UNIVERSIDADE FEDERAL DE PERNAMBUCO
CENTRO DE ARTES E COMUNICAÇÃO
DEPARTAMENTO DE CIENCIA DA INFORMAÇÃO
CURSO DE GESTÃO DA INFORMAÇÃO

ROSANNE EVEN DE MELO DIAS

AVALIAÇÃO DOS SINTAGMAS NOMINAIS NA RECUPERAÇÃO DE TESES E DISSERTAÇÕES

RECIFE
2017

Rosanne Even De Melo Dias

AVALIAÇÃO DOS SINTAGMAS NOMINAIS NA RECUPERAÇÃO DE TESES E DISSERTAÇÕES

Trabalho de Conclusão de Curso do Curso de Gestão da Informação do Departamento de Ciência da Informação do Centro de Artes e Comunicação da Universidade Federal de Pernambuco como requisito para obtenção do Grau em Bacharel em Gestão da Informação.

Professor Orientador Dr: Renato Fernandes Correa

RECIFE
2017

Catalogação na fonte
Bibliotecário Jonas Lucas Vieira, CRB4-1204

D541a Dias, Rosanne Even de Melo

Avaliação dos sintagmas nominais na recuperação de teses e dissertações / Rosanne Even de Melo Dias. – Recife, 2017.

77 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.

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

FOLHA DE APROVAÇÃO

Título do TCC

AVALIAÇÃO DOS SINTAGMAS NOMINAIS NA RECUPERAÇÃO DE TESES E DISSERTAÇÕES

Rosanne Even de Melo Dias
(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 05 de dezembro de 2017

Banca Examinadora:

Orientador – Renato Fernandes Corrêa
DCI/Universidade Federal de Pernambuco

Examinador 1 – Márcia Ivo Bráz
DCI/Universidade Federal de Pernambuco

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

AGRADECIMENTOS

Primeiramente a Deus, por ter alcançado meus objetivos, e aos meus familiares pelo apoio.

Agradeço ao meu orientador, professor Renato Fernandes Corrêa, pela confiança que me foi dada, pelas correções e incentivos no tempo que lhe coube conduzindo este trabalho.

E a todos que direta ou indiretamente fizeram parte da minha formação durante toda minha jornada na universidade, em especial à Brenna, Nathally, Sthéfanie, o meu muito obrigado.

RESUMO

Este trabalho avalia os sintagmas nominais na recuperação de teses e dissertações no sistema denominado Mapeador Temático de Teses e Dissertações (MTTD). Nesse contexto, medidas para a avaliação de Sistema de Recuperação da Informação (SRI) foram utilizadas, como a mensuração dos coeficientes de revocação e precisão, para expressões de busca envolvendo sintagmas nominais expressos por palavras isoladas e frases, respectivamente. Considerando a hipótese de que os sintagmas nominais consistem numa melhor unidade de conhecimento para a indexação e recuperação de informação que as palavras isoladas, os sintagmas nominais são avaliados através de um experimento empírico no qual é realizada a comparação dos resultados de buscas no MTTD sobre os polos econômicos do estado de Pernambuco, através de expressões de busca constituídas por sintagmas nominais expressos como frases e palavras isoladas. Devido os sintagmas nominais serem melhores descritores de assunto, a recuperação da informação baseada em sintagmas nominais permitiu um mapeamento preciso das temáticas ou assuntos tratados nas teses e dissertações da UFPE, aumentando a eficácia da recuperação da informação e diminuindo a sobrecarga de informação. Os resultados são apresentados através de tabelas e gráfico, indicando o desempenho dos dois métodos de busca, recuperação por sintagmas nominais através de frases e por palavras isoladas, na busca por informação relevante sobre os polos de confecção, gesseiro e de fruticultura, em teses e dissertações. Foram obtidos diferentes resultados nos dois métodos de recuperação, mas o melhor índice de precisão nas teses e dissertações foi obtido através da recuperação por sintagmas nominais expressos por frase.

Palavras-chave: Sintagmas nominais. Recuperação da Informação. Teses e Dissertações.

ABSTRACT

This work evaluates the noun phrases in the retrieval of theses and dissertations in the system called the Thematic Mapper of Theses and Dissertations (MTTD). In this context, measures for evaluation of Information Retrieval System (SRI) such as recall and precision coefficients, were used for search expressions involving noun phrases expressed by isolated words and phrase respectively. Considering the hypothesis that the noun phrases consist of a better unit of knowledge for indexing and retrieval of information than the isolated words, the noun phrases are evaluated through an empirical experiment that search are performed by nominal phrases expressed as phrases and isolated words about economic poles of the state of Pernambuco, and the search results are compared. Because the noun phrases are better subject descriptors, the retrieval of the information based on the noun phrases allowed a precise mapping of the topics or subjects treated in the theses and dissertations, increasing the effectiveness of information retrieval and reducing information overload. The results are presented through tables and graph, indicating the performance of the two methods of search, by noun phrases through phrases and isolated words, economic poles of clothing, gypsum and fruticulture, in the search for theses and dissertations. Both obtaining different results in retrieval, but better index of precision in retrieval of theses and dissertations is get searching by noun phrases expressed by phrase.

Keywords: Noun phrases.Information Retrieval. Theses and Dissertations.

LISTA DE ILUSTRAÇÕES

Figura 1 - Procedimentos de interação usuário−protótipo 46

Figura 2 - Detalhes da tela inicial (Mapa) 51

LISTA DE TABELAS

Tabela 1 - Expressões de busca por sintagmas nominais na forma de frase 56

Tabela 2 - Expressões de busca por sintagmas nominais na forma de palavras isoladas 57

Tabela 3 - Expressões de busca por sintagmas nominais na forma de frase 59

Tabela 4 - Expressões de busca por sintagmas nominais na forma de palavras isoladas 60

Tabela 5 - Expressões de busca por sintagmas nominais na forma de frase 62

Tabela 6 - Expressões de busca por sintagmas nominais na forma de palavras isoladas 63

Tabela 7 - Quantidade de documentos relevantes por sintagmas nominais 65

LISTA DE QUADROS

Quadro 1- Sintagmas nominais para a recuperação de informação sobre os polos. 55

LISTA DE GRÁFICOS

Gráfico 1- Avaliação da revocação e precisão na recuperação de informação por sintagmas nominais via frases ou palavras isoladas no MTTD. 66

LISTA DE SIGLAS

BDTD- Biblioteca Digital de Teses e Dissertações

CI- Ciência da Informação

MTTD- Mapeamento Temático de Teses e Dissertações

PI- Palavras Isoladas

RI - Recuperação da Informação

RNA- Rede Neural Artificial

SN- Sintagma Nominal

SRI - Sistema de Recuperação da Informação

TI - Tecnologias de Informação

UFPE- Universidade Federal de Pernambuco

SUMÁRIO

1. INTRODUÇÃO 13
2. RECUPERAÇÃO DA INFORMAÇÃO 18
2.1. Indexação intelectual e automática 24
2.2. Expressão de busca 29
2.3. Avaliação de sistemas de recuperação de informação 32
3. SINTAGMAS NOMINAIS NA RECUPERAÇÃO DA INFORMAÇÃO 38
3.1 Indexação automática por sintagmas nominais 42
3.2. Recuperação de informação por sintagmas nominais 45
3.3 Recuperação de teses e dissertações no MTTD 48
4. METODOLOGIA 52
5. ANÁLISE DE RESULTADOS 55
5.1 Elaboração de expressão de busca 55
5.2 Recuperação de teses e dissertações no sistema MTTD por sintagmas nominais na forma de frase e na forma de palavras isoladas sobre os polos 56
5.4 Avaliação e comparação dos resultados obtidos por sintagmas nominais expressos por frases (frases) ou por palavras isoladas (PI) sobre os polos de acordo com as medidas de revocação e precisão 66
6. CONCLUSÃO 68
REFERÊNCIAS 70

1 INTRODUÇÃO

Segundo Oliveira (2011), a Ciência da Informação (CI) nasceu para resolver um grande problema, que é o de reunir, organizar e tornar acessível o conhecimento cultural, científico e tecnológico produzido em todo o mundo. Desde após a Segunda Guerra Mundial, marcada pelo aumento considerável de conhecimento no campo da ciência e tecnologia, fenômeno conhecido como a explosão de informação, a CI visa lidar com a disseminação de todo conhecimento registrado. Este crescimento exponencial da massa documental, que vem ocorrendo em proporções geométricas, trouxe preocupações quanto à organização, guarda, e a recuperação das informações disponibilizadas para a sociedade.

Sendo assim, a CI surgiu no bojo da revolução cientifica e técnica (OLIVEIRA, 2011), na qual, novas tecnologias foram surgindo e novos métodos foram adotados com o propósito de resolver os problemas informacionais existentes na tarefa de recuperação de informação.

A Recuperação da Informação (RI) se constituiu como uma área de pesquisa em 1950, e envolve a aplicação de métodos computacionais no tratamento da informação e na sua recuperação. O componente tecnológico, principalmente a “tecnologia da computação”, é visto por alguns autores como resultado da interdisciplinaridade da ciência da Informação. Segundo Saracevic (1996, p.47):

> A Ciência da informação é um campo dedicado às questões científicas e à prática profissional voltadas para os problemas da efetiva comunicação do conhecimento e de seus registros entre os seres humanos, no contexto social, institucional ou individual do uso e das necessidades de informação. No tratamento destas questões são consideradas de particular interesse as vantagens das modernas tecnologias informacionais.

Saracevic (1996) afirma que o desenvolvimento da relação entre CI e comunicação apresentam dimensões em interesse compartilhado na comunicação humana, juntamente com a crescente compreensão de que a informação como fenômeno e a comunicação como processo devem ser estudadas em conjunto.

Assim, a natureza interdisciplinar da Ciência da informação segundo Saracevic (1996), possibilita o surgimento de diferentes correntes em diversas áreas, havendo uma sistematização da ciência da informação. A Recuperação da

Informação se tornou para a ciência da informação uma área de pesquisa com um rápido desenvolvimento.

Segundo Saracevic (1996), a Recuperação da Informação pode ser considerada a vertente tecnológica da Ciência da Informação e é resultado da relação desta com a Ciência da Computação. Os processos desenvolvidos da ciência da informação estão relacionados a vários processos da recuperação da informação. Já a relação da Ciência da informação com Ciência da computação está na aplicação dos computadores e da computação na recuperação da informação, assim como nos produtos, serviços e redes associados (SARACEVIC,1996).

Sendo assim, pode-se afirmar que de acordo com Saracevic (1996), a Recuperação da Informação foi responsável pelo desenvolvimento de inúmeras aplicações bem sucedidas, pois, devido aos avanços tecnológicos, os computadores se tornaram instrumentos básicos para processar e fornecer informações, cabendo ao usuário o uso de seus diversos e distintos recursos, entre eles, os voltados para recuperação da informação.

Os métodos utilizados para recuperar a informação têm como unidade básica o uso da palavra. Segundo Probst, Raub e Romhardt (2000), esse método é importante para o usuário, para o conhecimento e uso de formas alternativas de expressar sua necessidade de informação no ato de recuperar informações.

A satisfação das necessidades de informação dos usuários diante de um grande volume de dados é essencial em se tratando da recuperação da informação. Os sintagmas nominais (SNs) por serem considerados melhores descritores de assuntos ou temas, podem contribuir para o alcance desse requisito num sistema de recuperação da informação.

Segundo Kuramoto (2002. p. 6) “O sintagma nominal é a menor parte do discurso portadora de informação”, em outras palavras, os SNs constituem dentro de uma oração uma representatividade informacional, formando um conjunto de palavras que possuem significado, possuem um sentido.

A expressão de busca por meio de um termo composto faz uma enorme diferença no que é retornado para o usuário em detrimento da busca por palavras

isoladas. Os SNs por conterem termos de uma linguagem de especialidade, podem ser usados para expressar as necessidades de informação e permitir maior precisão no encontro dos documentos relevantes.

Entretanto, nem todo SN que se encontra em um texto tem potencial para representar o conteúdo temático desse texto. Segundo Oliveira (2011. p. 21) “Nem todos os SNs possuem poder discriminante suficiente para representar o conteúdo informacional de um documento”. Uma vez que nem todo sintagma nominal tem seu valor como descritor, e para que ele possa funcionar como descritor documental, o processo de indexação é importante para uma melhor recuperação da informação.

Corrêa et.al. (2011) afirmam que a extração dos SNs:

> [...]não garante por si só a seleção de bons descritores, sendo necessário que a ferramenta de extração de sintagmas nominais possa fazer a análise dos textos e pontuar os sintagmas nominais com a potencialidade de serem bons descritores [...] (CORRÊA et al., 2011, p. 11).

Nesse contexto é perceptível a necessidade de uma avaliação dos SNs objetivando, mais especificamente, a recuperação de informação. A avaliação é definida como um processo que identifica e coleta dados sobre serviços e atividades específicas, estabelecendo critérios e possibilitando que o sucesso seja analisado, assim é determinado a qualidade do serviço e o grau com que o serviço atingiu o objetivo e metas traçadas. (Hernon; Maclure, 1993 apud HARTER; HERT, 1997).

Portanto, este trabalho tem como objetivo analisar a eficácia do uso de sintagmas nominais como expressão de busca para recuperar teses e dissertações da UFPE, depositadas na Biblioteca Digital de Teses e Dissertações (BDTD), avaliando os sintagmas nominais expressos por frase ou por palavras isoladas quanto à precisão na definição do assunto buscado, e na revocação de documentos relevantes à necessidade de informação dos usuários que buscam teses e dissertações sobre polos econômicos do estado de Pernambuco. Será utilizado o Mapeador Temático de Teses e Dissertações (MTTD), como sistema de busca.

A avaliação da eficácia do uso dos SNs como expressão de busca, são consideradas teses e dissertações que tratem, parcialmente ou totalmente, o assunto buscado.

O Mapeador Temático de Teses e Dissertações (MTTD) foi construído para prover a recuperação da informação por meio dos sintagmas nominais em teses e dissertações depositadas na BDTD-UFPE. O MTTD será usado como ambiente para a avaliação do uso dos SNs na recuperação da Informação.

Desta forma são vislumbrados os seguintes objetivos específicos:

1. Diferenciar a recuperação de informação por meio das palavras isoladas e dos sintagmas nominais;
2. Desenvolver expressões de busca utilizando sintagmas nominais para recuperar teses e dissertações sobre os polos econômicos do estado de Pernambuco, mais especificamente os polos de confecção, gesseiro e de fruticultura;
3. Recuperar Teses e Dissertações pelo MTTD usando os sintagmas nominais expressos em frase ou palavras isoladas como expressões de busca para a recuperação dos polos econômicos do estado de Pernambuco;
4. Avaliar a eficácia do uso dos sintagmas nominais na recuperação da informação comparando a precisão do sistema de recuperação de informação em função dos documentos retornados.

A BDTD-UFPE tem como propósito organizar e disponibilizar toda coleção produzida que cresce no ritmo muito acelerado devido a novos materiais incorporados ao final de cada turma de mestrado e doutorado, assim também aumentando o volume de conteúdo considerado rico em cultura, ciência e tecnologia.

A Biblioteca Digital de Teses e Dissertações da Universidade Federal de Pernambuco (BDTD-UFPE) é um repositório organizado em uma comunidade no repositório institucional da UFPE. De modo que, os alunos de doutorado e mestrado depositam na BDTD seu conhecimento científico e tecnológico em diversas áreas, portanto, se faz necessário que as potencialidades de acesso sejam efetivamente exploradas com maior precisão na recuperação das informações.

Para a literatura científica, teses e dissertações são importantes, pois mostram a trajetória dos pesquisadores quanto à configuração do campo em

períodos específicos ou ao longo da carreira. Os doutores credenciados desenvolvem atividades de pesquisa e possuem domínio de campos específicos da área de conhecimentos. A disponibilização de bases de dados de teses e dissertações na Internet tornou mais perceptível esses documentos.

2 RECUPERAÇÃO DA INFORMAÇÃO

Segundo Ferneda (2003), a Recuperação da informação (RI) se firmou como uma área de pesquisa autônoma no seio da Ciência da Informação, com um acelerado desenvolvimento, e se estabeleceu como área de pesquisa em 1951.

> A recuperação é o ato de investigar ou explorar com o fim de tornar a encontrar algo perdido. Para realizar tal procedimento é usado processo que compreende o arranjo ordenado dos registros de conhecimentos. (KENT, 1972, p. 23)

A organização dos dados é fundamental na recuperação da informação, a organização prévia de maneira criteriosa abrange um tratamento técnico na informação e é a partir dela que as informações são selecionadas, localizadas e recuperadas. Entretanto, Kent (1972) explica que processo mecânico apenas facilita o acesso para os futuros usuários e que a recuperação é na verdade a pesquisa dos papéis escritos realizada pelas máquinas.

De acordo com Calvin Mooers (1951 apud SARACEVIC, 1996, p. 44), o termo recuperar informação “engloba os aspectos intelectuais de descrição de informações e suas especificidades para a busca, além de quaisquer sistemas, técnicas ou máquinas empregados para o desempenho da operação”.

Para Choo (2006), recuperar uma informação é disponibilizá-la ao usuário / consulente, que a solicita por necessidades espontâneas e/ou induzidas, objetivando construir significado, produzir novo conhecimento e tomar decisões, podendo ser administrativas ou pessoais. Choo (2006) acrescenta que a informação recuperada pelo usuário, além de suprir necessidades informacionais, também tem como finalidade “construir significado” e “produzir novo conhecimento”. De modo, que seja possível facilitar o acesso à informação e melhorar a precisão do resultado de uma busca ou consulta, tendo em base o uso da palavra, com o auxílio de alguns modelos de recuperação da informação.

O processo de recuperação consiste na geração de uma lista de documentos recuperados para responder a consulta formulada pelo usuário. Os componentes do sistema incluem documentos, a consulta formulada, e finalmente o processo de recuperação que, a partir das estruturas de dados e da consulta

formulada, recupera uma lista de documentos considerados relevantes (BARION e LAGO, 2008).

A indexação tem uma extrema importância aos documentos que se pretende recuperar, tanto a indexação automática, como também a realizada pelos indexadores.

Através da indexação é possível se obter um aproveitamento melhor no processo de busca e recuperação da informação, uma vez que o elemento principal/fundamental estabelecido é a representação do conteúdo dos documentos (ARAÚJO JÚNIOR, 2007).

Os descritores, independente da maneira que forem extraídos ou atribuídos aos documentos, devem fazer referência a objetos, pois não devem ser só símbolos sem referência como são as palavras, esses elementos devem constituir-se de unidades extraídas do discurso, que possa servir de base a uma relação referencial autônoma. O sintagma nominal “é a menor parte do discurso portadora de informação” (KURAMOTO, 1995), que ao contrário das palavras não são símbolos sem referência, pois são portadores de uma estrutura lógico-semântica.

Nas palavras de Kuramoto (2002), a busca de informação em bases de dados contendo documentos textuais se dá tradicionalmente por cada palavra isolada como ponte de acesso aos documentos que a contem.

Segundo Kuramoto (2002), o processo de indexação extrai cada palavra do texto de um documento e o insere numa lista de palavras ordenadas, de forma a facilitar a recuperação da informação. As palavras que se inserem no texto de um documento assumem um significado específico, pois o autor exprime suas ideias através de uma combinação de um conjunto de palavras, as palavras enquanto unidades de um dicionário possuem um conjunto de predicados, não fazem referência a um objeto ou fato do mundo real.

A cada documento que atende a uma consulta é atribuído um grau de similaridade, possibilitando ao sistema de recuperação de informação apresentar resultados de uma consulta de forma ordenada, normalmente de forma decrescente dos respectivos graus de similaridade, trata-se de uma maneira de evidenciar a relevância de cada documento em relação a uma dada consulta (KUMAROTO, 2002).

A maior parte dos Sistemas de Recuperação da Informação utilizam como base dois tipos de modelos para tratar dos documentos (BAEZA-YATES e RIBEIRO NETO, 1999): Modelo Clássico - cada documento é descrito por um conjunto de palavras-chave representativas, também chamadas de termos de indexação, que buscam representar o assunto do documento e resumir seu conteúdo de forma significativa; Modelo Estruturado - podem-se especificar, além das palavras-chave, algumas informações acerca da estrutura do texto. Essas informações podem ser as seções a serem pesquisadas, fontes de letras, proximidade das palavras, entre outras características.

Disciplinas como lógica, estatística e teoria dos conjuntos, são usadas nos modelos de recuperação da informação por serem de natureza quantitativa. Segundo Robertson (1977), esclarece o predomínio pelo fato de que a determinação de um modelo matemático geralmente pressupõe uma cuidadosa análise formal do problema e especificações de hipóteses, além de uma formulação explícita da forma de como o modelo depende das hipóteses.

Os “modelos clássicos” de recuperação de informação comportam propostas que serviram de base para o desenvolvimento de diversos outros modelos e algumas técnicas. São eles: modelo booleano, modelo espaço vetorial e o modelo probabilístico (FERNEDA, 2012).

Entre as técnicas de busca, o modelo booleano se destaca dentre os modelos clássicos. O modelo booleano é baseado na teoria dos conjuntos, é um modelo de recuperação simples baseado na teoria da Álgebra Booleana (BAEZA e RIBEIRO, 1999), e possui consultas especificadas com termos em expressões booleanas. Nas consultas são utilizados operadores lógicos como E, OU, NÃO para filtragem do resultado.

Segundo Baeza e Ribeiro (1999), o modelo booleano fornece uma estrutura de fácil compreensão para o usuário comum de um sistema de recuperação de informação. As consultas são estabelecidas como expressões booleanas com semânticas precisas. Além disto, um documento pode ser considerado relevante ou não relevante a uma consulta. Não existe resultado parcial e não há informação que permita a ordenação do resultado da consulta.

O Modelo Booleano é um dos mais utilizados nos sistemas de recuperação de informação, por sua simplicidade e formalismo claro, o que o torna mais facilmente implementável. (KURAMOTO, 2002). De acordo com Baeza-Yates e Ribeiro-Neto (1999), apesar de ser um modelo bastante simples e muito utilizado ele apresenta as seguintes desvantagens:

- A recuperação é baseada numa decisão binária sem noção de combinação (casamento) parcial;
- Nenhuma ordenação de documentos é fornecida;
- A passagem da necessidade de informação do usuário para a especificação expressão booleana é considerada complicada;
- As consultas booleanas formuladas pelos usuários são frequentemente simplistas;
- Em consequência, o modelo booleano permite retorno de poucos ou muitos documentos em resposta às consultas;
- O uso de pesos binários é limitante;

Já o modelo vetorial pressupõe um ambiente no qual é possível obter documentos que respondem parcialmente a uma expressão de busca, também é conhecido como Modelo Espaço Vetorial e “baseia-se na comparação parcial entre a representação dos documentos e a da consulta do usuário” (KURAMOTO, 2002). Ou seja, o que somente é possível através da associação de pesos tanto aos termos de indexação, como aos termos de expressão de busca. Esses pesos são utilizados para calcular o grau de similaridade entre a expressão de busca formulada, pelo usuário e cada um dos documentos do corpus.

Segundo Salton (1983), um documento é representado por um conjunto de termos de indexação, cada qual associado a um valor numérico entre 0 e 1, que representa a relevância do respectivo termo na representação do conteúdo informacional do documento.

Conforme Baeza-Yates e Ribeiro-Neto (1999), uma expressão de busca é também representada por um conjunto de termos e seus respectivos pesos, que representam a importância do termo na expressão de busca.

A homogeneidade na forma das representações dos documentos nas expressões de busca permite criar um sistema do qual é possível calcular o “grau de similaridade” entre uma expressão de busca e cada um dos documentos do corpus, e verificar a semelhança entre dois documentos. Como resultado é obtido um conjunto de documentos ordenado pelo grau de similaridade entre a expressão de busca do usuário e cada um dos documentos do corpus em relação à expressão de busca (FERNEDA, 2012).

Uma característica do modelo vetorial é que os termos de indexação são independentes, não são considerados os relacionamentos existentes entre eles. Alguns autores apontam essa característica como desvantagem, segundo Baeza-Yates e Ribeiro-Neto (1999, p.30), declaram que não há evidências conclusivas que apontem que tais dependências afetam significativamente o desempenho de um SRI. No modelo vetorial é importante não permitir a formulação de buscas booleanas, para Ferneda (2003) restringe consideravelmente sua flexibilidade.

O modelo vetorial é capaz de definir um dos componentes essenciais de qualquer teoria científica: um modelo conceitual. Segundo Salton (1971), este modelo serviu como base para o desenvolvimento de uma teoria que alimentou uma grande quantidade de pesquisas e resultou no sistema System for the Manipulation and Retrieval of Text (SMART).

No modelo vetorial o conceito de relevância é tratado como um *continuum* representado numericamente por meio de um número real entre zero e um. Característica que permitiu o desenvolvimento de diversas técnicas de recuperação de informação utilizadas até hoje, tais como *clustering* (agrupamento), relevance feedback, classificação, reformulação da expressão de busca etc., que foram materializadas no sistema SMART desenvolvido por Salton nos anos 60 (FERNEDA, 2012).

O modelo probabilístico se pauta na teoria matemática das probabilidades. O modelo foi proposto inicialmente por Maron e Kuhns (1960) e posteriormente explorado por diversos outros pesquisadores. Souza (2006), afirma que esse modelo supõe que exista um conjunto ideal de documentos que atende a cada uma das possíveis buscas que podem ser feitas no sistema.

A partir de uma expressão de busca, o usuário expressa sua necessidade de informação e a submete ao sistema. Por meio de cálculos de probabilidade o sistema calcula, para cada documento do corpus, um valor numérico (similaridade), que representa a provável relevância do documento para a consulta. Esse valor é utilizado para ordenar os resultados da busca (FERNEDA; DIAS, 2013a).

No primeiro conjunto de documentos resultantes de uma busca são marcados alguns deles que são considerados verdadeiramente relevantes de acordo com sua necessidade. A expressão de busca, juntamente com os documentos que foram selecionados como relevantes, é submetida novamente ao sistema de informação, permitindo fornecer resultados mais precisos e procurando refinar a busca e tentando aproximar-se cada vez mais do conjunto ideal de documentos (FERNEDA; DIAS, 2013b).

Segundo Baeza-Yates e Ribeiro-Neto (1999) e Gonzalez (2000), o modelo probabilístico reconhece a tarefa do usuário, que é a atribuição de relevância, pois é o único modelo que incorpora explicitamente o processo de relevance feedback como base para a sua operacionalização. Esse processo pode ser repetido até que o usuário se sinta satisfeito com os resultados.

Segundo Ferneda (2012, p. 52):

> O processo de recuperação de informação é caracterizado por seu grau de incerteza no julgamento de relevância dos documentos em relação à expressão de busca. Assim sendo, é mais realístico pensar em uma probabilidade de relevância do que em uma pretensa relevância exata, como a utilizada nos modelos booleano e vetorial.

A relevância de um documento pode ser definida de forma exata, representado por um valor numérico, assim como definido nos três modelos clássicos. O processo de recuperação de informação é inerentemente impreciso. A modelagem matemática desse processo é somente possível através de simplificações teóricas e da adequação de conceitos tipicamente subjetivos como “necessidade de informação”, “relevância”, além do próprio conceito de “informação” (FERNEDA; DIAS, 2013a).

2.1 Indexação intelectual e automática

Segundo Fujita (2009), o termo indexação (indexing) pertence à corrente teórica inglesa e é 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. Trata-se da identificação do conteúdo do documento por meio de um processo de análise de assunto e a representação desse conteúdo por meio de conceitos, sua função é representar o assunto dos documentos através da elaboração de termos, mostrando aos usuários de uma forma geral, que assuntos, estão tratados nos documentos indexados (FUJITA, 2003).

Além disto, a indexação tem por finalidade representar da melhor maneira o documento a fim de que os usuários possam recuperar a informação através de um sistema de recuperação da informação.

Lancaster (2003) acredita que a representação da informação influencia no índice/coeficiente de recuperabilidade dos documentos. Não basta apenas fornecer informação, existem fatores necessários para que a informação seja encontrada pelo indivíduo que dela necessita e busca. A indexação tem como objetivo representar o conteúdo dos documentos para fins de organização e recuperação desses documentos.

Maia (2008, p. 27) define o objetivo primordial da indexação como:

> O processo de indexação produzindo uma lista de descritores visa à representação dos conteúdos dos documentos. Ou seja, esse processo tem como objetivo extrair as informações contidas nos documentos, organizando-as para permitir a recuperação destes últimos. Contudo, na maioria dos sistemas convencionais de recuperação de informação, os descritores não passam de uma simples lista de palavras extraídas dos documentos, que constituem a coleção.

A fim de tornar disponíveis recursos informacionais com base em seus assuntos em unidades de Informação, a indexação é essencial, principalmente no ambiente virtual, onde uma maior quantidade de informações é disponibilizada, o que acarreta em uma necessidade maior de procedimentos de organização dessas informações.

Segundo Chaumier (1998, p. 63) indexação é como uma “operação que consiste em descrever e caracterizar um documento, com o auxilio da representação dos conceitos nela contidos”. Ou seja, a indexação é um processo de análise

documentária que identifica o assunto que trata o documento para representá-los através de descritores de uma linguagem documentária com a finalidade de permitir a sua recuperação por meio de um sistema de informação.

Essa operação acontece de três maneiras, quando a indexação é feita manualmente, denominada na literatura como indexação manual, por um programa de computador, denominado de indexação automática, e a semi-automática onde um indexador humano revisa ou valida os termos propostos por um programa de computador (ROBREDO, 1982).

Com relação ao processo manual de indexação é realizada uma leitura documentária para identificar e selecionar os conceitos expressos no documento, a seguir, é representado, “traduzido”, esses conceitos selecionados em descritores da linguagem documentária adotada pelo sistema de informação (BOCCATO, 2005).

Segundo Lancaster (2004, p.9) o processo manual de indexação é constituído de duas etapas: análise conceitual e tradução. Onde a análise conceitual (identificação do conteúdo do documento), a seleção dos conceitos representativos desse conteúdo (nível de abordagem) e a tradução documental (da linguagem natural para o cunho artificial).

A etapa da análise conceitual objetiva determinar do que trata um documento, isto é, qual seu assunto. E a etapa de tradução objetiva, converter o conteúdo do documento determinado na etapa de análise conceitual, em um conjunto de termos de indexação.

“A indexação manual vem se revelando inadequada para minimizar a subjetividade inerente à indexação, além de ser caracterizada como um processo relativamente moroso e caro” (NASCIMENTO, 2008 p.24).

Nascimento (2008) cita vários fatores que podem ser apontados como a causa da inadequação da indexação manual, que esclarece o porquê de muitos pesquisadores se interessarem pela indexação automática, muito abordada e estudada pela Ciência da Informação. Entre os fatores destacam-se:

a) o conhecimento que o indexador tem sobre o assunto indexado determina o grau de consistência atingido;
b) a dinamicidade do conhecimento, que exige do indexador permanente atualização;

c) a inconsistência (diferentes indexadores atribuindo diferentes termos-índice a um mesmo conceito/documento e o mesmo indexador atribuindo diferentes termos-índice a um mesmo conceito/documento, em diferentes momentos);
d) a possibilidade do indexador não dominar o idioma do documento também é fator que prejudica a qualidade da indexação.

De acordo Vieira (1988) a indexação automática da informação é aquela realizada diretamente por sistemas de computador, que analisam, reconhecem e constroem índices para a recuperação do texto em pesquisas.

Na concepção de Santos e Ribeiro (2003) na indexação automática, um programa de computador, adotando critérios de frequências, extrai palavras, expressões ou radicais de palavras do texto para representar o seu conteúdo como todo.

"A indexação automática é qualquer procedimento que permita identificar e selecionar os termos que representem o conteúdo dos documentos, sem a intervenção direta do indexador" (ROBREDO, 2003, p. 96). Assim, a indexação é vista como ferramenta essencial nas unidades de informação, pois consiste no ato de identificar e descrever um texto informacional de acordo com o seu assunto, e cujo principal objetivo é orientar o usuário sobre esse conteúdo intelectual, permitindo, dessa forma, a sua recuperação de forma rápida e eficiente.

Lancaster (2004, p. 286-290) define dois tipos de diferentes de indexação automática. A indexação por extração automática e a indexação por atribuição automática.

Na indexação por extração automática, palavras ou expressões que aparecem no texto são extraídas, por um programa de computador, e utilizadas para representar o conteúdo do texto com um todo, adotando critérios de frequências, posição e contexto. (LANCASTER, 2004, p. 286). São contadas as palavras do texto pelo programa, a fim de eliminar palavras não significantes (artigos, preposições, conjunções etc.), para isso o programa tem que ser comparado com uma lista de palavras proibidas, e em seguida ordenar essas palavras segundo a frequência de sua ocorrência. Ou seja, as palavras com maior número de ocorrências são escolhidas para descrever os documentos.

Na indexação por atribuição automática é necessário desenvolver, para cada termo a ser atribuído, um perfil de palavras ou expressões que costumam ocorrer frequentemente nos documentos às quais um indexador humano atribuiria esse termo. Se a cada termo de um vocabulário controlado correspondesse um perfil desses, seria possível utilizar programas de computador para selecionar as expressões importantes num documento (essencialmente aquelas que fossem extraídas segundo critérios de frequência) com essa coleção de perfis, atribuindo um termo ao documento quando coincidisse com o perfil do termo (LANCASTER, 2004, p. 289).

Lancaster (2004, p. 312) considera que apesar da indexação automática não alcançar o nível de desempenho obtido pelos indexadores humanos, esse tipo de processo poderá reduzir a carga de trabalho desses indexadores ao realizar uma atribuição preliminar.

Na indexação semi-automática segundo Pinto (2001, p. 227), seria a combinação da indexação manual com a indexação automática. Inicialmente, o sistema realiza uma indexação automática dos documentos levando em conta as ocorrências das palavras mais frequentes num texto. Em um segundo momento, o indexador humano refina a lista dos descritores propostos pelo sistema fazendo os ajustes e/ou complementações necessárias. Os sistemas analisam os documentos de modo automático, mas os termos de indexação propostos são validados e editados por um profissional (indexação semi-automática).

As técnicas de indexação estão diretamente ligadas às questões relativas de precisão dos resultados dos sistemas de recuperação de informação. Nessa perspectiva, Cesarino (1985, p. 157) afirma que:

> Os sistemas de recuperação da informação podem ser definidos como um conjunto de operações consecutivas para localizar, dentro de uma totalidade de informações disponíveis, aquelas realmente relevantes. Para isso, executam-se as funções de seleção, análise, indexação e busca das informações.

Nesse sentido, a indexação assume uma posição crucial no âmbito da recuperação da informação. É um fator que auxilia a comunicação entre o sistema de recuperação da informação e o usuário que deseja satisfazer sua necessidade informacional. Para isso, é necessário que todo saber seja representado por meio da indexação para facilitar a recuperação.

De acordo com Robredo (1978, p. 73), a indexação é como “uma operação que permite representar o conteúdo de um documento, considerado como essencial, da maneira mais condensada possível, [...] com a finalidade de classificação ou recuperação”.

Assim, os sistemas de recuperação possibilitam que o usuário possa recuperar aquilo que realmente deseja. Portanto, indexar significa incluir um documento em um repositório a partir do seu assunto determinando, com palavras representativas de seu conteúdo, tendo como ferramenta uma linguagem de indexação.

Esse prévio procedimento que os documentos recebem antes de serem armazenados nas bases de dados, chamados de indexação automática é de suma importância, pois permite a extração dos descritores e sua estruturação com vistas a um acesso rápido às informações.

De acordo com Kumaroto (1995), as palavras, quando são extraídas de um documento, perdem valores atribuídos pela autoria do documento, o que leva a perda de qualquer referência da realidade extralinguística do autor. Sendo assim, para que isso não aconteça, é necessário que os documentos sejam contextualizados por descritores que melhor representem a informação sem descaracterizá-la.

O processo de indexação produzindo uma lista de descritores visa à representação dos conteúdos dos documentos (KURAMOTO, 1995). Esse processo busca extrair informações no documento, organizando para permitir a recuperação. Portanto, os descritores obrigatoriamente devem ser portadores de informação de maneira a relacionar um objeto da realidade extralinguística com o documento que traz informações sobre o objeto.

As tecnologias da informação no desenvolvimento de mecanismos que auxiliem as pessoas na busca por uma informação precisa, estão cada vez mais sendo abordada na comunidade científica. Portanto, criar novas ferramentas que possibilitem a realização da busca de um usuário e que esse se satisfaça é o objetivo dos estudiosos do objeto informação.

Sendo assim, para Lancaster (2004), a indexação é um processo com duas direções: de um lado os documentos e de outro, as necessidades de informação dos usuários. O surgimento e o desenvolvimento da indexação automática esta

relacionada com a tentativa de solucionar, ou talvez minimizar, alguns problemas encontrados na indexação baseada em palavras isoladas e também como uma forma de indexação que dê conta de todas as informações digitais que estão sendo produzidas no ambiente virtual.

Vieira (1988a) define a indexação como técnica de análise de conteúdo que possibilita a condensação da informação significativa de um documento por meio de termos, criando uma linguagem intermediária entre o usuário e o documento.

Já Borges (2009) conceitua o ato de indexar como atividade de selecionar ou definir palavras ou expressões que servirão como descritores de um conteúdo de um determinado documento, levando-se em conta as considerações da clientela específica.

2.2 Expressão de busca

Cada vez mais a internet está sendo considerada um recurso importante no desenvolvimento de pesquisas, de modo que as expressões de buscas se tornem meios para uma bem sucedida recuperação de informação.

Segundo Bertholino (1999, p. 151) “a formulação da estratégia de busca é fundamental para refinar a busca e poder obter resultados relevantes aos interesses do usuário”. A estratégia de busca foi conceituada por Bates (1987, 1988), como o “estudo da teoria, princípios e prática de planejar e executar táticas e estratégias de busca”, e esta foi a primeira autora a definir teoricamente o conceito de estratégia de busca e a tática para a sua execução.

A busca é a procura de informações, seja para solucionar problemas ou para adquirir novos conhecimentos. Para Choo (2006, p.99) “A busca da informação é o processo humano e social por meio do qual a informação se torna útil para um indivíduo ou grupo”.

Lancaster afirma que ao se realizar uma busca em uma base de dados, procura-se “encontrar documentos que sejam úteis para satisfazer a uma necessidade de informação, e evitar a recuperação de itens inúteis” (Lancaster, 2004, p. 3). Para autor o problema da recuperação da informação está em recuperar os itens relevantes como menor número de itens irrelevantes.

Para Baeza-Yates e Ribeiro-Neto (1999), alguns fatores dificultam uma solução para o problema de RI: (1) nem sempre o usuário consegue expressar o que quer de forma adequada, em virtude da construção da consulta ou da interpretação pelo sistema e, (2) a quantidade de documentos não relevantes pode ser, e geralmente é muito grande.

As expressões de busca surgem através de uma necessidade informacional, a partir de um problema/questionamento que precisa ser resolvido, quando o usuário reconhece que não tem informação, conhecimento ou compreensão suficiente para transpor a questão que deseja.

> As necessidades de informação são muitas vezes entendidas como as necessidades cognitivas de uma pessoa: falhas ou deficiências de conhecimento ou compreensão que podem ser expressas em perguntas ou tópicos colocados perante um sistema ou fonte de informação (CHOO 2006, p.99).

As informações são recuperadas das bases de dados, assim que, o usuário inicia uma relação de especificar o que se deseja, nessa interação ocorre à questão de como formular uma expressão de busca para assim concretizar as etapas da pesquisa do usuário. Rowley (1994, p. 129) menciona que a estratégia de busca é o "conjunto de decisões tomadas e de procedimentos adotados durante uma busca". Hartley et al. (1990, p. 153, tradução nossa) compartilha da mesma ideia e ainda acrescenta que a estratégia de busca é “[...] o conjunto total das decisões e ações tomadas durante todo o período de pesquisa, decisões que afetam os resultados em termos de itens recuperados e itens não recuperados”.

Portanto, como estratégia, o usuário inicia sua pesquisa delimitando o que procura expressando sua necessidade de forma que o sistema possa interpretá-lo e isso influência no resultado a ser recuperado.

De acordo com Lancaster (1979), a preparação da estratégia de busca envolve a análise e tradução dos conceitos. Num primeiro momento, é realizada a análise do pedido para determinar o que o usuário quer e após isso é realizada a tradução da análise conceitual para o vocabulário do sistema. Quando uma busca é realizada, o sistema compara os registros para encontrar quais termos pesquisados foram encontrados. Uma das formas do sistema fazer essa comparação é com o uso dos operadores booleanos (HARTLEY et al., 1990; ROWLEY, 2002).

Rowley (2002), afirma dizendo que os objetivos da formulação das estratégias de busca devem ser:

- Recuperar um número suficiente de registros relevantes;
- Evitar que sejam recuperados registros irrelevantes;
- Evitar recuperar um número excessivo de registros;
- Evitar recuperar um número insignificante de registros.

Segundo Rowley (2002), a lógica de buscas é utilizada para ligar os termos que descrevem os conceitos presentes no enunciado das buscas, permitindo a inclusão de todos os termos relacionados e as combinações aceitáveis e inaceitáveis de termos de busca. Os operadores lógicos booleanos são: E, OU, NÃO, além de suas variações como: E, NÃO.

A maioria dos mecanismos de busca utilizam dois níveis de especificação de expressão de busca: básico e avançado. Geralmente o nível básico utiliza janelas e menus que fazem a busca por lógica, buscas booleanas ou utiliza ainda a delimitação de frases utilizando aspas. O nível avançado oferece expressões booleanas mais complexas e também fornece recursos mais sofisticados.

Podem usar operadores de extensão que faz a busca em radicais de palavras, empregando um caractere indicativo de truncamento, um asterisco * por exemplo.

Segundo Rowley (2002), o tratamento mais importante é à direita, no qual são ignorados os caracteres situados à direita da sequência de caracteres. O truncamento à esquerda será útil nas situações onde ocorrem diversos prefixos. Este caractere instrui o sistema a fazer uma busca numa sequência de letras, independente dessa sequência formar ou não uma palavra completa.

Uma consulta em SRI expressa à necessidade do usuário por informação. Consulta é o nome dado a um comando que pode ser expresso em linguagem natural com a finalidade de especificar que informação deve ser recuperada (MEADOW, 2000).

Coulon e Kayser (1992) afirmam que podemos resumir um texto em palavras-chave e, através de uma expressão, identificar se um texto responde positiva ou negativamente a uma consulta.

2.3 Avaliação de sistemas de recuperação de informação

Pode-se avaliar um sistema de recuperação da informação (SRI) através de consultas que fazem parte de uma coleção de referência. Nesta coleção existe um conjunto de consultas e para cada consulta é fornecido um conjunto ideal de documentos resposta, criado por especialistas nos temas envolvidos. Ou seja, Segundo Salton (1983 apud FUJITA; LEIVA; 2015 p.52).

> Os sistemas de recuperação de informação consistem em um conjunto de itens de informação (documentos), um conjunto de petições de informação (perguntas) e algum mecanismo (comparação) para determinar quais documentos cumprem com as petições requeridas ao sistema.

Os sistemas de recuperação de informação, de acordo com Ferneda (2012, p. 13), têm por função “representar o conteúdo dos documentos do corpus e apresenta-los ao usuário de uma maneira que lhe permita uma rápida seleção dos itens que satisfazem total ou parcialmente a sua necessidade de informação [...]”.

Um sistema de recuperação de informação deve ser capaz de atender eficientemente as demandas e necessidades de seus usuários. Para verificar se este objetivo está sendo atingido, o SRI deve ser avaliado.

A literatura apresenta diversos modelos de avaliação de sistemas de informação, como o de Bailey e Pearson (1983), que tem como foco a satisfação do usuário, e o de Maia (2005), que desenvolveu um instrumento que mede o impacto causado pela adoção de um novo sistema informatizado sobre o trabalho e o desempenho dos funcionários.

A fundamentação para todo o trabalho de avaliação dos sistemas de recuperação da informação tem como base estudos nos anos cinquenta e sessenta. As referências fundamentais na avaliação dos SRIs são célebres Projetos de Cranfield I e II, desenvolvidos por Cyril Cleverdon entre 1957-1963. O Cranfield I foi utilizado uma coleção de 18 mil documentos em engenharia aeronáutica, essa coleção foi indexada usando-se 4 sistemas de indexação a serem comparados em sua eficiência de recuperação. Onde um conjunto de 1.200 perguntas de busca foi criado em base de documentos-fonte, foram analisados para se identificar se a causa do insucesso eram problemas relativos à formulação da pergunta de busca, à indexação, à busca ou ao sistema. A segunda etapa de testes, iniciada em 1963,

ficou conhecida como Cranfield II. Robertson (2008, p. 3) considera que embora o foco ainda estivesse em 'línguas de indexação' a transição de Cranfield I para Cranfield II, foi um grande salto, tanto em termos de metodologia quanto de conteúdo. Seu maior objetivo era investigar os componentes das linguagens de indexação e seus efeitos na performance dos SRIs (LANCASTER, 1979, p. 275).

O Projeto SMART, concebido por Gerard Salton entre 1965 e 1968, que devido o centro das preocupações da pesquisa estatística e probabilística ser o desenvolvimento de técnicas para a indexação, classificação e elaboração automática de resumos, bem como a busca automática, o sistema SMART foi o primeiro sistema em que esse tipo de pesquisa foi testado, essa tradição de pesquisa passou a ser intensivamente continuada nas pesquisas para a melhoria da recuperação da informação da internet, através dos mecanismos de busca.

E a avaliação do Sistema MEDLARS (Medical Literature Analysis and Retrieval System), levada a efeito por F. W. Lancaster, nos anos de 1966 e 1967. De acordo com Robertson (2008, p. 6) foi um dos primeiros sistemas de recuperação de informação baseado em computador e a pesquisa era realizada através de termos de indexação da linguagem documentária na área da Saúde - o MESH (Medical Subject Heading).

Esse experimento desenvolvido na National Library of Medicine (NLM), teve como objetivo avaliar a própria linguagem de indexação e os métodos e procedimentos utilizados para indexar documentos e formular perguntas de busca. Os usuários eram convidados a participar da pesquisa, a partir de suas necessidades reais de informação (ROBERTSON, 2008, p. 6).

Os aspectos mais significativos dos estudos de avaliação realizados nesta época, em síntese, foram da seguinte forma:

- as experiências de Cranfield foram importantes porque permitiram identificar os fatores que afetam o desempenho dos Sistemas de Recuperação de Informação e, sobretudo, porque desenvolveram métodos aplicáveis à avaliação destes sistemas e definiram os parâmetros e as medidas a utilizar nessa mesma avaliação;

- os resultados do teste ao sistema SMART confirmaram muitas das conclusões de Cranfield, embora a principal conclusão de Salton tenha sido a de que não haveria justificativa para realizar indexação manual, com uso de vocabulário controlado, pois a linguagem natural e a pesquisa em texto livre proporcionavam a

mesma eficácia, em termos de recuperação da informação; contudo, os resultados do teste de Salton foram criticados pelo facto de se considerar que não era possível aplicar as sofisticações do projeto a um sistema real, pois a complexidade do processamento automático era de tal ordem que se tornaria proibitivo, em termos econômicos, o seu uso em qualquer sistema operacional;

- a avaliação do MEDLARS, que adoptou os métodos e as medidas definidos em Cranfield, funcionou e, sobretudo como meio de quantificar o desempenho do sistema, proporcionando igualmente a correção de inúmeras deficiências detectadas.

De um modo geral, a avaliação tem procurado comparar o desempenho de diferentes sistemas. Segundo Silva e Ribeiro (2004), desde os anos setenta do século XX têm-se proliferado as experiências e os testes de avaliação do desempenho de sistemas de recuperação de informação, baseados na fundamentação e nos métodos definidos a partir de Cranfield, com o objetivo de testar a eficácia de sistemas concretos ou avaliar resultados da recuperação da informação para determinar a qualidade dos instrumentos de pesquisa, designadamente a qualidade das linguagens de indexação.

Esses testes forneceram as bases metodológicas para o desenvolvimento da disciplina de recuperação da informação, a abordagem para testar SRIs foi empregada como modelo para muitas outras avaliações experimentais e operacionais, suas conclusões representam os primeiros resultados científicos do campo (SILVA; RIBEIRO, 2004).

Avalia-se o sistema de recuperação da informação através da comparação das respostas geradas por um sistema e o conjunto ideal de respostas (CARDOSO, 2005).

De acordo com Oliveira (2011), as perguntas dos usuários passam por uma análise conceitual e são traduzidas para o vocabulário do sistema, é elaborada a estratégia de busca e formulada a expressão de busca, no qual os termos da busca são relacionados entre si através de operadores booleanos e não booleanos. Através da expressão de busca, o sistema compara, então, as representações dos documentos com as perguntas dos usuários. Na fase final, os documentos recuperados através das consultas ao sistema são apresentados ao usuário para que este julgue a sua relevância para as suas necessidades de informação.

Na recuperação da informação (RI) a relevância é um tópico extremamente importante, trata-se do julgamento da informação recuperada em relação à necessidade de informação do usuário, ou seja, é uma medida de desempenho no processo de recuperação da informação. Esta medida é aplicada de acordo com duas situações: aquela em que a relevância está relacionada a características do sistema e aquela em que a relevância está relacionada ao usuário. Estas duas situações têm sido designadas como 'relevância orientada ao sistema' e 'relevância orientada ao usuário'. (SCHAMBER, EISENBERG E NILAN, 1990).

A relevância orientada ao sistema refere-se a propriedades e mecanismos internos do sistema e caracteriza o resultado da correspondência entre os termos utilizados na consulta e os termos indexados e armazenados pelo sistema. A relevância orientada ao usuário inclui aspectos cognitivos, situacionais e psicológicos do usuário e refere-se aos contextos subjetivos do mesmo que são empregados para julgar os objetos informacionais (SARACEVIC,1996).

Considerando que cada SRI "classifica os documentos recuperados para cada consulta, de acordo com uma ordem de relevância gerando um vetor resultado" e que "o processo de recuperação consiste na geração de uma lista de documentos recuperados para responder a consulta formulada pelo usuário" (Cardoso, 2005, p. 33), se justifica a necessidade de aproximação entre a linguagem do usuário e a linguagem "entendida" pelo sistema, obtendo-se dois índices de avaliação: precisão e revocação. Medidas padrão da Recuperação da Informação, utilizadas para avaliar a qualidade dos resultados, e contribuir com a avaliação dos sistemas de recuperação de informação.

Precisão significa a quantidade de documentos recuperados e relevantes, no resultado da pesquisa, para o usuário. Piedade (1983, p. 11) esclarece que, "a precisão é a relação entre os documentos relevantes recuperados e o número total de documentos recuperados", enquanto que, "revocação é a relação entre os documentos relevantes recuperados e o número total de documentos relevantes sabidamente existentes na coleção".

Lancaster (2004, p. 4) explica que, "o coeficiente de precisão é a relação entre itens úteis e o total de itens recuperados". E que o coeficiente de revocação é "o índice empregado habitualmente para expressar a extensão com que todos os itens úteis são encontrados".

Precisão é uma medida muito importante para avaliação quando a busca é realizada por um intermediário. Porque na busca realizada pelo próprio usuário ele faz o julgamento de importância no momento que está realizando sua busca. A fórmula definida por Lancaster (1977):

$$PR = \frac{\text{N° de documentos relevantes recuperados pelo sistema}}{\text{N°. total de documentos recuperados na busca, A, B, C}} \times 100$$

Revocação e precisão tendem a variar inversamente. Para medir o coeficiente destes dois critérios é necessário levar em consideração estratégia de busca entre outros fatores.

De acordo com Foskett (1973), revocação é a quantidade de itens adicionais que encontramos ao ampliar a pesquisa, portanto, mede a proporção de documentos relevantes recuperados. O Coeficiente de revocação de uma busca é a proporção de todos os itens relevantes em uma coleção particular ou banco de dados que a busca é capaz de recuperar. Segue a fórmula definida por Lancaster (1977):

$$Re = \frac{\text{N° de documentos relevantes recuperados pelo sistema}}{\text{N°. total de documentos relevantes contidos no sistema}}$$

Para Lancaster (1977), relevância é uma consideração pessoal, cada usuário terá uma interpretação diferente para o que é e o que não é sua necessidade de informação o que torna a avaliação um processo complicado.

O desempenho de um sistema pode ser avaliado sob dois aspectos, de acordo com Araújo (1979, p.47) como sendo:

• Uma referência que foi julgada relevante para o usuário e não foi recuperada pelo sistema;

• Uma referência que foi julgada irrelevante pelo usuário e que foi recuperada pelo sistema.

Sobre “a qualidade de um SRI”, Cesariano (1985) afirma que depende da qualidade da análise conceitual tanto dos documentos quanto das questões propostas pelos usuários. Parte das falhas da recuperação da informação se deve a erros nas interpretações do conteúdo dos documentos e na percepção da demanda das pessoas a que se destina.

Na literatura, a palavra assunto pode ter várias interpretações. Em vista disso, “o processo de análise de assunto também pode ser denominado Análise temática,

Análise documentária, Análise conceitual, ou mesmo Análise de conteúdo" (FUJITA, 2003a, p. 68).

Conforme é apontado por Cesarino e Pinto (1980), a análise de assunto é a operação base para todo o procedimento de recuperação de informações.

O sistema de recuperação que não tiver como base uma eficiente análise de assunto, mesmo adotando procedimentos sofisticados, não conseguirá atingir seus objetivos.

O tratamento de conteúdo, por sua vez, objetiva proporcionar acessibilidade temática do conteúdo dos documentos por suas representações condensadas. Na descrição temática também ocorre à leitura técnica do documento, porém, neste momento ela tem o objetivo de identificar o assunto do documento.

3 SINTAGMAS NOMINAIS NA RECUPERAÇÃO DA INFORMAÇÃO

Várias pesquisas foram realizadas com o propósito de aumentar a precisão dos resultados dos sistemas de recuperação de informação, de forma que os usuários possam encontrar documentos que atendam às suas necessidades de informação, já que recuperar informação relevante na atualidade se tornou uma atividade difícil.

Kuramoto (1995) percebe os Sintagmas Nominais (SNs) como uma abordagem alternativa na recuperação da informação. Os SNs são tidos como essenciais descritores dos documentos por facilitar a Recuperação da Informação (RI). Os SNs podem se tornar descritores em Sistemas de Recuperação da Informação (SRIs) que são softwares buscadores de documentos, o que permitiria uma melhor recuperação.

De acordo Borges, Maculan e Lima (2008) “sintagmas” são expressões que definem uma relação de dependência, onde são estabelecidos elos de subordinação entre outros elementos, que por sua vez também são sintagmas.

Segundo Kuramoto (2002. p. 6) “O sintagma nominal é a menor parte do discurso portadora de informação”, ou seja, é o menor item de informação que possui significado necessário para representar o conteúdo dos documentos. O SN é definido como a menor unidade de sentido do discurso e que possui uma estrutura sintática e lógico-semântica (KURAMOTO, 1999).

Martins (2014, p. 42) afirma que “O SN é um conjunto de elementos que constituem uma unidade significativa dentro da oração e que mantêm a dependência e a ordem entre seus constituintes.” Os elementos a que se refere esse autor são palavras que giram em torno de outra fundamental, definida como núcleo do sintagma.

Os sintagmas nominais (SNs) são objetos de estudo de várias áreas do conhecimento como a Linguística, Ciência da Computação e Ciência da Informação (CI). Contudo, sob a perspectiva da CI, as pesquisas sobre SNs no Brasil são multidisciplinares e restritas.

Sob o ponto de vista da linguística, Perini (1996) define SN como uma classe gramatical com comportamento sintático de sujeito, de objeto direto e, se precedido de preposição, de adjunto adnominal ou de objeto indireto.

Segundo Liberato (1997) o SN é a parte do enunciado que representa um conceito ou referente. Os referentes podem ser entidades abstratas ou concretas; podem ser identificados por nomes próprios ou através de um SN descritivo; e podem ter uso referencial, representando uma entidade, ou uso atributivo, representando um papel.

As expressões de um texto podem ser classificadas de acordo com seu poder discriminatório. As de maior poder discriminatório são, em geral, aquelas de sentido substantivo que podem realizar funções temáticas, como sujeito e objeto, e certas funções semânticas, como agente e instrumento. As expressões desse tipo são em grande parte sintagmas nominais (SNs), onde surge a sua importância nos estudos linguísticos relacionados a RI. Os SNs são os indexadores mais promissores de um texto (KURAMOTO, 1995).

No contexto da Ciência da Computação, Miorelli (2001) vê o SN como uma informação que é composta de um conjunto de símbolos e que possui uma estrutura, assim, ele é muito significativo na identificação do tema central de um documento para a CI, especificamente na RI, o SN é tido como descritor no processo de classificação e recuperação da informação, assim como afirma Kuramoto (1995), é o conjunto que traz o tema do enunciado, por isso é considerado promissor por esse autor.

Kuramoto (2002) afirma que o SN permite que o usuário filtre com mais fidelidade os resultados de sua consulta, assim, os níveis da estrutura dos SNs vão do mais generalizado ao mais específico, resultando em uma maior interação do usuário com o sistema e em um maior retorno de documentos relevantes. A importância de tê-lo na organização da informação, é que são elementos identificadores de informação com alto poder discriminatório.

A maioria dos modelos de recuperação de informação usa a palavra como forma de acesso à informação (KURAMOTO, 2002). Entretanto, a palavra como uma unidade da língua constitui um conjunto de propriedades, sem referência a realidade

extralinguística. Kuramoto (1995), diz que as palavras passam a ter valor referencial a partir do momento que as mesmas se encontram dentro de um universo do discurso.

Perini (1996) já apontava que os SNs eram mais eficientes na classificação e recuperação da informação e que o uso dos lexemas (palavras) não era interessante. Desse modo Nascimento e Corrêa (2016), dizem que os SNs estão ligados à representatividade informacional, mas, é preciso saber que nem todo SN que se encontra em um texto tem potencial para representar o conteúdo temático desse texto. Ou seja, a extração automática de sintagmas nominais de um texto não necessariamente retornará em descritores documentais.

Os sintagmas nominais não sofrem dos problemas de polissemia e ambiguidade, diferente da utilização das palavras como representação temática de um documento, que segundo Kuramoto (2002), gera vários problemas devido às propriedades linguísticas das mesmas, atrapalhando na recuperação da obtenção de resultados precisos. Portanto, a utilização dos sintagmas nominais em substituição às palavras isoladas eliminaria estes inconvenientes.

Pesquisas sobre o uso dos sintagmas como recurso para recuperação da informação foram desenvolvidas com o propósito de avaliar o sintagma nominal quanto ao potencial discriminatório para representar o conteúdo informacional dos documentos e melhorar a recuperação da informação.

Desse modo, Kuramoto (1999), em sua tese de doutorado, desenvolveu uma pesquisa fundamental para a consideração da utilização de sintagmas nominais como descritores. Já em um trabalho anterior, KURAMOTO (1996) vislumbrou a maquete proposta na tese e já apontava o potencial natural de organização dos sintagmas nominais que, se explorado convenientemente, poderia propiciar aos usuários maior facilidade no uso de um SRI e resultados mais precisos em resposta a um processo de busca de informação.

Alguns trabalhos que contribuem para os estudos dos sintagmas nominais em língua portuguesa são os de: Souza (2005; 2006), Maia (2008), Maia e Souza (2010), Corrêa et al. (2011), Lopes (2012), Souza e Raghavan (2006, 2014), e Martins (2014).

Souza (2005; 2006), Maia (2008), Maia e Souza (2010) e Souza e Raghavan (2006, 2014) fizeram uso de diferentes metodologias para selecionar sintagmas nominais: frequência de ocorrência; estrutura e nível; descarte de sintagmas nominais não significativos; inverso da frequência no conjunto de documentos.

Martins (2014) contabiliza a frequência de ocorrência dos sintagmas nominais extraídos, com vistas a demonstrar a importância dessa métrica para atividades de indexação e classificação de documentos.

O trabalho de Lopes (2012) apresenta critérios para extração de termos a partir de sintagmas nominais extraídos automaticamente, sendo tais critérios equivalentes aos seguintes para seleção de sintagmas nominais: descarte dos sintagmas nominais que contêm numerais; descarte dos sintagmas nominais que possuem como núcleo um pronome; descarte dos sintagmas nominais que iniciam com advérbios; detecção de sintagmas nominais implícitos através da remoção sucessiva de adjetivos; e detecção de sintagmas nominais implícitos quando um substantivo é qualificado por mais de um adjetivo ligado por conjunção.

Nascimento e Corrêa (2016), em seu trabalho, consideram os sintagmas nominais (SNs) como melhores descritores de assunto ou temas no conteúdo de documentos. Constituem-se de estruturas gramaticais frasais que possuem substantivos como núcleo, capazes de descrever com maior exatidão os assuntos do que uma recuperação usando palavras isoladas.

Segundo Martins (2014, p. 42), “O SN é um conjunto de elementos que constituem uma unidade significativa dentro da oração e que mantêm a dependência e a ordem entre seus constituintes”. Para que se possa identificar um SN em uma oração é necessário examinar as relações de dependência e concordância, pois dentro dos SNs as palavras determinam outra principal, definida como núcleo do sintagma nominal.

Conforme Perini (1996), o sintagma nominal possui uma estrutura bastante complexa, pois é possível distinguir em sua composição várias funções sintáticas. Seu núcleo pode ser um nome (comum ou próprio) ou um pronome (pessoal, demonstrativo, indefinido, interrogativo ou possessivo). O sintagma nominal pode também ser constituído por determinantes e/ou modificadores, sendo que os modificadores antecedem ou sucedem o núcleo, enquanto os determinantes apenas o antecedem. (MIORELLI, 2001).

Kuramoto (1995) propõe os SNs como abordagem alternativa na recuperação da informação, por meio da construção de um protótipo, de sistema de recuperação de informação capaz de navegar uma estrutura em árvore de SNs. Os sintagmas nominais possuem uma estrutura sintática podendo ter diversas configurações em seus termos, são compostos de grupos nominais constituídos de uma organização hierárquica em árvore.

Kumaroto (2002) explica que ao percorrer a árvore de sintagmas nominais, de um baixo nível para o alto nível, o usuário está refinando sua demanda de informação, de outra forma, a navegação para um de nível mais baixo possibilita ao usuário a reformulação de sua consulta.

A forma com que o usuário navega à procura de informações, intensifica a interação entre o usuário e o computador. Mediante o apoio do computador, essa navegação traduz em um processo de refinamento de uma busca. Onde o usuário faz uma consulta e o computador imediatamente traz os documentos recuperados segundo a consulta submetida.

Em um estudo realizado por Corrêa et al (2011) na avaliação da indexação automática por sintagmas nominais dos documentos constituídos por título, resumo e palavras-chave de 30 teses e dissertações da 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, o autor afirma que a utilização dos sintagmas nominais como recurso de acesso à informação contida em uma base de dados textual se apresenta como uma forma alternativa aos SRIs tradicionais. Podendo aproximar um pouco mais a necessidade informacional do usuário.

3.1 Indexação automática por sintagmas nominais

A indexação automática por sintagmas nominais explora o potencial de uso dos sintagmas nominais como descritores de documentos em processos de indexação.

Le Guern (1991) é considerado pioneiro em propor os Sintagmas Nominais como descritores ao invés das palavras isoladas. O autor faz uma distinção relevante entre descritor e palavra, uma vez que o descritor utilizado para a

recuperação de informação deveria ser uma unidade do discurso e não uma unidade da língua (signo isolado sem significado). Esse autor é responsável pelo desenvolvimento conceitual acerca desse recurso como unidade portadora de significado para a indexação e recuperação de informação.

Kuramoto (1995; 1999), Souza (2005), Maia (2008), Lopes (2011), entre outros, afirma que os SNs são termos candidatos a descritor documental porque são portadores de informação conceitual, ou seja, a menor unidade de sentido do discurso (KURAMOTO, 1995).

Os sintagmas nominais se diferenciam das palavras, de acordo com KUMAROTO (2002), por que quando extraído do texto mantém o significado, o seu conceito.

Em um SRI, o usuário precisa recuperar informações condizentes com sua consulta, para efetivar um melhor resultado de busca, os documentos devem estar bem indexados, independentemente do tipo de consulta. O documento deve ser descrito por termos que tragam informações relevantes.

O estudo de Brito (1992) referente ao uso dos SNs na indexação automática pode ser considerado um dos primeiros estudos nessa vertente no Brasil. O estudo apresenta uma visão diferente sobre a análise e descrição linguística, fundada sobre uma descrição mais rica dos fenômenos linguísticos e que está na origem das reflexões sobre o tratamento automático da informação.

Kuramoto (1995) e Brito (1992) pode ser considerado como precursores no uso dos SNs na indexação automática, no contexto brasileiro. Vários autores se debruçaram em desenvolver métodos e instrumentos de extração e seleção de SNs de forma automática.

Miorelli (2001) propôs um método de extração de Sintagmas Nominais de textos em português. O Sistema Identificador de Sintagmas Nominais do Português – SISNOP, de Morellato (2007), que identifica e extrai SNs.

Outra ferramenta que identifica e extrai SNs, e que se encontra disponível online ao público, é o OGMA de Maia (2008). Esta, além de outras funções, etiqueta, identifica, extrai e seleciona os SNs. Essa ferramenta foi utilizada por alguns autores em pesquisas, como, por exemplo, Corrêa *et al.*(2011).

Segundo Souza (2005), algumas ferramentas são essenciais para que a identificação e extração dos SNs se desenvolvam de forma automática. Pode-se

dizer que as ferramentas necessárias para que haja a indexação automática por meio de SNs são: Etiquetadores, Identificadores de SNs, Extratores de SNs e Selecionadores de SNs.

Conforme Lapa e Corrêa (2010), na Linguística Computacional, etiquetagem (Tagging) consiste na atribuição de categorias a porções do texto objetivando se aproximar ao máximo do Processamento da Linguagem Natural (PLN).

O tagger, segundo Bick (1998), é um sistema que tem como meta identificar, por meio de uma tag (etiqueta) a categoria gramatical de cada léxico do texto analisado. Taggers são sistemas que analisam um texto e inserem etiquetas morfológicas, gramaticais ou sintáticas a cada item lexical.

Pinheiro (2009) aponta que um tagger marca, anota e rotula morfossintaticamente um texto escrito em uma determinada língua. Dessa maneira, marcam-se as palavras, os símbolos de pontuação, estrangeirismos e fórmulas matemáticas existentes dentro de um texto de acordo com seu contexto.

Segundo Morellato (2007, p. 52), um tagger pode ser definido como um software que realiza “o processo de encontrar uma etiqueta, marcar com uma etiqueta cada uma das palavras de um texto baseado em sua definição, assim como em seu contexto”. A etiquetagem consiste em uma descrição detalhada das categorias lexicais de forma que se observam as derivações e as inflexões das palavras.

Para Morellato (2007), o Sistema Identificador de Sintagmas Nominais (SIDSN) consiste em um conjunto de programas que tem por objetivo reconhecer e retornar os sintagmas nominais contidos em frases.

Para Santos (2005), a identificação de SNs é um problema de classificação, pois associa a cada item do corpus uma etiqueta adicional que o classifique como pertencente ou não a um SN.

De acordo com Morellato (2007), o Identificador de Sintagmas Nominais utiliza como entrada o conjunto de frases segmentadas fornecidas pelo Pré-processador de Textos. Cada frase é processada e verificada se, de acordo com as regras definidas da língua portuguesa, é uma sentença válida. É retornada uma lista contendo os sintagmas nominais encontrados com suas respectivas funções dentro da frase. Caso essa frase não obedeça às regras gramaticais

estabelecidas no identificador, uma mensagem é retornada informando que não foi possível reconhecê-la.

Os sintagmas extraídos são salvos em listas que podem conter tanto os SN na sua forma original no texto, como em sua forma canônica (LOPES et al., 2009). De acordo com mesmo autor, a ferramenta oferece algumas opções de manipulação usuais para listas de termos como a aplicação de pontos de corte, comparação de listas e cálculo de medidas usuais de precisão e abrangência.

Ferramentas de extração automática de SNs são softwares desenvolvidos para fazer o processo de etiquetagem que consiste em marcar as palavras de acordo com suas categorias gramaticais.

A seleção de SNs para fazê-los descritores é fundamental no processo de indexação, quanto à questão de relevância.

Kuramoto (1995; 1999) não trabalhou diretamente com a seleção dos melhores SNs para indexação e sim com a automação da extração desses. Por outro lado, Souza (2005) propõe uma maneira de escolher, dentre os SNs extraídos, os melhores descritores.

Para uma melhor seleção de descritores, Corrêa et al (2011) diz que a extração de sintagmas deveria ser acompanhada de estratégias de ordenação por relevância dos sintagmas, sendo levado em conta critérios de frequência e posicionamento.

Segundo Maia e Souza (2010), a utilização de sintagmas nominais na indexação automática é capaz de representar o conteúdo dos documentos, servindo como descritores ou características para o processo de classificação melhorando a precisão dos resultados.

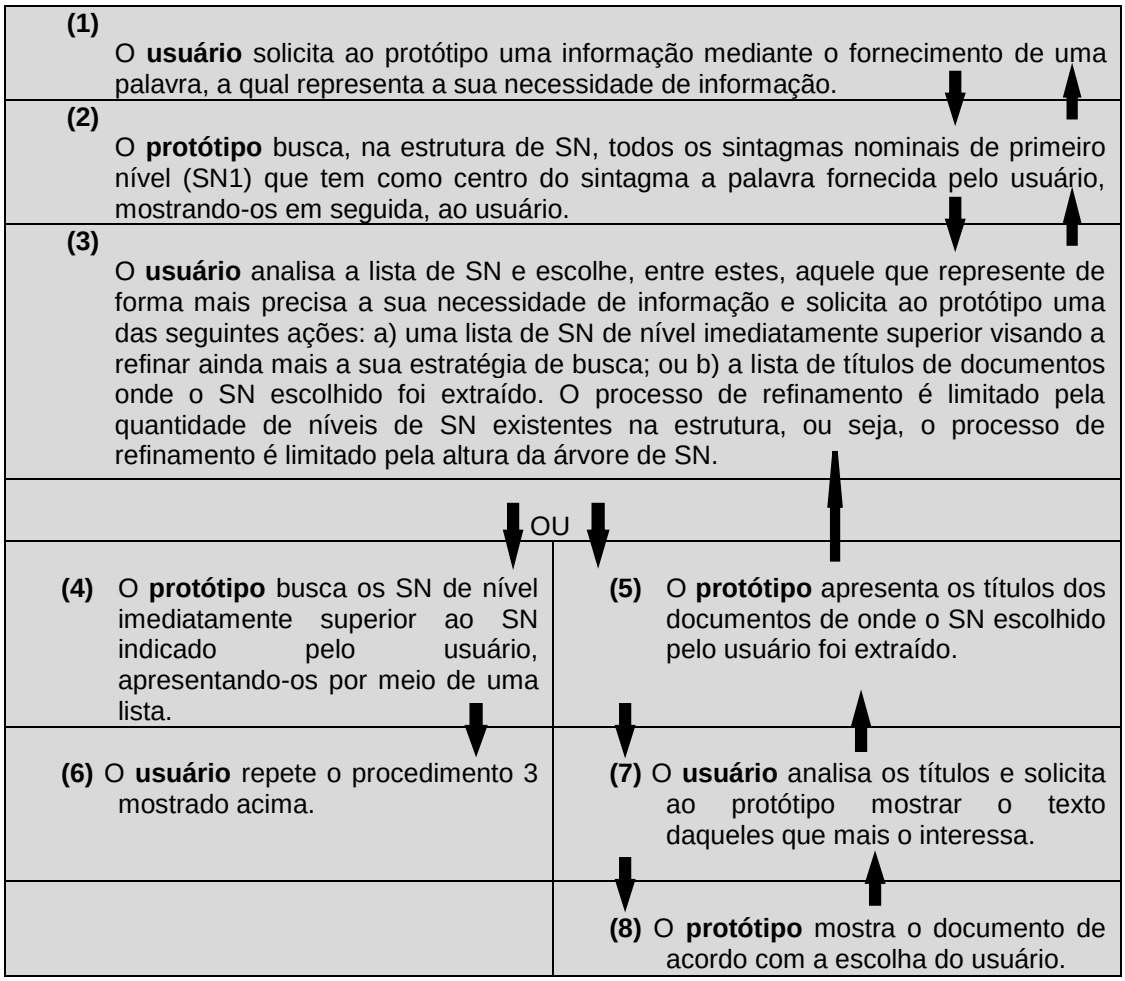
3.2 Recuperação de informação por sintagmas nominais

Kuramoto (1995) desenvolveu um protótipo de sistema de recuperação de informação baseado na navegação nas estruturas internas dos SNs. Esse protótipo tem a interface de busca baseada no encadeamento hierárquico existente entre os SNs, dessa maneira, a interface varre a estrutura arbórea dos SNs de modo interativo com o usuário.

A escolha para que houvesse a interação entre o usuário e a interface de busca do protótipo foi orientada por menus porque possibilita que a

implementação e ajustes sejam feitas facilmente. A Figura 1 descreve de maneira esquemática a interação entre o usuário e o protótipo de Kuramoto (1995).

Figura 1- Procedimentos de interação usuário−protótipo.



Fonte: adaptada de Kuramoto (1995)

Segundo Kumaroto (1995), o protótipo na primeira interação permite ao usuário fazer a sua solicitação de informação fornecendo uma palavra (centro de SN de nível um) que represente a sua necessidade de informação. O usuário faz a escolha do SN que mais lhe interessa e solicita ao sistema mostrar os SN relacionados de nível imediatamente superior, ou apresentar os documentos dos quais o SN escolhido foi extraído. A partir deste ponto, o esquema da figura 1 mostra toda a interação usuário-protótipo.

Conforme Kumaroto (1995), todas as funções se apresentam na tela para o usuário, ou seja, existem apenas duas operações básicas que o usuário deve dominar: a) informar ao protótipo o centro do SN que atende à sua necessidade de

informação; b) escolher uma função que se apresenta na tela em forma de ícones. Deve-se ressaltar ainda que o protótipo desenvolvido permite ao usuário um aprendizado maior do conteúdo da base de dados, tendo em vista que o usuário, na realidade, é guiado pelo protótipo na varredura da árvore dos SN. Este passeio por toda a estrutura acaba por permitir a visualização mais fácil do conteúdo de uma base de dados.

Para o desenvolvimento do protótipo, foi utilizado o sistema Microsoft Access 2.0. Trata-se de um sistema gerenciador de bancos de dados relacional. A escolha deste sistema foi feita em consideração os aspectos de rapidez e facilidade de desenvolvimento/ajustes oferecidas pelo mesmo, do que pela sua velocidade de processamento (KUMAROTO, 1995).

Nos trabalhos de Kuramoto (1995-2002), os resultados obtidos por meio do protótipo comprovaram que é possível tecnicamente implementar uma interface de busca que permite a navegação por meio das estruturas arbóreas construídas hierarquicamente dos SNs.

Kumaroto (1995) ressalta que a ideia, a priori, era construir uma interface capaz de criar menus dinamicamente à medida que os SNs fossem sendo recuperados e apresentados na tela. O autor propõe uma nova interface nos sistemas de buscas textuais e uma nova alternativa no tratamento de RI utilizando SNs.

Para fazer os testes, a extração de SNs foi feita manualmente devido à inexistência de ferramenta que fizesse esse processo automaticamente, a tarefa foi feita simulando uma extração feita automaticamente.

Os SNs foram extraídos de um corpus composto por uma amostragem de 15 artigos escritos em língua portuguesa e extraídos da revista Ciência da Informação.

Para selecionar os artigos, Kuramoto (1995) levou em consideração a importância da definição de um domínio já que isso está relacionado estritamente com o critério de obtenção de menor ocorrência de ambiguidades entre os SNs.

Para a extração de SNs, Kumaroto (1995) considerou algumas questões como resoluções de anáforas e de elipses, assim como problemas relacionados à

identificação de SNs sem determinação (quando não são precedidos por nenhum tipo de artigo), o que é comum na língua portuguesa.

Kuramoto (1999; 2002) considera os SNs como recursos na RI já que são oferecidas duas alternativas para seu uso. A indexação automática em que se reproduziria o método tradicional, mas substituindo os índices contendo palavras isoladas por índices contendo SNs, o que permite utilizar modelos de classificação e ranking. E a organização hierárquica em modelo arbóreo dos SNs, o que possibilita criar um novo conceito em termos de indexação e introduzir inovação quando se fala de interface de busca.

3.3 Recuperação de teses e dissertações no MTTD

O Mapeador Temático de Teses e Dissertações (MTTD) se constitui num sistema de recuperação de informação baseado em mapa de documentos, na forma de um provedor de serviço atrelado ao provedor de dados BDTD-UFPE nos moldes da arquitetura OAI (*Open Archives Initiative*).

Foi desenvolvido utilizando a tecnologia JavaEE, como aperfeiçoamento de um SRI desenvolvido em projeto de pesquisa anterior também fomentado pela FACEPE (CORREA, 2016).

A construção de um Sistema de Recuperação de Informação (SRI) baseado em sintagmas nominais para a (BDTD-UFPE), denominado MTTD teve por objetivo recuperar teses e dissertações na BDTD-UFPE por meio de navegação no mapa de documentos e busca por meio da autossugestão e seleção de sintagmas nominais.

O sistema MTTD realiza uma organização automática de documentos texto em uma estrutura de tabela com grupos definidos em células e relações entre tais grupos baseados no conceito de similaridade de conteúdo dos documentos.

Os documentos são agrupados pela similaridade de conteúdo através de uma Rede Neural Artificial (RNA) do tipo mapa auto organizável, para auxiliar na busca e navegação por teses e dissertações.

> RNAs são sistemas paralelos distribuídos compostos por unidades de processamento simples (nodos) que calculam determinadas funções

> matemáticas (normalmente não lineares). Tais unidades são dispostas em uma ou mais camadas e interligadas por um grande numero de conexões, geralmente unidirecionais. Na maioria dos modelos estas conexões estão associadas a pesos, os quais armazenam o conhecimento representado no modelo e servem para ponderar a entrada recebida por cada neurônio da rede. O funcionamento destas redes é inspirado em uma estrutura física concebida pela natureza: o cérebro humano. (Braga, Ludermir & Carvalho - Redes Neurais Artificiais Teoria e aplicações, 2000).

O objetivo da criação de sistemas baseados em inteligência artificial é desenvolver sistemas para realização de tarefas que geralmente são realizadas melhores por humanos do que por maquinas, ou não possuem uma solução algorítmica viável pela computação convencional.

O Mapa Auto Organizável, conhecido também como Self-Organizing Maps (SOM), é um modelo de rede neural artificial baseado em aprendizado competitivo e não supervisionado, desenvolvido por Teuvo Kohonen (KOHONEN, 1982).

A Recuperação é a etapa onde o usuário especificará a consulta selecionando um sintagma nominal sugerido pelo sistema à medida que digita palavras, e o sistema exibirá na tela documentos contendo o sintagma nominal selecionado, sendo estes últimos encontrados pelo servidor de busca através do índice previamente criado.

O MTTD realizará sugestão de busca exibindo sintagmas nominais existentes no sistema, relacionado ao que o usuário está digitando. Os dados exibidos pela autossugestão são sintagmas nominais que contém total ou parcialmente os caracteres digitados. A autossugestão funciona também com as Palavras-chave, dependendo da opção de sugestão selecionada. O processo desenvolvido permite gerar sugestões de busca baseado em palavras-chave e em sintagmas nominais.

Para o MTTD-UFPE os sintagmas nominais extraídos dos metadados das teses e dissertações servem como fonte para sugestões e para realização de buscas.

A autossugestão é uma estratégia bastante difundida em sistemas de recuperação de informação que consiste em mostrar termos existentes nos documentos do sistema que contém relação com a sequência de caracteres que estão sendo digitados pelo usuário.

O sistema disponibiliza os sintagmas nominais à medida que o usuário digita palavras, permitindo ao usuário descobrir assuntos tratados nos documentos que

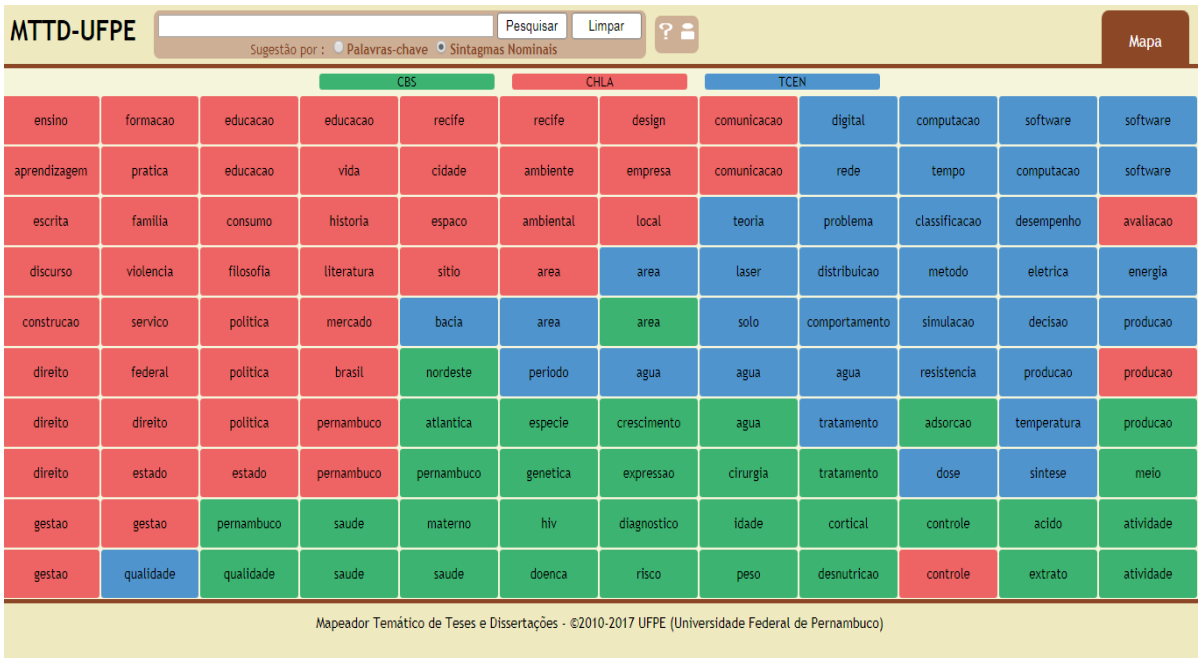
contém as palavras digitadas. Ao selecionar um sintagma nominal disponibilizado pelo sistema, o usuário terá acesso aos documentos retornados pelo sistema, onde nestes se encontram o sintagma nominal selecionado.

A recuperação de informação no MTTD pode ser feita também digitando palavras ou frase na caixa de pesquisa e clicando no botão Pesquisar. As Frases são especificadas envolvendo as palavras por aspas duplas. Serão retornados em ordem de relevância os documentos cujo texto dos metadados contenha alguma das palavras da busca ou a frase especificada, e os documentos são listados na aba retornados.

A interface web do MTTD permite ao usuário a especificação da consulta através da visualização e navegação pelos principais temas abordados no mapa de documentos, bem como a inspeção da lista de documentos selecionados pelo SRI como relevantes para uma consulta ou expressão de busca submetida.

O sistema MTTD possui atualmente 5 (cinco) telas: A tela (1) inicial Mapa: esta tela é acessada a partir da url inicial do sistema e exibe o mapa de documentos. A tela (2) Célula: exibe uma lista contendo todos os documentos da célula selecionada na tela Mapa dispostos em forma de listagem. A listagem é apresentada como uma tabela com as seguintes colunas de metadados: Autor, Título, Programa, Ano, Grau e URL. A tela (3) Retornados: exibe os documentos retornados por uma busca, esta tela exibe para os documentos retornados os mesmos campos da tela Célula, porém exibe apenas para os documentos retornados pela última busca realizada no sistema. As telas (4) Ajuda e (5) Contato, são exibidas como um popup, ficando sobre a tela que estiver aberta. A tela de Ajuda fica disponível durante todo o tempo no sistema, clicando no link ajuda representado pelo ícone de interrogação na parte superior do sistema. Nas telas Mapa, Célula e Retornados, a paleta de busca está sempre visível na parte superior da página. A figura 2 mostra a tela inicial do sistema MTTD.

Figura 2- Detalhes da tela inicial (Mapa)



Fonte: *print screen* do sistema MTTD, 2010-2017(Universidade Federal de Pernambuco).

Segundo Corrêa (2016):

> O MTTD-UFPE realiza a coleta de metadados de teses e dissertações contidas na BDTD-UFPE, cria a estrutura de índice invertido contendo sintagmas nominais e palavras isoladas como pontos de acesso aos documentos, gera arquivos de entradas para treinamento do mapa de documentos, incorpora o mapa de documento treinado à interface de busca, fornece sugestão de busca à medida que o usuário digita na caixa de texto, tendo como vocabulário palavras-chaves ou sintagmas nominais extraídos dos resumos dos documentos, permite visualizar e explorar o resultado da busca no mapa de documentos ou em lista de documentos retornados.

O sucesso em recuperar teses e dissertações por meio de consultas depende do casamento das palavras fornecidas pelo o usuário na consulta com os termos presentes no índice do sistema. No MTTD é possível realizar buscas baseada em coincidência exata, as palavras delimitadas por aspas duplas “”, devolvendo documentos que contenham exatamente a frase informada entre as aspas duplas, permitindo aos usuários recuperar teses e dissertações indexadas por sintagmas nominais expressas por frase ou palavras isoladas.

4 METODOLOGIA

A metodologia adotada para o desenvolvimento deste trabalho tem cmo base uma pesquisa exploratória, que segundo Gil (2010, p.27).

> As pesquisas exploratórias têm como propósito proporcionar maior familiaridade com o problema, com vistas a torná-lo mais explícito ou a construir hipóteses. Seu planejamento tende a ser bastante flexível, pois interessa considerar os mais variados aspectos relativos ao fato ou fenômeno estudado. Pode-se afirmar que a maioria das pesquisas realizadas com propósitos acadêmicos, pelo menos num primeiro momento, assume o caráter de pesquisa exploratória, pois neste momento é pouco provável que o pesquisador tenha uma definição clara do que irá investigar.

Entretanto, a pesquisa objetivou trazer um maior conhecimento do tema escolhido, proporcionando maior simplificação e assim torná-lo o mais claro possível, adotando os métodos de pesquisa bibliográfica e empírica.

Gil (2010, p.29) conceitua pesquisa bibliográfica da seguinte forma:

> A pesquisa bibliográfica é elaborada com base em material já publicado. Tradicionalmente, esta modalidade de pesquisa inclui material impresso, como livros, revistas, jornais, teses, dissertações e anais de eventos científicos. Todavia, em virtude da disseminação de novos formatos de informação, estas pesquisas passaram a incluir outros tipos de fontes, como discos, fitas magnéticas, CDs, bem como o material disponibilizado pela Internet.

Sendo assim, publicações científicas especializadas foram utilizadas como fonte de pesquisa sobre recuperação da informação por meio de sintagmas nominais e a indexação automática por sintagmas nominais, interligando-os com a parte experimental do trabalho.

Para isso, os principais autores foram mapeados para que fosse possível apresentar suas experiências destacando as melhorias que o uso dos sintagmas nominais proporciona na recuperação da informação para o usuário no ato da busca por informações.

A pesquisa apresenta caráter empírico, pois realiza um experimento computacional para coleta de dados, consistindo em um estudo de caso de avaliação e comparação dos resultados na recuperação da informação do uso das expressões de busca por sintagmas nominais expressas por frase ou palavras

isoladas, apresentando sua eficácia na recuperação de informação em teses e dissertações.

Segundo Demo (2000, p. 21) a pesquisa empírica trata da "face empírica e fatual da realidade; produz e analisa dados, procedendo sempre pela via do controle empírico e fatual". A valorização desse tipo de pesquisa é pela "possibilidade que oferece de maior concretude às argumentações, por mais tênue que possa ser a base fatual. O significado dos dados empíricos depende do referencial teórico, mas estes dados agregam impacto pertinente, sobretudo no sentido de facilitarem a aproximação prática" (Demo, 1994, p. 37).

O estudo de caso foi realizado através do no sistema MTTD na recuperação de teses e dissertações sobre três importantes polos econômicos do estado de Pernambuco sendo eles: o polo de confecções, o polo de fruticultura e o polo gesseiro. Sendo consideradas relevantes teses e dissertações que tratam de temas importantes para a estruturação e desenvolvimento desses polos nos níveis econômico, social, ambiental e político.

Foram elaborados cinco sintagmas nominais que referenciam o nome de cada polo, totalizando quinze expressões de busca consistindo em sintagmas nominais expressos por frase ou por palavras isoladas, que foram submetidas ao sistema MTTD-UFPE, com o objetivo de recuperar um conjunto relevante de teses e dissertações, de modo que, se fique claro a eficácia da aplicação de sintagmas nominais na expressão de busca por meio da relevância dos documentos recuperados.

As expressões de busca envolvendo sintagmas nominais na forma de frase e palavras isoladas serão avaliadas por meio das métricas de revocação e precisão na recuperação da informação, levando em conta os documentos retornados pelas buscas.

O julgamento de relevância será baseado no título, resumo e palavras-chave dos dez primeiros documentos retornados para cada expressão de busca ou consulta. Após o julgamento de relevância dos dez primeiros documentos retornados de cada consulta, foram geradas duas tabelas para cada polo, contendo as teses e dissertações recuperadas por sintagmas nominais expressos por frases (Frases) e

recuperadas por sintagmas nominais expressos por palavras isoladas (PI), apresentando as medias e desvio padrão da quantidade de documentos relevantes, total de documentos retornados, revocação e a precisão alcançadas nas dez primeiras das teses e dissertações recuperadas pelas expressões de busca de cada polo. Posteriormente foi criado e analisado um gráfico com as medias de revocação e precisão comparando os resultados da recuperação por sintagmas nominais utilizando frases e palavras isoladas obtidos pelo sistema MTTD.

As atividades da pesquisa foram desenvolvidas da seguinte forma:

1. Coleta e análise de bibliografias especializadas onde autores discutem sobre os sintagmas nominais para que sejam discutidas as contribuições científicas;
2. Descrição da estrutura dos sintagmas nominais e o uso de sintagmas nominais como expressões de busca;
3. Planejamento de experimento envolvendo a recuperação de teses e dissertações usando sintagmas nominais como expressões de busca;
4. Recuperação e julgamento de relevância de teses e dissertações utilizando cada expressão de busca envolvendo palavras isoladas e frase;
5. Análise da eficácia na recuperação de informação com a comparação da precisão e revocação do sistema MTTD em função dos documentos retornados por busca usando sintagmas nominais expressos por palavras isoladas ou frase.

5 ANÁLISE DE RESULTADOS

Através do experimento na busca de teses e dissertações submetendo os sintagmas nominais no sistema MTTD na forma de frase e forma de palavras isoladas para a recuperação da informação, serão apresentados os resultados obtidos em quadro, tabelas e gráfico a avaliação dos SNs com medidas de revocação e precisão.

5.1 Elaboração de expressão de busca

Visando o processo de avaliação dos sintagmas nominais, foram elaboradas cinco expressões de busca na forma de sintagmas nominais para cada um dos polos, para serem recuperadas teses e dissertações no sistema MTTD.

Quadro 1- Sintagmas nominais para a recuperação de informação sobre os polos.

PÓLO DE CONFECÇÕES
1. PÓLO DE CONFECÇÕES
2. PÓLO DE CONFECÇÕES DO AGRESTE
3. CONFECÇÕES DO AGRESTE
4. CONFECÇÕES DE PERNAMBUCO
5. POLO TÊXTIL

PÓLO GESSEIRO
1. PÓLO GESSEIRO
2. PÓLO GESSEIRO DO ARARIPE
3. PÓLO GESSEIRO DE PERNAMBUCO
4. ARARIPE PERNAMBUCO
5. GIPSITA DO ARARIPE

PÓLO DE FRUTICULTURA
1. PÓLO PETROLINA JUAZEIRO
2. PÓLO DE FRUTICULTURA IRRIGADA
3. PÓLO DE FRUTICULTURA
4. FRUTICULTURA IRRIGADA

5. PETROLINA JUAZEIRO

Fonte: a autora.

5.2 Recuperação de teses e dissertações no sistema MTTD por sintagmas nominais na forma de frase e na forma de palavras isoladas sobre os polos

De acordo com as expressões de busca de cada um dos polos submetidas no sistema, as Tabelas 1 a 6 apresentam os documentos relevantes entres os dez primeiros documentos retornados, o total de documentos retornados, a revocação e a precisão na décima posição de retorno pelos os dois métodos de busca, sintagmas nominais na forma de frase e na forma de palavras isoladas.

Tabela 1- Resultado da recuperação de teses e dissertações por sintagmas nominais na forma de frase sobre o polo de confecções no sistema MTTD-UFPE

POLO DE CONFECÇÕES				
CONSULTA	**DOCUMENTOS RELEVANTES ENTRE OS 10 PRIMEIROS RETORNADOS**	**TOTAL DE DOCUMENTOS RETORNADOS**	**REVOCAÇÃO**	**PRECISÃO**
PÓLO DE CONFECÇÕES	**09**	**11**	**53%**	**90%**
PÓLO DE CONFECÇÕES DO AGRESTE	**08**	**09**	**42%**	**89%**
CONFECÇÕES DO AGRESTE	**10**	**17**	**53%**	**100%**
CONFECÇÕES DE PERNAMBUCO	**06**	**07**	**35%**	**86%**
POLO TEXTIL	**04**	**06**	**25%**	**67%**
MÉDIA	**7,4**	**10**	**42%**	**88%**
DESVIO PADRAO	**2,4**	**4,4**	**12%**	**14%**

Para as teses e dissertações sobre o polo de confecções recuperadas por sintagmas nominais submetidos como frases no sistema MTTD-UFPE, de acordo

com a tabela 1, a média de documentos relevantes entre os 10 primeiros documentos retornados avaliados foi de 7,4 no total. O número de documentos relevantes entre os 10 primeiros documentos retornados por consulta é a partir de 4 teses e dissertações, obtido para a consulta “polo têxtil”, até o máximo de 10 documentos, obtido para a consulta “confecções do Agreste”.

De acordo com os 10 primeiros documentos avaliados a consulta “polo de confecções” retornou 11 documentos, mas apenas 9 documentos apresentavam relevância, a consulta “polo de confecções do agreste” retornou 9 documentos, onde apenas 8 documentos apresentaram relevância, a consulta “confecções de Pernambuco” retornou 7 documentos, mas apenas 6 documentos apresentaram relevância. Levando em consideração a relevância dos documentos pela avaliação do título, resumo e palavras-chave nem todos os 10 primeiros documentos retornados avaliados foram considerados relevantes, assim os considerados relevantes tratavam parcialmente ou totalmente o assunto buscado.

A revocação de documentos relevantes entre os 10 primeiros documentos retornados apresentam uma média de 42% no total, a consulta que menos obteve revocação foi “polo têxtil” com uma média de 25%, e as consultas que mais apresentaram revocação foram “polo de confecções” e “confecções do agreste” com 53%. A média de precisão dos documentos relevantes entre os 10 primeiros documentos retornados foi de 88%. A precisão mais baixa obteve-se na consulta “polo têxtil” com 67%, e a precisão mais alta foram com as consultas “polo de confecções” e “confecções do agreste” com 100% de precisão.

De acordo com os sintagmas nominais submetidos na forma de frase no sistema MTTD-UFPE a consulta que mais retornaram documentos foi “confecções do agreste” com 17 documentos e a consulta menos retornaram documentos foi “polo têxtil” com 6 documentos.

Tabela 2- Resultado da recuperação de teses e dissertações por sintagmas nominais na forma de palavras isoladas sobre o polo de confecções no sistema MTTD-UFPE

POLO DE CONFECÇÕES				
CONSULTA	**DOCUMENTOS RELEVANTES ENTRE OS 10 PRIMEIROS RETORNADOS**	**TOTAL DE DOCUMENTOS RETORNADOS**	**REVOCAÇÃO**	**PRECISÃO**

PÓLO DE CONFECÇÕES	**09**	**100**	**53%**	**90%**
PÓLO DE CONFECÇÕES DO AGRESTE	**10**	**100**	**53%**	**100%**
CONFECÇÕES DO AGRESTE	**10**	**100**	**53%**	**100%**
CONFECÇÕES DE PERNAMBUCO	**10**	**100**	**53%**	**100%**
POLO TEXTIL	**06**	**100**	**32%**	**60%**
MÉDIA	**9,0**	**100**	**49%**	**92%**
DESVIO PADRÃO	**1,5**	**0**	**9%**	**18%**

De acordo com a tabela 2, para a recuperação de teses e dissertações por sintagmas nominais expressos por palavras isoladas no sistema MTTD-UFPE sobre o polo de confecções, foi obtido na recuperação de documentos relevantes a média de 9,0 documentos entre os 10 primeiros documentos retornados. Observa-se na coluna “documentos relevantes” que a consulta polo têxtil é a consulta que menos apresenta documentos considerados relevantes entre os 10 primeiros resultados retornados pelo sistema, com 6 documentos, estando inferior aos demais resultados que apresentam o máximo de 10 documentos relevantes para três consultas. A consulta “polo de confecções” retornou 100 documentos, de acordo com 10 primeiros documentos avaliados dos 10 apenas 9 documentos apresentavam relevância, a consulta “polo textil” retornou 100 documentos, onde apenas dos 10 primeiros documentos avaliados apenas 6 documentos apresentaram relevância. Levando em consideração a relevância dos documentos pela avaliação do título, resumo e palavras-chave nem todos os 10 primeiros documentos retornados avaliados foram considerados relevantes, assim os considerados relevantes tratavam parcialmente ou totalmente o assunto buscado.

A revocação de documentos relevantes na décima posição de retornados apresentam a média de 49% no total, com o valor de revocação de 53% para a maioria das consultas. A média da precisão dos documentos na décima posição de retornados é de 92%, com uma precisão de 100% na maioria das consultas. Os

sintagmas nominais submetidos no sistema MTTD-UFPE na forma de palavras isoladas foram retornados 100 documentos para as 5 consultas.

Tabela 3- Resultado da recuperação de teses e dissertações por sintagmas nominais na forma de frase sobre o polo gesseiro no sistema MTTD-UFPE

POLO GESSEIRO				
CONSULTA	**DOCUMENTOS RELEVANTES ENTRE OS 10 PRIMEIROS RETORNADOS**	**TOTAL DE DOCUMENTOS RETORNADOS**	**REVOCAÇÃO**	**PRECISÃO**
PÓLO GESSEIRO	**09**	**19**	**82%**	**90%**
PÓLO GESSEIRO DO ARARIPE	**09**	**14**	**82%**	**90%**
PÓLO GESSEIRO DE PERNAMBUCO	**03**	**05**	**27%**	**60%**
ARARIPE PERNAMBUCO	**04**	**09**	**36%**	**44%**
GIPSITA DO ARARIPE	**03**	**05**	**27%**	**60%**
MÉDIA	**6,0**	**10**	**51%**	**69%**
DESVIO PADRÃO	**2,8**	**5,4**	**29%**	**20%**

De acordo com a Tabela 3, na recuperação das teses e dissertações por sintagmas nominais submetidos com frase sobre o polo gesseiro no sistema MTTD-UFPE, para cada consulta foi obtido a partir de 3 documentos relevantes como apresenta a consulta “gipsita do Araripe” com o menor número de documentos, até 9 documentos como indica as consultas “polo gesseiro” e “polo gesseiro do Araripe”, avaliando os 10 primeiros documentos retornados, com uma média de documentos relevantes de 6,0. De acordo com os 10 primeiros documentos avaliados as consultas “polo gesseiro”, “polo gesseiro do Araripe” apresentam apenas 9 documentos relevantes dos 10 documentos avaliados apenas 1 documento de cada consulta não apresentaram relevância. As consultas “polo gesseiro de Pernambuco” e “gipsita do Araripe” foram retornados 5 documentos para cada consulta, mas apenas 3 foram considerados relevantes. A consulta “Araripe Pernambuco” 9 documentos foram retornados, mas apenas 4 documentos foram relevantes.

Levando em consideração a relevância dos documentos pela avaliação do título, resumo e palavras-chave nem todos os 10 primeiros documentos retornados avaliados foram considerados relevantes, assim os considerados relevantes tratavam parcialmente ou totalmente o assunto buscado.

A média de revocação dos documentos relevantes na décima posição de retornados é 51% no total, com uma revocação mais baixa nas consultas “polo gesseiro de Pernambuco” e “gipsita do Araripe” com 27% e com uma revocação mais alta nas consultas “polo gesseiro” e “polo gesseiro do Araripe” com 82%. A média de precisão na décima posição de retornados foi de 69%, com uma precisão mais baixa de 44% na consulta “Araripe Pernambuco” e precisão mais alta nas consultas “polo gesseiro” e “polo gesseiro do Araripe” com 90% de precisão. Para os sintagmas nominais submetidos no sistema MTTD-UFPE na forma de frase foram retornados a partir de 5 documentos por sintagma nominal para os sintagmas “polo gesseiro de Pernambuco” e “gipsita do Araripe” e até 19 documentos para a consulta “polo gesseiro”.

Tabela 4- Resultado da recuperação de teses e dissertações por sintagmas nominais na forma de palavras isoladas sobre o polo gesseiro no sistema MTTD-UFPE

POLO GESSEIRO				
CONSULTA	**DOCUMENTOS RELEVANTES ENTRE OS 10 PRIMEIROS RETORNADOS**	**TOTAL DE DOCUMENTOS RETORNADOS**	**REVOCAÇÃO**	**PRECISÃO**
PÓLO GESSEIRO	**07**	**100**	**55%**	**60%**
PÓLO GESSEIRO DO ARARIPE	**08**	**100**	**64%**	**70%**
PÓLO GESSEIRO DE PERNAMBUCO	**07**	**100**	**55%**	**60%**
ARARIPE PERNAMBUCO	**04**	**100**	**36%**	**40%**
GIPSITA DO ARARIPE	**05**	**71**	**36%**	**40%**
MÉDIA	**6,2**	**94**	**49%**	**54%**
DESVIO PADRÃO	**1,5**	**12**	**13%**	**13%**

De acordo com a tabela 4, para a recuperação de teses e dissertações sobre o polo gesseiro por sintagmas nominais expressos na forma de palavras isoladas no sistema MTTD-UFPE, de acordo com os resultados apresentados na tabela 4, os documentos relevantes entre os 10 primeiros documentos retornados apresentam média 6,2 documentos relevantes, onde cada consulta obteve a partir de 4 documentos relevantes como apresenta a consulta Araripe Pernambuco, até 8 documentos relevantes como é visto para a consulta polo gesseiro do Araripe.

De acordo com os 10 primeiros documentos avaliados as consultas “polo gesseiro”, “polo gesseiro de Pernambuco”, apresentaram uma quantidade de documentos retornados e relevantes iguais, foram retornados 100 documentos e dos 10 primeiros documentos avaliados apenas 7 apresentavam relevância. A consulta “polo gesseiro do Araripe” foram retornados 100 documentos, mas dos 10 documentos avaliados, apenas 8 documentos apresentavam relevância. A consulta “Araripe Pernambuco”, 100 documentos foram retornados, mas apenas dos 10 documentos avaliados, 4 documentos foram relevantes. A consulta “Gipsita do Araripe”, 71 documentos foram retornados, dos 10 primeiros documentos avaliados, apenas 5 foram relevantes. Levando em consideração a relevância dos documentos pela avaliação do título, resumo e palavras-chave nem todos os 10 primeiros documentos retornados avaliados foram considerados relevantes, assim os considerados relevantes tratavam parcialmente ou totalmente o assunto buscado.

A média de revocação dos documentos relevantes entre os 10 primeiros documentos retornados é de 49%, com uma revocação mais baixa nas consultas Araripe Pernambuco e Gipsita do Araripe com 36%, e revocação mais alta de 64% para a consulta polo gesseiro do Araripe.

A média de precisão nos dez primeiros documentos retornados foi de 54%, com uma precisão mais baixa nas consultas Araripe Pernambuco e Gipsita do Araripe com 40% e precisão mais alta na consulta polo gesseiro do Araripe com 70% de precisão. Para os sintagmas nominais submetidos por palavras isoladas no sistema MTTD-UFPE foram retornados a partir 71 documentos como apresenta a consulta “gipsita do Araripe”, até 100 documentos para as demais consultas “polo gesseiro”, “polo gesseiro do Araripe”, “polo gesseiro de Pernambuco” e “Araripe Pernambuco”.

Tabela 5- Resultado da recuperação de teses e dissertações por sintagmas nominais na forma de frase sobre o polo de fruticultura no sistema MTTD-UFPE

POLO DE FRUTICULTURA				
CONSULTA	**DOCUMENTOS RELEVANTES ENTRE OS 10 PRIMEIROS RETORNADOS**	**TOTAL DE DOCUMENTOS RETORNADOS**	**REVOCAÇÃO**	**PRECISÃO**
PÓLO PETROLINA JUAZEIRO	**08**	**15**	**80%**	**80%**
PÓLO DE FRUTICULTURA IRRIGADA	**02**	**02**	**20%**	**100%**
PÓLO DE FRUTICULTURA	**02**	**03**	**20%**	**67%**
FRUTICULTURA IRRIGADA	**07**	**08**	**70%**	**88%**
PETROLINA JUAZEIRO	**07**	**19**	**70%**	**70%**
MÉDIA	**5,2**	**9,4**	**52%**	**81%**
DESVIO PADRÃO	**2,6**	**6,7**	**29%**	**13%**

De acordo com a Tabela 5, para as teses e dissertações recuperadas por sintagmas nominais submetidos com frase no sistema MTTD-UFPE, considerando a avaliação de documentos relevantes entre os 10 primeiros documentos retornados, o sistema recuperou uma média de 5,2 de documentos relevantes, o número mais baixo de documentos relevantes foram para as consultas “polo de fruticultura irrigada” e “polo de fruticultura” com dois documentos, a consulta que mais obteve documentos relevantes foi a consulta “polo Petrolina juazeiro” com 8 documentos relevantes.

De acordo com os 10 primeiros documentos avaliados, a consulta “polo Petrolina Juazeiro” foram retornados 15 documentos, mas apenas 8 documentos dos 10 avaliados foram relevantes, a consulta “polo de fruticultura irrigada” foram retornados 2 documentos e os 2 documentos apresentados foram considerados relevantes, a consulta “polo de fruticultura” retornou 3 documentos, mas apenas 2 documentos apresentaram relevância, a consulta “fruticultura irrigada” 8 documentos foram retornados mas apenas 7 documentos foram considerados relevantes, assim

como a consulta “Petrolina Juazeiro” com 7 documentos relevantes entre os 10 primeiros documentos avaliados embora foram retornados 19 documentos. Levando em consideração a relevância dos documentos pela avaliação do título, resumo e palavras-chave nem todos os 10 primeiros documentos retornados avaliados foram considerados relevantes, assim os considerados relevantes tratavam parcialmente ou totalmente o assunto buscado.

A revocação dos documentos relevantes entre os 10 primeiros documentos apresenta uma média de 52%, a consulta que apresentou uma revocação mais baixa foram as consultas “polo de fruticultura irrigada” e “polo de fruticultura” com 20% e a revocação mais alta com 80% na consulta “polo Petrolina Juazeiro”. A precisão média dos documentos retornados na décima posição de retornados apresenta uma média de 81%, com uma precisão mais baixa na consulta “polo de fruticultura” com 67% e precisão mais alta na consulta “polo de fruticultura irrigada” com 100% de precisão. Os documentos retornados pelo sistema por sintagmas nominais na forma de frase foram retornados a partir de 2 documentos como apresenta a consulta “polo de fruticultura irrigada”, até 19 documentos como na consulta “Petrolina Juazeiro”.

Tabela 6- Resultado da recuperação de teses e dissertações por sintagmas nominais na forma de palavras isoladas sobre o polo de fruticultura no sistema MTTD-UFPE

POLO DE FRUTICULTURA				
CONSULTA	**DOCUMENTOS RELEVANTES ENTRE OS 10 PRIMEIROS RETORNADOS**	**TOTAL DE DOCUMENTOS RETORNADOS**	**REVOCAÇÃO**	**PRECISÃO**
PÓLO PETROLINA JUAZEIRO	**10**	**100**	**77%**	**100%**
PÓLO DE FRUTICULTURA IRRIGADA	**07**	**100**	**54%**	**70%**
PÓLO DE FRUTICULTURA	**05**	**100**	**38%**	**50%**
FRUTICULTURA IRRIGADA	**07**	**44**	**54%**	**70%**

PETROLINA JUAZEIRO	**10**	**69**	**77%**	**100%**
MÉDIA	**8,0**	**83**	**60%**	**78%**
DESVIO PADRÃO	**1,9**	**23**	**17%**	**22%**

De acordo com a tabela 6, para as teses e dissertações sobre o polo de fruticultura recuperadas por sintagmas nominais na forma de palavras isoladas no sistema MTTD-UFPE, entre os 10 primeiros documentos retornados, foi obtido uma média de 8,0 de documentos relevantes, o menor número de documentos relevantes é apresentado pela consulta polo de fruticultura com 5 teses e dissertações relevantes entre os 10 primeiros documentos retornados, e o maior número de documentos relevantes foi obtido nas consultas polo Petrolina Juazeiro e Petrolina Juazeiro com 10 documentos.

De acordo com os 10 primeiros documentos avaliados os 10 documentos da consulta “polo Petrolina Juazeiro” foram relevantes e foram retornados 100 documentos, a consulta “polo de fruticultura irrigada” retornou 100 documentos, onde apenas 7 documentos dos 10 primeiros documentos avaliados foram considerados relevantes, a consulta “polo de fruticultura” apresentaram 100 documentos mas apenas 5 documentos dos 10 primeiros documentos avaliados apresentavam relevância, a consulta “fruticultura irrigada” foram retornados 44 documentos, dos 10 primeiros documentos avaliados apenas 7 documentos apresentavam relevância, e a consulta “Petrolina Juazeiro” os 10 primeiros documentos avaliados foram relevantes e foram retornados 69 documentos. Levando em consideração a relevância dos documentos pela avaliação do título, resumo e palavras-chave nem todos os 10 primeiros documentos retornados avaliados foram considerados relevantes, assim os considerados relevantes tratavam parcialmente ou totalmente o assunto buscado.

A revocação dos documentos relevantes apresenta uma média de 60% no total, com uma revocação menor na consulta polo de fruticultura com 38% e uma revocação maior nas consultas polo Petrolina Juazeiro e Petrolina Juazeiro com 77% de revocação. A precisão média dos documentos retornados na décima posição de retornados é de 78%, com uma precisão maior na consulta polo de fruticultura com 50% e uma precisão maior nas consultas polo Petrolina Juazeiro e Petrolina Juazeiro com 100% de precisão. Para os sintagmas nominais submetidos

por palavras isoladas no sistema MTTD-UFPE foram retornados a partir de 44 documentos como apresenta a consulta “fruticultura irrigada” até 100 documentos como apresenta na maioria das consultas: “polo Petrolina Juazeiro”, “polo de fruticultura irrigada” e “polo de fruticultura”.

5.3 Teses e dissertações relevantes no sistema MTTD-UFPE por sintagmas nominais na forma de frase e na forma de palavras isoladas sobre os polos

De acordo com 5 consultas na recuperação de teses e dissertações sobre cada um dos polos, a tabela 7 apresenta o total de documentos relevantes na comparação dos dois métodos de busca, sintagmas nominais submetidos por frases e por palavras isoladas ao sistema MTTD-UFPE.

Tabela 7- Quantidade de documentos relevantes por sintagmas nominais na forma de frase e na forma de palavras Isoladas.

SINTAGMAS NOMINAIS		
POLOS	**FRASES**	**PALAVRAS ISOLADAS**
POLO CONFECÇOES	19	19
POLO GESSEIRO	11	11
POLO FRUTICULTURA	10	13

Na avaliação do titulo, resumo e palavras-chave, das 10 primeiras teses e dissertações retornadas, de acordo com as 5 consultas submetidas de cada polo utilizando os dois métodos de busca no sistema MTTD, no polo de confecções foram obtidos 19 documentos relevantes submetendo as consultas pelos os dois métodos de busca no sistema, no polo gesseiro 11 documentos relevantes também foram encontrados submetendo as consultas pelos os dois métodos de busca e no polo de fruticultura foram encontrados 10 documentos relevantes por sintagmas nominais em forma de frases e 13 documentos relevantes por sintagmas nominais em forma de palavras isoladas.

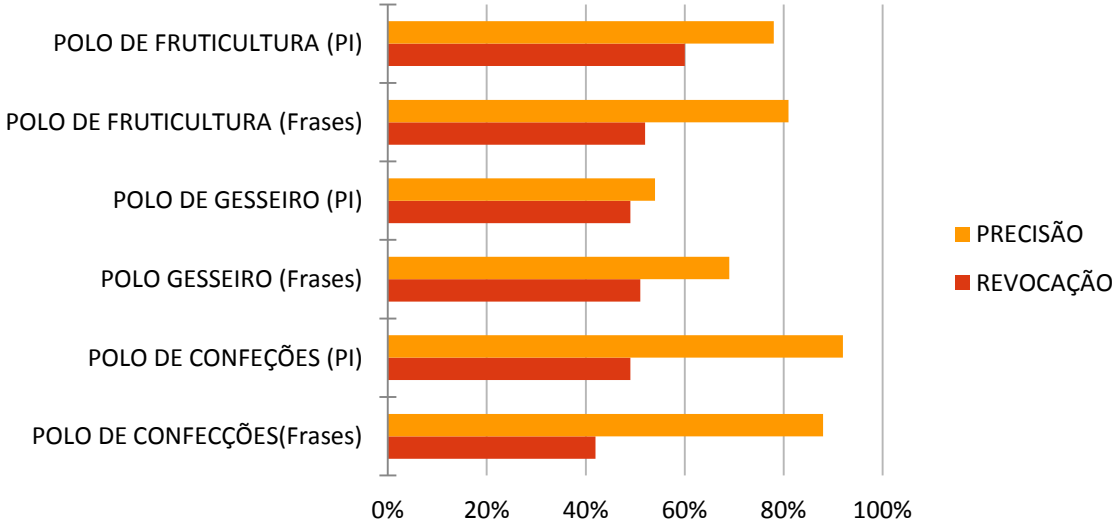
A diferença no polo de fruticultura de 3 documentos na quantidade em comparação aos dois métodos de busca por sintagma nominal é que, os 3 documentos a mais que apresentam na busca das teses e dissertações submetendo as consultas por sintagma nominal na forma de palavras isoladas é que, entre os 10 primeiros documentos avaliados por sintagma nominal na forma de frase, não

apresentaram esses 3 documentos que foram considerados relevantes, nas consultas submetidas por sintagmas nominais na forma de palavras isoladas.

5.4 Avaliação e comparação dos resultados obtidos por sintagmas nominais expressos por frases ou por palavras isoladas sobre os polos de acordo com as medidas de revocação e precisão

A partir das seis Tabelas 1 a 6 construídas (duas para cada polo) na comparação dos dois métodos de busca por sintagmas nominais expressos por frases (Frases) e por palavras isoladas (PI) no sistema MTTD, foi possível avaliar os sintagmas nominais através das medidas revocação e precisão, onde estes percentuais podem ser visualizados no gráfico 1

Gráfico 1- Avaliação da revocação e precisão na recuperação de informação por sintagmas nominais via frases ou palavras isoladas no MTTD



Fonte: o autora.

No Gráfico 1, pode se observar que apesar de uma diferença não estatisticamente significante entre os dois métodos de busca, recuperação por sintagmas nominais via frases ou palavras isoladas, através dos resultados obtidos foi possível observar que a revocação utilizando sintagmas nominais expressos como palavras isoladas se sobressaiu a revocação por sintagmas nominais expressos por frase, visto que para os três polos avaliados, dois apresentaram uma revocação maior por sintagmas expressos por palavras isoladas, sendo eles: polo de

fruticultura e polo de confecção, com uma variação menor de desvio padrão na revocação do polo de confecções. Isto pode ser explicado pelo fato de que, a busca por palavras isoladas torna o casamento do sintagma nominal com expressões similares nos documentos mais flexível, a tendência do sistema MTTD é recuperar mais documentos, gerando uma revocação maior.

Já a precisão por sintagmas nominais expressos por frases se sobressaiu a precisão por palavras isoladas, onde dos três polos, dois deles polos obteve-se uma maior precisão na recuperação por sintagmas nominais expressos por frases, sendo eles: polo de fruticultura e polo gesseiro, apresentando uma variação menor de desvio padrão na precisão do polo de fruticultura. Isto se dá pelo fato de que os sintagmas nominais na forma de frase mantém o sentido estabelecido pela memória discursiva do documento, ou seja, mantém o significado, o seu conceito.

A busca por sintagmas nominais na forma de frase no sistema MTTD-UFPE, retornavam documentos que mencionavam os polos, sendo possível encontrar o sintagma nominal dentro do texto, aumentando a satisfação da necessidade de informação do usuário durante a busca pela informação. Entretanto, para alguns casos, mesmo tendo a expressão do sintagma nominal alguns documentos não são relevantes para o polo.

Dessa forma, na comparação entre a recuperação por sintagmas nominais na forma de frase e na forma de palavras isoladas sobre os polos no sistema MTTD, percebe-se que apesar dos resultados por medias estarem próximos, a precisão dos documentos por sintagmas nominais na forma de frase obteve melhor desempenho em comparação com o uso de sintagmas nominais expressos por palavras isoladas. Utilizando-se os sintagmas nominais como frase consegue-se melhor precisão pelo numero menor de erros e a possibilidade do sistema somente retornar teses e dissertações com os sintagmas nominais.

6 CONCLUSÃO

Um dos principais objetivos dos usuários na recuperação da informação é garantir o acesso rápido e preciso à informação solicitada, com o intuito de chegar mais próximo do ideal, ou seja, retornar o maior número de documentos relevantes à necessidade de informação do usuário.

Os usos dos sintagmas nominais na recuperação da informação cada vez mais veem sendo pesquisado, devido ao fato de que os SNs são essenciais descritores dos documentos por possuírem uma semântica mais bem definida que as palavras isoladas, para tanto, acredita-se que a pesquisa tenha atingido seu objetivo na avaliação dos sintagmas nominais na recuperação de teses e dissertações no sistema MTTD-UFPE, através da comparação da recuperação por sintagmas nominais expressos como frases e através de palavras isoladas.

A contribuição para Ciência da Informação dada por esta pesquisa é a avaliação dos sintagmas nominais na recuperação das teses e dissertações corroborando que os sintagmas nominais consistem em boa fonte de expressões de busca.

Os pressupostos da pesquisa foram positivamente confirmados, os sintagmas nominais se apresentam como uma alternativa ao uso de palavras isoladas nos sistemas de recuperação de informação, os problemas de sinonímia e polissemia são eliminados e são retornados documentos com sintagmas os nominais. Todavia, qualquer que seja a forma utilizada para expressar os sintagmas nominais, esses se configuram melhores pontos de acesso aos documentos, levando em conta os dez primeiros documentos retornados (avaliados pelo titulo, resumo e palavras-chave), para buscas envolvendo sintagmas nominais na forma de frases e na forma de palavras isoladas.

A revocação foi maior na recuperação dos sintagmas nominais por palavras isoladas, e a precisão foi maior na recuperação de sintagmas nominais na forma de frases. Porém os resultados não apresentam uma diferença muito grande nas médias de documentos relevantes, revocação e precisão. Um fato motivador na comparação dos dois métodos de busca por sintagmas nominais diferente da busca

expressa por palavras isoladas, os sintagmas nominais expressos por frases tratam totalmente ou parcialmente o assunto buscado, se mostrando promissor.

Diante dessa constatação, os SNs na forma de frase, seriam alternativas para que os usuários pudessem navegar pelos níveis sintagmáticos nominais até encontrar a informação que procura, como sugeriu Kuramoto (1995) com o seu protótipo de interface de busca que se baseava no encadeamento hierárquico existente entre os SNs, fazendo com que haja interatividade entre a estrutura arbórea e o usuário.

Espera-se que a metodologia usada para o alcance da análise dos sintagmas nominais na recuperação da informação seja utilizada em situações nas quais se deseje avaliar resultados em sistemas de busca quanto ao valor dos sintagmas nominais, como expressão de busca com base na mensuração do nível de revocação e precisão nas buscas.

REFERÊNCIAS

ARAÚJO JÚNIOR, R. H. **Precisão no processo de busca e recuperação da informação.** Brasília: Thesaurus, 2007. 175 p.

ARAÚJO, E.E.R. Revocação (recall) e precisão (precision) no SDI/CIN/CNEN. **Ciência da Informação.** Rio de Janeiro, v.8, n.1 p.47-50, 1979.

BAEZA-YATES, R.; RIBEIRO-NETO, B. **Modern information retrieval.** New York: ACM Press, 1999.

BAILEY, J. E.; PEARSON, S. Development of a tool for measuring and analyzing computer user satisfaction. **Management Science**, n. 29, v. 05, p. 530-45, 1983.

BARION, E. C. N.; LAGO, D.. **Mineração de Textos.** Revista de Ciências Exatas e Tecnologia, São Paulo, v. 3, n. 3, p. 123-140, 8 Dez. 2008.

BATES, M. E. The making of a super searcher. **Searcher**, v. 7, n. 10, p. 33-35, 1999.

BDTD-UFPE. **Biblioteca Digital de Teses e Dissertações da UFPE.** Disponível em:<http://www.bdtd.ufpe.br>. Acesso em: 10 ago. 2017.

BERTHOLINO, Maria Luzia Fernandes. **Buscas em bases de dados.** In: TECNOLOGIA e novas formas de gestão em bibliotecas universitárias. Org. por Maria Etelvina Madalozzo Ramos. Ponta Grossa : UEPG, 1999. 249p., p. 145- 155.
BESSA. O SN em português: A hipótese mórfica. Belo Horizonte: **Revista de Estudos de Linguagem** - UFMG, Julho/Dezembro 1996. p.43-56.

BICK, Eckhard. Structural lexical heuristics in the automatic analysis of portuguese. In: NORDIC CONFERENCE ON COMPUTATIONAL LINGUISTICS, 11. **Proceedings…** Copenhagen: Nodalida ‘98, 1998. p. 44 - 56.

BOCCATO, V. R. C. **Avaliação de linguagem documentária em fonoaudiologia na perspectiva do usuário:** estudo de recuperação da informação com protocolo verbal. 2005. 239 f. Dissertação (Mestrado em Ciência da Informação)- Faculdade de Filosofia Ciências, Universidade Estadual Paulista, Marília, 2005.

BORGES, Graciane Silva Bruzinga. **Indexação automática de documentos textuais: proposta de critérios essenciais.** 2009. 111 f. Dissertação (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, Graciane Silva Bruzinga; MACULAN, Benildes Coura Moreira dos Santos; LIMA, Gercina Angela Borem de Oliveira. Indexação automática e semântica: estudo da análise do conteúdo de teses e dissertações. **Informação & Sociedade:** Estudos, v. 18, n. 2, p.181-193, 2008. Disponível em: <http://www.ies.ufpb.br/ojs2/index.php/ies/article/download/1759/2129 > Acesso em: 06/07/2017.

BRAGA, Antônio; LUDERMIR, Teresa; CARVALHO, André. **Redes neurais artificiais: teoria e aplicações.** LTC, 2000.

BRITO, Marcilio de. Sistemas de Informação em linguagem natural: em busca de uma indexação automática. **Ci. Inf.** , Brasília, 21(3): 223-232, set./dez. 1992.

CARDOSO, O. N. P. Recuperação de informação. **Infocomp,** v.2 p.33-38, 2005. Disponível em: <http://www.dcc.ufla.br/infocomp/artigos/v2.1/art07.pdf>. Acesso em: 24 out. 2017.

CESARINO, Maria Augusta da Nóbrega. Sistemas de recuperação da informação. **Revista da Escola de Biblioteconomia da UFMG.** Belo Horizonte, v. 14, n. 2, p. 157-168, set. 1985.

CHAUMIER J. Indexação: conceitos, etapas e instrumentos. **Revista Brasileira de Biblioteconomia e Documentação**, São Paulo, v. 21, n.1/2, p.63-79, jan./ jun.1998.

CHOO, Chun. **A organização do conhecimento**. Como as organizações usam a informação para criar significado, construir conhecimento e tomar decisões. 2 ed. São Paulo: Editora SENAC, 2006.

CORRÊA, R. F. et al. 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.

CORREA, R. F. **Mapeador Temático de Teses e Dissertações.** Relatório Final do Projeto de Pesquisa. 2016.

CORRÊA, R. F. **Sistemas Baseados em Mapas Auto-organizáveis para Organização Automática de Documentos Texto.** Tese de Doutorado. Centro de Informática da UFPE, Recife, 2008.

COULON, Daniel, KAYSER, Daniel. **Informática e linguagem natural:** uma visão geral dos métodos de interpretação de textos escritos. Brasília: IBICT, 1992.

DANTAS, Marcos. Sistemas de informação: a evolução dos enfoques. **Ciência da Informação**, [S.l.], v. 21, n. 3, dec. 1992. ISSN 1518-8353. Disponível em: <http://revista.ibict.br/ciinf/article/view/431>. Acesso em: 26 jun. 2017.

DATTA, Suman. A organização de conceitos para recuperação da informação. **Ciência da Informação**, [S.l.], v. 6, n. 1, jun. 1977. ISSN 1518-8353. Disponível em: < http://revista.ibict.br/ciinf/article/view/88>. Acesso em: 26 jun. 2017.

DEMO, Pedro. **Pesquisa e construção do conhecimento:** metodologia científica no caminho de Habermas. Rio de Janeiro: Tempo Brasileiro,1994.

FERNEDA, E. ; DIAS, G. A. Um método de expansão automática de consulta baseada em ontologia. **Encontro Nacional de Pesquisa em Ciência da Informação**, v. 14, 2013b.

FERNEDA, E. ; DIAS, Guilherme Ataíde. A lógica Fuzzy aplicada à recuperação da informação. **InterScientia,** João Pessoa, v.1, n.1, p. 51-65, jan./abr. 2013a. Disponível em: <https://periodicos.unipe.br/index.php/interscientia/article/download/24/21>. Acesso em: 20 jun.2017.

FERNEDA, Edberto. Aplicando algoritmos genéticos na recuperação da informação. **DataGramaZero - Revista de Ciência da Informação,** Rio de Janeiro, v. 10, n. 1, fev. 2009.

FERNEDA, Edberto. **Introdução aos Modelos Computacionais de Recuperação de Informação.** Rio de Janeiro: Ciência Moderna, 2012.

FERNEDA, Edberto. **Recuperação da informação:** análise sobre a contribuição da Ciência da computação para a Ciência da Informação. Tese (doutorado em comunicação) – USP. Escola de Comunicação e Artes, São Paulo, 2003.

FERNEDA, Edberto; DIAS, Guilherme Ataíde. OntoSmart: um modelo de recuperação de informação baseado em ontologia. **Perspectivas em Ciência da Informação**, [S.l.], v. 22, n. 2, p. 170-187, jun. 2017. ISSN 19815344. Disponível em: <http://portaldeperiodicos.eci.ufmg.br/index.php/pci/article/view/2081>. Acesso em: 19 nov. 2017.

FOSKETT, A. C. **A abordagem temática da informação.** São Paulo: Polígono, 1973.

FUJITA, Mariângela Spotti Lopes A identificação de conceitos no processo de análise de assunto para indexação. **Revista Digital de Biblioteconomia e Ciência da Informação,** 2003b, vol. 1, n. 1.

FUJITA, Mariângela Spotti Lopes; LEIVA, Isidoro Gil-. Avaliação da indexação por meio da recuperação da informação. **Ciência da Informação**, [S.l.], v. 43, n. 1, jun. 2015. ISSN 1518-8353.

FUJITA, Mariângela Spotti. Lopes **A análise documentária no tratamento da informação:** as operações e os aspectos conceituais interdisciplinares. Marília: Departamento de Ciência da Informação, FFC/UNESP, 2003a. 15f.

GIL, Antonio Carlos. **Como Elaborar Projetos de Pesquisa**. 5ª ed. São Paulo: Atlas, 2010.

GONZALEZ, Marco; LIMA, Vera L. S. de. Sintagma Nominal em Estrutura Hierárquica Temática na Recuperação de Informação. **Anais**. ENIA 2001, Fortaleza: 2001. Disponível em: http://www.inf.pucrs.br/~gonzalez/docs/sneht.pdf Acesso em: 02 nov. 2017.

HARTER, S.; HERT, C.. Evaluation of Information Retrieval Systems: Approaches, Issues, and Methods. **Annual Review of Information Science and Technology (ARIST),** v. 32, Medford: Martha, 1997.

HARTLEY, R. J. et al. **Online searching:** principle and practice. Bowker-Saur: London, 1990.

KENT, A. **Manual da recuperação mecânica da informação.** São Paulo: Polígono, 1972.

KOHONEN, T., Kaski, S., Lagus, K., Salojärvi, J., Honkela, J., Paatero, V., Saarela, A. Self Organization of a Massive Document Collection. IEEE Transaction on Neural Networks, v. 11, n. 3, May 2000, pp. 574-585.

KUMAROTO, Hélio. Uma abordagem alternativa para o tratamento e recuperação de informação textual: os sintagmas nominais. **Ciência da Informação,** Brasília, v.25, n. 2, 1995.

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

KURAMOTO, Hélio. **Proposition d'um Système de Recherche d'Information Assistée par Ordinateur.** Tese (Doutorado). L'Université Lumière – Lyon - França, 1999.

LANCASTER, F. W. **Avaliação de serviços de bibliotecas.** Brasilia: Briquet de Lemos, 1996. 356 p.

LANCASTER, F. W. **Construção e uso de tesauro:** curso condensado. Trad. César Almeida de Meneses Silva. Brasília: IBICT, 1987.

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

LANCASTER, F. W. **Information retrieval systems:** characteristics, testing and evaluation. 2. ed. New York: J. Wiley, 1979.

LAPA, Remi Correia; Corrêa, Renato Fernandes. Indexação automática de teses e dissertações da UFPE. In: ENCONTRO DE ESTUDOS SOBRE TECNOLOGIA, CIÊNCIA E GESTÃO DA INFORMAÇÃO, 1, 2010, Recife**. Anais....** Recife, 2010. 1 CD-ROM.

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

LIBERATO, Yara G. **A estrutura do SN em português:** uma abordagem cognitiva. 1997. 203f. Tese (Doutorado em Letras) Faculdade de Letras, Universidade Federal de Minas Gerais, Belo Horizonte.

LIBERATO, Yara Goulart. **A estrutura do SN em português:** uma abordagem cognitiva. 1997. 203 f. Tese (Doutorado) - Curso de Doutorado em Letras, Faculdade de Letras, Universidade Federal de Minas Gerais – UFMG, Belo Horizonte, 1997.

LOPES, L. **Extração automática de conceitos a partir de textos em língua portuguesa.** 2011. 156 f. Tese (Doutorado em Ciência da Computação), Faculdade de Informática, PUCRS, Porto Alegre, 2011.

LOPES, L., Vieira, R., Finatto, M. J., Zanette, A., Martins, D., and Ribeiro Jr, L. C. **Automatic extraction of composite terms for construction of ontologies: an experiment in the health care area.** RECIIS, 3 (1):72–84, 2009.

LOPES, Lucelene. **Extração automática de conceitos a partir de textos em língua portuguesa.** 2012, 156 f. Tese (Doutorado em Ciência da Computação).

Faculdade de Informática, Pontifícia Universidade Católica do Rio Grande do Sul, Porto Alegre, 2012.

MAIA, Evandro. **Medindo a satisfação dos clientes,** disponível em < http://www.itcom.com.br/pdf/medindo_a_satisfacao_dos_clientes.pdf>, acesso em 29/08/2017.

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, Luiz Cláudio; SOUZA, Renato Rocha. Uso de sintagmas nominais na classificação automática de documentos eletrônicos. **Perspect. ciênc. inf.**, Belo Horizonte, v. 15, n. 1, p. 154-172, 2010.

MARON, M. E.; KUHNS, J. L. **On relevance, probabilistic indexing, and information retrieval. Journal of the ACM,** 7(3): 216-244,1960.

MARTINS, Agnaldo Lopes. **O uso do sintagma nominal na recuperação de documentos [manuscrito]**: proposta de um mecanismo automático para classificação temática de textos digitais. 2014, 192 f. Tese (Doutorado em Ciência da Informação) – Escola de Ciência da Informação, Universidade Federal de Minas Gerais Minas Gerais, 2014.

MEADOW, Charles T. **Text Information Retrieval Systems.** San Diego: Academic Press,1992.

MEADOW, Charles T.; BOYCE, Bert R.; KRAFT, Donald H. Text **Information Retrieval Systems.** Academic Press, 2000. 364p.

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

MORELLATO, Luana Vieira. **SIDSN:** sistema identificador de sintagmas nominais, 2007. 58 f. Monografia (Bacharelado em Ciência da Computação) – Departamente de Informática, Centro Tecnológico, Universidade Federal do Espírito Santo, Vitória, 2007.

MTTD-UFPE. **Mapeador Temático de Teses e Dissertações da UFPE.** Disponível em: <http://logic2.ddns.net:8080/mttd/> Acesso em: 10 ago. 2017.

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

NASCIMENTO, Geysa Flávia Câmara de Lima. **Folksonomia como estratégia de indexação dos bibliotecários no Del.icio.us.** 2008. 104f. Dissertação (Mestrado em Ciência da Informação) – Programa de Pós-Graduação em Ciência da Informação, Universidade Federal da Paraíba, João Pessoa, 2008. Disponível em: Acesso em: 22 jul. 2017.

NASCIMENTO, Gustavo Diniz do. **Dos Sintagmas Nominais aos Descritores Documentais:** estudo de caso na indexação de teses e dissertações da área de direito. 2015, 198 f. Dissertação (Mestrado) – Mestrado em Ciência da Informação, Universidade Federal de Pernambuco, Recife-PE, 2015.

OLIVEIRA, Marlene de. (Coord.). **Ciência da informação e biblioteconomia:** novos conteúdos e espaços de atuação. Belo Horizonte: UFMG, 2011.

PERINI M, A. O Sintagma nominal em português: estrutura, significado e função, Revista de Estudos da Linguagem, n. esp., 1996. Disponível em: <http://relin.letras.ufmg.br/revista/upload/Relin_NEspecial_1996.pdf >. Acesso em: 01 nov. 2017.

PERINI, M. A. **Gramática descritiva do Português**. 3 ed. São Paulo: Ática, 1998.
PIEDADE, M. A. R. **Introdução à teoria da classificação.** 2. ed. Rio de Janeiro: Interciência, 1983. p. 9-15.

PINHEIRO, Marcello Sandi. **Uma abordagem usando sintagmas nominais como descritores no processo de mineração de opiniões.** 2009. 110 f. Tese (Doutorado em Engenharia Civil) – COPPE, Universidade Federal do Rio de Janeiro, Rio de Janeiro, 2009.

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.

PROBST, Gilbert; RAUB, Steffen; ROMHARDT, Kai. **Gestão do conhecimento:** os elementos construtivos do sucesso. Porto Alegre: 2000.

RIBEIRO, Fernanda. **Novos caminhos da avaliação de informação.** (2005). Disponível em: <http://www.brapci.inf.br/_repositorio/2009/10/pdf_a1e62b7264_0006595.pdf>, acesso em29/08/2017.

RIBEIRO, Fernanda; SILVA, Armando Malheiro da. A Avaliação de informação: uma operação metodológica. **Páginas a&b:** arquivos e bibliotecas, Lisboa, p. 7-37, v. 14, 2004

RIBEIRO, Fernanda; SILVA, Armando Malheiro da. A avaliação de informação: uma operação metodológica. **Páginas A&B:** arquivos e bibliotecas, Lisboa, n. 14, p. 7-37, 2004.

ROBERTSON, S.E. Theories and models in information retrieval. **Journal of Documentation,** 33, p.126-148. 1977.
ROBERTSON, Stephen. On the history of evaluation in IR. **Journal of Information Science,** Thousand Oaks, v. 34, n. 4, p. 439-456, 2008.

ROBREDO, Jaime. **Da ciência da informação revisitada aos sistemas humanos de informação.** Brasília: Thesaurus, 2003.

ROBREDO, Jaime. Otimização dos processos de indexação dos documentos e de recuperação da informação mediante o uso de instrumentos de controle terminológico.**Ci. Inf**., Brasília, v. 11, n. 1, p. 3-18.1982.

ROWLEY, Jennifer. **Informática para bibliotecas.** Brasília: Briquet de Lemos/Livros, 1994.

ROWLEY,Jenifer. **A biblioteca eletrônica**. Tradução de Antônio Agenor Briquet de Lemos. 2. ed. Brasília: Briquet de Lemos/Livros, 2002. 399 p. Segunda edição de Informática para bibliotecas; Título original: The eletronic library. ISBN 858563720X.

SALTON, G. (ed.) (1971).**The SMART retrieval system:** experiments in automatic document processing. Prentice-Hall.

SALTON, G.; McGILL, M. J. **Introduction to Modern Information Retrieval.** Computer Science Series, USA: McGraw-Hill, 1983.

SANTOS, Cícero Nogueira dos. **Aprendizado de máquina na identificação de sintagmas nominais:** o caso do português brasileiro. 2005. 104 f. Dissertação (Mestrado em Sistemas e Computação) – Instituto Militar de Engenharia, Rio de Janeiro, 2005.

SANTOS, G. C.; RIBEIRO, C. M. **Ancrônimos, siglas e termos técnicos:** Arquivística, Biblioteconomia, Documentação, Informática. Campinas: Àtomo, 2003. 277p.

SARACEVIC, T. Ciência da informação: origem, evolução e relações. **Perspectivas em Ciência da Informação**, v. 1, n. 1, p. 41-62, 1996. Disponível em: <http://www.brapci.inf.br/v/a/3224>. Acesso em: 10 Dez. 2017.

SARACEVIC, T. Modeling Interaction in Information Retrieval (IR): a review and proposal. In: **Proceedings of The American Society for Information Science,** v.33, p. 3-9, 1996a.

SARACEVIC, Tefko. A natureza interdisciplinar da ciência da informação. Ciência da Informação, [S.I.], v. 24, n. 1, apr. 1995. ISBN 1518-8353. Disponível em: <http://revista.ibict.br/ciinf/article/view/608>. Acesso em: 17 nov. 2017.

SCHAMBER, L.; EISENBERG, M.; NILAN, M. A Re-examination of relevance: toward a dynamic, situational definition. **Information Processing and Management,** v.26 (6), p.755-766,1990.

SOUZA, R. R. **Uma proposta de metodologia para escolha automática de descritores utilizando sintagmas nominais**. 2005. 197 f. Tese (Doutorado em Ciência da Informação) - ECI, UFMG, 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.; ALVARENGA NETO, R. C. D. de; MENDES, K. C. I. Mapeamento semântico através da análise de ocorrência de descritores sobre gestão do conhecimento. **Transinformação**, v. 19, n. 1, p. 19-30, 2007.

SOUZA, Renato Rocha. Sistemas de recuperação de informações e mecanismos de busca na web: panorama atual e tendências. **Perspectivas em Ciência da Informação,** Belo Horizonte, v. 11, n. 2, set. 2011. ISSN 19815344. Disponível em:

<http://portaldeperiodicos.eci.ufmg.br/index.php/pci/article/view/320>. Acesso em: 21 set. 2017.
TEIXEIRA, Iberê L. R.. Uma linguagem de busca para sistemas de recuperação de informação. **Ciência da Informação**, [S.l.], v. 3, n. 1, june 1974. ISSN 1518-8353. Disponível em: < http://revista.ibict.br/ciinf/article/view/39>. Acesso em: 26 jun. 2017.

VIEIRA, Simone Bastos. Análise comparativa entre indexação automática e manual da literatura brasileira de ciência da informação. **Revista de Biblioteconomia de Brasília,** v. 16, n. 1, p. 83-94, 1988.

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