do artigo.

No presente estudo é utilizado o mesmo software e conjunto de documentos, porém, diferentemente, retirou-se a secção palavras-chave do processo de indexação automática, intuindo primeiramente observar quais palavras-chave indicadas pelos autores dos documentos apareciam nos títulos e resumos dos trabalhos, para então observar se as palavras-chave presentes nos títulos e resumos foram extraídas ou não como sintagmas nominais e descrever as características de cada grupo de palavras-chaves: as presentes e extraídas e as presentes e não extraídas como sintagmas nominais.

Mais precisamente, o presente trabalho avalia a revocação das palavras-chaves informadas pelos autores dos documentos no processo de indexação automática por sintagmas nominais utilizando o software OGMA.

O software OGMA (MAIA, 2008) (MAIA; SOUZA, 2010) realiza análise de texto, cálculo de similaridade de documentos, a extração dos sintagmas nominais, a identificação da classe do sintagma nominal e o cálculo da pontuação do mesmo como descritor, tudo isso automaticamente, a partir da análise das classes gramaticais das palavras e de sua frequência no texto. Para isto, faz uso de léxico da língua portuguesa construído a partir do vocabulário utilizado pelo dicionário BR.ISPELL, a fim de realizar a etiquetagem das palavras quanto à classe gramatical, e uma lista de 475 palavras irrelevantes criada tendo a gramática de Tufano como base (CORRÊA et al. 2011).

2 METODOLOGIA

O estudo em questão aborda o uso da indexação automática por meio de sintagmas nominais, semelhantes a palavras-chave contidas nos títulos e resumos de 30 teses e dissertações da UFPE, divididas igualmente em grupos correspondentes a três programas de pós-graduação, Ciência da Computação, Direito e Nutrição.

Os documentos foram tomados de um estudo anterior realizado por Corrêa et al (2011), onde foram analisadas a indexação automática por sintagmas nominais dos títulos, resumos e palavras-chave de 30 teses e dissertações do BDTD da UFPE a partir dos metadados no formato MTD-BR das primeiras teses e dissertações depositadas. Diferentemente, no presente estudo, os valores dos seguintes campos de metadados das teses e dissertações foram utilizados na seguinte ordem: título e resumo.

As palavras-chave não foram incluídas, pois avalia-se a capacidade do OGMA em extrair sintagmas nominais semelhantes as palavra-chave presentes no titulo ou no resumo, buscando assim entender o comportamento de extração da ferramenta para as palavras-chave presentes nos documentos.

Na primeira etapa, foi quantificado o número de palavras-chave definidas pelos autores em cada documento e sequencialmente foi verificado se essas mesmas palavras-chave apareciam no titulo ou resumo das teses e dissertações. No uso da ferramenta todos os três grupos de documentos passaram pelo processo de extração dos sintagmas nominais. Foi analisado principalmente a capacidade do OGMA de extrair as palavras-chave dos documentos, naturalmente caracterizadas como sintagmas nominais.

Na segunda etapa do processo, com as tabelas resultantes do processo de extração foi realizada uma análise dos padrões de sequências de etiquetas das palavras-chave presentes e extraídas, e das palavras-chave presentes e não extraídas como sintagmas nominais. Buscou-se identificar com isso, o que a ferramenta considerava ou não como sintagma nominal. Para facilitar a identificação dos documentos no corpus, estes foram denotados por sigla e um número em ordem crescente para cada um dos três grupos de documentos: Ciência da Computação (cc1, cc2, cc3, ..., cc10), Direito (d1, d2, d3, ..., d10) e Nutrição (n1, n2, n3, ..., n10).

Não foi analisado se todos os sintagmas nominais extraídos eram relevantes ou não, visto que o propósito deste trabalho é identificar o comportamento do OGMA na extração de sintagmas nominais semelhantes às palavras-chave, ou seja, o quão próximo se pode chegar da indexação intelectual via palavras-chave por meio da extração automática dos sintagmas nominais. Considerando as palavras-chave como os melhores descritores para os documentos.

No processo de extração dos sintagmas nominais pela ferramenta OGMA, buscou-se observar seu desempenho na extração das palavras-chave que estavam presentes nos documentos através da identificação delas como sintagma nominal. Foram considerados todos os sintagmas nominais que possuíam alguma das palavras-chave na forma integral.

A principal métrica utilizada para avaliar a indexação automática por sintagmas nominais neste estudo é a revocação das palavras-chaves presentes no título ou resumo das teses e dissertações. Que pode ser definida como o percentual de palavras-chaves extraídas automaticamente como sintagmas nominais, dividido pelo total de palavras-chaves presentes no título e resumo das teses e dissertações.

3 RESULTADOS

3.1 Palavras chave presentes e ausentes

Antes do processo de extração dos sintagmas nominais, foram identificadas as palavras-chave atribuídas pelos autores das teses e dissertações que estavam presentes no titulo ou resumo dos documentos. Foram criadas tabelas para quantificar as palavras-chave informadas, as que estavam presentes e as que não estavam presentes.

A Tabela 1 exibe os números totais para as palavras-chave nos três subconjuntos: Ciência da Computação, Direito e Nutrição.

Tabela 1: **Palavras-chaves presentes e palavras-chave ausentes no corpus**

Subconjunto	Palavras-chaves	Presentes	Ausentes	Percentual de Presentes	Percentual de ausentes
Computação	41	27	14	66%	34%
Direito	34	26	8	76%	24%
Nutrição	51	33	18	65%	35%
TOTAL	126	86	40	68%	32%

Fonte: os autores.

Observa-se que no subconjunto de Ciência da Computação, de um total de 41 palavras-chave informadas pelos autores apenas 27 estavam presentes no titulo ou resumo. Apenas em três casos de documentos, todas as palavras-chave descritas pelos autores estavam no titulo ou resumo dos documentos. Assim 14 palavras-chave não estavam presentes nos documentos deste subconjunto.

No subconjunto de Direito, das 34 palavras-chave indicadas pelos autores das teses e dissertações, 26 palavras-chave de fato estavam presentes nos documentos. Foi identificada uma lacuna de apenas 8 palavras-chaves não presentes, um fator positivo para o processo de indexação automática pois, o OGMA poderá extrair sintagmas nominais contendo mais palavras-chaves.

No subconjunto de Nutrição, das 51 palavras-chave atribuídas pelos autores das teses e dissertações, 33 palavras-chave ocorriam no titulo ou resumo dos documentos, em quatro casos todas as palavras-chave estavam contidas nos documentos, mas 19 palavras-chave não estavam presentes nos documentos.

Nessa primeira etapa foi contabilizado 126 palavras-chave no corpus, onde 86 estavam presentes e 40 delas estavam ausentes nos títulos e nos resumos das teses e dissertações selecionadas para o estudo. Percebe-se que a ausência das palavras-chave nos documentos faz com que parte importante dos descritores documentais não serão encontrados e extraídos automaticamente, ou seja, a extração automática dos sintagmas nominais não permitirá a revocação de todas as palavras-chave definidas pelos autores.

3.2 Extração das palavras-chaves presentes

Os resultados alcançados pela ferramenta OGMA na extração de sintagmas nominais contendo as palavras-chaves presentes nos documentos são mostrados na Tabela 2.

Tabela 2: **Palavras-chaves presentes e extraídas no corpus**

Subconjunto	Palavras-chaves	Presentes	Extraídas	Não extraídas	Percentual de Presentes Extraídas	Percentual de Presentes Não Extraídas
Computação	41	27	20	7	74%	26%
Direito	34	26	21	5	81%	19%
Nutrição	51	33	16	17	48%	52%
TOTAL	126	86	57	29	66%	34%

Fonte: os autores.

No subconjunto de Ciência da Computação, das 27 palavras-chave presentes o OGMA extraiu 20 sintagmas nominais contendo-as. Pouco menos que 1/4 das palavras-chave presentes não foram extraídas pelo OGMA. Apenas no décimo documento de Ciência da Computação (cc10), todas as palavras-chave atribuídas estavam presentes e foram também extraídas como sintagmas nominais e em cinco documentos a ferramenta extraiu todas as palavras-chave presentes.

No subconjunto de Direito, como esperado devido ao maior número de palavras-chaves presentes, a ferramenta mostrou o melhor desempenho, das 26 palavras-chave presentes no titulo ou resumo dos documentos, 21 palavras-chave foram extraídas pelo OGMA. Nos documentos de Direito (d3, d5, d6, d7, d8 e d10) todas as palavras-chave presentes foram extraídas pela ferramenta, restando apenas 5 palavras-chave que não foram extraídas nos quatro documentos restantes.

Nos documentos de Nutrição, das 33 palavras-chave presentes nos títulos e resumos dos documentos, 16 palavras-chave foram extraídas pela ferramenta. No documento (n2) todas as palavras-chave foram extraídas pelo OGMA. Em vários casos como no documento n1, metade das palavras-chave não foram consideradas sintagmas nominais pela ferramenta. Neste subconjunto, o OGMA obteve o pior desempenho em termos de revocação das palavras-chaves presentes.

A Figura 1 ilustra o resultado geral dessa etapa, das 86 palavras-chave presentes, o OGMA extraiu 57 e deixou de extrair 29.

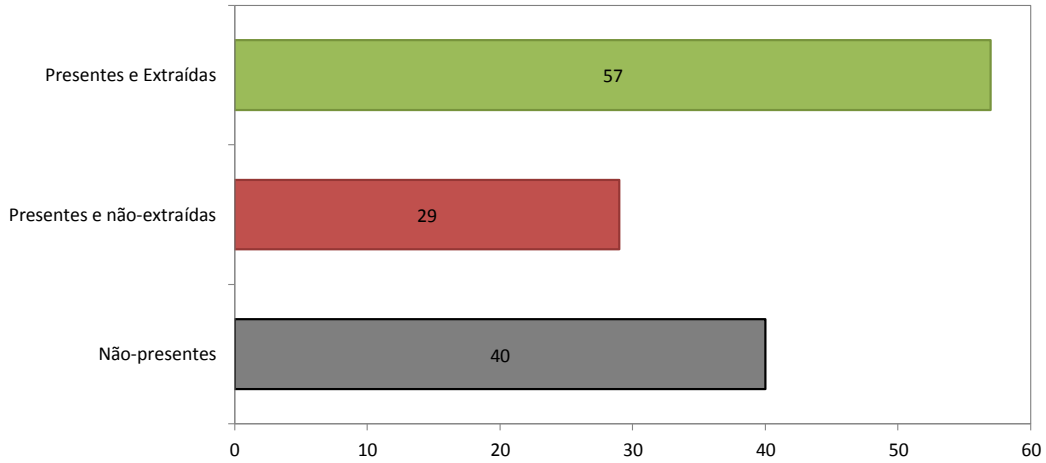

Figura 1: **Números totais das extrações das palavras-chaves**
Fonte: os autores.

Nos valores totais para o corpus percebe-se um bom desempenho da ferramenta OGMA, levando em consideração que aproximadamente 1/3 das palavras-chave atribuídas pelos autores das teses e dissertações não estavam presentes nem nos títulos e nem nos resumos dos documentos, e que das 86 palavras-chave presentes, 57, ou seja, 66% foram extraídas como sintagmas nominais.

Assim, o OGMA obteve 66% de revocação das palavras-chaves presentes. Sendo que 34% do total de palavras-chaves presentes não foram extraídas pela ferramenta, o que abre precedentes para ajustes na ferramenta. Como todas as palavras-chaves presentes são sintagmas nominais por natureza, o desempenho esperado de um software de extração de sintagmas nominais seria de 100% de revocação das palavras-chaves presentes, assim o desempenho desta ferramenta pode ser incrementado em até 34%, se após os ajustes for possível recuperar todas as palavras-chaves presentes.

Visando uma análise mais profunda do desempenho da ferramenta, fez-se necessário observar e descrever a natureza das palavras-chave que estavam presentes nos títulos e resumos e foram extraídas pelo OGMA, bem como das palavras-chave que estavam presentes nos títulos e resumos dos documentos e não foram extraídas pela ferramenta.

3.3 Natureza das palavras-chaves presentes não extraídas

Para o entendimento dos motivos pelos quais o OGMA não extraiu algumas palavras-chaves, foi necessário analisar o padrão de etiquetas das palavras-chaves que estavam presentes e não foram extraídas. No corpus total de documentos, 27 palavras-chaves que estavam presentes não foram consideradas sintagmas nominais, analisou-se a estrutura das palavras-chave que não foram consideradas sintagmas nominais e a frequência da ocorrência dos padrões de etiquetas.

Para a descrição das estruturas gramaticais das palavras-chaves extraídas ou não extraídas como sintagmas nominais, são utilizadas as siglas das classes gramaticais atribuídas pelo OGMA na etiquetagem das palavras. Por exemplo, AD para adjetivo, SU para substantivo e assim por diante, como exibido no Quadro 1.

Quadro 1: **Siglas das etiquetas atribuídas às palavras pelo OGMA**

ETIQUETA OGMA	CLASSE GRAMATICAL
AD	Artigo definido
AI	Artigo indefinido
AJ	Adjetivo
AV	Advérbio
CJ	Conjunção (Aditiva, adversativa, alternativa)
IT	Interjeição
NC	Números cardinais
NM	Números multiplicativos
NO	Números ordinais
NP	Nome próprio
NR	Número romano
PS	Pronome possessivo
PD	Pronome demonstrativo
PI	Pronome indefinido
PL	Pronome relativo
PN	Pontuação (exceto vírgula)
PP	Pronome pessoal
PR	Preposições
SU	Substantivo
VB	Verbos
VG	Vírgulas
VP	Verbos no Particípio

Fonte: Adaptado de (Maia 2008).

As regras de formação de um sintagma nominal apresentam um substantivo como núcleo, seguido de algum adjetivo, ou sintagma preposicionado contendo substantivo. As combinações mais comuns na formação de sintagmas nominais são: SU; SU+AJ; e SU+PR+SU.

Os quadros a seguir mostram as palavras-chave presentes e não extraídas e sua etiquetagem.

Quadro 2: **Palavras-chaves presentes não extraídas no subconjunto de Ciência da Computação**

Palavras-chaves não extraídas etiquetadas (Computação)
sistemas/SU embarcados/SU
simulação/SU single-pass/SU
ambiente/VBSU www/SU
educação/SU à/PR distância/SU
ambiente/VBSU híbrido/SU
previsão/SU de/PR acordes/VB musicais/VBAJ
tempo/SU real/SU

Fonte: os autores.

No Quadro 2 observa-se erros de etiquetagem do OGMA. No documento cc1, a palavra-chave "sistemas/SU embarcados/SU" foi erroneamente etiquetada, a ferramenta caracterizou a palavra "embarcados" como substantivo, quando na verdade se trata nesse caso de um adjetivo. Ocorreu etiquetagem incorreta também nas palavras-chaves "ambiente/VBSU híbrido/SU" e "previsão/SU de/PR acordes/VB musicais/VBAJ", onde a ferramenta etiquetou a palavra "híbrido" como substantivo ao invés de adjetivo e "acordes" como sendo uma flexão do verbo "acordar" quando na verdade se tratava de um substantivo. Tais erros de etiquetagem influenciam a não extração das palavras-chave presentes como sintagmas nominais.

Para as outras palavras-chaves a etiquetagem está correta, mas em três casos, a estrutura do sintagma nominal é SU+SU, se tratando de um nome próprio, o que não é previsto nas regras na formação de sintagmas nominais implementadas pelo OGMA. Em um caso envolvendo "educação à distância", a etiquetagem está correta, mas o sintagma é extraído incorretamente como "www de educação à distância", o que revela um problema conhecido na área de processamento de linguagem natural como desambiguação de sintagma preposicionado (do inglês *prepositional phrase attachment*).

Quadro 3: **Palavras-chaves presentes não extraídas no subconjunto de Direito**

Palavras-chaves não extraídas etiquetadas (Direito)
capacidade/SU contributiva/SU
tratados/AJ internacionais/AJ
entidades/SU familiares/VB
tutela/VB jurisdicional/AJ
mandado/VPAJ de/PR segurança/SU

Fonte: os autores.

No subconjunto de Direito houve um número de palavras etiquetadas incorretamente, cinco no total, correspondendo aos cinco casos de palavras-chaves presentes não extraídas. Por exemplo, no documento d2 a palavra-chave "tratados internacionais" foi etiquetada como se houvessem dois adjetivos seguidos AJ+AJ, porém a palavra "tratados" nesse caso se trata de um substantivo, sendo o padrão de etiquetagem correto SU+AJ, se constituindo em um sintagma nominal. Erros na etiquetagem de adjetivos e substantivos são as causas principais da não revocação das palavras-chaves deste grupo.

Quadro 4: **Palavras-chave presentes não extraídas no subconjunto de Nutrição**

Palavras-chaves não extraídas etiquetadas (Nutrição)
hipertensão/SU arterial/AJ sistêmica/SU
idosos/AJ
idoso/AJ
cálcio/SU sérico/SU
controle/VB de/PR qualidade/SU laboratorial/AJ
deficiência/SU de/PR vitamina/SU A/AD
consumo/VB alimentar/VB
consumo/VB alimentar/VB
depressão/SU alastrante/SU
ratos/AJ
ratos/AJ jovens/AJ e/CJ adultos/AJ
sistemas/SU serotoninérgico/SU e/CJ nitrérgico/SU
origanum/SU vulgare/SU
índice/SU glicêmico/SU
pão/SU light/SU
vitamina/SU A/AD
efeito/SU prebiótico/SU

Fonte: os autores.

No subconjunto de Nutrição a ferramenta OGMA obteve o maior número de etiquetagens incorretas, 17 no total. A etiquetagem das palavras-chaves compostas por sucessão de substantivos "SU+SU" inviabilizou o encontro de sete palavras-chaves que estavam presentes no titulo ou resumo e não foram extraídas, diminuindo a capacidade de representação automática da informação por parte do OGMA. Destas sete, seis foram causadas por erros de etiquetagem das palavras e uma estava corretamente etiquetada como "SU+SU", se constituindo num nome próprio: origanum vulgare. Substantivos isolados como "idoso", "idosos" e "ratos" foram incorretamente categorizados como adjetivos em três casos. O padrão "VB+VB" ocorreu incorretamente duas vezes na etiquetagem da palavra-chave "consumo alimentar". Em outros dois casos, a etiquetagem de "A" como adjetivo em "Vitamina A", somente será possível com regras específicas de tratamento de letras isoladas. Nos demais casos, as palavras-chaves não são extraídas por problemas de etiquetagem em uma palavra.

No Quadro 5 apresenta-se o padrão de etiquetagem e a frequência de ocorrência para palavras-chaves presentes e não extraídas em todo o corpus.

Quadro 5: **Padrão de etiquetagem das palavras-chaves presentes não extraídas**

Padrão de Etiquetagem	**Frequência**
SU+SU	13
AJ	3
VB+VB	2
SU+AD	2
VB+AJ	2
AJ+AJ	2
SU+VB	1
SU+AJ+SU	1
VB+PR+SU	1
SU+PR+SN	1
VPAJ+PR+SU	1
TOTAL	29

Fonte: os autores.

Levando em conta o padrão de etiquetagem das 29 palavras-chave não extraídas como sintagmas nominais pelo OGMA, percebeu-se que o motivo da não extração foi em grande parte causada por erros de etiquetagem gramatical das palavras que compõem o sintagma nominal. O erro de etiquetagem mais frequente foi categorização como substantivo de palavras que na verdade seriam adjetivos (10 ocorrências), seguido da atribuição da categoria verbo a palavras que seriam substantivos (5 ocorrências).

3.4 Natureza das palavras-chaves presentes extraídas

Os quadros a seguir mostram as palavras-chaves presentes extraídas e seu respectivo padrão de etiquetagem. Tais palavras-chaves foram consideradas sintagmas nominais por seguirem as regras de formação dos sintagmas nominais implementadas pela ferramenta OGMA.

Quadro 6: **Palavras-chaves presentes e extraídas no subconjunto de Ciência da Computação**

Palavras-chaves extraídas etiquetadas (Computação)
hierarquias/SU de/PR memória/SU
análise/SU de/PR dados/AJSU
casca/SU convexa/AJ
dados/AJSU simbólicos/AJ
introspecção/SU
systemc/SU
peer-to-peer/SU
compartilhamento/SU de/PR arquivo/VBSU
corba/SU
engenho/SU de/PR busca/VBSU
crawling/SU e/CJ indexação/SU
escalabilidade/SU
chaveamento/SU por/PR pacotes/SU
cenários/SU de/PR codificação/SU
redes/SU neurais/AJ
base/SU de/PR casos/AVSU
qualidade/SU de/PR serviço/SU
serviços/SU diferenciados/AJ
controlador/SU de/PR banda/AJSU
roteamento/SU dinâmico/AJ

Fonte: os autores.

De acordo com o Quadro 6, no subconjunto de Ciência da Computação, cinco palavras-chave formadas unicamente por um substantivo "SU" foram extraídas pelo OGMA. Nota-se que dois termos estão em inglês e foram considerados pela ferramenta corretamente como substantivos. Outras cinco palavras são substantivos seguidos de adjetivo, tendo o padrão SU+AJ de etiquetas, como em "casca/SU convexa/AJ" e, nove palavras-chave seguem o padrão SU+PR+SU (substantivos ligados por preposição). Ambos os padrões de etiquetagem para termos compostos indicam no geral sintagmas nominais considerados bons descritores.

Quadro 7: **Palavras-chave presentes e extraídas do subconjunto de Direito**

Palavras-chaves extraídas etiquetadas (Direito)
contribuições/SU previdenciárias/AJ
mínimo/AJ existencial/SU
direitos/AJSU fundamentais/VBAJ de/PR o/AD contribuinte/SU
direitos/AJSU humanos/AJ
vontade/SU
dignidade/SU humana/AJ
filosofia/SU
hermenêutica/SU filosófica/AJ
verdade-causalidade (verdade/SU e/CJ causalidade/SU)
cláusulas/SU gerais/VBAJ
dignidade/SU de/PR a/AD pessoa/SU humana/AJ
personalidade/SU
Reintegração/SU judicial/AJ em/PR o/AD emprego/VBSU
Efetividade/SU de/PR o/AD modelo/VBSU processual/AJ
justiça/SU criminal/AJ
satisfação/SU de/PR a/AD vítima/SU
duplo/NM grau/SU obrigatório/AJ
isonomia/SU
direito/AJSU natural/AJ
direitos/AJSU de/PR o/AD homem/SU
constituições/SU hodiernas/AJ

Fonte: os autores.

No subconjunto de Direito grande parte das palavras-chave extraídas pelo OGMA foram etiquetadas com substantivo seguido de adjetivo (padrão de etiquetas SU+AJ), totalizando oito palavras-chave. Substantivos isolados (padrão de etiqueta SU) constituíram quatro sintagmas nominais extraídos. O padrão substantivo seguido de preposição, artigo e outro substantivo (SU+PR+AD+SU) como em "satisfação/SU de/PR a/AD vítima/SU", ocorreram duas vezes. No caso da palavra-chave "verdade-casualidade" a ferramenta encontrou o termo "verdade/SU e/CJ causalidade/SU" como sintagma nominal.

Quadro 8: **Palavras-chave presentes e extraídas no corpus de Nutrição**

Palavras chave ocorrentes etiquetadas (Nutrição)
pressão/SU arterial/AJ
prevalência/SU
índice/SU de/PR massa/SU corporal/AJ
anemia/SU
diagnóstico/SU
cinética/AJ microbiana/SU
métodos/SU de/PR análise/SU
escolares/SU
crianças/SU menores/AJSU de/PR cinco/NC anos/SU
crescimento/SU somático/AJ
desnutrição/SU
óleo/SU essencial/AJ
leveduras/SU
inulina/SU
flocos/SU de/PR abóbora/SU
inulina/SU

Fonte: os autores.

No subconjunto de Nutrição, como pode ser observado no Quadro 8, substantivos isolados representam oito das 16 palavras-chave presentes extraídas pela ferramenta, 50% do total, dando-se a perceber uma certa tendência por palavras isoladas atribuídas pelos autores nos trabalhos de Nutrição. Três palavras-chave extraídas tinham como padrão substantivo seguido de adjetivo (SU+AJ), e duas palavras-chaves possuíam o padrão SU+PR+SU.

O quadro 9 exibe as frequências totais para cada padrão de etiquetagem das palavras-chaves presentes e extraídas como sintagmas nominais pelo OGMA.

Quadro 9: **Padrão de etiquetagem das palavras-chave encontradas**

Padrão de Etiquetagem	**Freq.**
SU	17
SU+AJ	16
SU+PR+SU	11
SU+PR+AD+SU	2
SU+CJ+SU	2
AJ+SU	2
SU+PR+AD+SU+AJ	2
SU+PR+SU+AJ	1
SU+AJ+PR+AD+SU	1
NM+SU+AJ	1
SU+AJ+PR+NC+SU	1
TOTAL	57

Fonte: os autores.

Entre as palavras-chaves presentes e extraídas mais comuns estão os substantivos isolados, termos únicos que servem de descritor documental e que foram extraídos como sintagmas nominais. Outro padrão de etiquetas muito comum nas palavras-chaves presentes e extraídas como sintagmas nominais é o padrão substantivo seguido por adjetivo que especifica mais o termo ou descritor documental. Sabendo-se que os substantivos indicam coisas e nomeiam, e que os adjetivos qualificam esses conceitos, os padrões SU+AJ e SU+PR+SU norteiam bem o perfil dos descritores que permitem uma recuperação da informação com maior precisão.

4 CONSIDERAÇÕES FINAIS

O uso dos sintagmas nominais para a indexação automática e recuperação da informação veem sendo cada vez mais pesquisado, principalmente motivado pelo fato dos sintagmas nominais possuírem semântica mais bem definida que as palavras isoladas. Outro fato motivador é que os sintagmas nominais se constituem em boa fonte de palavras-chaves para a descrição documental, tendo em vista que qualquer palavra-chave por natureza se constitui em um sintagma nominal.

Para que as pesquisas sobre esta temática avancem, é de fundamental importância analisar e avaliar como os softwares realizam esse processo de extração de descritores em documentos eletrônicos, já que os descritores constituem ponto de acesso aos documentos e consequentemente definem os meios pelos quais os documentos podem ser encontrados.

Esse trabalho contribui para as pesquisas sobre indexação automática por sintagmas nominais, por meio da analise da capacidade do software OGMA em indexar automaticamente teses e dissertações da UFPE para fins de recuperação da informação, com foco na revocação das palavras-chaves presentes no texto dos documentos, compostos pelo título e resumo dos trabalhos.

Da totalidade de palavras-chaves informadas pelos autores, 68% se encontravam no título ou resumo da tese ou dissertação, destas 66% foram extraídas como sintagmas nominais, correspondendo ao nível de revocação de palavras-chaves presentes alcançado pelo software OGMA.

As palavras-chaves presentes e não extraídas na grande maioria apresentavam substantivos ou adjetivos etiquetados com classe gramatical errada pelo software, e por isso não foram extraídas. As palavras-chaves presentes e extraídas eram na maioria substantivos isolados (30%), substantivos seguidos de adjetivo (28%) e substantivo seguido de preposição e substantivo (19%).

Conclui-se que o OGMA obteve um bom desempenho na revocação de palavras-chaves presentes no título e resumo de teses e dissertações, extraindo cerca de 66% das palavras-chave presentes no texto dos documentos. E que para se alcançar uma maior revocação, que pode ser incrementada em até 34%, torna-se necessária a expansão da lista de palavras associadas às respectivas categorias gramaticais utilizada pelo módulo etiquetador do OGMA, afim de permitir a correta categorização de palavras ausentes e evitar a etiquetação de forma errônea de palavras presentes.

Como trabalhos futuros, vislumbra-se: a análise do impacto do uso de diferentes etiquetadores gramaticais para o português do Brasil na extração e revocação das palavras-chaves presentes nos documentos; e a investigação de métodos para seleção ou ranqueamento de sintagmas nominais quanto ao valor como descritor, permitindo a mensuração do nível de precisão alcançado na indexação automática por sintagmas nominais.

AGRADECIMENTOS

Os autores agradecem o fomento da Fundação de Amparo à Ciência e Tecnologia de Pernambuco (FACEPE) ao projeto intitulado "Mapeador Temático de Teses e Dissertações".

REFERÊNCIAS

CORRÊA, Renato Fernandes; LAPA, Remi Corrêa; Panorama de estudos sobre indexação automática no âmbito da ciência da informação no Brasil (1973-2012). **Ciência da Informação**, Brasília, v. 42, n. 2, p.255-273, 2013.

CORRÊA, Renato Fernandes; MIRANDA, Darliane Goes de; LIMA, Camila Oliveira de Almeida; SILVA, Tiago José da. Indexação e recuperação de teses e dissertações por meio de sintagmas nominais. **AtoZ: Novas Práticas em Informação e Conhecimento**, Curitiba, v. 1, n. 1, p. 11-22, 2011.

KURAMOTO, Hélio. Sintagmas nominais: uma nova proposta para a recuperação de informação. **DataGramaZero**, Rio de Janeiro, v. 3, n. 1, fev. 2002.

MAIA, Luiz Cláudio Gomes. **Uso de sintagmas nominais na classificação automática de documentos eletrônicos**. 2008. Tese (Doutorado em Ciência da Informação) – Universidade Federal de Minas Gerais, Belo Horizonte-MG, 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, 2010.

ROBREDO, Jaime. Otimização dos processos de indexação dos documentos e de recuperação da informação mediante o uso de instrumentos de controle terminológico. **Ciência da Informação**, Brasília, v. 11, n. 1, 1982.

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

SILVA, T. J. da; CORRÊA, R. F. Ferramentas Para Indexação Automática: uma análise comparativa entre o OGMA, Parser PALAVRAS, LX-Parser e a extração manual de sintagmas nominais. 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.

SOUZA, Renato Rocha; RAGHAVAN, K. S. A extração de palavras-chave a partir de textos: um estudo exploratório utilizando sintagmas. **Informação & Tecnologia (ITEC)**: Marília / João Pessoa, v.1, n.1, p.5-16, 2014.

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