UNIVERSIDADE FEDERAL DE PERNAMBUCO
CENTRO DE ARTES E COMUNICAÇÃO
DEPARTAMENTO DE CIÊNCIA DA INFORMAÇÃO
GRADUAÇÃO EM GESTÃO DA INFORMAÇÃO

NATHALLY STEPHANIE DE MELO TORRES

AVALIAÇÃO DE SISTEMAS DE INDEXAÇÃO AUTOMÁTICA APLICADOS A ARTIGOS CIENTÍFICOS NA ÁREA DE CIÊNCIA DA INFORMAÇÃO

RECIFE
2017

Nathally Stéphanie de Melo Torres

AVALIAÇÃO DE SISTEMAS DE INDEXAÇÃO AUTOMÁTICA APLICADOS A ARTIGOS CIENTÍFICOS NA ÁREA DE CIÊNCIA DA INFORMAÇÃO

Trabalho de Conclusão de Curso apresentado ao Curso de Gestão da Informação do Departamento de Ciência da Informação da Universidade Federal de Pernambuco como requisito parcial para obtenção do grau de Bacharel em Gestão da Informação.

Orientador: Renato Corrêa

RECIFE
2017

Catalogação na fonte
Bibliotecário Jonas Lucas Vieira, CRB4-1204

T693a Torres, Nathally Stéphanie de Melo
Avaliação de sistemas de indexação automática aplicados a artigos científicos na área de Ciência da Informação / Nathally Stéphanie de Melo Torres. – Recife, 2017.
49 f.: il., fig.

Orientador: Renato Fernandes Corrêa.
Trabalho de Conclusão de Curso (Graduação) – Universidade Federal de Pernambuco. Centro de Artes e Comunicação. Ciência da Informação, 2017.

Inclui referências.

1. Avaliação de sistemas. 2. Ciência da Informação. 3. Indexação automática. 4. Sistemas de indexação automática. I. Corrêa, Renato Fernandes (Orientador). II. Título.

020 CDD (22.ed.) UFPE (CAC 2017-252)

Serviço Público Federal
Universidade Federal de Pernambuco
Centro de Artes e Comunicação
Departamento de Ciência da Informação

FOLHA DE APROVAÇÃO

Título do TCC

AVALIAÇÃO DE SISTEMAS DE INDEXAÇÃO AUTOMÁTICA APLICADOS A ARTIGOS CIENTÍFICOS NA ÁREA DE CIÊNCIA DA INFORMAÇÃO

Nathally Stephanie de Melo Torres
(Autor)

Trabalho de Conclusão de Curso submetido à Banca Examinadora, apresentado no Curso de Biblioteconomia, do Departamento de Ciência da Informação, da Universidade Federal de Pernambuco, como requisito parcial para obtenção do título de Bacharel em Gestão da Informação.

TCC aprovado 07 de dezembro de 2017

Banca Examinadora:

Renato Fernandes Corrêa
Orientador – Renato Fernandes Corrêa
DCI/Universidade Federal de Pernambuco

André Anderson C. Felipe
Examinador 1 – André Anderson Cavalcante Felipe
DCI/Universidade Federal de Pernambuco

Márcio Henrique W. Ferreira
Examinador 2 - Márcio Henrique Wanderley Ferreira
PPGCI/Universidade Federal de Pernambuco

AGRADECIMENTOS

O mais bonito de concluir uma etapa é ver que todo caminho percorrido para chegar até ela valeu a pena.

Agradeço a Deus por ter me dado os subsídios necessários para chegar até aqui.

Aos meus pais e minha irmã por ser minha base e por toda força e apoio durante a graduação.

Ao meu noivo João Henrique por estar sempre presente nos momentos bons, ruins e não me deixar desistir dos meus sonhos.

Agradeço a minha turma 2013.2 por todo companheirismo, especialmente Amanda, Brenna, Rosanne, Loredana e Sthéfanie.

Agradeço ao Professor Dr. Renato Corrêa, por aceitar ser meu orientador, pela paciência e pelos projetos que conquistamos desde o começo do curso.

RESUMO

Avalia sistemas de indexação automática em um corpus em português de artigos científicos na área de Ciência da Informação. Como objetivos específicos: investiga a avaliação de software que realize a indexação automática; constrói um corpus em português de artigos científicos na área de Ciência da Informação; propõe método de avaliação de sistemas de indexação automática associado ao corpus criado; avalia comparativamente sistemas de indexação automática presentes na literatura científica quanto à qualidade na indexação automática do corpus proposto. A pesquisa tem caráter exploratório quanto aos objetivos, e quanto aos métodos a pesquisa tem caráter bibliográfico e experimental. O experimento computacional consistiu em analisar comparativamente a consistência, precisão, revocação e medida F obtidas na indexação automática do corpus pelos softwares SISA e OGMA. A partir da comparação dos resultados obtidos pelos sistemas de indexação automática avaliados, pode-se perceber que o SISA se sobressaiu ao OGMA em consistência, precisão e medida F, já o OGMA apresentou melhor revocação. Conclui-se que o SISA é o melhor sistema de indexação automática, não descartando a necessidade de aperfeiçoamento de ambos os sistemas.

Palavras-chave: Avaliação de sistemas; Ciência da informação; Indexação automática; Sistemas de indexação automática.

ABSTRACT

This work evaluates automatic indexing systems in a Portuguese corpus of scientific articles of the area of Information Science. As specific objectives: investigates the evaluation of software that performs automatic indexing; constructs a corpus in Portuguese of scientific articles in the area of Information Science; proposes method of evaluation of automatic indexing systems associated with the created corpus; evaluates comparatively automatic indexing systems present in the scientific literature regarding the quality in the automatic indexing of the proposed corpus. The research has an exploratory character regarding the objectives, and the methods the research has bibliographic and experimental character. The computational experiment consisted of comparing the consistency, precision, recall and f-measure obtained in the automatic indexing of the corpus by the softwares SISA and OGMA. From the comparison of the results obtained by the automatic indexing systems evaluated, it can be seen that the SISA is superior to OGMA in consistency, precision and f-measure, beside OGMA has better recall than SISA. It is concluded that SISA is the best automatic indexing system, not ruling out the need for improvements of both systems.

Keywords: Evaluation of systems; Information Science; Automatic indexing; Automatic indexing systems.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE QUADROS

SUMÁRIO

1. INTRODUÇÃO

Os atuais sistemas de publicação científica eletrônica dão suporte ao funcionamento das bibliotecas digitais de teses e dissertações, dos repositórios institucionais, dos periódicos eletrônicos de acesso aberto, bem como de importantes bases de dados de artigos de periódicos como a Scielo e Redayc. Na arquitetura da maioria dos sistemas deste tipo está presente o modelo de arquivos abertos (Open Archives), o que torna possível a construção de agregadores ou bases de dados que reúnam e facilitem o acesso às publicações científicas de diversas tipologias documentais e áreas de conhecimento.

A integração destas fontes de conhecimento é essencial para o desenvolvimento nacional das pesquisas em qualquer área do conhecimento. Entretanto, a simples agregação de todas elas, envolvendo todas as áreas do conhecimento, sem o devido tratamento da indexação dos assuntos abordados em cada documento (seja tese, dissertação, publicação em evento ou artigo de periódico), intensifica o problema da sobrecarga de informação sobre os usuários, causado pelos fenômenos linguísticos da variação morfológica, polissemia, sinonímia, homonímia e ambiguidade das palavras.

Assim, para que as potencialidades de uma base centralizada de publicações científicas em determinada área do conhecimento sejam efetivamente exploradas, um Sistema de Recuperação de Informação eficaz deve ser disponibilizado. Um Sistema de Recuperação de Informação (SRI), segundo (BAEZA-YATES e RIBEIRO-NETO 2011), é um software que trata essencialmente de indexação, busca e classificação de documentos (textuais), com o objetivo de satisfazer necessidades de informação dos usuários, geralmente expressa através de consultas composta por palavras ou expressões lógicas envolvendo as mesmas.

No momento da busca dos usuários, a grande maioria dos Sistemas de Recuperação de Informação recupera documentos utilizando palavras isoladas como ponto de acesso a documentos. Esta metodologia de busca apresenta sérias limitações já que os SRIs na sua grande maioria trabalham apenas no nível léxico e não dispõem de meios para identificar o sentido de cada palavra isolada na consulta do usuário e nos documentos, levando a uma baixa precisão e excesso de documentos retornados.

Embora, interfaces mais amigáveis possam auxiliar os usuários de um SRI (CORRÊA, R. F.; VIEIRA 2013), uma maneira de tornar o SRI mais eficaz é dotá-lo de métodos mais sofisticados de indexação automática baseados na extração de sintagmas nominais (MIORELLI 2001) (SOUZA 2006) (SOUZA; RAGHAVAN, 2014) (KURAMOTO 2006) (MAIA 2008) (MAIA; SOUZA 2010) (CORREA et al. 2011); e baseado na atribuição de descritores de vocabulários controlados (NARUKAWA; GIL LEIVA; FUJITA, 2009). Tais métodos permitem a atribuição automática de termos compostos aos documentos de acordo com o conteúdo dos mesmos, sendo tais termos, quando utilizados como ponto de acesso aos documentos, capazes de descrever com maior exatidão os assuntos e permitirem uma maior precisão na recuperação da informação, superando em qualidade a representação de assunto por palavras isoladas.

Entretanto, embora existam semelhanças nos métodos de indexação apontados, tais métodos não têm sido avaliados simultaneamente num único corpus.

O presente trabalho visa investigar a avaliação de sistemas de indexação automática, bem como avaliar sistemas de indexação automática através de um corpus em português de artigos científicos na área de Ciência da Informação.

Como objetivos específicos, tem-se: investigar a avaliação de software que realize a indexação automática; construir um corpus em português de artigos científicos na área de Ciência da Informação; propor método de avaliação de sistemas de indexação automática associado ao corpus criado; avaliar comparativamente sistemas de indexação automática presentes na literatura científica através do corpus de artigos científicos na área de Ciência da Informação, realizando um estudo comparativo da eficácia dos sistemas de indexação automática presentes na literatura científica, através do recurso e método de avaliação proposto.

2. FUNDAMENTAÇÃO TEÓRICA

Neste capítulo, são apresentados os marcos teóricos necessários para o entendimento dos experimentos propostos neste trabalho.

2.1 Indexação

A recuperação da informação é uma área da Ciência da informação que lida com o armazenamento e a recuperação automática da informação, facilitando o acesso a documentos importantes de acordo com a necessidade do usuário.

Para tornar esse processo mais ágil e eficaz, é preciso que os termos de indexação dos documentos estejam de acordo com o conteúdo dos mesmos.

Segundo Fujita (2009), o termo indexação (do inglês *indexing*) pertence à corrente teórica inglesa, onde este pode ser conceituado como a etapa da representação temática que tem o objetivo de reportar ao conteúdo do documento de modo que possa ser recuperado quando for solicitado em outro momento. Neste sentido, a função da indexação é representar o assunto dos documentos através da elaboração de termos, mostrando “ao pesquisador e ao leitor de uma forma geral, que tópicos, fatos ou outros itens de informação estão tratados nos documentos indexados” GASPAR (2011 *apud* CORREA; LAPA, 2013).

Ainda em termos de definição, o processo de indexação conceitua-se como a realização de uma atividade que consiste em selecionar e definir termos ou expressões que serão usados para descrever (descritores) o conteúdo de um documento. Na seleção e definição destes termos ou descritores devem-se considerar também os possíveis usuários que necessitam acessar as informações contidas nestes documentos realizando o acesso e a busca através de Sistemas de Recuperação da Informação (BORGES, 2009).

Para Souza (2005) há duas fases independentes na indexação. Fujita (2013) corrobora com Souza, ao dizer que basicamente

> “existem dois momentos ou etapas no ato de indexar. Um primeiro momento na análise do assunto, quando se identifica e se faz a seleção de conceitos via associação a conceitos universais, como ação, objeto que sofreu a ação e o agente que praticou a ação. No segundo momento, quando os termos são traduzidos segundo os termos de uma linguagem documental e, compatibilizando ou

> adequando os termos indexados com os termos que serão úteis ao atendimento das necessidades informacionais dos usuários.”

Segundo Lancaster (2004), existe três dimensões de indexação: a exaustividade que está relacionada ao adicionar mais termos a indexação; a seletiva que se refere ao processo de quando menos termos são incluídos; e a especificidade que é referida a tarefa de se usar termos mais específicos que façam com que o documento seja compreendido integralmente.

Pode-se dizer que existem três formas de se fazer a indexação de documentos, a primeira é a manual, caracterizada por ser desenvolvida pelo homem, a segunda é a automática, onde se realiza o processo de indexação através de uma máquina, e a terceira é a semiautomática que consiste na utilização das duas primeiras.

A indexação manual, é realizada pelos humanos, sejam eles profissionais da informação ou especialistas do domínio das bases de dados. Este tipo de indexação se baseia, sobretudo, no julgamento, normalmente intuitivo, dos indexadores, em função do texto e do interesse para a sua comunidade de usuários. Para realizar essa indexação, é preciso, inicialmente, analisar o conteúdo do documento, lendo-o não do inicio ao fim, mas por partes, ou seja, lendo suas estruturas lógicas (PINTO, 2000). A indexação manual requer um esforço intelectual muito intenso do indexador que, por sua vez, acaba deixando uma carga de subjetividade na escolha dos termos e em todo processo de indexação de um documento.

Na seção 2.2 serão apresentados os fundamentos teóricos necessários para o entendimento acerca da indexação automática.

2.2 Indexação Automática

De acordo com Araújo Júnior (2007), a indexação automática é qualquer procedimento realizado por computador que permita identificar e selecionar os termos que representam o assunto dos documentos sem a intervenção direta do homem.

Lapa *et al* (2014), definem a indexação automática como um conjunto de operações, basicamente matemáticas, linguísticas, de programação, destinadas a

selecionar termos como elementos descritivos de um documento pelo processamento de seu conteúdo.

O processo de indexação automática é similar ao processo de leitura-memorização humano, sendo seu princípio geral baseado na comparação de cada palavra do texto com uma relação de palavras vazias de significado, previamente estabelecida, que conduz, por eliminação, a considerar as palavras restantes do texto como palavras significativas. (ROBREDO 1982 *apud* BORGES, 2009).

É importante ressaltar que, a indexação automática atende as necessidades de comunicação, disponibilização e acesso rápido a produção científica via uso do computador no processamento de dados e informações (ROBREDO, 2005 *apud* BANDIM, 2017).

Segundo Silva (2014) a indexação automática pode ser tomada como o tratamento prévio que os documentos devem passar para serem armazenados em uma base de dados, no intuito de que cada documento possa ser posteriormente recuperado.

A indexação automática extrai os descritores para representar os documentos a partir da análise do texto pelos computadores, dotados de programas elaborados pela área de Inteligência Artificial ou Inteligência Computacional, no escopo da Ciência da Computação, e pela área de Organização da Informação, no escopo da Ciência da Informação (SILVA, 2014).

Desta forma, segundo Gil Leiva (1997), a inclusão do computador na realização da indexação buscou: tornar mais ágil o processo da análise de informação, a obtenção de melhores índices de consistência, a redução dos custos, e uma maior qualidade nos sistemas de informação.

Lancaster (2004) afirma que é possível citar alguns tipos de indexação automática, são estes:

- ✓ Indexação por extração automática: de acordo com Lancaster (2004) é o procedimento remete aos mesmos princípios utilizados pela extração manual, pois pode ser feito observando a frequência da palavra dentro do texto, posição dessa palavra no texto (no título, nas palavras-chave, no resumo, etc.) e seu contexto.

- ✓ Indexação por atribuição automática: trabalha da mesma lógica que a indexação por extração automática, contudo faz uso de um controle terminológico. Assim, desenvolve-se previamente para cada termo a ser atribuído um perfil de palavras ou expressões possivelmente presentes nos documentos. Tais perfis condicionam a atribuição de cada termo ao documento.

Ambos os tipos de indexação automática serão discutidos, em detalhes, nas próximas subseções.

2.2.1 Indexação Automática Por Atribuição

Segundo Lancaster (2004), a maior parte da indexação realizada por seres humanos é a indexação por atribuição, utilizando um vocabulário controlado como ferramenta normalizadora. O vocabulário controlado é especificamente uma lista de termos autorizados, que pode também incluir uma estrutura semântica, especialmente criada para controlar sinônimos optando por um termo padrão, diferenciar homógrafos, reunir ou ligar termos que apresentem o mesmo significado.

Os vocabulários controlados podem ser disponibilizados para os usuários de um sistema de informação, permitindo que tenham acesso à terminologia empregada na indexação dos documentos. Isto possibilita compatibilizar a linguagem dos usuários à linguagem utilizada na representação dos documentos, resultando em uma recuperação mais eficiente.

Nicolino (2014) atesta que a indexação por atribuição automática é realizada por meio da comparação entre termos extraídos dos textos de um *corpus* e um vocabulário do domínio. Portanto, é necessário existir uma coincidência entre os termos extraídos de um documento e os termos do vocabulário controlado.

No processo de indexação automática por atribuição, os vocabulários controlados atuam no próprio processo de análise automática do documento e na representação, ou seja, condicionam os resultados na atribuição de descritores, assim, há sempre a necessidade de se levar em consideração todos os seus atributos, além da necessidade de considerar a sua atuação associada aos métodos de indexação automática pelos quais se realiza a análise do conteúdo temático dos documentos (NARUKAWA, 2009).

De acordo com Bruzinga (*et al*, 2007), o processo de indexação automática por atribuição é mais complexo de ser realizado com maior eficiência que o processo de indexação por extração automática. Pode ser considerada uma atividade difícil, pois para a representação do conteúdo temático é necessário um controle terminológico. Deve-se desenvolver, para cada termo atribuído, um 'perfil' de palavras ou expressões que costumam ocorrer nos documentos.

Sendo assim, a indexação automática por atribuição se dá em duas fazes: em uma primeira etapa extraem-se palavras ou expressões do texto por meio de técnicas estatísticas. Em uma segunda fase, partindo desse conjunto de palavras/expressões, seleciona-se no vocabulário controlado o termo cujo perfil possui certo nível de coincidência (LANCASTER, 2004).

Uma metodologia de indexação automática por atribuição é utilizada no *Sistema de Indización Semi Automatica* (SISA), no qual se efetua a indexação por comparação entre o documento - constituído por título, resumo e texto - e um vocabulário controlado, partindo de critérios de frequência preestabelecidos pelo sistema para propor os termos de indexação.

2.2.2 Indexação Automática Por Sintagmas Nominais

A Palavra, apesar de ter sido a primeira unidade base para a indexação automática, aos poucos foi se mostrando ineficiente para os fins de representação e recuperação da informação. Isto se deve ao fato de que, as palavras sofrem problemas semânticos relacionados a polissemia, quando um significante tem vários significados, e sinonímia, quando o mesmo significado pode ser representado por vários significantes. Isso faz com que na recuperação da informação haja um aumento das taxas de silêncio (devido à sinonímia) e ruídos (devido à polissemia e combinação de palavras).

Segundo Kuramoto (1996), o sintagma nominal (SN) é a menor parte do discurso portadora de informação, ao contrário das palavras que são símbolos sem referência, os sintagmas nominais são considerados melhores descritores de conteúdo, capazes de descrever com maior exatidão os assuntos tratados nos documentos e permitir uma recuperação de informação com maior precisão que as palavras isoladas.

Os SNs constituem estruturas gramaticais frasais que possuem substantivo como núcleo, sofrem menos dos problemas de sinonímia, polissemia e ambiguidade que as palavras isoladas, possuem estrutura sintática, assim como são portadores de uma estrutura lógico-semântica e quando extraídos do texto mantém o seu conceito (KURAMOTO, 2002).

De acordo com Perini (1998 *apud* CORRÊA *et al* 2011) os sintagmas nominais possuem duas estruturas, uma estrutura à esquerda do núcleo e outra à direita do núcleo. O núcleo é o elemento essencial à existência do SN, esse núcleo pode ser um substantivo, um pronome substantivo, um numeral ou uma palavra substantivada. A estrutura à esquerda do núcleo é composta por determinantes, como, artigos, possessivos etc., e na estrutura à direita se encontram modificadores ou até outros sintagmas nominais. Assim, um SN pode ser constituído por apenas uma estrutura, o núcleo, (um nome), como também por três estruturas, como, por exemplo, determinantes + núcleo + modificadores (incluindo outros SNs).

A ideia de utilizar os sintagmas nominais ao invés das palavras isoladas foi de Michel Le Guern (1991), considerado o pioneiro nas pesquisas sobre o uso dos sintagmas nominais na indexação automática. O autor 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, uma vez que o descritor utilizado para a recuperação de informação deveria ser uma unidade do discurso (como o SN) e não uma unidade da língua (signo isolado sem significado como a palavra).

Desta forma, segundo Miorelli (2001 *apud* Silva 2014) pode-se dizer que os documentos possuem uma enorme variedade de frases que podem ser usadas na construção de um índice, porém é necessário selecionar um conjunto delas.

Os SNs podem ser usados no processo de indexação substituindo as palavras isoladas, objetivando a construção de um índice hierárquico o qual poderia ser usado pelo usuário em uma interface de busca (KURAMOTO, 2002).

A indexação automática por sintagmas nominais precisa seguir algumas etapas para ser realizada. Nascimento (2015) apresenta um quadro que descreve as etapas do processo de indexação automática por sintagmas nominais (Quadro 1).

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"

Quadro 1 - Etapas da indexação automática por sintagmas nominais. Fonte: Nascimento (2015)

Segundo Nascimento (2015), essas etapas são fundamentais para que ocorra uma seleção eficiente de SNs tidos como Descritores, visto que o propósito é que se selecionem os SNs mais representativos do tema tratado no documento. Da mesma forma que o indexador humano seleciona os termos mais representativos de um determinado documento, a máquina também deve ser capaz de selecionar os SNs mais apropriados para a descrição do conteúdo de um documento.

Um exemplo de software que executa todas essas três etapas mencionadas no quadro anterior é o OGMA. Além realizar análise de texto e calcular a similaridade entre documentos, executa a extração e a seleção de SNs, esta ferramenta será mais bem descrita na seção 4.2.

2.3 Avaliação Da Indexação Automática

Bandim (2017) menciona Leiva (1997) ao dizer que no processo de indexação tem-se, basicamente, a avaliação intrínseca e a avaliação extrínseca. Na avaliação intrínseca é medido o grau de consistência da indexação, que de acordo com Leiva é definida como o grau de concordância entre indexadores de um mesmo grupo ou entre indexadores de grupos diferentes, quando da representação da informação essencial de um documento, por meio de um conjunto de termos de indexação selecionados por estes indexadores. Na avaliação extrínseca quantitativa são

medidos os índices de precisão, revocação e medida F, dos termos propostos para indexação.

Segundo Leiva (1999), para obter os valores para o índice de precisão calcula-se a relação entre os termos relevantes recuperados e o total de termos recuperados, já o índice de revocação é obtido através da relação entre os termos relevantes recuperados e o total de termos relevantes existentes para cada artigo. A medida F é a média harmônica ponderada entre o índice de precisão e o índice de revocação, sendo uma maneira de combinar a precisão e revocação em um único número. Se nenhum termo relevante foi recuperado a medida F assume valor 0 e, quando todos os termos recuperados são relevantes e foram exaustivamente recuperados todos os termos relevantes, assume valor 1.

2.4 Softwares De Indexação Automática

Os softwares de indexação apresentados a seguir são descritos na literatura da Ciência da informação, apesar de que segundo Narukawa (2009) são identificadas poucas iniciativas no Brasil a respeito disto. Nesta seção será dado destaque aos softwares utilizados na metodologia de pesquisa: o Sistema de indexação semiautomática (SISA); e o OGMA.

2.4.1 SISA

Desenvolvido na Espanha por Gil Leiva em 1999, o Sistema de Indexação Semiautomática (SISA), foi proposto inicialmente para a área de Biblioteconomia, mas a sua flexibilidade permite que o sistema seja adaptável para qualquer área. O objetivo é aplicá-lo a artigos científicos, considerando as estruturas do documento analisadas pelo sistema, delimitadas por marcadores específicos que indicam o título (#CTI# e #FTI#), o resumo (#CR# e #FR#) e o texto (#CTE# e #FTE#) (NARUKAWA, 2011).

A Figura 1 apresenta o Diagrama de fluxos do algoritmo do SISA.

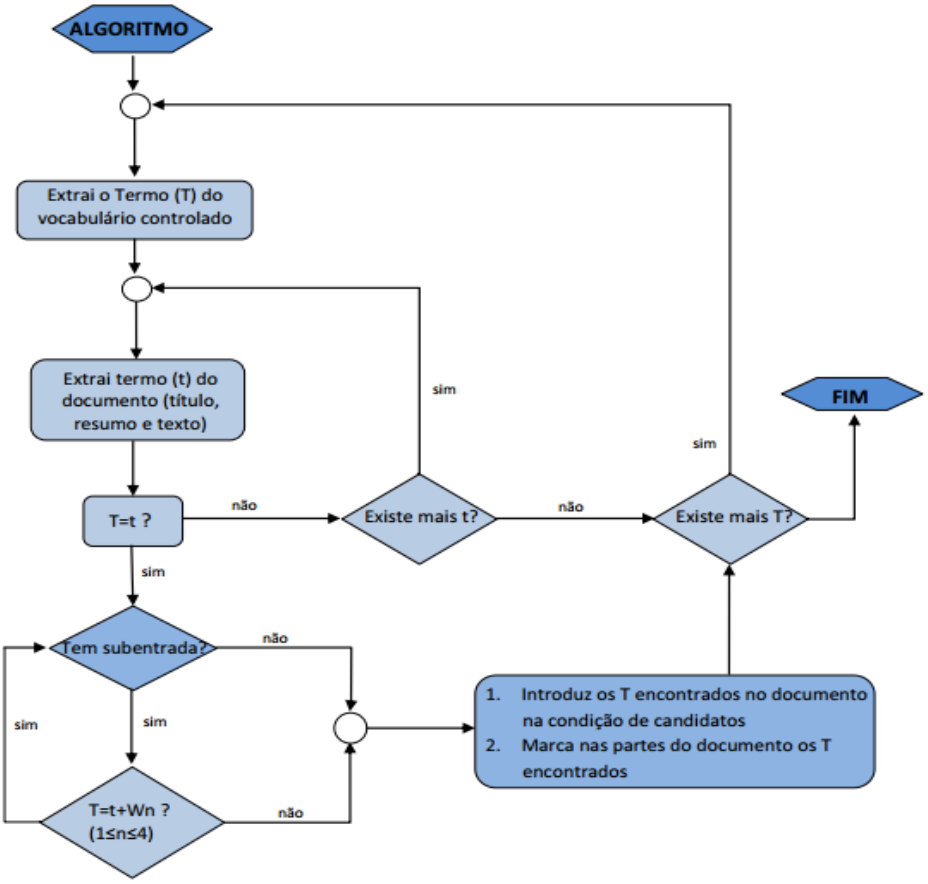

Figura 1 – Diagrama de fluxos do algoritmo SISA. Fonte: Gil Leiva (2008)

Segundo Narukawa (2011), o SISA exige para sua aplicação, como fonte de entrada, um documento em texto completo, um vocabulário controlado e uma lista de palavras vazias. A seguir é descrito o processo de análise automática do SISA que ocorre em três módulos:

- Módulo 1 – fase de pré-processamento do documento inserido no sistema, este é inicialmente sinalizado com os marcadores exigidos pelo SISA que correspondem a título, resumo e texto, este preparo é realizado para que posteriormente os cálculos de ponderação possam ser realizados a partir da identificação da frequência nessas estruturas. Nesta etapa ocorrem dois processos: eliminação de palavras vazias, através do confronto das mesmas com a lista de palavras vazias, e a horizontalização onde as frases e orações compreendidas entre os sinais de pontuação (. , ; :) são dispostas em forma horizontal, ou seja, são separadas em cada linha do texto.

- Módulo 2 – é a etapa de análise de conteúdo, fase em que o algoritmo busca e seleciona os termos preferidos que sejam coincidentes com os termos da linguagem documentária; os sinônimos que são os termos não

preferidos, não podem ser utilizados e remetem aos preferidos; e os termos que são construídos sintaticamente diferentes dos termos preferidos que sejam as palavras semivazias, aquelas em que o sistema julga importante, mas não se enquadra nas anteriores.

- Módulo 3 – fase em que os termos considerados pelos sistemas são ponderados, pois do contrário seriam selecionados e propostos como termos de indexação todos os termos do vocabulário controlado que coincidem com os das fontes do documento. Desta forma o sistema considera os seguintes critérios para propor os termos de indexação:

1) Se um termo autorizado aparece na fonte-título e na fonte-resumo;

2) Se um termo autorizado aparece na fonte-título e na fonte-texto;

3) Se um termo autorizado aparece na fonte-resumo e na fonte-texto;

4) Se o termo candidato a descritor aparece no título, no resumo e no texto;

5) Se um termo candidato a descritor aparece no texto dez vezes ou mais, além de aparecer em oito parágrafos diferentes ou mais, e não está incluído em nenhum dos termos propostos, apresenta-se como termo candidato a termo de indexação.

A Figura 2 apresenta a interface do software SISA.

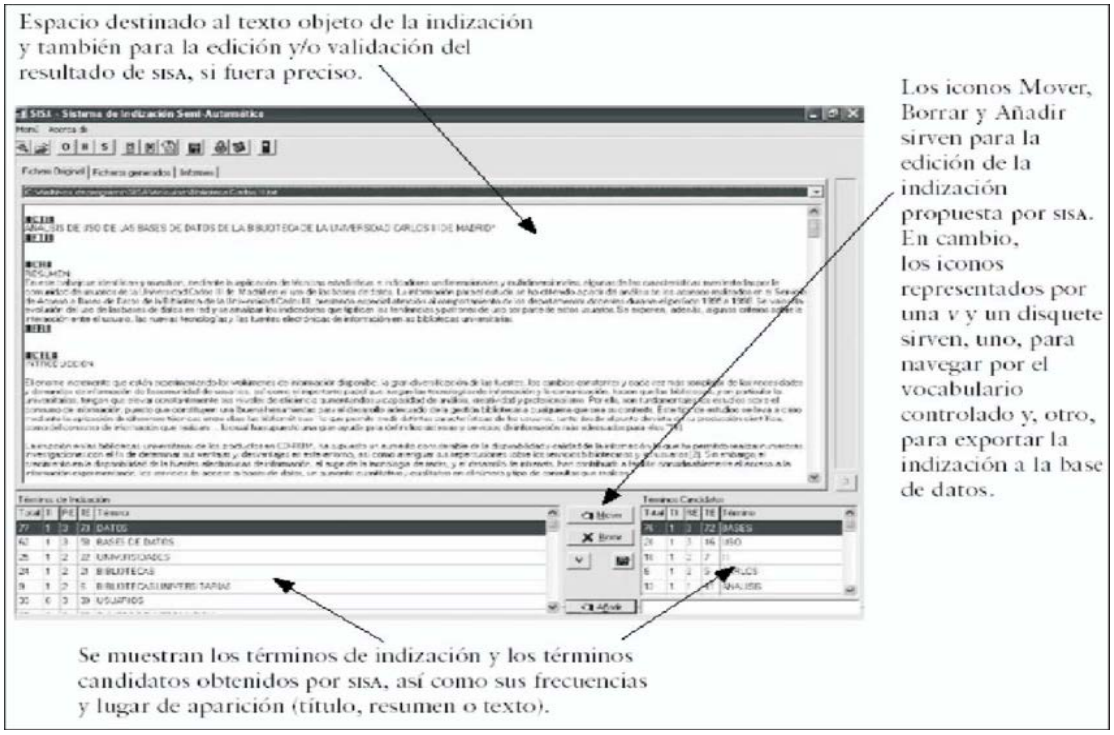

Figura 2 – Interface do SISA. Fonte: Gil Leiva (2008)

2.4.2 OGMA

O software OGMA foi desenvolvido por Maia (2008), é uma ferramenta que objetiva analisar texto, realizar cálculo da similaridade de documentos e extração de sintagmas nominais. Este foi desenvolvido na ferramenta *Visual Studio*. NET em linguagem C#. Por se tratar de uma ferramenta para análise de texto optou-se pelo desenvolvimento em modo texto, porém foi desenvolvida interface gráfica posteriormente.

É possível realizar a identificação da classe do sintagma nominal e o cálculo da pontuação do mesmo como descritor. Para a extração dos sintagmas nominais o OGMA faz o uso de um conjunto de regras, aplicando uma por uma em ordem de leitura até obter o sintagma nominal formado pelo o símbolo SN, assim como mostra o Quadro 2.

AR ← AD	de ← AR	AV ← AV ad	re ← SU	NS ← MD NS
AR ← AI	de ← PD	MD ← AV MD	de ← PP	NS ← NS pr NS
AJ ← VP	de ← PI	MD ← MD co MD	re ← NP	NS ← NS pr de NS
NU ← NR	qu ← AJ	NS ← NS MD	NS ← re	NS ← NS co NS
NU ← NC	qu ← NU	co ← CO	MD ← qu	NS ← NS co de NS
CO ← VG	qu ← PS	pr ← PR	SN ← NS	NS ← AV NS
CO ← CJ	ad ← AV		AV ← ad	SN ← de SN

Quadro 2 - Regras de extração de SN do método OGMA. Fonte: Maia (2008)

Estas regras atuam sobre as etiquetas (que representam as classes gramaticais) atribuídas às palavras, visando marcar o início e fim do sintagma em cada sentença do texto. Para extrair os sintagmas nominais o OGMA utiliza um léxico da língua portuguesa construída a partir do vocabulário do dicionário BR.ISPELL, e uma lista de 475 palavras irrelevantes criada baseada através da gramática de Tufano (1990). O léxico é utilizado para etiquetar cada palavra do texto com as possíveis classes gramaticais correspondentes.

Para lidar com a ambiguidade na etiquetagem das palavras, o OGMA forma uma lista com todas as combinações encontradas de etiquetas gramaticas para uma frase e submete cada frase etiquetada às regras para extração dos SNs, e depois submete os sintagmas nominais encontrados a uma lista geral de sintagmas nominais da frase, eliminado a duplicidade (MAIA, 2008).

Ainda segundo Maia (2008), o OGMA foi projetado para ser capaz de:

a) Extrair os Sintagmas Nominais.
b) Atribuir pesos aos Sintagmas Nominais extraídos de acordo com a freqüência que aparecem no texto.
c) Atribuir pesos aos Sintagmas Nominais extraídos de acordo com a freqüência que aparecem no texto e dentro de outros Sintagmas Nominais.
d) Identificar a classe do Sintagma Nominal (CSN) extraído de acordo com a metodologia proposta por SOUZA (2005) e explicada no item 6 deste trabalho.
e) Calcular a pontuação de cada Sintagma Nominal extraído (relevância como descritor) utilizando a mesma metodologia.

f) Extrair termos e atribuir pesos de acordo com sua freqüência no texto.
g) Extrair termos, exceto os constantes na lista de Stopwords, e atribuir pesos de acordo com sua freqüência no texto.
h) Calcular a similaridade entre duas listas de termos (extraídas do documento) utilizando o cosseno.

A figura 3 mostra a interface do software OGMA.

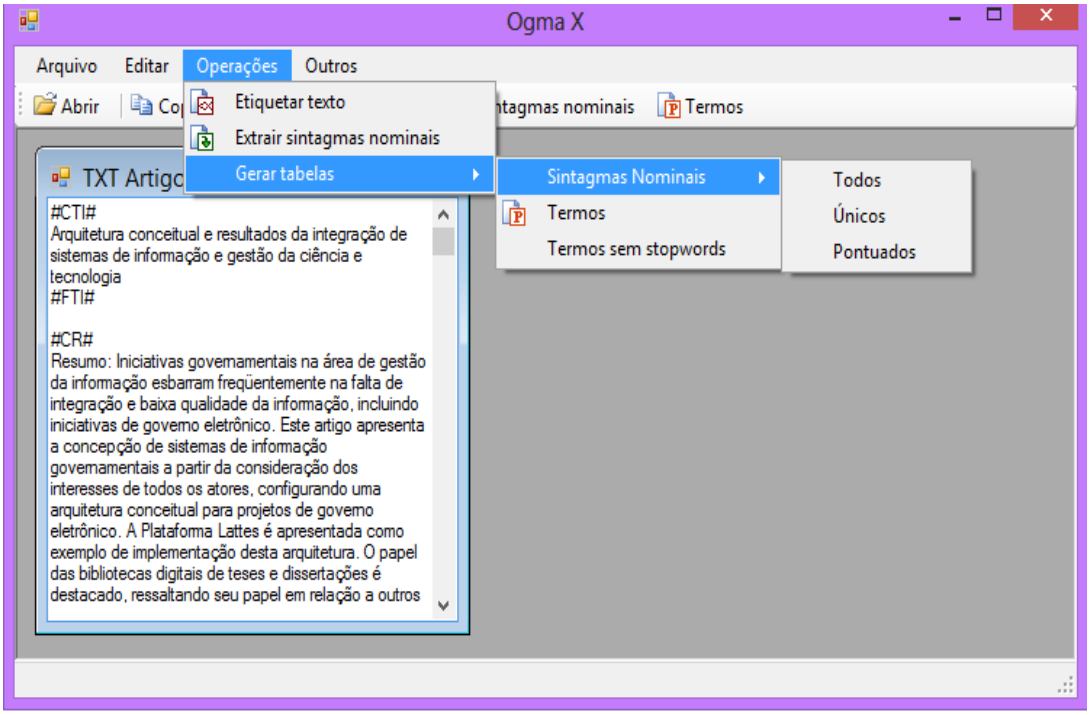

Figura 3 – Interface do OGMA

2.5 Trabalhos Relacionados

Com a finalidade de entender e levantar o estado da arte da avaliação de sistemas de indexação automática foi realizado um levantamento e síntese de trabalhos da literatura científica de Ciência da Informação no Brasil. Como base para o presente trabalho de pesquisa, selecionou-se doze trabalhos para subsidiarem a proposta de método de avaliação de indexação automática. Os trabalhos relacionados são apresentados em ordem cronológica e alfabética para facilitar o entendimento sobre a evolução das abordagens existentes.

Em (ALCAIDE *et al*, 2001), o objetivo foi analisar a consistência dos produtos de indexação automática em domínios científicos de Saneamento Básico e de Educação e averiguar se essa via se assemelha em resultados, aos processos de indexação manual que utilizam a metodologia da Análise Documentária. O corpus foi composto por quatro textos dentro dos domínios científicos Educação e Saneamento básico, que correspondem respectivamente às áreas de Ciências Humanas e de Ciências Exatas. Sendo três científicos e um de divulgação para cada domínio. Para a indexação automatizada foi utilizado o protótipo de FERNEDA. O protótipo consegue efetuar a extração de palavras através da comparação a um dicionário morfológico. Primeiramente foram extraídos do texto os termos significativos existentes no dicionário e realizada contagem total de termos, no segundo passo o protótipo cria uma matriz relacional entre os termos, onde são avaliadas a freqüência que um termo aparece no texto e sua relação com os demais termos, por último o protótipo extrai as palavras-chaves baseando-se nas forças relacionais calculadas na matriz. Para a indexação manual os textos foram impressos e enumerados. Excluíram-se o resumo e as palavras-chave. Os vocabulários controlados foram entregues aos indexadores especialistas nas áreas dos domínios tratados. Logo após foi realizada uma análise comparativa e de consistência dos processos e produtos documentários, onde foram comparados os valores das palavras-chaves extraídas do texto e as atribuídas ao texto. Pode-se constatar que se não fossem as ocorrências de sinonímias, os produtos automáticos se aproximariam potencialmente dos manuais e que indexação automática está direcionada para um sistema que permita a pós-coordenação mais efetiva, com alto grau de extensividade, já a indexação manual tratou de conceitos mais gerais sem muitas inter-relações entre as palavras-chave. O autor relata que, em determinadas representações, a indexação automática conseguiu remeter aos significados dos discursos, e a manual, em outras, necessitava ser precisa, não em termos quantitativos, mas em aspectos semânticos.

Souza (2005, 2006) desenvolveu uma metodologia para seleção automática de descritores para documentos textuais digitalizados, em língua portuguesa, utilizando como descritores sintagmas nominais extraídos do texto completo. A metodologia foi aplicada a um corpus de 60 documentos, particionado em dois conjuntos: o primeiro com 30 textos, sendo que 29 provenientes da Revista

DataGramaZero, e 1 proveniente da Revista Ciência da Informação; o segundo com 30 textos, todos provenientes da Revista Ciência da Informação. Apresentou-se uma metodologia de indexação automática, viabilizando um processo de atribuição de descritores a documentos digitalizados. Estes descritores foram escolhidos através da extração de sintagmas nominais (SNs) e análise de fatores como a freqüência de ocorrência destes SNs no texto de cada documento, no conjunto dos documentos; a estrutura e nível dos SNs; e a ocorrência destes em um tesauro. A consideração destes fatores de forma conjunta permitiu a criação de um ranking dos SNs candidatos a descritores, a partir dos SNs extraídos e pontuados de acordo com uma fórmula definida. Os SNs foram tratados estatisticamente utilizando o software MICROSOFT Excel. O número de descritores escolhidos para cada documento foi calculado tendo como base 1% dos SNs únicos identificados no documento, e levando em conta e os limites inferior de 8 e superior de 15 descritores por documento. Este valor é limitado apenas por uma conveniência metodológica, não havendo limitações reais para a escolha do número de descritores, excetuando o total de SNs extraídos, no final foram apresentadas as médias e os valores percentuais relativos de freqüência de SNs extremamente relevantes como descritores (SNs***), razoavelmente relevantes como descritores (SNs**), moderadamente relevantes como descritores (SNs*) e não relevantes como descritores (SNs-); além da média e o valor percentual dos “stopwords” (SW) em relação ao total dos SNs que foram eliminados. Concluiu-se que a metodologia aplicada foi um sucesso contrariando estudos feitos anteriormente declarados malsucedidos, estudos estes que buscavam a extração de descritores baseando-se em estruturas sintáticas das orações. Os resultados foram considerados eminentemente positivos pelo autor.

No artigo (BORGES; MACULAN; LIMA, 2008) são apresentados os critérios teóricos da semântica e da estrutura sintática no processo de indexação automática, e como o triângulo semântico de Ogden e Richards (1972) exposto na Teoria do Conceito de Dahlberg (1978) pode ser relacionado com esse contexto. Neste trabalho os autores apresentam um parser de extração automática que utiliza os critérios sintático-semânticos, esta pesquisa está restrita apenas a uma avaliação teórica da importância do uso desses critérios tanto sobre seu aspecto sintático quanto semântico. Como o objetivo do artigo era a análise de conteúdo, não foi feita

uma avaliação da indexação automática. Os autores afirmam que indexação é o elo forte entre o que é disponibilizado no sistema e a necessidade do usuário e que a fase de análise de conteúdo é a mais importante para o indexador. Eles acreditam que a adoção de uma taxonomia como cenário semântico usado no parser de teste é essencial para um resultado satisfatório. Para o desenvolvimento do trabalho, foram utilizadas as seguintes ferramentas: o parser Tropes e a taxonomia da área de Ciência da Informação construída por Hawkins, Larson e Caton, elaborada em 2003, como cenário semântico. Não foi realizada uma análise de qualidade de indexação visto que o objetivo desse trabalho foi o de investigar o processo de indexação automática e as teorias nas quais ele se baseia.

Em (LIMA; BOCCATO, 2009), buscou-se avaliar o desempenho terminológico, nos processos de indexação manual, automática e semi-automática, dos descritores do Vocabulário Controlado do SIBi/USP que representam o domínio da Ciência da Informação. Para a avaliação nos processos de indexação manual, automática e semi-automática, foi utilizado como corpus 70 resumos das dissertações e teses, defendidas no Programa de Pós-Graduação em Ciência da Informação da Escola de Comunicações e Artes da Universidade de São Paulo, cadastradas no Banco de Dados Bibliográficos da USP (DEDALUS), no período de janeiro de 2002 a dezembro de 2007. Para indexação automática e semi-automática foi utilizado o software “Sistema de Indización Automático” (SISA). O primeiro passo da metodologia de indexação foi o de preparar e inserir no SISA as listas em português a partir dos descritores e termos genéricos da Ciência da Informação retirados do Vocabulário Controlado do SIBI/USP. Foram coletados os resumos do DEDALUS, salvando-o no formato txt e inseridas cada uma das marcas exigidas pelo SISA (delimitando título, resumo e texto). O resumo foi repetido no campo texto, pois o SISA só realiza a indexação quando existe texto em todos os campos. Após a preparação das listas e do corpus, foi realizada a indexação automática e salva em um arquivo os descritores atribuídos pelo software a cada um dos resumos. Posteriormente, realizou-se a indexação semi-automática. A partir da leitura do resumo, avaliaram-se os descritores atribuídos automaticamente a cada um pelo SISA, assim a partir dos termos candidatos à indexação indicados pelo software, foram escolhidos de acordo com o entendimento dos autores sobre o resumo, os que representavam mais adequadamente o conteúdo informacional. A partir dos

dados obtidos foi possível elaborar um quadro comparativo onde para cada resumo foram indicados: os descritores atribuídos na indexação manual realizada durante o cadastramento no DEDALUS; os descritores atribuídos automaticamente pelo SISA e os descritores atribuídos pelos autores, configurados como indexação semi-automática. Foi calculada a coincidência entre os diferentes processos de indexação. Foram relacionadas algumas considerações, não apenas sobre os processos de indexação realizados, como também sobre o objeto do processo de indexação, ou seja, os resumos, e ainda sobre a ferramenta SISA. Os resultados indicam também as áreas que são objeto do maior número de dissertações e teses nas linhas de pesquisa do Programa de Pós-graduação de Ciência da Informação da ECA/USP.

No artigo (NARUKAWA; LEIVA; FUJITA, 2009), o objetivo foi investigar a consistência da indexação e a exaustividade e precisão na recuperação da informação fazendo uma comparação entre a indexação automática do Sistema de Indización Semiautomático (SISA) e a indexação manual do Centro Latino-Americano e do Caribe de Informação em Ciências da Saúde (BIREME). O corpus utilizado foi composto por 100 artigos científicos de pesquisadores brasileiros de odontologia estruturados com marcadores de título, resumo, texto e, convertidos em formato txt, os mesmos já se encontravam indexados manualmente com o uso da linguagem documentária DeCS por bibliotecários (indexadores experientes) de instituições que participam e cooperam com a BIREME. A avaliação da indexação automática foi dividida em três partes: primeiro foram feitas duas listas, uma dos descritores e suas respectivas remissas, e a outra com palavras vazias, ambas organizadas em ordem alfabética e convertidas em formato txt; depois foram realizados testes com vinte artigos para verificar a adequação das fontes utilizadas no SISA; após os testes, foi realizada a indexação dos artigos científicos pelo sistema. Os descritores obtidos pelo SISA e os obtidos pela BIREME foram organizados em um quadro comparativo, através deste quadro foi possível verificar quais e quantos descritores iguais aos descritores da BIREME foram atribuídos pelo SISA. Foram realizados cálculos para verificar o índice de consistência em relação aos termos obtidos, logo após seguiu-se a avaliação da exaustividade e precisão na recuperação da informação por meio de buscas em bases de dados constituídas pelos descritores dos dois tipos de indexação. O índice de exaustividade na

recuperação pode ser obtido através da relação entre os artigos científicos relevantes recuperados e o total de artigos científicos relevantes que se encontram na coleção total de artigos, e o índice de precisão na recuperação se obtém da relação entre os artigos científicos relevantes recuperados e o total de artigos recuperados. A análise dos resultados realizou-se a partir da análise quantitativa e análise qualitativa e como resultado pode-se observar que a indexação automática possui critérios diferentes da indexação manual, mesmo que as duas sigam o mesmo objetivo, isso porque a indexação manual requer uma reflexão sobre o assunto, se baseando no significado dos conceitos do assunto, já a automática é baseada na frequência de palavras que respeitam uma regra determinada.

O trabalho (MAIA; SOUZA, 2010) buscou investigar a utilização de sintagmas nominais pontuados como elementos para classificação por similaridade e aglomerados de documentos eletrônicos. Duas ferramentas foram utilizadas no experimento e foram de fundamental importância na pesquisa: o OGMA e o WEKA. Utilizaram-se dois Corpus: o primeiro constituído de 50 artigos selecionados do Encontro Nacional de Pesquisa em Ciência da Informação - ENANCIB 2005 e o segundo formado por textos menores e de conteúdo jornalístico com conteúdo bem definido em relação aos assuntos, para isso extraiu-se do site do jornal Hoje em Dia todas as notícias de 2004. O experimento contou com as seguintes etapas: extração dos sintagmas nominais para documentos em ambos os corpus utilizando a ferramenta OGMA; a construção de uma tabela com descritores, uma lista de palavras extraídas dos documentos, gerada pelo OGMA, esta recebeu como parâmetro um arquivo com o resultado da extração dos sintagmas nominais e a partir dele gerou uma tabela com a quantidade de vezes que cada um dos sintagmas apareceu no texto, também foi calculado a semelhança entre as tabelas, para isso, utilizou-se o OGMA que recebeu dois parâmetros que correspondessem ao nome dos arquivos das duas tabelas que se pretendiam comparar; automação das etapas pelo OGMA, para realizar todo o procedimento foi criado um arquivo de lote para a realização de grande parte das etapas; para realizar o uso dos dados obtidos pelo software OGMA e aplicá-los aos algoritmos fornecidos pelo WEKA, foram necessários convertê-los para um formato próprio, o ARFF. Logo após ocorreu a geração da matriz de similaridade entre os documentos para cada corpus e conversão da mesma para utilização no software WEKA; por fim as matrizes foram

submetidos aos algoritmos *Naive Bayes* e *simplekmeans* implementados pelo Weka. A primeira etapa do experimento prospectivo correspondeu à geração das tabelas de indexação, contendo todos os descritores, para cada documento do corpus, sendo uma tabela para cada método. As tabelas foram utilizadas posteriormente no cálculo de similaridade entre os documentos. Para obter valores que permitam uma comparação entre os métodos, o desvio, a média, o valor máximo (eliminando o valor 1) e o valor mínimo foram calculados. Apesar de resultados similares, o uso de sintagmas nominais envolveu um processamento computacional muito maior que o uso de termos isolados sem stopwords, os métodos com melhores resultados foram o de termos isolados sem stopwords e o de sintagmas nominais classificados e pontuados como descritores.

Em (CORREA et al, 2011), foi proposta a utilização dos sintagmas nominais no processo de indexação automática de teses e dissertações da Biblioteca de teses e dissertações da UFPE, para isso, foi utilizado um Corpus com trinta resumos de três programas de pós-graduação: Ciências da computação, Nutrição e Direito, dez de cada área foram selecionados através do sistema TEDE da BDTD-UFPE. Os estudos de caso foram realizados com o auxílio da ferramenta OGMA, de onde se extraiu os sintagmas nominais dos resumos de teses e dissertações. A avaliação do processo de extração automática de sintagmas nominais dos resumos de cada programa foi realizada através do cálculo e análise dos percentuais de precisão em extrair sintagmas nominais relevantes como descritores para os resumos (sintagmas que correspondem à necessidade de informação do usuário no momento da busca por um documento), a taxa de erro ao extrair cadeias de caracteres que não constituem sintagmas nominais e o percentual de sintagmas nominais extraídos não relevantes como descritores. Para cálculo dos percentuais foi necessária a classificação dos possíveis sintagmas nominais extraídos como verdadeiros sintagmas nominais ou falsos sintagmas nominais, sendo esta classificação realizada com base nas definições de sintagmas nominais descritas na revisão de literatura. Bem como a classificação dos verdadeiros sintagmas nominais extraídos em relevantes ou não como descritores para o resumo de onde foram extraídos, sendo esta última realizada com base na análise de assunto do resumo e as palavras-chaves (também constantes no resumo) da respectiva tese ou dissertação. Estes dados foram organizados em tabelas para cada programa de pós-graduação,

o resultado do processo de extração obteve diferentes desempenhos para cada programa, sendo que para os resumos de Direito foi obtida melhor precisão, depois Computação e Nutrição. Também se observou que nem todos os sintagmas extraídos pelo OGMA são de fato sintagmas nominais e que nem todos os sintagmas nominais são extraídos pelo OGMA. Foi possível concluir que os sintagmas nominais extraídos classificados como relevantes se constituem em bons descritores para os resumos, e que os sintagmas nominais se apresentam como uma alternativa ao uso de palavras isoladas nos sistemas de recuperação de informação, pois se configuram como melhores descritores e pontos de acesso aos documentos, já que estes não possuem os problemas causados pela ambiguidade e a polissemia das palavras isoladas.

No artigo (CORREA; LAPA, 2013) se desenvolve uma pesquisa exploratória apresentando um panorama dos estudos sobre a indexação automática no âmbito da ciência da informação no Brasil, através do mapeamento e análise da produção acadêmica e científica nacional no período de 1973 a 2012. Foram analisados 69 documentos brasileiros sobre a indexação automática, como livros e capítulos de livros, dissertações e teses, publicações em periódicos e comunicações em anais de congressos e de seminários, destinados à área de ciência da informação publicados/comunicados nos últimos 40 anos (1973-2012). O estudo se formou por meio do mapeamento e discussão da produção acadêmica e científica através de uma abordagem qualitativa sobre a indexação automática no campo da ciência da informação e por uma abordagem quantitativa, oriunda de investigação dos resultados das análises bibliométricas sobre os autores que desenvolveram pesquisas no âmbito nacional sobre a indexação automática, as instituições acadêmicas envolvidas na elaboração da obra, os tipos de fontes de informação utilizadas para publicar os trabalhos, as instituições responsáveis pela publicação do trabalho, e o ano em que o trabalho foi publicado. Na análise de conteúdo procurou-se traçar o panorama sobre as principais ideias pesquisadas ao caracterizar o objetivo dos trabalhos e aspectos metodológicos, como nome do sistema/método/ fórmula aplicados e/ou propostos, como ocorreu à avaliação, qual a natureza e a tipologia do corpus, como ocorreu à validação e a identificação dos termos, qual o tratamento no texto no momento da entrada de dados, a linguagem de indexação e o método/processo usado na identificação/ponderação/seleção dos termos. Os

resultados indicaram que as pesquisas sobre indexação automática estão concentradas em certos autores, instituições acadêmicas, fontes de publicação e instituições publicadoras, além de existir tendência de aumento das publicações sobre esse assunto nos próximos anos. É um trabalho tem muita contribuição para a indexação automática por trazer uma analise da produção científica nacional acerca do tema, apontando as avaliações, sistemas, métodos e fórmulas mais utilizadas, permitindo assim distinguir as tendências e obter subsídio para a realização de pesquisas futuras.

O trabalho (NICOLINO; FERNEDA, 2014) apresenta diretrizes técnicas para a construção e utilização de ontologias no processo de indexação automática, para a realização do estudo utilizou-se uma ontologia de termos de pediatria (PedTerm), disponível no BioPortal (repositório de ontologias na área biomédica), no qual apresenta informações relacionadas à saúde e ao desenvolvimento infantil. O método proposto para utilização de ontologias no processo de indexação automática consistiu nas seguintes etapas: Ontologia para indexação automática – caracterizada como uma indexação por atribuição, na qual um único documento ou um conjunto de documentos é vinculado a uma estrutura terminológica; Extração de termos – onde um sistema automatizado para a extração de informação é utilizado para extrair um conjunto inicial de termos que representem o conteúdo informacional dos documentos (indexação automática por extração); Atribuição de conceitos – o termo extraído do texto deve coincidir com um termo que foi definido na propriedade *Label* de uma das classes da ontologia para atribuição da classe ou conceito ao documento; Atribuição de termos sinônimos – esta etapa é realizada através da propriedade que fornece uma maneira legível (por humanos) de descrever ou identificar uma classe, podendo relacionar sinônimos para um mesmo idioma; Indexação de um corpus multilíngue – a propriedade deve possuir o parâmetro XML: *Lang*, através disso é possível fazer com que uma mesma ontologia seja utilizada na indexação de um corpus contendo documentos de diferentes idiomas. por meio de exemplos, este trabalho propôs um método de utilização de ontologias no processo de indexação automática, não foi feita uma avaliação de qualidade da indexação proposta. Os autores acreditam que sistemas automatizados de indexação baseados em ontologias podem se mostrar mais competentes que os que não se baseiam,

melhorando assim os recursos do processo de indexação e avanços na funcionalidade e sistemas de recuperação da informação.

Em (LAPA; CORREA, 2014) é apresentado o panorama da pesquisa no âmbito da CI no Brasil referentes aos estudos sobre a Indexação Automática no período 1973 – 2012. Para isto 69 documentos foram examinados através do método de análise de conteúdo que parte de uma perspectiva quantitativa, analisando numericamente a frequência de ocorrência de determinados termos, construções e referências em um dado texto. Os documentos analisados foram categorizados da seguinte forma: em relação ao objetivo dos documentos – analisou-se como estava a distribuição dos trabalhos que realizaram proposição, aplicação e a proposição e aplicação; Natureza do corpus – avaliou-se a preferência em realizar pesquisas quanto à indexação automática de texto completo, resumo, título, verbetes ou citação; Tipologia documental – foi analisado qual o tipo de documento foi o mais pesquisado; Validação dos termos – analisou-se o tipo de indexação que os trabalhos aplicaram ou propuseram para validar os termos de suas pesquisas; Identificação dos termos – buscou avaliar como o processo de identificação foi realizado, por meio da extração ou por meio da atribuição. Após a análise dos documentos de acordo com cada categorização, os autores perceberam que há uma tendência em estudos sobre a indexação automática por meio dos sintagmas nominais e que com o uso de novas tecnologias procura-se desenvolver uma identificação automática dos termos por meio da atribuição. Este trabalho deixou contribuições para a indexação automática por abordar os diversos aspectos referentes ao tema e a análise de conteúdo através de um corpus levantado, mostrando que a maioria dos trabalhos a cerca deste assunto procuram avaliar se a implantação do sistema automático traz benefícios, obtendo resultados equivalentes em menos tempo e também incentivando pesquisadores a continuarem produzindo novas pesquisas.

No artigo (SILVA; SOUZA, 2014) é proposto um método automatizado que utiliza uma heurística determinística denominada Heudet que visa extrair bigramas do texto. O objetivo principal é extrair o significado do texto através de um conjunto de expressões multipalavras (EM), ou *n*-gramas, que são termos adjacentes encontrados a partir da frequência de co-ocorrências observadas, e dependências significativas. O método Heudet, foi implementado em um programa de computador

desenvolvido na linguagem C++ pelos autores. O corpus utilizado na avaliação foi composto por 194 artigos publicados no Enancib de 2010 no formato texto, totalizando 682.537 termos normalizados, sendo 46.888 distintos. Para que seja possível extrair as EM, o algoritmo Heudet percorre as sentenças verificando para cada palavra, que ainda não foi processada, quais são as suas adjacentes. As EM englobam diversos fenômenos distintos como compostos nominais, expressões idiomáticas e termos compostos, são necessariamente compostos por mais de uma palavra e as *stopwords* são descartadas. Logo após, verifica-se a frequência em que a repetição dos termos adjacentes ocorre. Os termos, com a frequência de repetição maior ou igual a uma quantidade informada em parâmetro de entrada para o processamento do algoritmo, serão adicionados na lista dos bigramas identificados. Os demais são descartados. Um ponto destacado pelos autores é que, apesar de esse processamento extrair apenas bigramas, não significa que expressões com *n*-gramas não sejam consideradas. Após finalizar a implementação do método Heudet, foi necessário comparar o resultado dos bigramas obtidos com o produzido por outras técnicas a fim de avaliar o seu desempenho, para isto foi utilizado o pacote *Ngram Statistics Package* (NSP), escolhido por disponibilizar um processo de identificação de bigramas através do processamento de treze medidas de associação estatísticas distintas. Para todo o processamento descrito anteriormente foi utilizado quatro como sendo a frequência de co-ocorrências do bigrama no documento, para todas as técnicas empregadas. Após os resultados foram realizadas consultas SQL a fim de comparar os conteúdos obtidos por cada uma das técnicas utilizadas. Percebeu-se que as EM identificadas por todas as técnicas do pacote NSP foram as mesmas, a única diferença no resultado produzido por essas técnicas foi a ordenação dos bigramas em função da relevância calculada. Após todo o processo, conclui-se que o algoritmo Heudet apresenta vantagens em relação ao uso das técnicas estatísticas. Isso se dá pelo fato de ele levar em consideração a estrutura do documento. Para os autores o método Heudet pode ser empregado com vantagens como parte de um Processamento de Linguagem Natural que demande a identificação de Expressões Multipalavras em um documento.

O trabalho de (CORREA; BASILIO, 2017) apresenta uma análise da indexação automática através de sintagmas nominais, mais especificamente avalia a revocação das palavras-chaves informadas pelos autores dos documentos no

processo de indexação automática por sintagmas nominais utilizando o software OGMA. O corpus utilizado consistia em 30 teses e dissertações da UFPE, divididas igualmente em grupos correspondentes a três programas de pós-graduação, Ciência da Computação, Direito e Nutrição. Esse corpus foi tomado de um estudo 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. Nesta avaliação as palavras chaves não foram incluídas, pois se buscou avaliar apenas a capacidade do OGMA em extrair sintagmas nominais semelhantes às palavras chaves presentes no título ou resumo dos documentos. O método consistiu nas seguintes etapas: primeiramente quantificou-se 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, todos os três grupos passaram pela extração de sintagmas nominais realizadas pelo OGMA, na segunda parte, 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, Foram considerados todos os sintagmas nominais que possuíam alguma das palavras-chave na forma integral. A métrica utilizada para avaliar a indexação automática por sintagmas nominais nesta pesquisa foi 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. Pode-se concluir que o OGMA obteve um bom desempenho visto que na revocação de palavras-chaves presentes no título e resumo de teses e dissertações, consegui extrair cerca de 70% das palavras-chave presentes no texto dos documentos.

3. METODOLOGIA

Quanto aos objetivos a pesquisa tem caráter exploratório. Quanto aos métodos a pesquisa tem caráter bibliográfico e experimental. O caráter bibliográfico se justifica pela necessidade de revisão e análise da literatura, bem como no

levantamento do estado da arte da avaliação de sistemas de indexação automática para documentos escritos em português. O caráter experimental se consolida na avaliação e comparação dos resultados dos sistemas de indexação automática encontrados na literatura.

As avaliações dos sistemas de indexação automática serão pautadas em estudo de caso único, com base em corpus construído a partir de artigos de periódicos da área de Ciência da Informação. A escolha desta área se justifica pela maior proximidade com a terminologia, bases de dados e vocabulários controlados da mesma, o que auxilia na concepção e avaliação dos sistemas de indexação automática.

Na seção 3.1 é descrita a construção do corpus utilizado na avaliação; a 3.2 descreve a metodologia utilizada no experimento. A seção 4 apresenta a discussão dos resultados encontrados no experimento, e na seção 5 são apresentadas as considerações finais.

3.1 Construção de corpus

O corpus consiste do conteúdo textual de 60 artigos de periódicos científicos da área da Ciência da Informação no Brasil utilizados por Souza (2005), se constituindo de textos científicos na área de Ciência da Informação escritos em português do Brasil, que foram convertidos para o formato texto. Os originais pdf e HTML foram convertidos para o formato txt e formatados para deixar as sentenças contíguas. O arquivo no formato texto foi estruturado através de marcações em três campos como requisitado pelo SISA: título, resumo e texto. O campo texto começa na introdução e termina antes do início da seção referências, que não foi incluída no campo. Através destes arquivos foram realizados os seguintes procedimentos a fim de compilar o corpus:

- ✓ Renomear os arquivos originais para ficar Artigo X.pdf ou Artigo X.htm, onde X é o número do artigo no anexo de Souza (2005);
- ✓ Adequar a ordem dos artigos originais e txt com o anexo da tese de Souza (2005);
- ✓ Completar tabela em planilha eletrônica com as palavras-chaves de cada trabalho, as palavras-chaves presentes no título ou resumo do documento, e

os sintagmas nominais relevantes (A) que Souza (2005) obteve para cada artigo, sendo considerado na presente pesquisa como um método semiautomático de indexação automática. Posteriormente, após a execução do experimento, os termos de indexação atribuídos pelos softwares de indexação automática podem ser incluídos nesta tabela para permitir a comparação;

- ✓ Para cada arquivo texto, foi aberto no Word, solicitado a visualização de marcas de parágrafo, e depois:
  - o Substituído a sequência de caracteres “°.” por espaço em branco;
  - o Substituído o caracter “°” por espaço em branco;
  - o Salvo o arquivo;
  - o Marcado título, resumo e texto com as marcas exigidas pelo SISA.
- ✓ Validar o arquivo, abrindo o arquivo original e arquivo texto relacionado, e comparar o conteúdo dos dois textos que devem ser o mesmo. O campo texto deve começar na introdução e somente a seção referências deve ficar de fora do arquivo texto;
- ✓ Remover quebra de linha no meio de sentenças, e mover parte do texto das sentenças para torná-las contíguas caso estejam sendo quebradas por outras estruturas como notas, cabeçalho, rodapé, fim de página, legendas, figuras, quadros e tabelas.

3.2 Experimento e método de avaliação

Após a revisão da literatura, foi possível analisar as metodologias de avaliação da indexação automática utilizadas nas pesquisas, resultando na escolha das métricas de consistência com a indexação manual, e precisão e revocação das palavras-chaves como base para a metodologia que será aplicada. O método de avaliação envolverá as etapas descritas a seguir:

- ✓ Obter os descritores atribuídos pelo software SISA utilizando o Tesauro Brasileiro de Ciência da Informação (TBCI) a cada arquivo texto do corpus;
- ✓ Extrair os sintagmas nominais dos arquivos texto através do software OGMA e ordená-los em ordem decrescente de pontuação, selecionando os 15 primeiros;

- ✓ Numa tabela contendo as palavras-chave dos autores, marcar de negrito os termos propostos pelos sistemas de indexação automática que casam com as palavras-chaves, e em negrito e itálico o que casam parcialmente com as palavras-chaves;
- ✓ Comparar e contabilizar as palavras-chaves dos autores com os termos comuns propostos pelo SISA, OGMA e método semiautomático baseado em (SOUZA, 2005). Para o método semiautomático baseado em (SOUZA, 2005), foram utilizados os descritores escolhidos para cada texto que possuíam o valor de relevância nomeado como A (maior relevância) conforme julgamento do próprio autor, estes descritores podem ser encontrados no Anexo C em (SOUZA, 2005);
- ✓ A avaliação do processo de indexação automática dos artigos científicos será realizada através do cálculo e análise dos percentuais de consistência na indexação, de precisão em extrair descritores (equivalentes às palavras-chaves dos autores), de revocação dos descritores (equivalentes às palavras-chaves dos autores), e a medida F que é a média harmônica de precisão e revocação. Os índices de consistência, precisão, revocação e medida F serão utilizados como indicadores da qualidade na indexação automática obtida pelos sistemas OGMA e SISA e o método semi-automático baseado em (SOUZA, 2005).

4. RESULTADOS E DISCUSSÃO

A partir dos dados obtidos para os sistemas analisados, foi possível elaborar tabelas comparativas onde para cada indicador foi calculado o valor mínimo, a média, o valor máximo e o desvio das porcentagens. Os indicadores de qualidade na indexação automática foram a consistência, precisão, revocação e medida F dos descritores atribuídos pelo OGMA, SISA e indexação semiautomática baseada em (SOUZA, 2005). A última serve como padrão máximo de qualidade para sistemas de indexação automática por extração no presente corpus.

Através das Tabelas 1, 2 e 3 é possível observar o desempenho dos sistemas indexadores e analisar comparativamente os resultados quantitativos de cada sistema utilizando a média dos resultados. As tabelas apresentam o mínimo, a

média, o máximo e o desvio dos seguintes indicadores: Número de palavras-chaves (Nº de PC); Número de palavras-chaves presentes (Nº de PC presentes); Número de termos propostos pelo sistema (Nº de termos propostos); Número de palavras-chaves comuns (Nº de PC comuns); Consistência; Precisão; Revocação; e Medida F (ou F-Measure).

A Tabela 1 representa o resultado da análise dos termos extraídos pelo sistema OGMA, a média da consistência foi de 10, 2%, da precisão foi de 11,8%, da revocação foi de 41,2% e da medida F foi de 17,7%.

Tabela 1 – Resultado da extração de descritores pelo OGMA.

	Nº de PC	Nº de PC Presentes	Nº de Termos propostos	Nº de PC comuns	Consistência	Precisão	Revocação	Medida F
Mínimo	2	0	15	0	0,0%	0,0%	0,0%	0,0%
Média	4,5	2,8	15,0	1,8	10,2%	11,8%	41,2%	17,7%
Máximo	9	6	15	4	23,5%	26,7%	100,0%	38,1%
Desvio	1,6	1,3	0,0	1,1	6,3%	7,1%	25,4%	10,6%

Tabela 1 - Resultado da análise dos termos extraídos pelo sistema OGMA. Fonte: os autores

A Tabela 2 apresenta o resultado da análise dos termos atribuídos pelo sistema SISA, a média da consistência foi de 11, 3%, da precisão foi de 14,9%, da revocação foi de 34,4%, e da medida F foi de 19,0%.

Tabela 2 – Resultado da atribuição de descritores pelo SISA

	Nº de PC	Nº de PC Presentes	Nº de Termos propostos	Nº de PC comuns	Consistência	Precisão	Revocação	Medida F
Mínimo	2	0	3	0	0,0%	0,0%	0,0%	0,0%
Média	4,5	2,8	9,8	1,4	11,3%	14,9%	34,4%	19,0%
Máximo	9	6	24	5	41,7%	57,1%	100,0%	58,8%
Desvio	1,6	1,3	5,0	1,1	10,1%	13,6%	28,8%	15,3%

Tabela 2 - Resultado da análise dos termos atribuídos pelo sistema SISA. Fonte: os autores

A Tabela 3 representa o resultado da análise dos termos extraídos através da indexação semiautomática baseada em (SOUZA, 2005). Tal desempenho será

utilizado como referência de qualidade na indexação para os sistemas avaliados. A média foi de 22,6% para a consistência, 32,9% para a precisão, 42,6% para a revocação e 34,4% para a medida F.

Tabela 3 – Resultado da extração de descritores por sistema semiautomático baseado em (SOUZA, 2005).

	Nº de PC	Nº de PC Presentes	Nº de Termos propostos	Nº de PC comuns	Consistência	Precisão	Revocação	Medida F
Mínimo	2	0	2	0	0,0%	0,0%	0,0%	0,0%
Média	4,5	2,8	5,8	1,8	22,6%	32,9%	42,6%	34,4%
Máximo	9	6	13	4	66,7%	100,0%	100,0%	80,0%
Desvio	1,6	1,3	2,5	1,0	14,5%	18,5%	25,9%	18,8%

Tabela 3 – Resultado da análise dos termos extraídos através da indexação semiautomática baseada em (SOUZA, 2005). Fonte: os autores

O gráfico na Figura 4 representa os valores médios das medidas de qualidade na indexação para o OGMA, SISA e sistema semiautomático baseado em (SOUZA, 2005). Nota-se que a indexação semiautomática, apresentou maior consistência, precisão, revocação e medida F. Já entre os dois sistemas de indexação automática avaliados, o SISA se destacou em três quesitos e só perdeu para o OGMA no percentual médio de revocação das palavras-chaves.

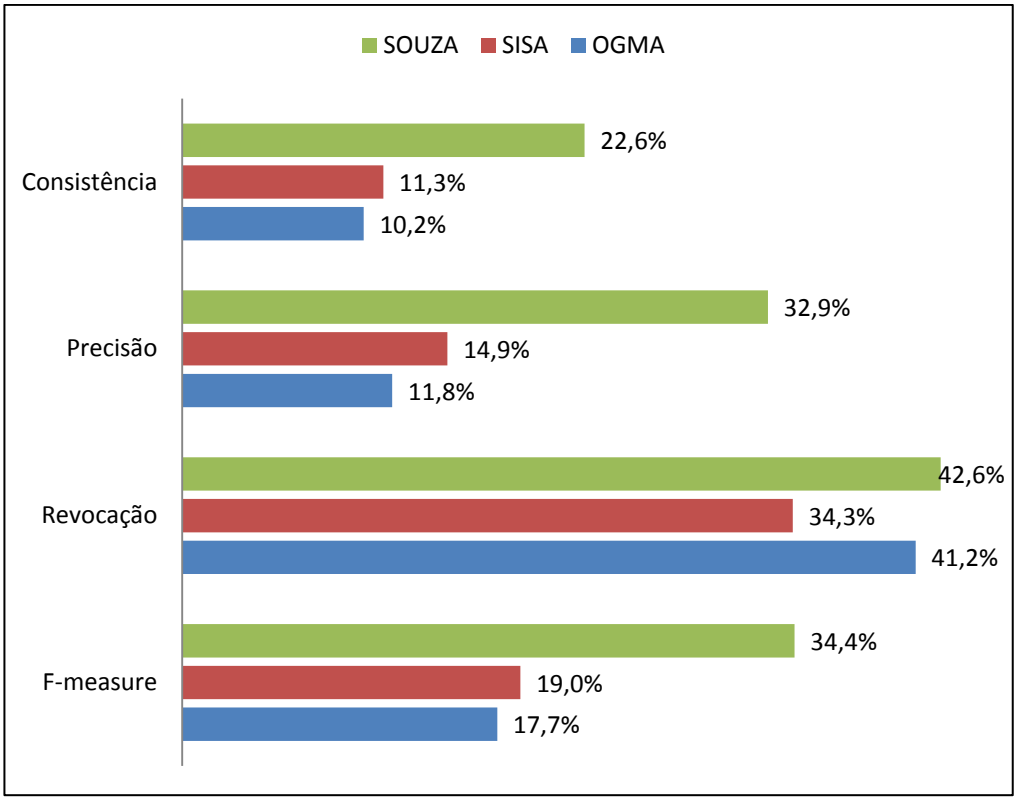

Figura 4 – Gráfico dos valores médios dos índices de qualidade na indexação

A partir da comparação dos resultados obtidos, pode-se perceber que a indexação semi-automática se sobressaiu aos dois sistemas, uma vez que houve um processo cognitivo humano na seleção dos sintagmas nominais relevantes. Entre os sistemas avaliados, o SISA se sobressaiu ao OGMA como melhor sistema de indexação automática, o que pode ser explicado pela escolha do vocabulário controlado (TBCI) que melhor atende os requisitos dos termos que irão representar o conteúdo dos documentos utilizados. Por outro lado, o fato do SISA atribuir apenas termos do artigo científico que estão no vocabulário controlado, e não atribuir como descritor termos que estão somente na estrutura do "texto", fez com que este perdesse para o OGMA no quesito revocação.

No entanto, ambos os sistemas precisam ser aperfeiçoados para alcançar o desempenho do sistema semiautomático, que sugerimos ser utilizado como padrão de qualidade para sistemas de indexação automática sobre o corpus compilado.

5. CONSIDERAÇÕES FINAIS

A procura por um bom sistema de indexação automática parte do pressuposto de tornar a representação temática o mais próximo possível do conteúdo textual, uma vez que as técnicas manuais não conseguem atender a grande demanda informacional que vem surgindo. Um sistema de indexação automática eficaz pode reduzir de forma significativa à subjetividade encontrada no processo feito manualmente, e possibilita a indexação de grandes volumes de informações em um tempo curto. O uso de um bom sistema de indexação automática é uma garantia para que tenhamos no futuro facilidade no acesso à memória institucional, científica e tecnológica.

O principal problema enfrentado nesta pesquisa foi o fato do tema “indexação automática de textos científicos escritos em português do Brasil” ser pouco explorado, resultando em poucos trabalhos convergentes, principalmente no aspecto da avaliação e comparação do desempenho dos sistemas de indexação automática. Sem recursos e métodos de avaliação padronizados, fica difícil mensurar e comparar a eficácia dos métodos de indexação automática implementados, bem como identificar métodos mais promissores e as características mais promissoras dos métodos de indexação automática.

Após avaliar os sistemas de indexação automática SISA e OGMA foi possível concluir que o SISA é o melhor sistema, visto que este através dos resultados obteve melhor consistência, precisão e medida F, enquanto o OGMA obteve apenas melhor revocação, o que não descarta a necessidade de melhoria de ambos para que estes alcancem o desempenho do sistema semiautomático, que sugerimos ser utilizado como padrão de qualidade.

Através desta pesquisa, contribui-se fornecendo embasamento a futuros trabalhos que se proponham a aplicar o método de avaliação de sistemas de indexação automática, assim como enriquecer a área da Ciência da informação acerca do tema que ainda é pouco explorado. A presente pesquisa cria subsídios para trabalhos que tenham como objetivo comparar dois sistemas utilizando medidas reconhecidas na literatura e uso do corpus elaborado, ou até mesmo para o aperfeiçoamento da metodologia de avaliação proposta. Como sugestão para trabalhos futuros, apontamos: analisar profundamente a razão dos valores obtidos

para as métricas de cada sistema; e propor aperfeiçoamentos nos sistemas avaliados.

REFERÊNCIAS

ALCAIDE, G. S. et al. Análise comparativa e de consistência entre representações automática e manual de informações documentárias. **Transinformação**, v. 13, n. 1, p. 23-41, 2001.

ARAÚJO JÚNIOR, R. H. de. **Precisão no processo de busca e recuperação da informação***.* Brasília: Thesaurus, 2007.

BAEZA-YATES, R.; RIBEIRO-NETO, B. **Modern Information Retrieval**. Addison-Wesley. 2011.

BANDIM, M. A. S. **Indexação automática por atribuição de artigos científicos da área de Ciência da Informação**. Recife, 2017.

BORGES, G. S. B. **Indexação automática de documentos textuais: proposta de critérios essenciais.** Dissertação de Mestrado em Ciência da Informação. Programa de Pós-Graduação em Ciência da Informação, Escola de Ciência da Informação, Universidade Federal de Minas Gerais - UFMG, Belo Horizonte, 2009.

BORGES, G. S. B.; MACULAN, B. C. M. S.; LIMA, G. N. B. M. O. Indexação automática e semântica: estudo da análise do conteúdo de teses e dissertações. **Informação & Sociedade: Estudos***,* v. 18, n. 2, p. 181-193, 2008.

BRUZINGA, G. S.; MACULAN, B. C.; LIMA; G. A. Indexação automática e semântica: estudo da análise do conteúdo de teses e dissertações. **VIII ENANCIB – Encontro Nacional de Pesquisa em Ciência da Informação**. Salvador, 31 out. 2007.

CORREA, 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.

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

CORREA, R.F.; BAZÍLIO, H.T. Análise da extração de descritores como sintagmas nominais através do software OGMA. **Encontros Bibli: revista eletrônica de biblioteconomia e ciência da informação,** v. 22, n. 50, p. 44-58, set./dez., 2017.

FUJITA, M. S. L. A representação documentária no processo de indexação com o modelo de leitura documentária para textos científicos e livros: uma abordagem cognitiva com protocolo verbal. **Ponto de Acesso**, Salvador, v. 7, n. 1, p. 42-66, abr. 2013.

FUJITA, Mariângela S. Lopes (Org.). **A indexação de livros: a percepção de catalogadores e usuários de bibliotecas universitárias. Um estudo de observação do contexto sociocognitivo com protocolos verbais.** São Paulo: Cultura Acadêmica, 2009.

KURAMOTO, H. Sintagmas nominais: uma nova abordagem no processo de indexação. In: Madalena Martins Lopes Naves; Hélio KURAMOTO (org.). **Organização da informação: princípios e tendências**. Brasília: Briquet de Lemos/Livros, 2006.

KURAMOTO, Hélio. Proposta de um Sistema de Recuperação de Informação Assistido por Computador - SRIAC. **Revista de Biblioteconomia de Brasília** v. 21, n. 2, jul./dez. 1997.

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

KURAMOTO, Hélio. Uma abordagem alternativa para o tratamento e a recuperação de informação textual: os sintagmas nominais. **Ciência da Informação** v. 25, n. 2, maio/ago. 1996.

LANCASTER, F. W. **Indexação e resumos: teoria e prática.** 2. ed. Brasília: Briquet de Lemos Livros, 2004.

LAPA, R.; CORREA, R. Indexação automática no âmbito da ciência da informação no brasil. **Informação & Tecnologia**, v. 1, n. 2, p. 59-76, 2014.

LEIVA, I. G. **La automatización de la indización, propuesta teórico-metodológica: aplicación al área de Biblioteconomía y Documentación**.1997. 268f. Tese – Universidad de Murcia, Murcia, España, 1997.

LEIVA, I. G. **Manual de indización: teoría y práctica**. Gijón: Trea, 2008. 429 p.

LIMA, V. N. M. A.; BOCCATO, V. R. C. O desempenho terminológico dos descritores em Ciência da Informação do Vocabulário Controlado do SIBi/USP nos processos de indexação manual, automática e semi-automática. **Perspectivas em Ciência da Informação**, v. 14, n. 1, 2009.

LOPES, I. L.. Uso das linguagens controlada e natural em bases de dados: revisão da literatura. **Ciência da Informação**, Brasília, v. 31, n. 1, p.41-52, jan-abril. 2002.

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

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**. Dissertação (Mestrado em Ciência da Computação) – Pontifícia Universidade Católica do Rio Grande do Sul, Porto Alegre, 2001.

NARUKAWA, C. M. **Estudo de Vocabulário Controlado na Indexação Automática: Aplicação no Processo de Indexação do Sistema de Indización Semi-automática (SISA).** 2011. 222 f. Dissertação (Mestrado) - Faculdade de Filosofia e Ciências, Universidade Estadual Paulista, Marília, 2011.

NARUKAWA, C. M.; LEIVA, I. G.; FUJITA, M. N. S. L. Indexação automatizada de artigos de periódicos científicos: análise da aplicação do software sisa com uso da terminologia DeCS na área de odontologia. **Informação & Sociedade***: Estudos*, v. 19, n. 2, p. 99-118, 2009.

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.

NICOLINO, M. E. V. P.; FERNEDA, E. Um método para a utilização de ontologias na indexação automática. **Informação & Tecnologia**, v. 1, n. 2, p. 13-33, 2014.

PINTO, V. B. Indexação documentária: uma forma de representação do conhecimento registrado. **Perspect. cienc. Inf***.,* Belo Horizonte, v. 6, n. 2, p. 223 – 234, jul./dez. 2001.

SILVA, E. M.; SOUZA, R. R. Fundamentos em processamento de linguagem natural: uma proposta para extração de bigramas. **Encontros Bibli: Revista Eletrônica de Biblioteconomia e Ciência da Informação,** v. 19, n. 40, 2014.

SILVA, T. J. **Indexação automática por meio da extração e seleção de sintagmas nominais em textos em língua portuguesa**. Recife, 2014.

SOUZA, R. R. **Uma proposta de metodologia para escolha automática de descritores utilizando sintagmas nominais.** 2005. 215 f. Tese (Doutorado) - Curso de Ciência da Informação, Escola de Ciência da Informação, Universidade Federal de Minas Gerais, Belo Horizonte, 2005.

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

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.