Graciane Silva Bruzinga Borges

INDEXAÇÃO AUTOMÁTICA DE DOCUMENTOS TEXTUAIS:

PROPOSTA DE CRITÉRIOS ESSENCIAIS

Belo Horizonte
Escola de Ciência da Informação da UFMG
2009

Graciane Silva Bruzinga Borges

INDEXAÇÃO AUTOMÁTICA DE DOCUMENTOS TEXTUAIS:

PROPOSTA DE CRITÉRIOS ESSENCIAIS

Dissertação apresentada ao Programa de Pós-graduação em Ciência da Informação da Escola de Ciência da Informação da Universidade Federal de Minas Gerais como requisito para obtenção do Grau de Mestre em Ciência da Informação.

Linha de pesquisa: Organização e Uso da Informação

Orientadora: Profª. Dra. Gercina Ângela Borém de Oliveira Lima.

Belo Horizonte
Escola de Ciência da Informação da UFMG
2009

Borges, Graciane Silva Bruzinga.
B732i Indexação automática de documentos textuais [manuscrito] : critérios essenciais / Graciane Silva Bruzinga Borges. – 2009.
111 f. : il., enc.

Orientadora: Gercina Ângela Borém de Oliveira Lima.
Inclui anexos.
Dissertação (mestrado) – Universidade Federal de Minas Gerais, Escola de Ciência da Informação.
Referências : f. 91-99.

1. Ciência da informação – Teses. 2. Indexação – Teses. 3. Indexação automática – Teses. 4. Representação da informação – Teses. 5. Semântica – Teses. 6. Lingüística computacional – Teses. 7. Taxonomia – Teses. I. Título. II. Lima, Gercina Ângela Borém de Oliveira. III. Universidade Federal de Minas Gerais, Escola de Ciência da Informação.
CDU: 025.4

Ficha catalográfica: Biblioteca Profª Etelvina Lima, Escola de Ciência da Informação da UFMG

Universidade Federal de Minas Gerais
Escola de Ciência da Informação
Programa de Pós-graduação em Ciência da Informação

Dissertação intitulada “Indexação automática de documentos textuais: critérios essenciais” de autoria da mestranda Graciane Silva Bruzinga Borges, aprovada pela banca examinadora constituída pelos seguintes professores:

_____
Prof. Dra. Gercina Ângela Borém de Oliveira Lima – ECI/UFMG – Orientadora

_____
Prof. Dr. Eduardo Wense Dias – ECI/UFMG

_____
Prof. Dra. Júlia Gonçalves da Silveira

_____
Prof. Dra. Cíntia Lourenço – ECI/UFMG

_____
Prof. Dra. Gercina Ângela Borém de Oliveira Lima
Coordenadora do Programa de Pós-graduação em Ciência da Informação
PPGCI/ECI/UFMG

Belo Horizonte, 14 de agosto de 2009.

Av. Antônio Carlos, 6627 – Belo Horizonte, MG – 31270-901 – Brasil – tel.: (031) 3409-6103 – fax: (031) 3409-5207

Este trabalho é dedicado aos meus pais, José Bruzinga e Maria das Graças Silva Bruzinga, por verdadeiramente deixarem os seus sonhos para que pudéssemos, meus irmãos e eu, sonhar os nossos.

Em especial, ao meu marido, Christiano Limp Borges, que faz minha vida ter razão, que é minha família e o "objetivo geral" de todos os meus esforços.

Ainda...

Àqueles que acreditam na interdisciplinaridade das áreas em prol da evolução científica.

AGRADECIMENTOS

A DEUS, por ter permitido que eu desenvolvesse este trabalho e por ter iluminado meu caminho todo esse tempo.

À querida professora Gercina, a quem devo esta conquista na academia. OBRIGADA! Exemplo de professora! Exemplo de profissional apaixonada pelo que faz! Amiga e mãe! Mestre para a vida inteira!

Aos meus pais, a quem não posso expressar com palavras minha gratidão por todo sacrifício dedicado a mim e aos meus irmãos. Por todo amor e fé em prol do nosso estudo.

A Christiano, meu amor, minha alegria, minha vida!

A Geordani, meu irmão, que foi o primeiro a me apontar a universidade como o melhor caminho a seguir. Que sempre esteve comigo em BH para dividir a saudade da família e para me pedir *cinquim...* Que uma pessoa excepcional, de um coração que não tem tamanho!

A Josiane, minha irmã querida que simplesmente me completa como ser humano e me faz ter um orgulho danado de ser sua irmã caçula. Exemplo de pessoa simples e de pesquisadora que ama o que faz!

A Joélcio, meu irmão querido que sempre será meu *padrim Joãoélcio.* Que faz nossa família ter sentido com a presença do Luquinha e da Bia!

A Benildes, por ter participado e contribuído imensamente com esta pesquisa. Por ser uma pessoa generosa e capaz de grandes sacrifícios para ajudar um amigo. Aluna exemplar! Amiga sem igual!

A Érica Sarsur, por ter se dedicado, de maneira excepcional, na correção e melhoramento deste trabalho quanto aos aspectos lingüísticos.

A Ana Paula, por ser minha irmã do coração. Por ter me deixado usar seu computador tantas madrugadas, pelo companheirismo nos tempos de república. Por SEMPRE ter me incentivado e acreditado no meu trabalho.

A Fernanda, pelo convívio na ECI e no Grupo de Pesquisa MHTX. Sempre amável e disposta a ajudar. Por ser uma grande amiga.

A Rafael, por sua companhia todos esses anos de estudo. Por ser um grande amigo.

A Neia, por seus puxões de orelha e por seu amor. Por cuidar da minha casa, para que eu me dedicasse aos estudos, como se fosse a sua. Por cuidar de mim e do Chris como se fôssemos seus filhos.

Aos Colegas do Grupo de Pesquisa MHTX, pelos proveitosos encontros e discussões, pelos congressos e é claro, pelas *MHTXiroscas*...

Às professoras da Escola Municipal Afonso Pena, Maria Auxiliadora (*in memoriam*) e Tia Mabel.

Aos professores do Colégio Nossa Senhora do Carmo, Euza, Genilson e Paulo Vicente.

À Profª Lídia Alvarenga, por ter me orientado nas etapas de qualificação e de defesa do trabalho. Por ser uma professora exemplar para a ECI.

À Profª Maria Luiza de Almeida Campos, por ter me orientado nas etapas de qualificação e de defesa do estudo e por contribuir sempre para a evolução do projeto de pesquisa MHTX.

À Profª Maria Guiomar da Cunha Frota, pelo apoio e orientação nas duas etapas de apresentação do trabalho. Por sua objetividade e disposição para contribuir com a parte metodológica do estudo.

À Profª Madalena Martins Lopes Naves, por sua inestimável orientação no Estudo Especial desenvolvido durante esta pesquisa e por suas valiosas contribuições ao Grupo de Pesquisa MHTX.

À Profª Júlia Gonçalves da Silveira, por sua atenção e colaboração na avaliação final do trabalho.

Aos professores da graduação, Renato Souza e Maurício Barcellos, que me incentivaram a seguir a vida acadêmica.

Aos profissionais do PPGCI pelo trabalho constante em prol do desenvolvimento do programa de mestrado e doutorado da ECI.

À CAPES pelo apoio financeiro que facilitou o desenvolvimento deste trabalho.

"Se fui capaz de ver mais longe foi apenas porque eu estava apoiado sobre ombro de gigantes."

Isaac Newton

RESUMO

Este estudo visa à avaliação de critérios de indexação automática para o desenvolvimento de *softwares* destinados à extração de termos representativos do conteúdo de documentos textuais. Estudam-se as maneiras de se realizar o processo de indexação – manual e automática – e discute-se a aplicação desses critérios para a otimização da primeira etapa do processo, que é a análise de assunto. O volume de documentos publicados na atualidade demanda grandes esforços para se adquirir técnicas alternativas para sua indexação, realizar esse processo manualmente tem sido considerado um processo lento. Identificaram-se os critérios de indexação automática mais utilizados através de estudo de artigos técnico-científicos da área, para, então, analisar o grau de satisfação obtido pelos pesquisadores por meio de sua combinação. Entre os objetivos alcançados, encontram-se: (1) listagem dos critérios encontrados na literatura, (2) caracterização de cada critério, (3) listagem dos critérios mais recorrentes. Além desses, encontra-se o objetivo principal deste trabalho, com a obtenção de um conjunto de critérios selecionados, formando o conjunto considerado ideal pela pesquisadora.

Palavras-chave: Indexação automática. Indexação manual. Representação da informação. Critérios de indexação automática.

ABSTRACT

This study aims to evaluate the criteria for automatic indexing in order to develop software that will be responsible by the extraction of terms which represent the textual content of documents. It is studied the ways to realize the indexing process - manual and automatic - and discussed the application of these criteria for the optimization of the first stage of the process, that is the subject analysis. The volume of documents published in the current literature shows that great efforts to acquire alternative techniques for indexing are required. The manual process has been considered a slow process. Through the study of technical and scientific articles in the area, it was identified the criteria for automatic indexing that are most used. Then we analyze the degree of satisfaction obtained by researchers through their combination. Among the goals achieved, there are: (1) list of the criteria available in literature, (2) characterization of each criterion, (3) list of the criteria most used. Besides these, it is the main objective of this work, with the obtaining of a set of selected criteria, form a set considered ideal by the researcher.

Keywords: Automatic indexing. Manual indexing. Representation of information. Criteria of automatic indexing.

LISTA DE ILUSTRAÇÕES

LISTA DE ABREVIATURAS E SIGLAS

ABNT	Associação Brasileira de Normas Técnicas
ACM	Association for Computing Machinery
ARIST	Annual Review of Information Science and Technology
BCI	Biblioteconomia e Ciência da Informação
BNB	British National Bibliography
BTDECI	Biblioteca de Teses e Dissertações do Programa de Pós-Graduação da Escola de Ciência da Informação da UFMG
CIS	Faculdade de Computação e de Ciência da Informação da Cornell University
IDF	Indexação Estatística de Termos por Freqüência
KWIC	Keyword in Context – Palavra-chave no contexto
KWOC	Keyword out of Context – Palavra-chave fora do contexto
LC	Linguagem Controlada
LISA	Library and Information Science Abstracts
LN	Linguagem Natural
MC	Mapa Conceitual
Mesh	Medical Subject Headings
MHTX	Modelagem Conceitual para Organização Hipertextual de Documentos
NEPHIS	Nested Phrase Indexing System – sistema de indexação de frase encaixada
PLN	Processamento de Linguagem Natural
PRECIS	Preserved Context Indexing System
RI	Recuperação de informação
SE	Sumário Expandido
SLIC	Selective listing in Combination – Listagem seletiva em combinação
SMART	Sistema de Manipulação e Recuperação de Texto
SRI	Sistema de Recuperação da Informação
TFIDF	Term frequency, inverse document frequency
WAIS	Wide Area Information Server

SUMÁRIO

1 INTRODUÇÃO

Neste capítulo, apresentam-se o contexto e a proposta da pesquisa, suas justificativas, seu problema, seus objetivos e a metodologia utilizada.

Ao organizar a informação, o ser humano tem como meta tornar a recuperação dessa informação um processo simples e eficaz. Para isso, ele conta com processos cognitivos e estruturas informacionais que vêm sendo estudadas ao longo dos anos. Entre esses estudos, encontram-se aqueles desenvolvidos no campo da Biblioteconomia e da Ciência da Informação – BCI.

A recuperação bem sucedida de um documento depende diretamente da qualidade com que este foi tratado no momento de sua inserção em Sistema de Recuperação da Informação – SRI. Em outras palavras, a maneira como os dados são processados e armazenados no SRI reflete diretamente nas possibilidades de o usuário final obter um documento relevante.

Segundo Araújo (1995, p.1-2), os SRIs são aqueles sistemas que, entre outras funções, objetivam dar acesso às informações contidas em documentos neles registrados. Os documentos, nesses sistemas, contêm informação potencial e são formalmente organizados, processados e recuperados com a finalidade de maximizar o uso da informação.

Assim, no contexto deste trabalho, assume-se que um SRI pode ser entendido como um sistema que possibilita a entrada de informações, efetua o processamento/tratamento dessas informações e permite sua recuperação pelo usuário. Na entrada, os documentos são representados através da atividade de indexação; no processamento, os documentos e suas representações, obtidas a partir da indexação, são armazenados; finalmente, na saída, as informações podem ser recuperadas pelo usuário.

Indexar é a atividade de representar um documento através de uma descrição abreviada de seu conteúdo. Essa representação é feita a partir da análise do conteúdo do texto-fonte, que, necessariamente, deveria ser feita por especialistas, que tivessem um olhar atento para metodologias e procedimentos. Existem, pelo menos, duas maneiras de se realizar esse processo: indexação manual e indexação automática. Ambas serão descritas detalhadamente nos capítulos dedicados ao referencial teórico.

No contexto da BCI, os estudos sobre essas duas formas de indexação têm sido continuamente abordados por pesquisadores da área: Vieira (1988); Fugmann (1993); Svenonius (1993; 2000); Guedes (1994); Gil Leiva (1999); Moens (2000), Lancaster (2004); Silva, Fujita (2004); Loehrlein, *et al*. (2005); Dias e Naves (2007); Oliveira, *et al*. (2007); Robredo (1982; 1999; 2005); Lima, Boccato (2009). As investigações sobre a indexação tradicional apontam para alternativas que objetivam melhorar a capacidade de abstração do profissional e tornar a representação temática o mais próximo possível do conteúdo tratado pelo autor. Em relação aos estudos sobre a indexação automática, nota-se que seu surgimento se deu devido à necessidade de serem resolvidos problemas como a morosidade trazida pela indexação manual. Por isso, a indexação automática é vista como uma alternativa para agilizar esse processo, através dos recursos oferecidos pela tecnologia.

Os estudos nessa área começaram a surgir a partir da década de 1950, apontando critérios necessários para aumentar a eficiência dos *softwares* destinados à realização do processo da indexação automática. Nota-se que, apesar de ter quase seis décadas de estudo, a indexação automática não se desenvolveu tanto quanto poderia, devido à sua dependência em relação ao desenvolvimento tecnológico. Assim, esta pesquisa procurou estudar os critérios de indexação automática baseados na literatura da área publicada do final da década de 1950 ao ano de 2008 e verificar se eles têm sido úteis no desenvolvimento de aplicativos ao longo desse período.

1.1 **Justificativa**

Esta pesquisa foi motivada pela percepção de um problema vislumbrado através de discussões realizadas no grupo de pesquisa “Modelagem Conceitual para Organização Hipertextual de Documentos – MHTX”. Esse grupo tem como temática o estudo de um modelo hipertextual para organização de teses e dissertações, com navegação e recuperação em contexto, através de mapas conceituais com interfaces hiperbólicas (LIMA, 2004) .

O grupo teve sua origem a partir da pesquisa de doutorado intitulada "Mapa Hipertextual (MHTX): um modelo para organização hipertextual de documentos”. O protótipo desenvolvido nessa pesquisa deverá ser implementado na Biblioteca de Teses e Dissertações do Programa de Pós-Graduação da Escola de Ciência da Informação da

UFMG – BTDECI. No momento, várias pesquisas têm sido desenvolvidas para que esse protótipo se torne um *software* aplicável ao gerenciamento de bibliotecas digitais de qualquer área do conhecimento humano (LIMA, 2007).

Na implementação tecnológica do MHTX, tornou-se notável o problema de indexar manualmente todas as teses e dissertações a serem inseridas na base de dados. Essa questão serviu de incentivo para a realização deste estudo, que tem como intuito estudar melhorias para o processo automático de indexação.

Além disso, pode-se discorrer sobre outras dificuldades na indexação manual que justificam esta pesquisa. Entre elas, é possível citar o tempo limitado que o indexador tem para fazer a análise de assunto manualmente sem diminuir a qualidade da indexação. Para isso, propõem-se as técnicas de indexação automática. Embora algumas delas não sejam totalmente satisfatórias, elas podem agilizar o processo realizado manualmente, fazendo uma extração inicial de termos, deixando para o indexador o trabalho de selecionar aqueles mais adequados para representar o documento. Outro benefício que pode ser trazido pela técnica automática de indexação é a redução da subjetividade, característica inerente à realização manual desse processo.

A literatura da área aponta outros problemas práticos da indexação manual, tais como: (1) quando diferentes indexadores atribuem diferentes termos a um mesmo documento; (2) quando o mesmo indexador atribui diferentes termos a um mesmo documento, em momentos distintos; (3) o conhecimento do indexador sobre o assunto tratado, que influenciará no nível de consistência atingido na atividade; (4) as possibilidades do indexador em acompanhar a dinamicidade do conhecimento e (5) a capacidade de compreensão do idioma do documento tratado.

As dificuldades inerentes ao processo de indexação manual apresentadas anteriormente; a grande quantidade de documentos publicada diariamente em forma impressa e digital, no Brasil e no mundo; e a necessidade de indexação desses documentos com rapidez justificam esta pesquisa.

Assim, é possível a definição do **problema de pesquisa**, a partir do qual questiona-se: q*uais são os critérios de indexação automática necessários para o desenvolvimento de um software capaz de extrair termos representativos do conteúdo de documentos textuais, aproximando-se ao máximo do trabalho realizado pelo profissional indexador?*

1.2 **Objetivos**

O objetivo geral desta pesquisa é propor um conjunto de critérios de indexação automática para o desenvolvimento de *softwares* que sejam capazes de automatizar o processo de extração de termos representativos de documentos armazenados em bibliotecas digitais.

Especificamente, objetiva-se:

- Fornecer parâmetros comparativos para a melhoria da indexação automática de documentos textuais;
- Melhorar o processo da indexação automática através de critérios relevantes para a extração de termos representativos do conteúdo do documento;
- Auxiliar os profissionais da ciência da computação e de áreas afins no desenvolvimento de *softwares* de indexação automática;
- Contrubuir com os profissionais e pesquisadores da BCI em suas rotinas de trabalho, como indexação de documentos textuais e realização de pesquisas bibliográficas.

1.3 **Apresentação da dissertação**

A organização do texto resultante desta pesquisa baseia-se em capítulos assim divididos:

Capítulo 1 – Introdução: apresenta os fatores que impulsionaram o desenvolvimento desta pesquisa, suas justificativas, o problema central, ou seja, a pergunta que deve ser respondida ao final do trabalho, além dos objetivos que devem ser atingidos e a metodologia, ou seja, o método utilizado para a investigação realizada.

Capítulo 2 – Fundamentos históricos, teóricos e metodológicos: esse capítulo apresenta duas partes principais, nas quais são focalizadas duas das formas de indexação existentes, a manual e a automática. Na primeira parte, sobre a indexação manual, são definidos o profissional responsável por executar a atividade e as etapas envolvidas na execução. Na segunda parte, são apresentados os tipos de indexação

automática, o histórico de desenvolvimento dessa área e sua relação com a lingüística computacional, especificamente a lingüística de *corpus* e o Processamento de Linguagem Natural – PLN.

Capítulo 3 – Combinação de critérios de indexação para a automatização do processo: apresenta o processo de identificação, sistematização, análise e interpretação dos dados.

Capítulo 4 – Apresentação dos resultados: apresenta a conclusão que pretende solucionar o problema de pesquisa proposto no capítulo 1.

Capítulo 5 – Considerações finais: apresenta o entendimento conclusivo do trabalho e aponta a possibilidade de estudos futuros.

2 FUNDAMENTOS HISTÓRICOS, TEÓRICOS E METODOLÓGICOS

Neste capítulo, apresentam-se os aportes históricos, teóricos e metodológicos sobre os quais esta pesquisa se sustenta. A revisão de literatura abordou os temas centrais de interesse deste trabalho para uma discussão a respeito das maneiras de se realizar a indexação de documentos textuais. O objetivo desta etapa da pesquisa é demonstrar como estas duas formas de indexação (a manual e a automática) evoluíram ao longo das últimas décadas.

2.1 Indexação manual

Indexar é representar um documento por uma descrição abreviada de seu conteúdo, com o intuito de sinalizar sua essência. É a atividade de selecionar ou definir termos (palavras ou expressões) que descreverão o conteúdo de um determinado documento, sempre levando em consideração uma clientela específica. Essa representação é feita a partir da análise do conteúdo do texto-fonte e, necessariamente, deveria ser feita por especialistas, que devem seguir metodologias e procedimentos específicos.

Segundo Navarro (1988), o ato de indexar é definido pela tradução do conteúdo de um documento em palavras que tornem possível sua recuperação. Entretanto, nesse processo, são observadas significativas dificuldades na interseção entre lingüística e indexação.

A capacidade íntima de reconhecer sobre o que trata o documento em análise é a questão central do procedimento de indexação. Para fins de indexação, os termos selecionados são a correlação comportamental sobre o que se pensa e 'sobre o que o documento trata', pois seria o termo usado para se procurar por tal documento (MARON, 1977 *apud* GUEDES, 1994).

Segundo o UNISIST (1981), o processo utilizado para descrever e identificar um documento de acordo com seu assunto é denominado indexação. Embora algumas vezes a complexidade desse processo possa ser minimizada, essa é, sem dúvida, umas das atividades mais apuradas do cotidiano do bibliotecário. Para Naves (2004, p. 3-4),

> No campo do tratamento da informação, o termo indexação apresenta dois sentidos: um mais amplo, quando se refere à atividade de criar índices, seja de autor, de título, de assunto, tanto de publicações (livros, periódicos), quanto de catálogos ou de banco de dados, em bibliotecas ou centros de informação. O outro sentido, mais restrito, refere-se à indexação, classificação ou catalogação de assuntos das informações contidas em documentos.

Para a execução da atividade, faz-se necessária uma formação especializada voltada para o tratamento do conteúdo de documentos. A experiência e o conhecimento prévio deste profissional diferenciarão o resultado do tratamento realizado por outro indexador menos preparado. Naves (2004), afirma que:

> O profissional da informação que desenvolve a atividade de indexar assuntos de documentos é chamado de *indexador*, catalogador de assuntos ou classificador. A maioria desses profissionais é graduada em Biblioteconomia, e deve conhecer os fundamentos teóricos e técnicos do tratamento temático da informação.

Contudo, há alguns autores, como Barbosa, que acreditam que "não há uma definição universalmente aceita a respeito de quem se constitui o profissional da informação, podendo este ser um pesquisador, um engenheiro, projetista, desenhista industrial, gerente, contador" (BARBOSA, 1998).

Neste trabalho, em concordância com Naves (2004), será utilizado o termo 'indexador' para designar o profissional que realiza a atividade de indexação. E acredita-se que o bibliotecário seja o profissional mais qualificado para exercer tal função.

Por se tratar de uma atividade intelectual, é natural que, no cotidiano dos indexadores, sejam percebidas divergências entre termos atribuídos a um mesmo documento por diferentes profissionais de instituições e em contextos diferentes. Pode-se imaginar uma publicação que aborde, por exemplo, vinte assuntos de maneira significativa, e que a biblioteca cujo acervo a obra compõe determine um limite de atribuição de dez assuntos para cada publicação. Logo, alguns assuntos terão que ser desconsiderados ou englobados por assuntos semelhantes sob a indicação de um termo mais genérico correspondente.

Assim, segundo com Lancaster (2004), uma mesma publicação poderá apresentar conjuntos diferentes de termos de indexação, dependendo do grupo de usuários ao qual se destina e dos interesses particulares desse grupo, ou seja, há várias formas 'corretas' de se indexar.

De acordo com Fujita (2003, p. 65), a partir do texto *Indexing Principles,* escrito pelo UNISIST em 1976, elaborou-se a primeira norma para análise, identificação de assuntos e seleção de termos de indexação, publicada pela International

Standardization for Organization em 1985. A norma apresenta-se sob o número 5693 e é intitulada *Documentation - methods for examining documents, determining their subjects, and selecting indexing terms*. Assim, de acordo com a ISO 5693, a indexação é vista como "[...] a representação do conteúdo dos documentos por meio de símbolos especiais, quer retirados do texto original, quer escolhidos numa linguagem de informação ou de indexação".

Segundo Silva e Fujita (2004, p. 136-137), "o conceito de indexação surgiu a partir da elaboração de índices e atualmente está mais vinculada ao conceito de análise de assunto". Com a necessidade de se recuperar a informação de uma maneira cada vez mais rápida, precisa e especializada, a prática de elaboração de índices passou a priorizar o contexto de cada documento. As autoras acreditam que a atividade de indexação, como um processo, passou a ocorrer com maior intensidade devido ao aumento das publicações periódicas e da literatura técnico-científica. Considerando a evolução do tratamento da informação, a indexação pode ser entendida como uma operação de tratamento temático, que comporta as atividades de análise, síntese e representação do conteúdo do documento.

Neste trabalho, contudo, considera-se que o processo de indexação manual compreende duas etapas principais: a análise de assunto e a tradução dessa análise, ou seja, do conteúdo do documento em termos de indexação.

2.1.1 **Análise de assunto**

A etapa de análise de assunto determina de que trata um documento, isto é, qual seu assunto. Para tanto, a leitura e a compreensão do texto são primordiais, porém, o tempo restrito do indexador e a quantidade cada vez maior de documentos que demandam tratamento são fatores preocupantes, porque podem comprometer a qualidade da atividade realizada. Segundo Lancaster (1993, p. 20-21), "ao indexador raramente é dado o luxo de poder ler um documento do começo ao fim", realidade que pode ser comprovada no cotidiano de indexadores de instituições nacionais e internacionais.

Para a execução desta primeira etapa, é preciso considerar o domínio no qual o documento está inserido, identificando as características específicas do campo de conhecimento, sejam elas de ordem cultural, terminológica, históricas ou lingüísticas. Para

tanto, o conhecimento do indexador sobre esse domínio é importante para a qualidade da análise. Assim, a atividade será feita de acordo com o contexto, pois o documento não será considerado como uma parte isolada, mas como parte de um todo (HJORLAND, 1992).

A análise de assunto pode ser considerada a etapa mais importante do processo de indexação. De acordo com Cesarino (1980), a análise de assunto é a operação-base para todo procedimento de recuperação de informação. Isso justifica o fato de todo SRI utilizar essa técnica, tanto na etapa de entrada de dados no sistema quanto no momento de busca da informação pelo usuário final.

De acordo com algumas pesquisas, como UNISIST (1981, p.83) e Fujita (2003, p. 64), a análise de assunto é dividida em três estágios:

1 - Compreensão do conteúdo do documento como um todo;
2 - Identificação dos conceitos que representam esse conteúdo;
3 - Seleção dos conceitos válidos para recuperação.

De acordo com o UNISIST (1981), “na prática, esses três estágios se superpõem". Para Fujita (2003, p. 64), esta superposição ocorre no momento da leitura do documento.

Para a *compreensão do conteúdo do documento*, de acordo com as recomendações do texto do UNISIST (1981), a leitura integral é ideal, embora seja impraticável. Assim, delimita-se um roteiro das partes entendidas como mais importantes para que o indexador possa se nortear durante a leitura:

- Título;
- Introdução e as primeiras frases de capítulos e parágrafos;
- Ilustrações;
- Tabelas;
- Diagrama e suas explicações;
- Conclusão;
- Palavras ou grupos de palavras sublinhadas ou impressas com tipo diferente.

O texto ressalta a importância de o indexador não se dedicar exclusivamente às partes preliminares do texto, uma vez que estas demonstram a intenção do autor, enquanto as partes finais indicam o alcance dessas intenções.

Para Naves (2004, p. 5),

> O texto é o ponto de partida para operações analítico-documentárias, sendo constante a presença dos elementos conteúdo e forma, como partes essenciais do mesmo. Para ter uma competência textual é preciso que, além de conhecer o texto que tem em mãos para análise sob todos os aspectos, o indexador faça dele uma leitura adequada [...].

É preciso destacar que, para a assimilação e compreensão do conteúdo do texto, é necessário o processamento de informações na mente do indexador. De acordo com Naves (2004), "parece ser consenso entre os especialistas em leitura que o processamento do ato de ler se dá interativamente, dependendo dessa interação a não compreensão ou a compreensão de um texto".

NAVES (2004) apresenta "dois tipos de processamento mental da informação: o *top-down* e o *bottom-up*, que parecem ocorrer simultaneamente na mente humana ao fazer a leitura de um texto". Estes são inversos e complementares e chamados por alguns autores de *modelos de leitura*. Define-se o modelo *Bottom-up* como ascendente, guiado por dados, indutivo, no qual a leitura é linear, com origem nas partes em direção ao todo textual. Por sua vez, o modelo *top-down* define-se da seguinte maneira: descendente, dedutivo, no qual se move na direção inversa, obtendo vantagem da base de conhecimento do leitor. Os dois modelos tratam-se de uma dupla ação: percepção e compreensão.

Após o estágio de compreensão do conteúdo do documento como um todo, passa-se para o segundo momento, que corresponde à identificação dos conceitos do documento que melhor podem representar seu conteúdo.

Dentro do estágio de *identificação dos conceitos que representam o conteúdo do texto*, o indexador deve abordar o texto de maneira lógica, selecionando os conceitos que representem adequadamente o assunto do documento. Para isso, recomenda-se que seja obedecido um esquema de categorias da área coberta pelo documento, como o indicado a seguir:

- O fenômeno;
- O processo;
- As propriedades;
- As operações;
- O material;

•O equipamento, entre outros aspectos (UNISIST 1981, p. 87).

Contudo, antes do detalhamento da atividade propriamente dita, faz-se necessária a definição do termo *conceito*:

> Conceitos são unidades do conhecimento identificadas através de enunciados verdadeiros sobre um item de referência, representados por um termo ou palavra. É uma idéia, uma representação mental que nos permite categorizar objetos. Existem tipos de conceitos, como os individuais (representados por nomes de coisas individuais, em linguagem simples), os gerais (representados por nomes de classes de coisas e podem ser expressos em uma multiplicidade de expressões lexicais e não lexicais), os científicos (podem ser aprendidos na vida acadêmica, e, depois, são estabelecidas conexões entre eles e os eventos da vida diária), os cotidianos (são objetos usados normalmente, como roupas, carros) (NAVES, 2004, p. 6).

O relacionamento entre os conceitos de um documento corresponde a um aspecto relevante neste estágio da análise de assunto. De acordo com Naves (2004, p. 7-8), essas relações podem ser de três tipos: (1) relação de equivalência; (2) relação hierárquica e (3) relação associativa (afinidade ou de coordenação).

São exemplos de relações de equivalência:

•Abreviaturas e acrônimos: CRB X Conselho Regional de Biblioteconomia
•Grafias diferentes: contato X contacto
•Quase sinônimos: datiloscopia X impressão digital
•Sinônimos: trivial X comum
•Traduções: delete X apagar

São exemplos de relações hierárquicas:

•Gênero/espécie
•Quase genérica: cachorro X animal de estimação
•Todo/parte: computador X teclado
•Verdadeiramente genérica: cachorro X mamífero

São exemplos de relações associativas:

•Concorrente: técnica X prática

- Coordenação: sol X lua
- Genética: avó X neto
- Instrumental: pintura X pincel
- Material: livro X papel
- Similaridade de processo: indexação X classificação

Destacam-se, neste estágio, a *exaustividade* e a *especificidade*. A exaustividade relaciona-se à capacidade do sistema de indexar o documento em profundidade, ou seja, além do assunto principal, são indexados também os assuntos secundários. Já a especificidade é a capacidade do sistema de permitir ao indexador ser preciso ao especificar o assunto de um documento (NAVES, 2004, p. 7-8).

Para Naves (2004), esses aspectos dependem da *política de indexação* adotada pela biblioteca. Essa política é composta de um manual, que é próprio de cada biblioteca e contém regras bem definidas para a realização do processo de indexação. São critérios pré-estabelecidos, que consideram os objetivos do sistema, os recursos disponíveis e o tipo de usuário que tal sistema pretende atingir.

Também é possível o detalhamento deste estágio abordando os tipos de assuntos a serem definidos pelo indexador. Naves (2004, p. 7-8) aponta para três principais tipos de assuntos: *assunto simples; assunto composto* e *assunto complexo.* O assunto simples é formado por apenas um conceito, ex: biblioteconomia. O assunto composto é formado por mais de um conceito, porém, que pertençam a uma mesma área do conhecimento, ex: indexação em bibliotecas digitais, que são dois conceitos da biblioteconomia. O assunto complexo é formado por conceitos de áreas diferentes, ex: administração de bibliotecas universitárias, que são conceitos da administração e da biblioteconomia.

O processo cognitivo está intrinsecamente ligado ao processo de identificação da idéia principal apontada em um texto. Para Fujita (2003, p. 69),

> O processo de análise de assunto reveste-se de uma subjetividade característica, dadas as circunstâncias e elementos envolvidos, pois, a partir da leitura do documento pelo indexador, é realizado um processo de comunicação interativo entre três variáveis: leitor, texto e contexto. Cada uma dessas variáveis estará sujeita a diferentes condições, mas é o indexador como leitor a variável mais influente nessa interação para análise de assunto, porque precisa realizar a compreensão da leitura mediante sua cognição.

Após o estágio de identificação dos conceitos representativos do conteúdo do texto, passa-se para o terceiro estágio da análise de assunto, que corresponde à seleção daqueles conceitos que são julgados como úteis para a recuperação do documento indexado.

A atividade de *seleção de conceitos* é necessária, tendo em vista os objetivos para os quais as informações são indexadas. Assim, nem todos os conceitos identificados serão necessariamente selecionados. São exemplos de objetivos da atividade da indexação: (1) elaboração de índices alfabéticos impressos e (2) armazenamento mecânico dos dados para recuperação em meios computadorizados (UNISIST 1981, p. 87).

Há pesquisadores, como Lancaster (2004) e Naves (2004), que definem a atividade de seleção de conceitos como a *determinação da atinência*. Em tradução para o inglês, a expressão corresponde ao termo *aboutness,* porém, encontram-se na literatura outras traduções: concernência, sobrecidade, temacidade.

O final do estágio é indicado com a definição da chamada *frase de indexação*. Esta é elaborada pelo indexador em Linguagem Natural – LN. Após todo o processo intelectual de leitura e compreensão do texto, de identificação e seleção de conceitos representativos do documento em foco, o indexador deve afirmar: *Este documento trata de.....* . A partir dessa definição, o indexador pode passar para a etapa final do processo de indexação, a tradução da análise de assunto em termos de indexação.

2.1.2 **Tradução da análise de assunto**

A tradução da análise de assunto tem o objetivo de converter o assunto do documento em um conjunto de termos de indexação. Essa análise vai acontecer mesmo em casos nos quais não houver prescrição de regras formais. Tais regras podem ser estipuladas em função dos interesses da instituição ou do instrumento de controle terminológico. Esse controle pode ser feito a partir do uso de um *vocabulário controlado*, sendo que, muitas vezes, essa tarefa é feita de forma intuitiva.

Vocabulário controlado é definido por Naves (2004, p. 10-11), como um tipo de *linguagem artificial*. Esta é construída cuidadosamente, através do processo minucioso da escolha dos termos a serem usados, o significado de cada um e as relações que apresentam.

Alguns dos principais vocabulários controlados utilizados no âmbito da Biblioteconomia são: Taxonomia, Tesauro, Lista de Cabeçalho de Assunto, Classificações Bibliográficas.

Por se tratar de um processo intelectual realizado por um indivíduo, mesmo este sendo especializado para tal função, a indexação é uma atividade complexa em que é possível perceber significativas dificuldades. A seção seguinte tratará especificamente dessas dificuldades.

2.1.3 **Dificuldades do processo manual de indexação**

A atividade de leitura e de interpretação do documento-fonte pelo indexador representa uma atividade extremamente complexa. Fujita (1999 e 2003b) realizou pesquisas no âmbito da leitura documentária no intuito de observar o processo de interpretação do indexador. A autora verificou que o profissional geralmente apresenta dificuldades no que se refere à identificação e seleção de conceitos representativos do assunto de documentos. A preocupação da autora era a elaboração de diretrizes para a composição de um modelo de leitura documentária para auxílio do indexador.

A atividade mental da leitura é complexa e realizada pelo bibliotecário não apenas para o processo de representação temática da informação (indexação), mas também para a classificação e representação descritiva (catalogação). Para Neves, Dias e Pinheiro (2006, p. 143),

> [...] a leitura do indexador não está relacionada apenas à identificação das superestruturas e dos esquemas textuais; vai além, pois exige a incorporação de uma série de atitudes, normas e habilidades que podem vir a ser automatizadas em nível de atitudes, mas que exigem do profissional uma adaptação freqüente [...].

Durante a atividade da leitura, é necessário um monitoramento do processo, em que aspectos cognitivos são evidenciados. Para Neves; Dias; Pinheiro (2006), é relevante o monitoramento da compreensão e do uso de estratégias metacognitivas no processamento do texto. Segundo Pressley (1995), os indexadores preparam-se antecipadamente para ler um texto. Eles observam aspectos como o grau de dificuldade e a extensão do texto. Identificam também as partes que serão lidas mais atentamente e as

que serão ignoradas ou apenas esquadrinhadas visualmente na busca de elementos prospectivos.

É necessária uma organização das atividades a serem realizadas no momento da leitura. Neves *et al.* (2006), aponta para algumas dessas atividades:

- Identificação de informações relevantes;
- Leitura das partes aparentemente mais importantes;
- Realização de inferências;
- Leitura em voz alta;
- Repetição e reformulação de uma idéia na busca de sua correspondência na memória de trabalho;
- Registros de notas;
- Pausas para reflexão sobre o texto;
- Elaboração de paráfrases;
- Busca de padrões textuais;
- Realização de predições;
- Relacionamento de partes do texto na busca de esclarecimento de dúvidas;
- Interpretação do texto;
- Emissão de juízos de valor sobre a qualidade do texto e a veracidade do relato, entre outras.

Pode-se perceber, então, que a atividade de leitura documental demanda grandes esforços cognitivos, sendo considerada por alguns pesquisadores como a parte mais exaustiva da indexação. Assim, o desafio do processo de indexação e, conseqüentemente, do indexador, é perceber e revelar a parte essencial do conteúdo de um documento, ou seja, expor o esqueleto semântico que sustenta toda a estrutura textual, sem a qual todo o restante do texto não faria sentido.

Percebe-se que, no exercício de sua atividade, o indexador dispõe de um tempo extremamente limitado, o que pode representar um considerável prejuízo à qualidade do trabalho. Embora determinadas técnicas de indexação automática não sejam totalmente satisfatórias, elas podem agilizar o processo realizado manualmente, fazendo uma extração inicial de termos.

Percebe-se também que a forma manual de se indexar documentos textuais vem se revelando inadequada para minimizar a subjetividade inerente a essa atividade.

Segundo Carvalho (1995), na análise em um sistema documental, é elevadíssimo o grau de envolvimento humano. Todo texto escrito é, simultaneamente, veículo de informação e meio de comunicação e utiliza a linguagem como ligação entre os homens.

Eco (1992) ratifica a tese de Carvalho, explicando que o universo das palavras permite interpretações múltiplas, e que várias leituras são válidas. Para o autor, ler é criar de novo o texto, e analisá-lo é recriá-lo, fazendo-o renascer. Diante disso, questiona-se: como encontrar a objetividade em um processo em que o sujeito e o objeto se misturam na recriação textual? Assim, a subjetividade representa um dos principais problemas da indexação manual. Para minimizá-la, são necessários significativos avanços computacionais.

Para o processo de indexação ser considerado consistente, é indispensável que se observe criteriosamente a política de indexação adotada pela instituição. Faz-se necessário, também, que exista um sistema adequado, que permita o intercâmbio de informações com outros centros de documentação. Além disso, é importante que o profissional faça registros das decisões por ele tomadas e que tenha um olhar criterioso durante toda a atividade. Neste contexto, como indicado na introdução deste trabalho, segundo Borko (1977), dois problemas práticos podem ser apontados: (1) a inconsistência interindexadores, que ocorre quando diferentes indexadores atribuem diferentes termos a um mesmo documento, e (2) a inconsistência intraindexador, que ocorre quando o mesmo indexador atribui diferentes termos a um mesmo documento, em momentos distintos.

Outra dificuldade observada é o conhecimento do indexador sobre o assunto tratado, o que influenciará no nível de consistência atingido na atividade. No Brasil, o curso de graduação em Biblioteconomia normalmente apresenta duração de quatro anos, e, após esse período, o graduado pode exercer sua profissão em diversas áreas do conhecimento, como direito, medicina, engenharia ou qualquer outra biblioteca. Dessa maneira, grande parte desses profissionais começa a atuar em ambientes completamente distintos de sua área de origem, o que faz com que ele precise de determinado tempo para se familiarizar com a terminologia daquela área, conhecer suas principais disciplinas, enfim, para dominar os mecanismos necessários para o desenvolvimento de um trabalho de qualidade.

Como visto, a indexação é uma atividade mental complexa. Para Neves, Dias e Pinheiro (2006, p.142),

> Para compreender um texto, os indivíduos lançam mão de todo o conhecimento prévio armazenado na memória de longo prazo, demandando, inclusive, possíveis esquemas de procedimento existentes na memória semântica. O conhecimento anterior facilita o processamento do texto e a compreensão, por oferecer uma estrutura na qual o conteúdo do material lido possa ser relacionado. A integração do conhecimento passado com o texto que está sendo lido permite aos leitores formar o que é chamado por Van Dijk e Kintsch (1983) e Kintsch (1998) de "modelo situacional". Este consiste na combinação das informações (ou proposições – unidades abstratas de significado) retiradas do texto com as proposições formadas a partir de conhecimentos gerais preestabelecidos e da experiência pessoal dos leitores.

Assim, naturalmente, a experiência do indexador refletirá diretamente no processo executado.

As possibilidades que o indexador possui de acompanhar a evolução do conhecimento humano também pode influenciar na qualidade da indexação. O conhecimento é extremamente dinâmico, o que exige do indexador permanente atualização. Infelizmente, não é isso que ocorre com grande parte dos casos. Mesmo nos casos em que o indexador já possui grande domínio em sua área de atuação, é necessário que ele acompanhe as transformações desse ambiente, reconhecendo, por exemplo, termos ultrapassados, termos novos, ambíguos, sinônimos, entre outros aspectos.

A capacidade de compreensão do idioma do documento tratado que o profissional apresenta é fundamental para o processo de indexação. Ao se realizar tal atividade, o indexador deve ter clareza e objetividade no entendimento do conteúdo do texto. Não se julga essencial que o profissional tenha fluência no idioma, mas é necessário um mínimo de familiaridade com a língua e uma habilidade avançada para leitura, compreensão e tradução do idioma. Infelizmente, ter domínio de outras línguas, em especial da língua inglesa, não é a realidade de grande parte dos profissionais da área.

As dificuldades inerentes ao processo de indexação manual apresentadas nesta seção; a grande quantidade de documentos publicada diariamente no Brasil e no mundo; e a necessidade de indexação desses documentos, justificam estudos que buscam soluções alternativas para auxiliar o indexador no exercício de sua atividade.

Assim, na década de 1950, tiveram início os estudos sobre o processo de indexação automática, em que recursos computacionais foram pensados, tendo em vista agilizar a etapa de análise de assunto do processo de indexação.

2.2 **Indexação automática**

Também chamada de *indexação assistida por computador* e de *indexação semi-automática*, esse tipo de indexação é considerada um modelo de extração com características estatísticas e probabilísticas. Sua origem coincide com as tentativas iniciais de junção da informática e da estatística com a área de documentação. Para Moreiro González (2004, p. 3 *apud* BUFREM, 2005),

> [...] A essência do processo é a identificação automática de palavras-chave no texto pela freqüência com que aparecem e sua fundamentação teórica tem origem na lei de Zipf. Novas formulações desta Lei originaram outras técnicas de discriminação dos termos, sobre as quais discorre o autor, destacando a indexação estatística de termos por freqüência, conhecida pela sigla IDF, a *Term frequency, inverse document frequency* (TFIDF), o método *N-grams*, que modifica a lei de Zipf para possibilitar o tratamento de palavras compostas e os *Stemmers*, que utilizam a freqüência com que aparecem seqüências de letras no corpo de um texto para extrair a raiz das palavras. Além dessas possibilidades, as relações semânticas entre os termos lingüísticos podem ser estabelecidas por métodos de agrupamento e classificação.

A indexação automática visa à mecanização das atividades descritas anteriormente na seção 2.1, com o objetivo de agilizar e auxiliar o processo intelectual realizado pelos profissionais da área. Segundo Robredo (1982), “o processo de indexação automática é similar ao processo de leitura-memorização humano, sendo seu princípio geral baseado na comparação de cada palavra do texto com uma relação de palavras vazias de significado”. Essa relação deve ser previamente estabelecida e o resultado dessa comparação conduz, por eliminação, a considerar que as palavras restantes do texto são palavras significativas.

O processo de leitura realizado por um indivíduo qualquer apresenta semelhança com o processo de indexação. Para Robredo (1982), ao ler um texto, o leitor não apresenta interesse pelas letras, mas pela idéia que elas representam quando organizadas em palavras ou em conjuntos de palavras. O olho é considerado a janela do cérebro: ele reconhece as palavras significativas e suas associações, fixando-se nelas um

tempo necessário para assegurar a memorização das idéias, ignorando, praticamente, as palavras não significativas.

Pode-se separar o processo de memorização humana em duas etapas principais: (1) memorização temporária e inconsciente – nesta etapa, há a conservação das palavras significativas passando por uma modificação ou aperfeiçoamento a partir da detecção de novos conceitos significativos; (2) memorização permanente dos conceitos assim trabalhados, à qual se atribui o nome de *memória*. Segundo Robredo (1982), depois de ocorridas essas duas etapas, o leitor fixa em sua memória uma série de *palavras-conceitos-descritores* que representam as idéias básicas do documento por ele lido. A leitura, através de um processo de análise-indexação, leva ao armazenamento dos descritores que representam o conteúdo dos documentos.

O histórico da indexação automática pode ser associado ao uso de programas computacionais para geração de índices pré-coordenados (ROBREDO, 1982). Para Naves (2004), são exemplos de linguagens pré-coordenadas: “listas de cabeçalhos de assunto (Library of Congress, Rovira, Wanda Ferraz), os índices permutados, os índices em cadeia e as classificações bibliográficas (Classificação Decimal de Dewey – CDD, Classificação Decimal Universal – CDU)”.

Segundo Lancaster (2004, p. 52), “vários programas de computador foram desenvolvidos para gerar, automaticamente, um conjunto de entradas de índices a partir de uma seqüência de termos”. Como exemplos, existem os modelos KWIC, o KWOC, o SLIC, o PRECIS e o NEPHIS, descritos a seguir.

No final da década de 1950, desenvolveram-se métodos relativamente simples para a construção de índices a partir de textos, especialmente a partir de palavras que ocorrem nos títulos dos documentos. São exemplos desses métodos o *Keyword in Context* – KWIC (Palavra-chave no Contexto) – e o *Keyword out of Context* – KWOC (Palavra-chave fora do Contexto) (LANCASTER, 2004, p.54).

O KWIC foi desenvolvido por H. P. Luhn em 1959 e corresponde a um índice rotativo em que cada palavra-chave que aparece nos títulos dos documentos torna-se uma entrada do índice. Cada palavra-chave é destacada de alguma forma e as palavras restantes do título aparecem “envolvendo-a”. O critério usado para selecionar as palavras que irão compor o índice é chamado de processo “reverso”, ou seja, o programa reconhece as palavras que não são palavras-chave, baseando-se em uma lista de palavras proibidas, e impede que elas sejam adotadas na entrada. Os vocábulos dessa lista de palavras proibidas têm função sintática (artigos, preposições, conjunções, etc.),

mas em si mesmos não representam conteúdo temático. O KWIC é um método simples, barato e que obtém, em certo nível, acesso temático ao conteúdo de uma coleção. É nítido, porém, que sua qualidade está diretamente relacionada à qualidade dos títulos, no sentido em que estes devem ser bons indicadores do conteúdo dos textos (LANCASTER, 2004, p.54 e 55).

Para Robredo (1982, p. 238), a primeira aplicação generalizada da indexação automática de documentos técnicos, a partir de palavras significativas dos títulos, se deu com o modelo KWIC, produzindo índices de títulos permutados.

O método KWOC é semelhante ao KWIC, porém as palavras-chave que se tornam pontos de acesso são repetidas fora do contexto, normalmente destacadas no canto esquerdo da página ou usadas como cabeçalhos de assunto.

O *Selective Listing in Combination* – SLIC (Listagem Seletiva em Combinação) – foi criado por J. R. Sharp em 1966. O programa organiza a seqüência de termos de um documento em ordem alfabética e elimina as seqüências redundantes. Modelos como o SLIC pressupõem o emprego de termos de indexação e não de texto livre. Já o método PRECIS produz índice impresso baseado na ordem alfabética e na "alteração" sistemática de termos para que ocupem a posição de entrada (LANCASTER, 2004).

De acordo com Belluzzo *et al.* (1990), o Preserved Context Indexing System – PRECIS foi criado pelo Dr. Derek Austin em 1968 e possui uma metodologia de indexação que é utilizada desde então pela British National Bibliography – BNB na produção automática de seus índices de assunto. Este sistema tem sido usado por outras instituições que abrigam sistemas de informação em diferentes países, comprovando sua eficiência para diferentes línguas e documentos.

Ainda segundo Belluzzo *et al.* (1990), no Brasil, a primeira iniciativa de estudo do PRECIS foi tomada em 1978, pela professora Jandira Baptista Assumpção, tendo em vista o desenvolvimento de sua tese de doutorado. Assumpção iniciou uma experiência com bibliotecários da Biblioteca Central da Universidade Federal de Minas Gerais, tendo sido criado um índice de assunto, que foi divulgado na Revista da Escola de Biblioteconomia da UFMG, especialmente dedicado ao assunto. Entretanto, esse trabalho foi interrompido em 1983 devido ao falecimento da pesquisadora.

Conforme afirmado por Lancaster (2004, p.56-59), outro importante sistema desenvolvido foi o *Nested Phrase Indexing System* – NEPHIS (Sistema de Indexação de Frase Encaixada). Este sistema foi criado em 1977 e corresponde a um índice articulado de assunto e foi criado por T. C. Craven. Esse índice foi descrito minuciosamente por

Armstrong e Keen em 1982. Nesse modelo, os termos de entrada são reordenados de tal modo que cada um deles se liga a seu vizinho original por meio de uma palavra funcional ou pontuação especial, conservando-se, assim, estrutura similar à de uma frase, mesmo que muitas vezes disposta em ordem diferente.

De acordo com autores como Edmundson (1969), Garvin (1969 *apud* SALTON, 1973) e Salton (1973), já nesta época, percebia-se a intrínseca relação entre processamento da informação e aspectos lingüísticos, especificamente com a lingüística computacional. Segundo os autores mencionados, os esforços deviam ser voltados para estudos das propriedades estruturais e semânticas das línguas naturais. Contudo, percebe-se que grande parte das metodologias lingüísticas da época geralmente produzia resultados decepcionantes. A importância dos aspectos lingüísticos para o adequado tratamento da informação será apresentada em seção posterior.

Segundo Salton (1970; 1973) e Swanson (1960), a indexação automática apresenta relativos méritos em relação às técnicas manuais. Os pesquisadores afirmavam que era possível extrair automaticamente de textos palavras-chave relevantes, e que, quando estas eram comparadas com aquelas atribuídas por indexadores, constatava-se um acordo entre 60 e 80% dos termos atribuídos. Swanson (1960) acreditava que, embora as máquinas nunca pudessem alcançar um sucesso mais que parcial no processo de indexação, as pessoas eram ainda menos promissoras nessa atividade.

A partir da década de 1970, percebe-se uma intensificação das pesquisas na área de indexação automática de documentos textuais. A fim de se determinar métodos inteiramente automáticos de transformação de textos que poderiam competir em eficácia com a obtenção manual convencional e de se identificar técnicas de operações automáticas mais úteis nesse aspecto, desenvolveram-se inúmeros estudos nessa época. Dois dos mais importantes experimentos baseavam-se no desempenho do SRI MEDlars, que operava no National Library of Medicine, em Washington, e do SRI experimental SMART, criado por Gerard Salton enquanto trabalhava na universidade de Cornell (SALTON, 1973).

O MEDlars opera no modo convencional por ter treinado peritos para atribuir palavras-chave, ou termos-índice, a todos os pedidos recebidos de pesquisa de documentos (THE PRINCIPLES, 1970 *apud* SALTON, 1973). O vocabulário controlado utilizado para dar consistência a esta atribuição é o Medical Subject Headings – Mesh, onde a a recuperação é efetuada através da comparação entre uma lista de palavras-chave determinada para os documentos com os termos de busca das formulações

booleanas de pesquisa. Especificamente, todos os documentos que apresentam combinação adequada de palavras-chave são recuperados em resposta à consulta correspondente, enquanto os documentos que não apresentam essa combinação permanecem no arquivo.

Já o sistema SMART, de acordo com Salton (1968 e 1971), funciona sem qualquer análise manual do conteúdo. Trechos dos documentos – normalmente resumos – e consultas de textos são introduzidos em um computador, e uma variedade de procedimentos automáticos de análise de texto é utilizada para produzir para cada item um 'conceito vetor', constituído por ponderação de termos ou conceitos representativos do conteúdo do documento. Normalmente, cerca de cem diferentes conceitos poderiam ser utilizados para identificar um determinado documento. Seguindo uma comparação entre documento e consulta de vetores, é calculado, para cada par de documentos consultados, um coeficiente de semelhança ou de correlação. Assim, os documentos são submetidos ao usuário em ordem decrescente dos respectivos coeficientes de similaridade.

Uma comparação entre MEDlars e SMART pode ser especialmente adequado na avaliação da eficácia de métodos de processamento automático de texto em um ambiente de recuperação. O MEDlars, porque representa um conhecido sistema convencional que já funcionava há muitos anos em uma grande base de dados de várias centenas de milhares de documentos. O SMART, porque exclui análises técnicas e linguagem automática iterativa, uso controlado de estratégias de busca para substituir a contribuição intelectual dos indexadores e dos usuários em ambientes convencionais. Assim, até meados da década de 1970 estes eram os sistemas utilizados para a recuperação automática de documentos (SALTON, 1973).

Anderson (2000) relata que, no ano de 1982, a indexação assistida por computador era pouco conhecida. O autor realiza uma análise de três artigos de sua autoria sobre indexação automática, publicados entre os anos de 1983 e 1987, em que faz reflexões sobre a evolução da área e as ferramentas utilizadas em prol de seu desenvolvimento.

Para Lancaster (2004), a indexação automática tradicionalmente começa com unidades léxicas individuais. Auto (2007, p.38), afirma que "a tokenização, ou seja, a separação do texto em palavras é definida pelo reconhecimento do texto entre marcas de pontuação". Para Auto, tokenização deriva da palavra *token* e representa uma cadeia de caracteres separada por determinadas marcações. Pode-se considerar como caracteres marcações do tipo hífen, parêntese, apóstrofo, ponto final e até mesmo um espaço em

branco. O autor acredita que a definição dessas marcações não é uma tarefa fácil. Para ilustrar tal dificuldade, Auto aponta para os seguintes exemplos: utilização do hífen para separar uma oração explicativa e para separação de uma unidade lexical composta. No primeiro caso, o hífen deve ser considerado uma marca de tokenização; já no segundo, deve ser ignorado, pois faz parte da palavra e a unidade lexical deve ser considerada como um todo.

Pode-se observar na literatura o apontamento para alguns tipos de indexação automática. A *indexação por extração automática* é um deles. Nesse processo, palavras ou expressões que aparecem no texto são extraídas para representar seu conteúdo como um todo. Considerando uma versão eletrônica desse documento, é possível utilizar um programa computacional para extrair os termos a partir dos mesmos princípios utilizados por seres humanos. São eles: freqüência da palavra dentro do texto, posição da palavra no texto (no título, nas legendas, no resumo etc.) e seu contexto (LANCASTER, 2004).

Na década de 1950, teve início a indexação automática baseada em freqüência de ocorrência de palavras no texto com os trabalhos de Luhn, em 1957, e de Baxendale, em 1958. Baxendale (1958 *apud* LANCASTER, 2004) sugere que, em substituição ao processo que analisa todo o texto, sejam analisados apenas o "tópico frasal" e as "palavras sugestivas". Seus estudos demonstraram que era necessário o processamento apenas da primeira e da última frase de cada parágrafo, pois, em 85% das vezes, a primeira frase era o tópico frasal e em 7% dos casos a última frase o era. Considera-se como tópico frasal a parte do texto que provê o máximo de informações relativas ao conteúdo do texto.

Os sistemas baseados em indexação por extração automática realizam, basicamente, as seguintes tarefas:

- Contar palavras num texto;
- Cotejá-las com uma lista de palavras proibidas;
- Eliminar palavras não significativas (artigos, preposições, conjunções, etc.);
- Ordenar as palavras de acordo com sua freqüência.

Percebe-se que esse tipo de indexação, por ser baseado unicamente em critérios estatísticos, apresenta limitações para a realização do processo de maneira consistente. Semelhante a esse processo específico, porém com uma preocupação

quanto aos aspectos semânticos do texto, pode-se indicar a *indexação por atribuição automática.*

O processo de indexação por atribuição automática é mais complexo de ser realizado com maior eficiência do que o processo de indexação por extração automática. Em geral, é considerada uma atividade difícil, pois, para a representação do conteúdo temático, é necessário um controle terminológico. Deve-se desenvolver, para cada termo atribuído, um "perfil" de palavras ou expressões que costumam ocorrer nos documentos. Por exemplo, para o termo 'chuva ácida', poder-se-ia incluir, entre outras, as seguintes expressões: 'precipitação ácida', 'poluição atmosférica' e 'dióxido de enxofre'. Um problema relevante nesse processo pode ser ilustrado com a seguinte situação: a frase "dois dias depois de a substância haver sido ingerida surgiram diversos sintomas" pode ser legitimamente indexada por uma pessoa sob o assunto 'intoxicação', entretanto, para um programa computacional, essa tarefa é verdadeiramente difícil (O'CONNOR, 1965 *apud* LANCASTER, 2004).

Uma das ferramentas da BCI são os vocabulários controlados. No contexto da indexação automática, acredita-se que a utilização de um determinado tipo de vocabulário controlado, a *taxonomia,* pode trazer resultados positivos.

Para o tratamento da produção científica, pode-se utilizar a linguagem natural ou a linguagem controlada. Na linguagem natural, o termo (palavra ou expressão) será retirado do texto. Na linguagem controlada, há uma lista de termos escolhidos, cuja função é a de admitir somente uma forma de interpretação, ou seja, de significado, além de possibilitar uma maior padronização e rigor de utilização de termos. Nesse contexto, a BCI estuda as possibilidades de elaboração de linguagens documentais que possibilitem identificar o conteúdo, isto é, termos (palavras e expressões) mais significativos e estabelecer relações semânticas entre esses termos, por meio de hierarquias. O principal objetivo dessas investigações é facilitar a representação temática do conteúdo de um documento e indexá-lo.

Dessa forma, a taxonomia, sendo um instrumento de controle terminológico, torna-se uma ferramenta importante na representação semântica de um documento. Como ferramenta especializada, é construída por meio de um processo que tem o objetivo de arranjar hierarquicamente uma lista de conceitos que representam a temática de determinado domínio ou área. As taxonomias devem atender a diversos tipos de objetivos e podem ser apresentadas na forma de representações gráficas, facilitando a compreensão e exploração do conteúdo (FIGUEIREDO, 2006).

De acordo com Campos e Gomes (2008), é possível definir uma taxonomia:

> Taxonomia é, por definição, classificação sistemática e está sendo conceituada no âmbito da Ciência da Informação como ferramenta de organização intelectual. É empregada em portais institucionais e bibliotecas digitais como um novo mecanismo de consulta, ao lado de ferramentas de busca. Além destas aplicações, a taxonomia é um dos componentes em Ontologias. A organização das informações através do conceito de Taxonomia permite alocar, recuperar e comunicar informações dentro de um sistema de maneira lógica através de navegação.

Uma taxonomia, de maneira geral, corresponde à área do conhecimento que se ocupa das regras e dos princípios de sua terminologia, podendo ser vista como:

> [...] um sistema de classificação tendo por base, normalmente, uma hierarquia de termos e conceitos, na qual os termos localizados nos níveis mais baixos representam os aspectos mais específicos do conteúdo. Até recentemente, o seu interesse era restrito a profissionais da área de ciência da informação, biblioteconomia ou especialistas em determinadas ciências, como a biologia, mas agora é parte do interesse dos profissionais da gestão do conhecimento. A correta definição e classificação das bases de conhecimento de uma empresa, ou seja, uma estrutura adequada de termos e conceitos tornou-se fundamental para a gestão da intranet, portais, etc. Taxonomia, [200-, s/página].

Quando uma taxonomia assume interface gráfica, as informações que se encontram dispersas no texto são organizadas respeitando-se os temas, os assuntos e a hierarquia estipulados pela ferramenta. Dessa forma, será extraído o que há de mais relevante naquele contexto.

De acordo com Pickler (2007, p. 73), uma "taxonomia define classes, subclasses e as relações entre elas, e o conjunto de regras de inferência fornece o mecanismo de manipulação dos objetos das classes, utilizando raciocínio lógico". Assim, com base nas características mencionadas a respeito de uma taxonomia, percebe-se que ela corresponde a um instrumento importante no processo da indexação automática, atuando na contextualização da representação do conteúdo do documento indexado. O conteúdo de um documento pode ser representado por uma taxonomia a partir de termos e definições utilizados numa área específica do conhecimento humano.

A indexação automática por atribuição remonta a uma longa história. Tentativas iniciais não obtiveram muito êxito, porém, nos últimos 40 anos, obtiveram-se resultados melhores nessa área (LANCASTER, 2004).

É possível a comparação das indexações por extração e por atribuição automática às indexações derivativa e atributiva, respectivamente. De acordo com Mamfrim (1991), "a indexação é *derivativa* quando os termos de indexação são derivados

do próprio texto do documento analisado e é *atributiva* quando os termos de indexação são alocados independentemente dos termos do texto do documento".

Outro tipo de indexação automática apontada na literatura é a *identificação automática de palavras full text.* Este método é considerado mais simples. Através dele, analisa-se o documento na íntegra e não se considera a semântica do texto nem a posição sintática das palavras nas orações.

Existe também a *indexação automática sintática,* que objetiva a análise das palavras mais relevantes da oração. De acordo com Wives (1997, p. 9-10),

> [...] A linguagem do documento permite que este tipo de análise seja feita, já que as orações (a exemplo do português) possuem posições sintáticas pré-definidas para os termos (sujeito, predicado, local do verbo...), e alguns destes termos são mais importantes do que os outros (seus auxiliares). Somente os termos *importantes* são adicionados à estrutura de índice. Esta técnica exige uma *base de conhecimento* que contenha todas as combinações sintáticas possíveis, além de exigir mais poder computacional e tempo. Portanto, geralmente não é utilizada.

Há, ainda, a *indexação automática semântica,* que baseia-se no princípio de que o documento já possui estruturas de formatação para a indicação da semântica dos termos. De acordo com Wives (1997, p. 9-10),

> [...] Em HTML existem marcações (*tags*) que indicam onde encontram-se os títulos, as palavras-chave e algumas outras estruturas importantes ao documento. O processo de indexação deve identificar estas marcações e indexar os termos presentes entre estas marcações com maior importância. Podem surgir alguns problemas, como o da *indexação incerta*, onde a pessoa encarregada de demarcar o documento não utiliza palavras que identificam corretamente o documento.

Assim, tendo apresentado as características principais da indexação automática, vale destacar alguns aspectos que devem ser observados nesse processo. É consenso entre os pesquisadores da área que, para a obtenção de um tratamento automático adequado de documentos textuais, é necessário o desenvolvimento de algoritmos que levem em consideração a semântica e a sintaxe do conteúdo desses documentos. Assim, a seção seguinte é dedicada à discussão dos aspectos que devem ser considerados para a realização da indexação de documentos textuais baseada em seu contexto.

2.2.1 **A semântica e a sintaxe na indexação automática**

Dentro da área da lingüística, destaca-se a semântica como meio para a representação do significado dos enunciados e a sintaxe como determinante se uma expressão ou frase está adequada à gramática de uma língua. As relações semânticas são importantes na estruturação do conhecimento e na formação de conceitos para escolha de termos representativos de significado. A semântica é o meio utilizado para a representação do significado dos enunciados.

> [Do gr. semantiké, i. e., téchne semantiké, 'a arte da significação'.] S. f. 1. E. Ling. Estudo das mudanças ou translações sofridas, no tempo e no espaço, pela significação das palavras; semasiologia, sematologia, semiótica. 2. E. Ling. O estudo da relação de significação nos signos [v. signo (4 e 5)] e da representação do sentido dos enunciados. 3. E. Ling. P. us. V. semasiologia (1). (s.u. Aurélio Eletrônico). (SIMÕES, 2002).

Guiraud (1975, *apud* SIMÕES, 2002), levanta questões de três níveis relevantes na análise semântica:

> Psicológica – Por que e como nos comunicamos? Que é um signo, e que se passa em nosso espírito e no de nosso interlocutor quando nos comunicamos? Qual é o substrato e o mecanismo fisiológico e psíquico dessa operação? Etc.
>
> Lógico – Quais as relações do signo com a realidade? Em que condições será um signo aplicável a um objeto ou a uma situação que ele tem a função de representar? Quais são as regras que asseguram uma verdadeira significação?
>
> Lingüístico – (...) cada sistema de signos tem as suas regras específicas referentes à sua natureza e função.
>
> A semântica lingüística (...) estuda as palavras no seio da língua: que é uma palavra, quais são as relações entre a forma e o sentido de uma palavra, e as relações entre as palavras, como asseguram elas a sua função? Etc.

O foco deste estudo é a questão de nível lingüístico da análise semântica com objetivo de tratamento automático de textos.

Ainda seguindo o raciocínio de Simões (2002), segue quadro ilustrativo da função significativa das palavras (QUADRO 1).

QUADRO 1

A função significativa das palavras

	FORMA	FUNÇÃO		
SONS	**Fonologia**	Semiótica	REPRESENTATIVA	**Semântica**
PALAVRAS	**Morfologia**		COMUNICATIVA	**Pragmática**
CONSTRUÇÕES	**Sintaxe**		EXPRESSIVA	**Estilística**

forma	*CHAVE*	função		
Fonologia	/ʃaʊ/	SEMIÓTICA	**Semântica**	a) *chave* b) *chave* de roda c) *chave*-de-cadeia
Morfologia	chav-e		**Pragmática**	A porta precisa de uma "A" O mecânico precisa de uma "A" ou "B" Aquela pessoa é "C"
Sintaxe	emprego em contexto frasal		**Estilística**	*chave* de interpretação = pista interpretação-*chave* = idéia-base

Fonte: SIMÕES, 2002, p.2.

A ilustração do autor demonstra bem o panorama complexo do processo de significação das palavras. Se este contexto é denso em se tratando da linguagem escrita ou falada, para o tratamento automático as dificuldades são verdadeiramente mais significativas.

Simões (2002, p.1), afirma:

> As palavras estão sujeitas à análise em mais de uma dimensão. Geralmente, procede-se à investigação de sua forma e de suas funções. Quanto à forma, as disciplinas que se impõem são: *Fonologia, Morfologia e Sintaxe*; quanto às funções, outra ordem de estudos vem à tona, como: a *Semiótica, a Semântica, a Pragmática, a Estilística*, etc.

Sabe-se que toda língua tem seu próprio recorte e sua própria semântica, que, como dito anteriormente, estuda os significados das coisas. Essa língua pode ser repleta de regionalismos, metáforas, gírias, linguagem figurada, denotação e conotação. Tudo aquilo que está presente na vida das pessoas possui um nome, que é parte do léxico. A estrutura lexical compreende o conjunto de vocábulos de uma língua e abrange o conhecimento lingüístico partilhado pela sociedade na qual é falada, possuindo valor diferente de língua para língua. A análise sintática consegue determinar se uma

expressão ou frase está adequada à gramática dessa língua específica. Tem-se ainda, a unidade lexicalizada, que pode conter várias palavras com significado convencionado, como, por exemplo, *bater as botas* (morrer) ou *dar com os burros n'água* (fracassar). Nesse contexto, quando se entende que a expressão *bater as botas* é morrer, o que se faz é dar um novo valor semântico à expressão.

Para Rector e Yunes (1980, p. 14),

> Uma explicação de propriedades semânticas requer mais do que a análise do sentido das palavras apenas, isto é, para que se entenda o sentido de uma sentença e suas relações semânticas com outras expressões, é preciso saber não só o significado de suas unidades léxicas, mas, também, como estas se relacionam – a dependência da estrutura sintática da sentença.

Retomando a expressão *bater as botas*, apresenta-se na FIG. 1 o detalhamento de seus signos individuais para a composição de uma frase com sentido convencionado.

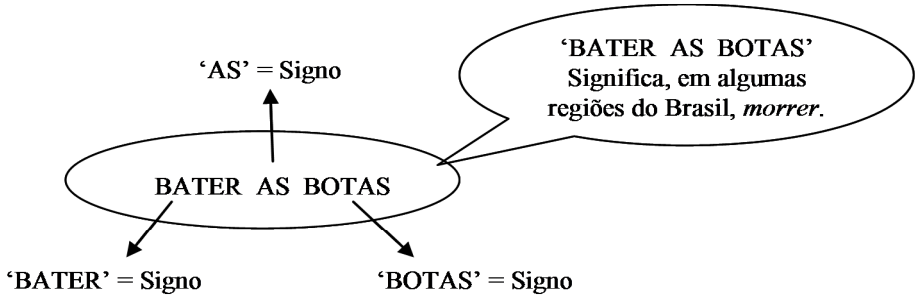

FIGURA 1 – A semântica de uma frase
Fonte: BORGES; MACULAN e LIMA, 2008, p. 185.

O processo de análise de assunto também demanda forte carga cognitiva, efetuada pela mente humana, que é a interpretação, a coordenação de signos e a abstração de significados. Signo é uma palavra que, isoladamente, pode ter sentido para determinado indivíduo, mas não possui significado. O significado de um signo (uma palavra) está estreitamente ligado à estrutura lexical, isto é, ao contexto no qual o signo está inserido. Já o sentido é uma abstração pessoal, ou seja, como cada indivíduo entende o signo. Sobre isso, afirma-se:

> Considerado isoladamente, signo algum tem significação [significado]. Toda significação de signo nasce de um contexto, quer entendamos por isso um contexto de situação ou um contexto explícito [...]. É necessário, assim, abster-se de acreditar que um substantivo está mais carregado de sentido do que uma preposição, ou que uma palavra está mais carregada de significação do que um sufixo de derivação ou uma terminação flexional (HJELMSLEV, 1975, p. 50 *apud* SILVA, [2004]).

A semântica e a sintaxe têm papéis importantes na indexação automática, na medida em que permitem ao *software* identificar a estrutura lexical das frases e o significado dos termos que representam o conteúdo do documento.

A sintaxe determina a forma correta de construção das frases de uma determinada língua, levando em consideração a seqüência de sujeitos, verbos, objetos, predicados, artigos, preposições etc. A semântica, por sua vez, encarrega-se do significado da frase construída. Dessa forma, pode haver frases sintaticamente corretas, mas sem um conteúdo semântico denotativo aceitável e vice-versa. Exemplos:

- *A chuva gosta de cair sobre meus cabelos ruivos* = Frase com sintaxe correta, porém, sem conteúdo semântico denotativo aceitável.
- *Fingimos que fumus e vortemos* = Frase semanticamente compreensível para um determinado grupo de indivíduos em determinada região do país, porém fora dos parâmetros da língua portuguesa formal.

A sintaxe permite apenas escrever frases corretas numa língua. Por exemplo, as frases "*o rato come o queijo"* e "*o queijo come o rato"* são sintaticamente corretas, porém, assume-se que apenas a primeira frase tem significado em nossa língua. Isso se deve ao conhecimento da maioria das pessoas de que ratos são animais que se alimentam de queijo e de que é impossível, dentro da realidade humana, um queijo comer um rato.

Durante a análise sintática, pode-se perceber se os sintagmas foram colocados na seqüência correta. Sintagmas são expressões que ditam uma relação de dependência, na qual um elo de subordinação é estabelecido, e cada um dos elementos é também um sintagma. Esse termo é geralmente empregado para designar cada parte de uma oração e pode ser:

- Sintagma nominal (nome);
- Sintagma adjetival (adjetivo);

- Sintagma verbal (verbo);
- Sintagma preposicional (preposição);
- Sintagma adverbial (advérbio).

Conseguir identificar os sintagmas é muito importante na análise sintática, porque isso facilitará a compreensão do papel exercido pelas palavras na frase. Como ilustração, temos a frase: *O Christiano acreditou na vitória,* analisada na FIG. 2.

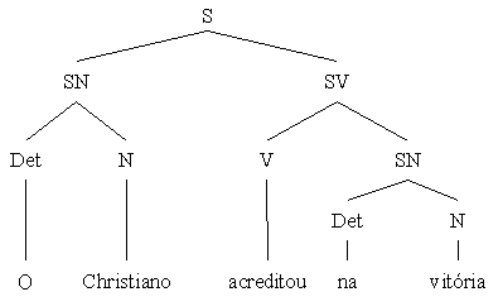

FIGURA 2 – Árvore sintagmática
Legenda: S = sentença (frase)
SN = sintagma nominal
Det = determinante
N = nome ou substantivo
SV = sintagma verbal
V = verbo
Fonte: adaptada de Othero e Menuzzi (2005, p. 49).

Para determinar um sintagma, é necessário identificar o elemento núcleo, uma vez que ele pode possuir mais de uma palavra. Ainda pode existir, numa mesma frase, mais de um sintagma de mesmo tipo, como é o caso do exemplo acima: *Christiano* e *vitória* são núcleos de sintagmas nominais. Nesses casos, é preciso estabelecer quais funções sintáticas esses núcleos desempenham. No caso desse exemplo, seria: *Christiano* tem função de sujeito e *vitória* tem a função de objeto indireto.

Em síntese, para que um *software* de indexação automática apresente resultados satisfatórios, entre vários outros critérios possíveis e importantes, é essencial que ele seja capaz de analisar um texto tanto sob seu aspecto sintático quanto semântico.

Trazendo a discussão para a prática da BCI, acredita-se que uma ferramenta de controle terminológico habitualmente utilizada pelos profissionais dessa área, a taxonomia, seja o instrumento ideal para a análise semântica de documentos textuais.

Percebe-se que a Lingüística Computacional é um importante aliado na construção de um *software* com essas características de análise. Este campo de

conhecimento tenta encontrar soluções tecnológicas para que as máquinas possam imitar o conhecimento lingüístico e semântico de um ser humano ou, pelo menos, auxiliar o indexador no processo de indexação.

2.2.2 **A lingüística computacional no ambiente da indexação automática**

A área de Lingüística Computacional preocupa-se com a compreensão da língua e de técnicas apropriadas à interpretação dessa língua, escrita ou falada, tentando imitar a capacidade humana de comunicação. Para tanto, essa área utiliza elementos de sintaxe, semântica, fonética e fonologia, pragmática e análise do discurso. De acordo com Othero e Menuzzi (2005), a lingüística computacional pode ser dividida em *Lingüística de Corpus e Processamento da Linguagem Natural – PLN*.

Othero e Menuzzi (2005) afirmam que a lingüística de *corpus* trabalha com "corpora eletrônicos", isto é, grandes bancos de dados que contêm amostras de linguagem natural, que podem ser de diferentes fontes. O objetivo é estudar os fenômenos lingüísticos que podem acontecer em grandes amostras de uma língua específica e um dos resultados desse estudo pode ser a produção de um *software*. Esses *corpora* podem ser compostos por linguagem falada ou por amostras de textos de várias fontes escritas. O PLN, por sua vez, preocupa-se com o estudo da linguagem diretamente voltado para a construção de *softwares*, como tradutores automáticos, *chatterbots*, reconhecedores automáticos de voz, geradores automáticos de resumos, *parsers,* entre outros.

Um *parser*,

> [...] no contexto da lingüística computacional, é um analisador automático (ou semi-automático) de sentenças [frases]. Esse tipo de programa é capaz de analisar uma sentença com base em uma gramática preestabelecida de determinada língua, verificando se as sentenças fazem parte ou não da língua, de acordo com o que autoriza a sua gramática. Um *parser* também analisa sintaticamente as sentenças, decompondo-as em uma série de unidades menores, primeiramente em nódulos não-terminais (os sintagmas), até chegar a nódulos terminais (os itens lexicais), atribuindo-lhes uma estrutura de constituintes. Essa estrutura de constituintes, que representa a representação hierárquica e sintática da frase, é apresentada comumente em árvores sintáticas ou através de colchetes rotulados (OTHERO; MENUZZI, 2005, p. 49).

Um *parser* utiliza linguagem do tipo "declarativa". Esse tipo de linguagem fornece proposições ao computador com as quais ele é capaz de analisar as frases de um

texto por meio de combinações lógicas. A maioria das linguagens de programação geralmente é do tipo “procedural”, que, ao contrário da linguagem do tipo “declarativa”, provê o computador de um algoritmo cujos passos são ações executadas pelo próprio computador até que ele chegue a determinado resultado (OTHERO; MENUZZI, 2005, p. 42).

Ainda na abordagem da lingüística computacional, a gramática é entendida como um conjunto relativamente pequeno de regras e vocábulos de uma língua, que possibilita reconhecer todas as frases possíveis dessa determinada língua, atribuindo a elas uma estrutura sintagmática. Essa gramática é denominada *gramática sintagmática*. Já no contexto da Lingüística, o termo *gramática* é entendido sob diferentes acepções, desde as normas que regem uma língua até o sentido de gramática histórica e de gramática comparada (OTHERO; MENUZZI, 2005).

Assim, percebe-se que o tratamento computacional para o processamento semântico das línguas naturais ainda se encontra em estágios iniciais de pesquisa para que possam trazer as soluções almejadas. Uma das soluções apontadas na literatura para reduzir os problemas de processamento da linguagem natural é o uso de linguagens controladas, nas quais existe uma estrutura semântica coesa, com relações terminológicas preestabelecidas, dentro de um determinado domínio.

Salton (1973, p. 258) simula um ambiente convencional de recuperação de documentos e verifica que uma simples duplicação de padrões, por meios automáticos, não irá produzir resultados aceitáveis. Propõe, então, algumas abordagens como: método de *ranking* de documentos, dicionário automático, geração de lista nominativa e *feedback* dos usuários. Para o autor, em condições normais de utilização, a metodologia totalmente automática é superior em efetividade quanto aos procedimentos convencionais.

Para Pack & Pratt (1971, *apud* SALTON 1973), infelizmente, as chances de se dominarem essas estruturas de linguagem natural, em futuro próximo, são excessivamente turvas. Até então, segundo os autores, nenhum sistema era capaz de lidar com mais do que um pequeno subconjunto da língua inglesa, e pouquíssimos sistemas tinham tentado ir além das fronteiras da análise das frases.

Para Salton (1973), a situação era sombria: por um lado, havia uma melhor compreensão da necessidade da sintaxe e da semântica, por outro lado, as probabilidades de se obter um progresso na lingüística computacional eram remotas.

No contexto da indexação automática, para o tratamento adequado de um documento textual, é necessário entender detalhadamente suas características.

2.2.3 **O tratamento de documentos textuais na indexação automática: características**

Até aqui, apresentaram-se os processos de indexação manual e automática, a importância da semântica e da sintaxe nesses processos, e, ainda, o papel da lingüística de *corpus* neste contexto. Assim, finalizando o embasamento teórico e metodológico desta pesquisa, julga-se necessária uma breve apresentação das principais características de um documento textual, que vem sendo abordado ao longo do trabalho.

Para Moens (2000, p. 27), um texto escrito possui três componentes principais: (1) *estrutura de layout*, composta por elementos extra-textuais, pelas fontes, estilo de fonte e cores; (2) *estrutura lógica,* baseada na organização de partes de informação, por instância, capítulos, parágrafos e nodos de informação. As estruturas lógicas e de *layout* referem-se à representação da estrutura e estão amarradas à intermediação e tecnologia do processo de comunicação; e, por último (3) o *conteúdo do texto*. Segundo Ellis (1992, p. 28), um texto é composto por unidades lingüísticas, sendo a lingüística um rigoroso estudo científico da linguagem natural formal.

De acordo com Halliday, (1989, p. 40 *apud* MOENS, 2000, p. 27), é necessária a distinção das funções do texto (QUADRO 2):

QUADRO 2

Funções do texto

Textos para empreender ações	•Signos públicos •Tabelas para instrução de produtos •Receitas •Mapas •Guias de rádio e televisão •Contas •Menus •Diretórios de telefone •Cédulas •Manuais de computadores
Textos para contatos sociais	•Cartas •E-mails •Cartão-postal
Textos escritos para prover informação	•Notícia •Artigos de revista •Artigos científicos •Reportagens

	•Panfletos •Livros •Anúncios •Brochuras para viagem
Textos para entretenimento	•Artigos de revistas •Textos poéticos e dramas •Novelas •Ensaios •Subtítulos de filmes

Fonte: elaborado pela autora com dados apontados por Halliday (1989, p. 40) retirados do texto de Moens (2000, p. 27).

Há vários tipos de texto: *texto expositivo*, *texto narrativo* e *textos amarrados a disciplinas específicas*. O texto narrativo foca uma história e consiste em várias ações. Como exemplo, podem ser citados novos artigos, novelas e pequenas histórias. No texto expositivo, o foco é nos tópicos e sub-tópicos do texto. Aqui, a organização do texto é importante para encontrar, de maneira eficiente, a informação a respeito dos tópicos no texto. Textos científicos são em grande parte textos expositivos. Exemplo desse tipo de texto são as enciclopédias, artigos científicos e documentos técnicos (MOENS, 2000, p. 29).

De acordo com Moens, (2000, p. 29), as unidades básicas de um texto são as palavras. Guiraud (1975, *apud* SIMÕES, 2002), parte do ponto da significação para perguntar:

> O que é uma palavra? Qual é a sua função? Como é essa função assegurada? Significação é, portanto, um termo tomado [...] em seu sentido ativo de substantivo verbal: **significação**; é um processo psicológico, enquanto **sentido** tem um valor estático, é a imagem mental que resulta do processo. Devemos evitar confundir os dois termos, tal como se faz na língua corrente, que fala indiferentemente da significação ou do sentido de uma palavra.

Tendo em vista o apresentado até aqui, acredita-se que todos os elementos teóricos e metodológicos essenciais para o desenvolvimento desta pesquisa foram explicitados. Assim, o capitulo seguinte é dedicado à apresentação do desenvolvimento prático deste estudo que, baseado na metodologia proposta inicialmente, pretende analisar os critérios de indexação automática encontrados na literatura utilizados para o tratamento de documentos textuais.

3 ANÁLISE DE CRITÉRIOS DE INDEXAÇÃO AUTOMÁTICA UTILIZADOS NO TRATAMENTO DE DOCUMENTOS TEXTUAIS

Neste capítulo, apresenta-se o processo de identificação, descrição, análise e interpretação dos dados, bem como as conclusões que pretendem solucionar o problema de pesquisa.

3.1 Metodologia

Tendo em vista os objetivos deste trabalho, fez-se necessária a escolha de determinados métodos de estudo.

Esta pesquisa apresenta, então, características de uma *exploração técnica e sistemática* com uma abordagem qualitativa para a análise dos resultados. Para tal, segue-se aqui o modelo de trabalho apresentado por Best (1972, *apud* MARCONI e LAKATOS, 1996), em que "o investigador, baseando-se em conhecimentos teóricos anteriores, planeja cuidadosamente o método a ser utilizado, formula problemas e hipóteses, registra sistematicamente os dados e os analisa com maior exatidão possível".

Para o início do trabalho, foi necessária uma *pesquisa bibliográfica.* Segundo Marconi e Lakatos (1996, p. 23-24), "a pesquisa bibliográfica é um apanhado geral sobre os principais trabalhos já realizados, revestidos de importância, por serem capazes de fornecer dados atuais e relevantes relacionados com o tema". Essa etapa da pesquisa teve como objetivo conhecer as informações sobre o que já foi discutido pelos teóricos na literatura sobre o assunto.

Neste trabalho encontram-se também características de uma *pesquisa descritiva*, definida por Best (1972, *apud* MARCONI e LAKATOS, 1996), como aquela que "delineia o que é", abordando quatro aspectos: descrição, registro, análise e interpretação de fenômenos atuais, buscando atuar diretamente no aperfeiçoamento desses fenômenos.

Por fim, e de acordo com a sua finalidade, esta é uma *pesquisa do tipo prática.* Ander-Egg (1978, *apud* MARCONI e LAKATOS, 1996) afirma que esse tipo de pesquisa, como "o próprio nome indica, caracteriza-se por seu interesse prático, isto é,

que os resultados sejam aplicados ou utilizados, imediatamente, na solução de problemas que ocorrem na realidade".

Para a análise dos dados, utilizou-se o método *analítico-sintético*, baseado em Dahlberg (1978), constituído neste trabalho, de duas principais etapas. Na etapa (1), identificaram-se os critérios de indexação automática encontrados na literatura. Essa etapa foi composta de dois estágios: (a) seleção de textos sobre indexação automática; (b) leitura dos textos e identificação dos critérios de indexação automática neles encontrados. Na etapa (2), foi proposto um conjunto de critérios ideal para o processo de indexação automática. Os estágios dessa etapa foram: (a) seleção de uma amostra dos textos utilizados na etapa 1; (b) análise da combinação dos critérios utilizados em cada texto e interpretação dos respectivos resultados.

Assim, levando-se em consideração o objetivo geral desta pesquisa, que é selecionar critérios de indexação automática e combiná-los em uma proposta que visa a automatizar a primeira etapa do processo de indexação manual, utilizou-se um método de estudo dividido em duas etapas principais: (1) *Identificação dos critérios*, que se subdivide em dois estágios: (a) Definição do universo e da amostra de estudo e (b) Definição do objeto empírico através da sistematização dos critérios; (2) *Análise das combinações dos critérios*, também subdividida em dois estágios: (a) Seleção de uma segunda amostra do universo de estudo e (b) Interpretação dos critérios.

3.2 **Identificação dos critérios – Etapa 1**

Nesta etapa, foram selecionados artigos técnico-científicos, que caracterizaram o *universo* e a *amostra de estudo*, e definido o *objeto empírico* desta pesquisa.

3.2.1 **Definição do universo e seleção da amostra de estudo nº 1**

O universo de estudo desta pesquisa é caracterizado por artigos técnico-científicos, dissertações, teses e livros sobre indexação automática que apresentam resultados de pesquisas da área. Os documentos deveriam conter, necessariamente,

metodologia de pesquisa e apontamento de resultados conclusivos quanto à pertinência dos critérios de indexação automática utilizados.

Com o objetivo de identificar os principais critérios de indexação automática apresentados na literatura, definiu-se uma amostra desse universo (amostra nº 1) a partir do exame das principais pesquisas nacionais e internacionais sobre o assunto publicadas entre a década de 1950 e o ano de 2008.

Delimitou-se o período indicado visto remontarem da década de 1950 os primeiros estudos sobre indexação automática. Em razão da Segunda Guerra Mundial, o mundo passou a vivenciar o início de um novo contexto informacional, onde registrou-se um crescimento exponencial da produção de documentos impressos. Chamado de "explosão bibliográfica", esse contexto impulsionou um esforço generalizado em prol do desenvolvimento de tecnologias para o controle dos documentos produzidos, e, posteriormente, com a Web, para a necessidade de tratamento adequado desse material em formato eletrônico. Desde então, o aumento no número de publicações de documentos eletrônicos é crescente, sem perspectiva de que haja um processo inverso do que se observa atualmente. Dessa maneira, esse assunto tornou-se foco de observação de vários pesquisadores, demandando também atenção quanto à sua organização, armazenamento e recuperação, e, como base de todos esses processos, a indexação de documentos textuais.

É digno de observação o fato de que, no processo inicial de seleção do material a ser lido e analisado, não se buscou, necessariamente, documentos publicados a partir da década de 1950. O critério usado na seleção dos textos foi unicamente que estes tivessem a indexação automática como assunto principal, visando ao maior número de documentos possível, para se compor um objeto empírico consistente. Em momento posterior, em que ocorreu a organização e a ordenação cronológica dos textos, observou-se que não foi recuperado um texto sequer que datasse de momento anterior à década de 1950, o que comprova a afirmação de que remontam dessa década os primeiros estudos sobre o assunto.

Para a realização da pesquisa bibliográfica definiu-se a seguinte estratégia de para seleção dos documentos.

1. **Delimitação do objetivo principal da pesquisa e de sua finalidade:** selecionar documentos que abordem como assunto principal, ou de maneira consistente, a indexação automática de documentos textuais apresentando, de preferência, abordagens metodológicas.

2. **Indicação das palavras-chave que utilizadas para delimitação do assunto nos idiomas inglês e português:**

Em português:

Algoritmos de radicalização
Aquisição de conhecimento
Categorização automática de documentos
Categorização de textos
Classificação automática de textos
Descoberta de conhecimento
Extração semi-automática de conhecimento a partir de textos
Identificação de conglomerados (*clustering*) de documentos
Indexação
Indexação assistida por computador
Indexação automática
Indexação automática de textos
indexação automática derivativa
Indexação automatizada
Indexação e resumos
indexação manual
Indexação semi-automática
Linguagem natural
Lógica difusa
Mineração de textos
Pré-processamento automático de texto
Programa para indexação automática
Representação de conteúdo
Sintagmas nominais
Sistemas de informação

Em inglês:

Abstracting of document texts
Analysis of descriptor suffixes

Automatic abstracting
Automatic analysis
Automatic book indexing
Automatic extracting
Automatic indexing
Automatic information organization
Automatic information retrieval
Automatic text analysis
Automatic text indexing
Automating survey coding
Classification of text automatic
Clustering to information system
Complex identifiers
Computational linguistics
Document retrieval
Generation of machine-readable texts
Hybrid approach to faceted classification
Indexing
Indexing systems
Indexing with a computer
Information extraction
Multiclass text categorization
Natural language
Natural-language processing
Probabilistic indexing
Random-walk models of term semantics
Subject analysis
summarization of machine-readable texts
Syntactic approaches
Term associations
Text categorization
Technology on indexing

3. **Determinação dos tipos de documentos que fariam parte da amostra:** (1) artigos técnico-científicos; (2) teses e dissertações; (3) trabalhos apresentados em eventos da área; (4) relatórios de pesquisas; (5) conteúdo de sites relevantes; (6) guias de pesquisa; (7) e-book's; (8) livros impressos; (9) capítulos de livros.

4. **Seleção das fontes de informação para pesquisa bibliográfica:**

Bases de dados

EBSCO – HOST
Illumina
SCIENCE CITATION INDEX
Scirus
SCIELO
Library and Information Science Abstracts - LISA
Science
Scopus
Biblioteca Digital Brasileira de Teses e Dissertações do IBICT
Biblioteca Digital de Teses e Dissertações da UFMG
Biblioteca Digital de Teses e Dissertações da UNB
Springer

Periódicos

Association for Computing Machinery - ACM
BIBLOS: Revista do Departamento de Biblioteconomia e História
Ciência da Informação
DataGramaZero: Revista de Ciência da Informação
Em Questão: Revista da Faculdade de Biblioteconomia e Comunicação / Universidade Federal do Rio Grande do Sul (UFRGS). Faculdade de Biblioteconomia e Comunicação
Encontros Bibli: Revista Eletrônica de Biblioteconomia e Ciência da Informação
Information Processing and Management
Information Systems Research
Journal of the American Society for Information Science and Technology

Perspectivas em Ciência da Informação
Revista ACB: Biblioteconomia em Santa Catarina
Revista Brasileira de Biblioteconomia e Documentação
Revista Digital de Biblioteconomia e Ciência da Informação
Science
The Indexer

Anuários

Annual Review of Information Science and Technology - ARIST

5. **Delimitação da quantidade da amostra:** levantamento exaustivo, acima de 50 documentos.

6. **Indicação do formato e do suporte dos documentos selecionados:** documentos eletrônicos e impressos, em formato pdf, htm ou doc.

7. **Determinação a estratégia de busca:** busca booleana (and, not, or).

8. **Definição das partes dos documentos a serem consideradas para leitura técnica:** título, resumo, palavras-chave, sumário, listas de ilustrações e siglas, prefácio, metodologia, considerações finais, glossário, índice, lista de referências bibliográficas.

9. **Análise da amostra selecionada:** realizar uma leitura técnica dos documentos e seleção daqueles potencialmente mais relevantes para a pesquisa.

Assim, após realização da pesquisa bibliográfica de acordo com a estratégia indicada, definiu-se a amostra de estudo nº 1 com a composição do conjunto de 103 (cento e três) documentos, referenciados no Anexo A. A partir da análise desta amostra, foi possível a realização dos procedimentos descritos no estágio a seguir.

3.2.2 **Definição do objeto empírico e sistematização dos critérios**

Primeiramente, os textos que se encontravam em versão eletrônica foram impressos, e os que estavam contidos em periódicos da área foram fotocopiados, permitindo a manipulação dos documentos de maneira única e facilitando o acesso a eles.

Posteriormente, os documentos foram ordenados cronologicamente, tendo sido a leitura iniciada pelo texto mais recente.

Em seguida, procedeu-se utilizando como instrumento de pesquisa um *guia de observação*, que é a definição de aspectos norteadores para uma determinada atividade. Embora esse instrumento apresente desvantagens, como a possibilidade de interpretação subjetiva do pesquisador e a dificuldade na generalização dos dados, ele foi a melhor opção encontrada na literatura. Isso se deu devido à natureza dos dados manipulados neste estudo, que demandam uma visualização clara de suas características para posterior análise. Neste estudo, o guia de observação nº 1 levou ao seguinte procedimento:

1) Salientar nos textos da amostra nº 1 os seguintes aspectos indicados no Quadro 3:

QUADRO 3

Definição do guia de observação nº 1

ASPECTO INDICADO NO QUADRO	DADOS PARA COMPOSIÇÃO
Critério:	Indicar o nome do critério de acordo com terminologia definida pelo(s) autor(es).
Propósito:	Indicar o objetivo principal de utilização e/ou desenvolvimento do critério.
Descrição:	Caracterizar o procedimento de utilização do critério.
Detalhamento/Exemplos:	Especificar características do critério e indicar exemplos de utilização.
Desvantagens:	Indicar desvantagem(s) observada(s) na utilização do critério de acordo com apontamento do(s) autor(es).
Vantagens:	Indicar vantagem(ns) observada(s) na utilização do critério de acordo com apontamento do(s) autor(es).
Citações indicadas:	Indicar os documentos que foram utilizados de forma direta para a elaboração da sistematização do critério.

Fonte: elaborado pela autora.

2) Elaborar, para cada critério, um quadro em que foram expostos os aspectos indicados no guia de observação nº 1 (QUADROS 4 a 20). Para a elaboração dos dezesseis quadros, utilizaram-se citações dos 103 documentos referenciados no Anexo A.

A partir desse procedimento, foram obtidos dois resultados, apresentados a seguir:

Resultado 1:

Lista dos dezesseis critérios identificados a partir da amostra de estudo nº 1, definindo-se, assim, o *objeto empírico* desta pesquisa:

- Formatação de frases-termo (*Word phrase formation*)
- Fórmula de transição de Goffman
- Freqüência absoluta de ocorrência da palavra no texto
- Freqüência de co-ocorrência relativa de termos
- Freqüência de co-ocorrência simples de termos
- Freqüência relativa de ocorrência da palavra no texto
- Identificação de palavras (Comparação com uso de dicionário)
- Identificação de radicais de palavras *(Word stemming)*
- Lista de palavras proibidas / Palavras proibidas *(Stop-list / stop-words)*
- Palavras destacadas no texto
- Peso numérico
- Posição do termo no texto (*Term weighting)*
- Primeira lei de Zipf
- Segunda lei de Zipf ou Lei de Zipf-Booth
- Tópico frasal
- Vocabulário semântico / Vocabulário de cabeçalhos conceituais / Tesauro

Resultado 2:

Sistematização das informações obtidas a partir do guia de observação nº 1

QUADRO 4

Formatação de frases-termo (*Word phrase formation*)

Propósito:	Formatar frases-termo, ou seja, unir as palavras adjacentes para formar novos termos, buscando solucionar o problema dos termos abrangentes, pois as idéias estão agrupadas em contextos, e palavras compostas, geralmente, categorizam melhor os assuntos. Assim, os termos passam a ser mais específicos.
Descrição:	"A utilização de palavras mais específicas consegue fazer com que o sistema recupere documentos de forma mais precisa, justamente pelo fato destas palavras aparecerem em um número menor de documentos (geralmente os documentos de contextos específicos, utilizam termos específicos)".
Detalhamento/	Imaginemos uma pessoa buscando informações sobre *programas de computador*.

Exemplos:	Essa pessoa poderia formular uma consulta utilizando a palavra *programa*, o que poderia ocasionar a recuperação de muitos documentos que contêm a palavra *programa*, mas que não pertencem ao contexto da computação. Uma solução para este problema seria utilizar o termo composto "programa de computador", ou simplesmente "programa computador" (pela eliminação da preposição). Em geral, não é necessário armazenar as palavras de forma composta, pois este processo de unificação das palavras exige tempo. Salton (1983) e Croft (1982), recomendam que ela não seja utilizada, pois não aumenta de forma considerável a eficiência do sistema. O que pode ser feito é o armazenamento da informação sobre as *distâncias entre as palavras* de um mesmo documento e deixar que a técnica de consulta avalie se as palavras são ou não adjacentes.
Desvantagens:	"Deve-se tomar o cuidado para não confundir o conceito de *frase-termo* com a utilização das duas palavras de forma independente. Ou seja, caso o usuário não tenha de alguma forma especificado que as duas palavras devem aparecer juntas, ou o sistema não possua alguma técnica que unifique as duas palavras, a consulta pode se tornar ainda mais abrangente. Isso significa que seriam retornados tanto documentos que tratam do assunto *computador* quanto documentos que tratam do assunto *programa*".
Vantagens:	"Esta *frase*, contextualiza melhor a palavra *programa*, tornando-a menos abrangente e mais específica. Agora os documentos retornados por esta *frase-termo*, fariam parte somente do contexto *programa de computador"*.
Citações indicadas:	CROFT e RUGGLES (1982); SALTON (1983); WIVES (1997, p. 8).

Fonte: desenvolvido pela autora com dados retirados dos documentos referenciados no Anexo A.

QUADRO 5

Formula de transição de Goffman

Propósito:	Identificar as palavras representativas do conteúdo do documento.
Descrição:	Baseado na primeira e na segunda lei de Zipf, Goffman observou que essas leis operavam apenas sobre os extremos da distribuição das palavras no texto. Assim, o pesquisador sugeriu um ponto do texto onde haveria a transição das palavras de alta freqüência para as palavras de baixa freqüência, ponto este onde o número de palavras tende para a unidade.
Detalhamento/ Exemplos:	A formula de Goffmam é uma lei bibliométrica, ou seja, que trabalha com a freqüência de palavras como instrumento de indexação em SRIs. Pretende encontrar um Ponto T e é representada matematicamente como: $$T = \frac{-1 + \sqrt{1 + 8\,I1}}{2}$$ Onde: •**I1** é o número de palavras que ocorrem uma única vez; •**8** é uma constante atribuída à língua inglesa; •**2** é uma constante matemática da fórmula de Baskara, para resolução de equação de 2º grau. "Operacionalmente Goffman propôs que, uma vez identificado o Ponto T, seria definida uma região dentro da qual estariam as palavras indicativas do conteúdo do documento. Esta região seria definida a partir de um ponto correspondente a uma freqüência aproximada. Assim, a partir desta freqüência são contidas as palavras entre o ponto T e a palavra de maior freqüência. Este mesmo número de palavras é projetado para baixo do Ponto T, definindo uma região".

Desvantagens:	O critério não considera qualquer tratamento semântico do texto, baseando-se exclusivamente em uma análise estatística de freqüência de termos no documento.
Vantagens:	Embora baseado exclusivamente em uma análise estatística, o critério pode ser considerado uma evolução da primeira e da segunda lei de Zipf, expandindo a análise puramente baseada na freqüência das palavras dispersas por todo o texto para uma análise onde se identifica uma região potencial para verificação de termos representativos do documento.
Citações indicadas:	LANCASTER (1993, p. 287-288).

Fonte: desenvolvido pela autora com dados retirados dos documentos referenciados no Anexo A.

QUADRO 6

Freqüência absoluta de ocorrência de termos

Propósito:	Ordenar as palavras de acordo com sua freqüência, considerando apenas o próprio documento indexado.
Descrição:	Palavras no topo da lista são candidatas mais fortes para a representação temática do documento.
Detalhamento/ Exemplos:	Para a análise, são considerados: o número absoluto de palavras, a extensão do texto e a freqüência acima de determinado limiar.
Desvantagens:	Dificuldade para se definir o ponto de corte da lista. Mesmo depois de usar listas de palavras vazais de significado (*stop-list)*, algumas palavras podem ocorrer freqüentemente em um texto, podendo não ser bons descritores, ou seja, que sirvam para diferenciar o documento em análise de outros da base de dados, porque essas palavras também ocorrem com muita freqüência na base de dados como um todo.
Vantagens:	Para o tratamento de documentos textuais, a análise absoluta da freqüência de termos é importante num primeiro momento, considerando-se que nesses documentos o tratamento do assunto principal é, na maioria das vezes, sustentado pela utilização sistemática de um determinado conjunto de termos que melhor o descrevem. Por exemplo, nesta pesquisa, em especial, o termo composto *indexação automática* aparece consideradas vezes ao longo do texto, sendo o principal descritor deste documento. (Análise da autora).
Citações indicadas:	LANCASTER (1993).

Fonte: desenvolvido pela autora com dados retirados dos documentos referenciados no Anexo A.

QUADRO 7

Freqüência de co-ocorrência relativa de termos

Propósito:	Identificar termos relacionados nos documentos indexados levando em consideração o total de vezes que tais termos ocorrem na base como um todo, a fim de recuperar textos que tratem de assuntos semelhantes.
Descrição:	“Na verdade, não se calcula o grau de associação entre dois termos com base na freqüência simples, mencionada no critério anterior, mas na freqüência de co-ocorrência relativa à freqüência de ocorrência de cada termo”.
Detalhamento/ Exemplos:	“Por exemplo, se os termos A e B coocorram 20 vezes na base de dados, enquanto A ocorra 10.000 vezes, e B ocorra 50.000 vezes, o *fator de associação* entre A e B

	será fraco". "Por outro lado, supondo que A ocorre 50 vezes, e B ocorra 25 vezes, e ambos coocorram 20 vezes, o fator de associação será grande, pois é muito improvável que B ocorra sem A e quase a metade das ocorrências de A coincida com as ocorrências de B. Portanto, a relacionalidade (R) de dois termos é comumente definida pela simples equação: $$R = \frac{a \; e \; b}{a \; ou \; b}$$ Quando R excede algum limiar preestabelecido, os dois termos são aceitos como se fossem relacionados".
Desvantagens:	Dificuldade para se definir o ponto de corte da lista. (Análise da autora).
Vantagens:	Para análise dos termos representativos, o critério considera não apenas o documento, mas a base de dados como um todo. (Análise da autora).
Citações indicadas:	LANCASTER (1993, p.294).

Fonte: desenvolvido pela autora com dados retirados dos documentos referenciados no Anexo A.

QUADRO 8

Freqüência de co-ocorrência simples de termos

Propósito:	Identificar termos relacionados nos documentos indexados a fim de recuperar textos que tratem de assuntos semelhantes.
Descrição:	"Quanto mais freqüentemente dois termos ocorrem juntos, mais provável será que eles tratem de assunto similar".
Detalhamento/ Exemplos:	"Se o termo A nunca ocorre sem B e o termo B nunca ocorre sem A (o que seria uma situação muito rara), os dois termos são completamente interdependentes e seriam completamente intercambiáveis nas buscas".
Desvantagens:	O critério considera apenas o documento para análise da ocorrência dos termos, e não a base de dados no qual o documento está armazenado. (Análise da autora).
Vantagens:	"O critério é capaz de identificar associações diretas (X e Y tendem a ocorrer juntos) e associações indiretas entre termos". "Suponhamos que o termo D quase nunca ocorra sem o termo W numa base de dados, e que o termo T também tenda a não ocorrer sem W, embora D e T jamais coocorram nos documentos. Conclui-se que há uma relação entre D e T; provavelmente são sinônimos".
Citações indicadas:	LANCASTER (1993, p.294).

Fonte: desenvolvido pela autora com dados retirados dos documentos referenciados no Anexo A.

QUADRO 9

Freqüência relativa de ocorrência de termos

Propósito:	Selecionar palavras ou expressões que ocorram num documento com mais freqüência do que sua taxa de ocorrência na base de dados com um todo.
Descrição:	Ordenar as palavras de acordo com sua freqüência de ocorrência no documento indexado e também na base de dados como um todo.
Detalhamento/ Exemplos:	A freqüência com que uma palavra ocorre na base de dados como um todo é ainda mais importante que a freqüência com que uma palavra ocorre num documento. Ou seja, as palavras que são melhores descritores são aquelas que são imprevisíveis e raras numa coleção. Por exemplo: o termo *amianto* em uma base de documentos

	da área de *biblioteconomia,* e o termo *biblioteca* em uma base de dados que armazene documentos sobre cimento-amianto. Em outras palavras, o termo *biblioteca* em uma base de dados sobre *biblioteconomia* não seria um bom descritor, mesmo que aparecesse, por exemplo, 15 vezes em um mesmo documento. Já a palavra *amianto*, nesta mesma base de dados, seria um ótimo descritor, mesmo que aparecesse apenas 4 vezes no documento tratado, pois seria um termo raro na base.
Desvantagens:	Possibilidade de haver documentos em que o assunto principal seja também um assunto que ocorre sistematicamente na base de dados como um todo. Contudo, nos demais documentos o termo ocorre geralmente na introdução, de maneira a contextualizar o assunto em uma área de conhecimento, e, no documento onde o termo é assunto principal, ele ocorre ao longo de todo o texto (introdução, desenvolvimento e conclusão). (Análise da autora). Por exemplo, uma tese ou dissertação que trate da história da BCI indexado em uma base de dados que contenha invariavelmente documentos desta área de assunto, porém, com outros temas centrais específicos. (Análise da autora).
Vantagens:	Uma lista de termos extraídos usando-se o critério de *freqüência relativa* será diferente de uma lista de termos onde se usou 'freqüência absoluta', porém não de forma radical. Provavelmente, desaparecerão os termos que ocorrem com muita freqüência num documento e também na base de dados com um todo.
Citações indicadas:	LANCASTER (1993, p. 287-288).

Fonte: desenvolvido pela autora com dados retirados dos documentos referenciados no Anexo A.

QUADRO 10

Identificação de palavras (Comparação com uso de dicionário)

Propósito:	"Identificar as palavras nos documentos a serem indexados".
Descrição:	"Identificação de palavras a partir da análise de seqüências de caracteres no texto".
Detalhamento/ Exemplos:	"Salton aconselha fazer um *dictionary lookup*, ou seja, comparar as seqüências de caracteres retiradas do texto com um dicionário a fim de validar se estas palavras realmente existem".
Desvantagens:	Possibilidade de o dicionário deixar de contemplar um termo relevante e este não ser analisado e/ou corrigido. (Análise da autora).
Vantagens:	"Este processo de validação torna-se bastante útil, especialmente quando o documento apresenta muitos caracteres inválidos ou palavras com erros gramaticais. As seqüências de caracteres inválidos devem ser eliminadas, e as palavras com erros, corrigidas. Pode-se aplicar ainda um processo de filtragem naqueles arquivos que possuem formatos de texto específicos, a fim de eliminar as seqüências de controle e/ou formatação de texto. O dicionário pode também auxiliar a identificação de termos específicos, quando se deseja utilizar palavras pré-definidas no índice, evitando que palavras desconhecidas sejam identificadas (ou seja, evita a utilização de um vocabulário não controlado). Um simples *analisador léxico* que identifique seqüências de caracteres e monte palavras pode ser utilizado".
Citações indicadas:	WIVES (1997, p. 6-7).

Fonte: desenvolvido pela autora com dados retirados dos documentos referenciados no Anexo A.

QUADRO 11

Identificação de radicais de palavras *(Word stemming)*

Propósito:	Aumentar o nível de recuperação de documentos através do processo de reduzir variações de uma mesma palavra a uma representação única. “Em teoria, essa representação tem a intenção de isolar o semantema das palavras dos seus morfemas, assim como na lingüística. Contudo, não existe obrigatoriedade nesse sentido, uma vez que as representações podem ser simplificações não preocupadas com a perfeição, mas, sim, com oferecer benefícios de recuperação sem onerar o sistema e impactar na rapidez de processamento, seja no momento da indexação, seja no momento da consulta”.
Descrição:	Para Sacconi (1991), "radical, lexema ou semantema é o elemento portador de significado, comum a um grupo de palavras da mesma família. Assim, na família de palavras terra, terrinha, terriola, térreo, terráqueo, terreno, terreiro, terroso, existe um elemento comum: terr-, que é o radical”. “Todas as palavras que possuem o mesmo radical, e, portanto, significados similares (mas categorias diferentes de linguagem: adjetivo, verbo, advérbio...) são reconhecidas pelo mesmo identificador (as palavras são armazenadas de uma só forma – o radical), facilitando a consulta”.
Detalhamento/ Exemplos:	“Uma maneira de identificar os radicais das palavras é remover seus sufixos e prefixos. Outro exemplo é a eliminação dos plurais das palavras”.
Desvantagens:	“A desvantagem deste método é que ele pode acabar utilizando palavras muito abrangentes, não recuperando documentos específicos (de termos específicos)”.
Vantagens:	Semelhante à *stop-list,* é possível a construção de uma lista de radicais proibidos que além de eliminar as palavras derivadas de tais radicas, possa, de maneira contrária, considerar determinadas palavras derivadas desse radical. Por exemplo, o radical *analis-.* Pode-se construir uma lista de radicas proibidos que exclua, a partir deste radical, as palavras analisando, analisado, análises, analisar, analisados, etc. Mas que, ao mesmo tempo, considere a palavra *análise,* quando esta for apresentada imediatamente anterior à palavra *conceitual,* formando o termo composto *análise conceitual.* (Análise da autora).
Citações indicadas:	WIVES (1997, p. 8); 2007] FREDDY; VIERA e VIRGIL (2007); SACCONI (1991).

Fonte: desenvolvido pela autora com dados retirados dos documentos referenciados no Anexo A.

QUADRO 12

Lista de palavras proibidas / Palavras proibidas *(Stop-list / stop-words)*

Propósito:	Impedir que as palavras que aparecem intensamente em todos os documentos da base de dados sejam indexadas, pois esses termos não são representativos dos conteúdos dos documentos.
Descrição:	“Construir uma lista contendo ‘todas’ as palavras que não devem ser indexadas. A esta estrutura foi atribuído o nome de s*top-list*, e as palavras presentes nesta lista são conhecidas como *stop-words”.* O critério pode ser considerado o comum entre os pesquisadores da área. (Análise da autora)
Detalhamento/ Exemplos:	“É possível a eliminação, por exemplo, de preposições, que são termos que servem para fazer o encadeamento de idéias e palavras, portanto, são termos inerentes à linguagem, e não ao conteúdo dos documentos”.
Desvantagens:	O tempo gasto para elaboração de uma *stop-list* consistente é relativamente grande. Existem também o risco de se omitir um termo relevante a esta lista e a possibilidade de se incluir um termo que seria um bom descritor de conteúdo.

Vantagens:	Com o uso de uma stop-list, torna-se possível a eliminação de palavras proibidas, como artigos, preposições, conjunções, etc. Essa eliminação reduz consideravelmente o tempo de processamento do restante do texto.
Citações indicadas:	WIVES (1997, p. 7).

Fonte: desenvolvido pela autora com dados retirados dos documentos referenciados no Anexo A.

QUADRO 13

Palavras destacadas no texto

Propósito:	Considerar o destaque atribuído pelo autor do texto a determinadas palavras ou expressões, que, possivelmente, são fortes candidatas a serem representativas do conteúdo do documento. (Análise da autora).
Descrição:	Com a indicação das palavras ou expressões destacadas em um texto, aumentam-se, consideravelmente, as possibilidades de se encontrar fortes candidatas para a representação do documento. Isso ocorre devido ao fato de que o destaque se deu, naturalmente, com a intenção de enfatizar, por algum motivo, aquele trecho do documento, sendo esta uma parte relevante do texto. (Análise da autora).
Detalhamento/ Exemplos:	São exemplos de destaques utilizados pelos autores: grifos, **negrito**, *itálico*, "aspas", sublinhado, MAIÚSCULAS, tamanho diferenciado da fonte, etc. Por exemplo: se a maior parte das palavras do documento encontra-se em fonte nº 12, e alguns termos apresentam-se em fonte nº 16, há uma significativa possibilidade de esses termos serem o título ou subtítulo do documento, ou seja, de serem representativos do documento. (Análise da autora).
Desvantagens:	Possibilidade de extração de termos que foram destacados com um enfoque negativo, e não positivo. (Análise da autora).
Vantagens:	Apostando que o destaque se deu de maneira positiva, ou seja, para enfatizar pontos fortes do texto, a análise desses termos será pertinente. (Análise da autora).
Citações indicadas:	LANCASTER (1993).

Fonte: desenvolvido pela autora com dados retirados dos documentos referenciados no Anexo A.

QUADRO 14

Peso numérico *(Term weighting)*

Propósito:	"Atribuir *pesos* ou *graus* de relação entre uma palavra e os documentos em que ela aparece". "Identificar a freqüência de determinada palavra em um documento (*term requency*) e o número de documentos em que esta palavra aparece (*inverse document frequency*). Com estas informações é possível atribuir um valor de relação entre esta palavra e o documento, e este valor é dado pela fórmula abaixo": $$Peso_{td} = \frac{Freq_{td}}{DocFreq_{t}}$$ Onde: • *"Peso td* é o grau de relação entre o termo *t* e o documento *d*; • *Freq td* é o número de vezes que o termo *t* aparece no documento *d*;

	• *DocFreq t* representa o número de documentos que o termo *t* aparece".
Descrição:	"Os itens da base de dados podem receber *peso numérico*, que reflita o número de termos que coincidam entre o item e a estratégia de busca e as forças de associação que existem entre esses termos (com base na co-ocorrência), e os itens recuperados podem ser ordenados por peso".
Detalhamento/ Exemplos:	Com o uso do critério, é possível que alguns itens que aparecem no alto da ordenação [*ranking*] não contenham nenhum dos termos com os quais se iniciou a busca. "Existem várias técnicas que buscam identificar o grau de relação entre um termo e um documento. Um estudo realizado por Viles (1995), indica que a maioria dos modelos de recuperação de informações utiliza estas técnicas. Salton (1983), cita algumas delas em seu livro. [...] A técnica baseia-se na teoria de que as palavras que aparecem com maior freqüência em um documento têm uma forte relação com seu conteúdo. A experiência indica também, que esta relação tende a diminuir quando este termo aparece em muitos documentos".
Desvantagens:	"Para cada termo do documento, é necessário calcular a sua relação utilizando-se a fórmula anteriormente mencionada. Este peso é armazenado na lista invertida. Quando a consulta for requisitada pelo usuário, estes valores são utilizados no processo de identificação dos documentos relevantes a esta consulta. Este processo de identificação da similaridade entre os termos e os documentos que utilizam algum método de relacionar as palavras com os documentos é conhecido como espaço de vetores (*vector space*). Agora cada documento possui um vetor com pares de elementos na forma {(palavra1, peso1), (palavra2, peso2), ... , (palavra n, peso n)}. Nestes pares, as *palavras* representam os termos utilizados na consulta, e o *peso*, seus respectivos valores de freqüência no documento. Caso uma palavra não exista em um documento, seu valor de freqüência é zero (0). Ao final, os pesos são somados, e os documentos, listados por ordem decrescente de pesos".
Vantagens:	"Havendo distinção entre os documentos, é possível obter um desempenho melhor, já que os itens relevantes podem ser recuperados isoladamente, sem que os seus *vizinhos* de menor importância sejam recuperados". "Refinamento do processo de indexação para se obter uma distinção entre os documentos, de maneira que os termos possam indicar seu grau de importância nos documentos".
Citações indicadas:	LANCASTER (1993); SALTON (1983); VILES e FRENCH (1995); WIVES (1997, p. 12-13).

Fonte: desenvolvido pela autora com dados retirados dos documentos referenciados no Anexo A.

QUADRO 15

Posição do termo no texto

Propósito:	Analisar, prioritariamente, determinadas partes do texto, para diminuir o tempo gasto com o processamento do documento na íntegra. (Análise da autora).
Descrição:	Substituição do processo que analisa de todo o texto, para a análise apenas das partes consideradas relevantes.
Detalhamento/ Exemplos:	Um termo que aparece no título ou no resumo de um texto tem mais possibilidades de ser um bom descritor do assunto daquele documento do que um termo que aparece nos anexos, por exemplo.
Desvantagens:	Naturalmente, a partir desse critério, termos relevantes podem deixar de ser indexados por estarem em posição diferente daquelas predeterminadas para análise do *software* e, no entanto, serem representativos.
Vantagens:	Enfatiza a análise de partes do documento que possuem maior probabilidade de conterem trechos significativos para a representação do documento.

Citações indicadas:	LANCASTER (1993).

Fonte: desenvolvido pela autora com dados retirados dos documentos referenciados no Anexo A.

QUADRO 16

Primeira lei de Zipf

Propósito:	Identificar a distribuição das palavras no texto.
Descrição:	Baseada em critérios estatísticos e desenvolvida por George Zipf, em 1948, a Primeira Lei de Zipf opera em relação às palavras de alta freqüência.
Detalhamento/Exemplos:	"De acordo com a lei, se as palavras de um texto suficientemente longo forem colocadas em ordem decrescente de freqüência, poder-se-á verificar que a ordem de série das palavras (R) multiplicada por sua freqüência (F) produz uma constante (K)". $R \times F = K$
Desvantagens:	O critério não considera aspectos semânticos para análise do documento.
Vantagens:	O critério, embora apresente limitações e, principalmente, por ser de natureza exclusivamente estatística, é a base para outros critérios que pretendem analisar o texto de maneira contextualizada.
Citações indicadas:	MAMFRIM (1991).

Fonte: desenvolvido pela autora com dados retirados dos documentos referenciados no Anexo A.

QUADRO 17

Segunda lei de Zipf ou Lei de Zipf-Booth

Propósito:	Identificar a distribuição das palavras no texto.
Descrição:	Também baseada em critérios estatísticos, a Segunda Lei de Zipf foi desenvolvida por George Zipf e aperfeiçoada por Booth, sendo conhecida como *Lei de Zipf-Booth.* A segunda lei opera sobre as palavras de baixa freqüência.
Detalhamento/Exemplos:	A lei é enunciada através da seguinte formula: $In = \frac{2}{I1 \quad n \times (n+1)}$ Onde: •**In** é o número de palavras que ocorrem n vezes para n <5 ou n <6; •**I1** é o número de palavras que ocorrem uma única vez; •**2** é uma constante atribuída à língua inglesa.
Desvantagens:	O critério não considera aspectos semânticos para análise do documento.
Vantagens:	O critério, embora apresente limitações e, principalmente, por ser de natureza exclusivamente estatística, é a base para outros critérios que pretendem analisar o texto de maneira contextualizada.
Citações indicadas:	MAMFRIM (1991).

Fonte: desenvolvido pela autora com dados retirados dos documentos referenciados no Anexo A.

QUADRO 18

Tópico frasal (Palavras sugestivas)

Propósito:	Diminuir o volume de palavras a serem processadas pelo sistema.
Descrição:	Substituição do processo que analisa todo o texto, para o processamento apenas do *tópico frasal* e das *palavras sugestivas* deste texto.
Detalhamento/Exemplos:	Os estudos de Baxendale (1958), demonstraram que seria necessário o processamento apenas da primeira e da última frase de cada parágrafo, pois, em 85% das vezes a primeira frase era o tópico frasal e em 7% dos casos a última frase o era. Considera-se como tópico frasal a parte do texto que provê o máximo de informações relativas ao conteúdo do texto.
Desvantagens:	Há a possibilidade de que partes diferentes daquelas definidas como tópico frasal sejam representativas.
Vantagens:	O tempo de análise do texto por meio dos *softwares* de indexação automática sofre considerável redução.
Citações indicadas:	BAXENDALE (1958); LANCASTER (1993).

Fonte: desenvolvido pela autora com dados retirados dos documentos referenciados no Anexo A.

QUADRO 19

Vocabulário semântico / Vocabulário de cabeçalhos conceituais / Tesauro

Propósito:	Cotejar os termos que ocorrem nos títulos dos documentos com um vocabulário semântico formado por termos de uma área específica, os quais são ligados a um vocabulário de cabeçalhos conceituais.
Descrição:	É possível cotejar, por exemplo, termos que aparecem nos títulos de artigos com um vocabulário semântico, os quais devem ser ligados a um vocabulário de cabeçalhos conceituais.
Detalhamento/ Exemplos:	Desse modo, os cabeçalhos conceituais podem ser atribuídos pelo computador com base em palavras/expressões que ocorrem nos títulos dos documentos;
Desvantagens:	*Subatribuição:* o programa poderá deixar de atribuir termos que deveriam ser atribuídos. *Superatribuição:* o programa poderá atribuir termos que não deveriam ser atribuídos, termos supérfluos.
Vantagens:	O programa é capaz de atribuir 61% dos cabeçalhos que um ser humano, possivelmente, atribuiria.
Referências utilizadas	LANCASTER (1993); Vleduts-Stokolov (1987, *apud* Lancaster, 2004, p.291).

Fonte: desenvolvido pela autora com dados retirados dos documentos referenciados no Anexo A.

De posse dos dados apresentados nesta etapa, parte-se para a análise da combinação dos critérios utilizados nos textos, para, então, interpretar os resultados obtidos em cada documento analisado.

3.3 **Análise das combinações dos critérios – Etapa 2**

3.3.1 **Seleção da amostra de estudo nº 2**

O primeiro estágio desta segunda etapa da pesquisa constitui-se da seleção de 12 (doze) textos a partir da amostra de estudo nº 1, obtendo-se, assim, a amostra de estudo nº 2.

De acordo com Marconi e Lakatos (1996); Lakatos (1991); Mattar (1996), a amostragem é o processo pelo qual se obtém informação sobre um todo – população –, examinando-se apenas uma parte do mesmo – amostra. A amostra deve ser representativa da população. Para uma amostra ser representativa, cada item da população deve ter a mesma chance de ser selecionado, ou seja, de ser incluído na amostra. A escolha da amostragem deve ser sempre imparcial, evitando-se preconceitos ou tendências.

Há tipos de amostragem pré-definidos, e, para este estudo, optou-se pela definição de uma *amostragem não-probabilística* – subjetiva. Diferentemente da *amostra probabilística* – estatística –, que se fundamenta na distribuição e na seleção aleatória, a amostra não-probabilística não tem base estatística, sendo definida por critérios pessoais decorrentes da experiência profissional e do conhecimento do setor em exame (MARCONI e LAKATOS, 1996; LAKATOS, 1991; MATTAR, 1996).

Desta forma, para a definição de amostragem não-probabilística, o tamanho da amostra é estabelecido sem nenhuma base de sustentação técnica, sendo usual que corresponda a 10% ou 15% da população alvo (MARCONI e LAKATOS, 1996; LAKATOS, 1991; MATTAR, 1996).

No contexto deste estudo, como a amostra nº 2 é decorrente da amostra nº 1, que totaliza 103 textos, assume-se, então, que esta amostra toma agora o lugar da *população,* sendo a base para a definição da amostra nº 2. Assim, visto o total de documentos da amostra nº 1, para se definir os textos que iriam compor a amostra nº 2, utilizaram-se alguns aspectos de seleção para composição de uma amostra. De acordo com Marconi e Lakatos (1996); Lakatos (1991); Mattar (1996), é possível citar alguns desses aspectos:

- Seleção com base em análise de vulnerabilidade, risco potencial inerente e importância relativa;
- Itens com características inovadoras, pouco usuais ou com grande complexidade;
- Itens em que ocorreram mudanças na natureza, extensão, prazo ou custo original;
- Indícios de irregularidade ou erro e
- Desejo de dispor de uma distribuição razoável em relação a órgãos, unidades responsáveis, modalidades operacionais, setores de atividade econômica, estados da federação etc.

Assim, o total de textos da amostra nº 2 foi definido levando-se em consideração a utilização de 2 (dois) textos publicados em cada década do período indicado, são elas: 1950, 1960, 1970, 1980, 1990 e 2000.

A decisão de utilização de dois textos por década justifica-se pelo fato de que do total de textos que compõem a amostra nº 1, encontrou-se na década 1950 um total de dois textos, o que, conseqüentemente e para fins estatísticos coerentes, forçou a escolha desta mesma quantidade de textos para as demais décadas.

Entende-se que a amostra selecionada nº 2 é consistente em quantidade e em qualidade dos textos para a refletir a realidade do universo de estudo. Sendo, assim, desnecessária a análise de todos os textos da primeira amostra de nº 1 para esta etapa da pesquisa. As referências dos textos selecionados podem ser encontradas no ANEXO B.

3.3.2 **Interpretação dos critérios**

Para a verificação da utilização prática dos critérios identificados na etapa 1 realizou-se uma análise das pesquisas que compõem a amostra de estudo nº 2 que fizeram uso de tais critérios e foram realizadas, anteriormente, por outros pesquisadores da área. Para fins comparativos, apresenta-se uma síntese das pesquisas, feita através de tabulação dos dados obtidos. A tabulação dos dados seguiu o guia de observação nº 2 detalhado no Quadro 20.

QUADRO 20

Definição do guia de observação nº 2

ASPECTO INDICADO NO QUADRO	DADOS PARA COMPOSIÇÃO
Pesquisa	Indicar o título da pesquisa.
Objetivo	Apontar os objetivos indicados pelo(s) autor(es) da pesquisa.
Pesquisador(es)	Indicar o(s) nome(s) do(s) autor(es).
Período	Indicar o intervalo de anos em que a pesquisa foi realizada.
Local	Indicar o país em que a pesquisa foi realizada.
Critério(s) utilizado(s)	Listar os critérios utilizados na pesquisa, utilizando a mesma nomenclatura e numeração indicada no 3.2.1.
***Software(s)* utilizado(s)**	Indicar o(s) nome(s) do(s) *software(s)* utilizado(s) para a realização da pesquisa.
Comparação com indexação manual	Indicar se ocorreu a comparação: [sim] em caso afirmativo e [não] em caso negativo.
Tipo de documento	Indicar a natureza do documento analisado.
Área de cobertura	Indicar a área do conhecimento enfocada no documento.
Resultado	Indicar se o resultado foi satisfatório ou insatisfatório, de acordo com a avaliação do(s) pesquisador(es).
Numeração do texto constante na amostra nº 1	Indicar o número correspondente ao texto de acordo com o indicado na amostra nº 1.

Fonte: desenvolvido pela autora.

Não se objetivou a exaustividade do assunto, pois seria um trabalho além do necessário, tendo em vista o foco do estudo. Abrangeram-se somente os elementos suficientes para apoio no processo de escolha dos melhores critérios para alcance dos objetivos deste trabalho. Assim, de acordo com detalhamento apresentado, segue a sistematização dos dados observados nas pesquisas que servirão de base para a interpretação dos critérios (Quadros 21 a 32).

QUADRO 21

Pesquisa 1: “Machine-made index for technical literature: an experiment”

Objetivo	Utilização de computadores para redução de documento através da elaboração de índice correspondente ao conteúdo do texto tratado.
Pesquisador(es)	BAXENDALE, P. B.

Período	1958.
Local	USA.
Critério(s) utilizado(s)	nº 1: Formatação de frases-termo (*Word phrase formation*) nº 15: Tópico frasal
***Software(s)* utilizado(s)**	Desenvolvido pelo pesquisador.
Comparação com indexação manual	Sim.
Tipo de documento	Documentos técnicos textuais.
Área de cobertura	Não indicado.
Resultado	Satisfatório.
Numeração do texto constante na amostra nº 1	Texto nº 1.

Fonte: desenvolvido pela autora.

QUADRO 22

Pesquisa 2: “Probabilistic indexing: a statistical approach to the library problem”

Objetivo	Contribuir para solução do problema de busca e recuperação automática de documentos.
Pesquisador(es)	MARON, M. E.; KUHNS, J. L.; RAY, L. C.
Período	1959
Local	USA
Critério(s) utilizado(s)	nº 4: Freqüência de co-ocorrência relativa de termos nº 6: Freqüência relativa de ocorrência da palavra no texto nº 11: Peso numérico
***Software(s)* utilizado(s)**	Desenvolvido pelo pesquisador.
Comparação com indexação manual	Sim.
Tipo de documento	Documentos técnicos textuais.
Área de cobertura	Não indicado.
Resultado	Satisfatório.
Numeração do texto constante na amostra nº 1	Texto nº 2

Fonte: desenvolvido pela autora.

QUADRO 23

Pesquisa 3: “Searching natural language text by computer”

Objetivo	Apresentar e discutir uma abordagem fundamental para a investigação da indexação automática e recuperação e para relatar os resultados preliminares de experiências sobre a busca de texto.
Pesquisador(es)	SWANSON, Don R.
Período	1960
Local	EUA
Critério(s) utilizado(s)	nº 1: Formatação de frases-termo (*Word phrase formation*) nº 7: Identificação de palavras (Comparação com uso de dicionário) nº 11: Peso numérico nº 16: Vocabulário semântico / Vocabulário de cabeçalhos conceituais / Tesauro
***Software(s)* utilizado(s)**	Desenvolvido pelo pesquisador.
Comparação com indexação manual	Sim.
Tipo de documento	Documentos textuais armazenados em biblioteca técnica.
Área de cobertura	Não indicado.
Resultado	Satisfatório.
Numeração do texto constante na amostra nº 1	Texto nº 3.

Fonte: desenvolvido pela autora.

QUADRO 24

Pesquisa 4: "New methods in automatic extracting”

Objetivo	Descrever novos métodos de extração automática de documentos para seleção automática de frases com o maior potencial para transmitir ao leitor o conteúdo do documento.
Pesquisador(es)	EDMUNDSON, H. P.
Período	1969
Local	EUA
Critério(s) utilizado(s)	nº 1: Formatação de frases-termo (*Word phrase formation*) nº 3: Frequência absoluta de ocorrência da palavra no texto nº 7: Identificação de palavras (Comparação com uso de dicionário) nº 10: Palavras destacadas no texto nº 11: Peso numérico

	nº 12: Posição do termo no texto (*Term weighting)*
***Software(s)* utilizado(s)**	Desenvolvido pelo pesquisador.
Comparação com indexação manual	Sim.
Tipo de documento	Documentos textuais armazenados em biblioteca técnica.
Área de cobertura	Não indicado.
Resultado	Satisfatório.
Numeração do texto constante na amostra nº 1	Texto nº 6.

Fonte: desenvolvido pela autora.

QUADRO 25

Pesquisa 5: “Automatic text analysis”

Objetivo	Analisar os principais estudos sobre análise automática de textos para oferecer uma indicação do desenvolvimento futuro da área.
Pesquisador(es)	SALTON, G.
Período	1970
Local	EUA
Critério(s) utilizado(s)	nº 1: Formatação de frases-termo (*Word phrase formation*) nº 3: Freqüência absoluta de ocorrência da palavra no texto nº 4: Freqüência de co-ocorrência relativa de termos nº 5: Freqüência de co-ocorrência simples de termos nº 7: Identificação de palavras (Comparação com uso de dicionário) nº 8: Identificação de radicais de palavras (*Word stemming*) nº 12: Posição do termo no texto (*Term weighting*) nº 16: Vocabulário semântico / Vocabulário de cabeçalhos conceituais / Tesauro
***Software(s)* utilizado(s)**	Desenvolvido pelo pesquisador.
Comparação com indexação manual	Sim.
Tipo de documento	Documentos textuais.
Área de cobertura	Não indicado.
Resultado	Satisfatório.
Numeração do texto constante na amostra nº 1	Texto nº 7.

Fonte: desenvolvido pela autora.

QUADRO 26

Pesquisa 6: “Recent studies in automatic text analysis and document retrieval”

Objetivo	Identificar os procedimentos automáticos mais utilizados em substituição aos processos de análise lingüística convencionais.
Pesquisador(es)	SALTON, G.
Período	1973
Local	EUA
Critério(s) utilizado(s)	nº 1: Formatação de frases-termo (*Word phrase formation*) nº 7: Identificação de palavras (Comparação com uso de dicionário) nº 8: Identificação de radicais de palavras *(Word stemming)* nº 9: Lista de palavras proibidas / Palavras proibidas *(Stop-list / stop-words)* nº 11: Peso numérico nº 16: Vocabulário semântico / vocabulário de cabeçalhos conceituais / Tesauro
***Software(s)* utilizado(s)**	Desenvolvido pelo pesquisador.
Comparação com indexação manual	Sim.
Tipo de documento	Documentos textuais.
Área de cobertura	Não indicado.
Resultado	Satisfatório.
Numeração do texto constante na amostra nº 1	Texto nº 9.

Fonte: desenvolvido pela autora.

QUADRO 27

Pesquisa 7: “Otimização dos processos de indexação dos documentos e de recuperação da informação mediante o uso de instrumentos de controle terminológico”

Objetivo	Estudar técnicas de indexação automática de textos, ou mesmo as técnicas de indexação automática simulada, para facilitar na preparação de instrumentos de controle terminológico que reflitam, com razoável fidelidade, o conteúdo dos documentos que integram os arquivos processados e assegurem a obtenção de bons resultados no processo de recuperação da informação, ao permitirem a atualização do vocabulário controlado, e, ao mesmo tempo, que atualizem as bases de dados.
Pesquisador(es)	ROBREDO, Jaime.
Período	1982

Local	Brasil
Critério(s) utilizado(s)	nº 2: Fórmula de transição de Goffman nº 6: Freqüência relativa de ocorrência da palavra no texto; nº 13: Primeira lei de Zipf nº 14: Segunda lei de Zipf ou Lei de Zipf-Booth nº 16: Vocabulário semântico / vocabulário de cabeçalhos conceituais / Tesauro
***Software(s)* utilizado(s)**	Desenvolvido pelo pesquisador.
Comparação com indexação manual	Sim.
Tipo de documento	Documentos textuais
Área de cobertura	Ciências da terra (pesquisa agrícola) e política científica e tecnológica.
Resultado	Satisfatório, porém, é uma área complexa que necessita de contínuo desenvolvimento.
Numeração do texto constante na amostra nº 1	Texto nº 16.

Fonte: desenvolvido pela autora.

QUADRO 28

Pesquisa 8: “On the Application of Syntactic Methodologies in Automatic Text Analysis”

Objetivo	Sintetizar várias abordagens lingüísticas propostas para a análise de documentos e recuperação da informação. Apresentar método de análise sintática para geração de conteúdo complexo identificador, bem como utilizar-se da semântica, obtida a partir de leitura óptica de dicionários especialmente construídos a partir de conhecimentos de base. Contribuir para construção automática de índices de livro.
Pesquisador(es)	SALTON, Gerard; SMITH, Maria.
Período	1989
Local	EUA
Critério(s) utilizado(s)	nº 1: Formatação de frases-termo (*Word phrase formation*) nº 3: Freqüência absoluta de ocorrência da palavra no texto nº 7: Identificação de palavras (Comparação com uso de dicionário) nº 8: Identificação de radicais de palavras (Word stemming) nº 12: Posição do termo no texto (*Term weighting)* nº 13 Primeira lei de Zipf nº 14: Segunda lei de Zipf ou Lei de Zipf-Booth
***Software(s)* utilizado(s)**	Desenvolvido pelo pesquisador.
Comparação com indexação manual	Sim.

Tipo de documento	Documentos textuais
Área de cobertura	Não indicado.
Resultado	Satisfatório, porém, é uma área complexa que necessita de contínuo desenvolvimento.
Numeração do texto constante na amostra nº 1	Texto nº 24.

Fonte: desenvolvido pela autora.

QUADRO 29

Pesquisa 9: "Automatic abstracting of magazine articles: the creation of 'highlight' abstracts"

Objetivo	Elaborar, automaticamente, resumos de artigos de revista baseando-se em gramáticas embutidas, a partir de frases e declarações de extratos relevantes do texto.
Pesquisador(es)	MOENS, Marie-Francine; DUMORTIER, Jos.
Período	1998
Local	EUA
Critério(s) utilizado(s)	nº 1: Formatação de frases-termo (*Word phrase formation*) nº 3: Freqüência absoluta de ocorrência da palavra no texto nº 7: Identificação de palavras (Comparação com uso de dicionário) nº 8: Identificação de radicais de palavras (Word stemming) nº 12: Posição do termo no texto (*Term weighting)*
***Software(s)* utilizado(s)**	Desenvolvido pelo pesquisador.
Comparação com indexação manual	Sim.
Tipo de documento	Documentos textuais: notícias e artigos de revistas.
Área de cobertura	Ciências sociais aplicadas: jornalismo.
Resultado	Satisfatório,
Numeração do texto constante na amostra nº 1	Texto nº 54.

Fonte: desenvolvido pela autora.

QUADRO 30

Pesquisa 10: “Aplicação de técnicas infométricas para identificar a abrangência do léxico básico que caracteriza os processos de indexação e recuperação da informação”

Objetivo	Otimizar os processos de indexação automática e recuperação da informação, passando pela possibilidade de acompanhar a evolução dos temas de interesse da pesquisa científica. Apresentar uma aplicação da análise das co-ocorrências de pares de palavras-chave para identificação do âmbito e da abrangência do léxico básico, que caracteriza os processos de indexação e recuperação da informação.
Pesquisador(es)	ROBREDO, Jaime; CUNHA, Murilo Bastos da.
Período	1998
Local	Brasil
Critério(s) utilizado(s)	nº 4: Freqüência de co-ocorrência relativa de termos nº 5: Freqüência de co-ocorrência simples de termos nº 7: Identificação de palavras (Comparação com uso de dicionário) nº 8: Identificação de radicais de palavras *(Word stemming)* nº 9: Lista de palavras proibidas / Palavras proibidas *(Stop-list / stop-words)* nº 12: Posição do termo no texto (*Term weighting)*
***Software(s)* utilizado(s)**	InfoDoc
Comparação com indexação manual	Não.
Tipo de documento	Técnico-científico: artigos.
Área de cobertura	Ciências sociais aplicadas: Ciência da Informação.
Resultado	Satisfatório.
Numeração do texto constante na amostra nº 1	Texto nº 55.

Fonte: desenvolvido pela autora.

QUADRO 31

Pesquisa 11: “Utilização da indexação automática para auxílio à construção de uma base de dados para a extração de conhecimento aplicada a doenças pépticas”

Objetivos	Extração de dados de base de dados para a descoberta de conhecimento a partir de um conjunto de dados, o qual poderá ser utilizado em um processo decisório.
Pesquisador(es)	Daniel de F. Honorato; Huei D. Lee; Renato B. Machado; Feng C.Wu; Antonio P. Neto.
Período	[2003]
Local	Brasil.

Critério(s) utilizado(s)	nº 1: Formatação de frases-termo (*Word phrase formation*) nº 7: Identificação de palavras (Comparação com uso de dicionário) nº 8: Identificação de radicais de palavras *(Word stemming)* nº 9: Lista de palavras proibidas / Palavras proibidas *(Stop-list / stop-words)*
***Software(s)* utilizado(s)**	Linguagem Delphi 6.0; Linguagem de consulta SQL; Sistema Gerenciador de Banco de Dados mySQL.
Comparação com indexação manual	Não.
Tipo de documento	Médico. Documentos onde são armazenadas informações sobre o exame de Endoscopia Digestiva Alta tais como data, médico que atendeu, idade, entre outros, assim como o laudo, que armazena o resultado do exame de EDA em formato texto.
Área de cobertura	Ciências da saúde.
Resultado	Satisfatório.
Numeração do texto constante na amostra nº 1	Texto nº 80.

Fonte: desenvolvido pela autora.

QUADRO 32

Pesquisa 12: "Um modelo algébrico para representação, indexação e classificação automática de documentos digitais"

Objetivo	Agrupar documentos semelhantes. Exemplificar metodologias algébricas de representação e de indexação automáticas de documentos textuais como mais uma ferramenta para o profissional da informação. Mostrar que este conjunto de técnicas pode ser utilizado para classificar documentos de forma automática (ou semi-automática em certas circunstâncias em que a máquina não conseguir ter certeza) e, em conseqüência disso, muito mais rápido do que faria um ser humano.
Pesquisador(es)	Elias Oliveira; Patrick Marques Ciarelli; Marcos Hercules Santos; Bruno Oliveira da Costa.
Período	[2007]
Local	Brasil.
Critério(s) utilizado(s)	nº 3: Freqüência absoluta de ocorrência da palavra no texto nº 7: Identificação de palavras (Comparação com uso de dicionário) nº 9: Lista de palavras proibidas / Palavras proibidas *(Stop-list / stop-words)* nº 11: Peso numérico
***Software(s)* utilizado(s)**	Linguagem Java;
Comparação com indexação manual	Sim.
Tipo de documento	Jornalístico: repositório de notícias RSS do UOL.
Área de cobertura	Ciências Sociais.
Resultado	Satisfatório.
Numeração do texto constante na amostra nº 1	Texto nº 101.

Fonte: desenvolvido pela autora.

A partir dos dados obtidos, foi composta a TAB 1, que apresenta de forma simplificada, a relação das pesquisas com os critérios por elas utilizados. Os dados para composição desta tabela estavam contidos nos Quadros de 19 a 40 e serão assim apresentados para uma melhor visualização do número de textos que utilizou cada critério.

Tabela 1 – Utilização dos critérios de indexação em cada texto da amostra de estudo nº 2

	Pesq. 1	Pesq. 2	Pesq. 3	Pesq. 4	Pesq. 5	Pesq. 6	Pesq. 7	Pesq. 8	Pesq. 9	Pesq. 10	Pesq. 11	Pesq. 12	Quantidade de pesquisas que utilizaram o critério	Porcentagem	Nome do critério
	[Década de 1950]		[Década de 1960]		[Década de 1970]		[Década de 1980]		[Década de 1990]		[Década de 2000]				
Critério 1	X		X	X	X	X		X	X		X		8	66,67%	Formatação de frases-termo (*Word phrase formation*)
Critério 2							X						1	8,33%	Fórmula de transição de Goffman
Critério 3				X	X			X	X			X	5	41,66%	Freqüência absoluta de ocorrência da palavra no texto;
Critério 4		X			X					X			3	25,00%	Freqüência de co-ocorrência relativa de termos
Critério 5					X					X			2	16,66%	Freqüência de co-ocorrência simples de termos
Critério 6		X					X						2	16,66%	Freqüência relativa de ocorrência da palavra no texto
Critério 7			X	X	X	X		X	X	X	X	X	9	75,00%	Identificação de palavras (Comparação com uso de dicionário)
Critério 8					X	X		X	X		X		5	41,66%	Identificação de radicais de palavras *(Word stemming)*
Critério 9						X				X	X	X	4	33,33%	Lista de palavras proibidas / Palavras proibidas *(Stop-list / stop-words)*
Critério 10				X									1	8,33%	Palavras destacadas no texto
Critério 11		X	X	X		X						1	5	41,66%	Peso numérico
Critério 12				X	X			X	X	X			5	41,66%	Posição do termo no texto (*Term weighting)*
Critério 13							X	X					2	16,66%	Primeira lei de Zipf
Critério 14							X	X					2	16,66%	Segunda lei de Zipf ou Lei de Zipf-Booth
Critério 15	X												1	8,33%	Tópico frasal
Critério 16			X		X	X	X						4	33,33%	Vocabulário semântico / vocabulário de cabeçalhos conceituais / Tesauro

Fonte: desenvolvida pela autora.

De maneira a unificar os critérios, confrontando-os quantitativamente, foi elaborado o GRÁFICO 1, que apresenta a porcentagem de utilização de cada critério no universo total dos textos contidos na amostra de nº 2.

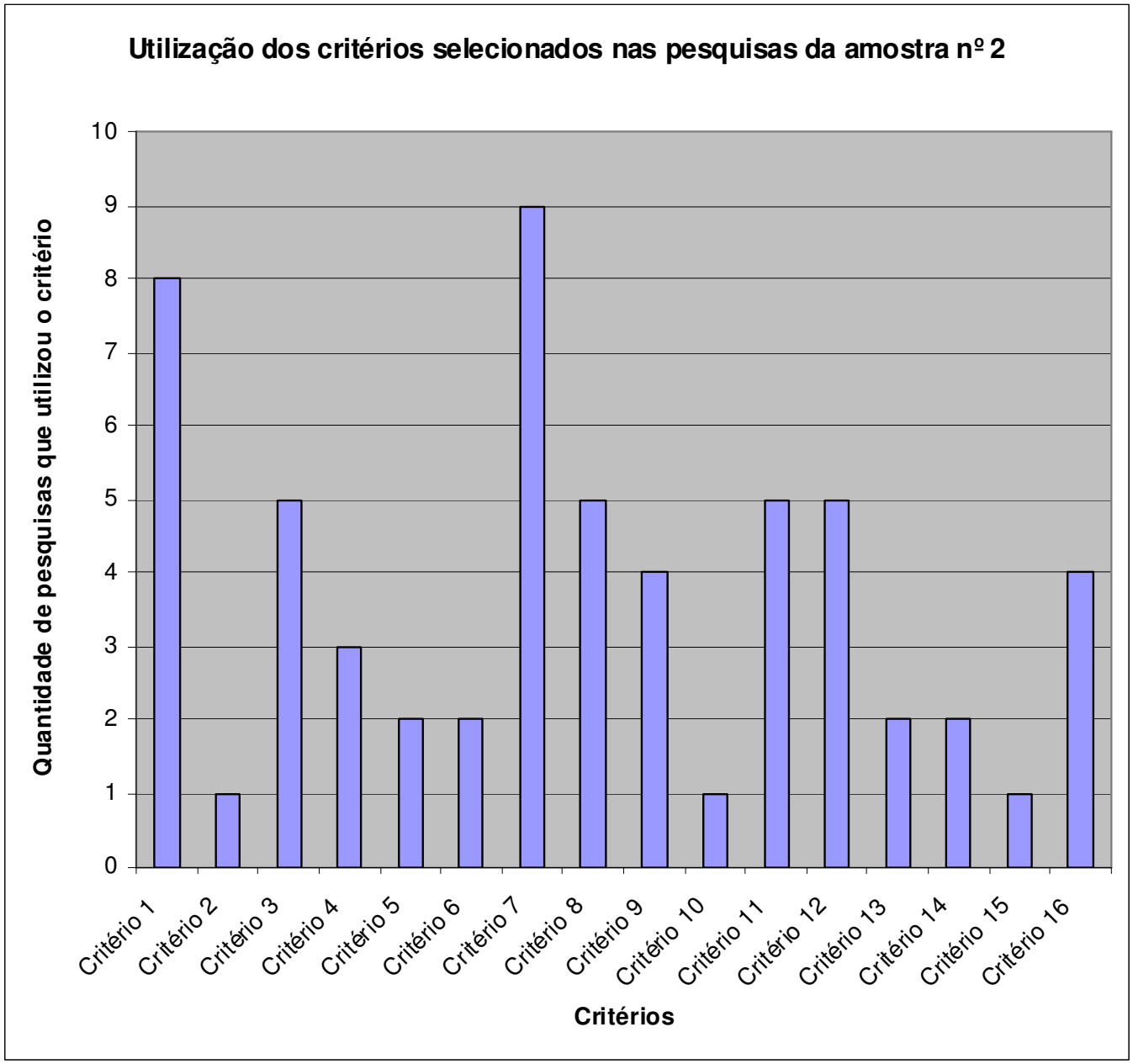

GRÁFICO 1 – Relação texto X critério utilizado
Fonte: desenvolvido pela autora.

A partir da análise das pesquisas realizadas, foi possível identificar o seguinte resultado.

Resultado 3:

Os critérios mais utilizados no processo de indexação automática

Com essa informação, é possível observar quais são os critérios mais utilizados e que são, conseqüentemente, combinados o maior número de vezes com outros critérios. Observa-se que tais critérios são relevantes para o processo.

Levando em consideração que a maior parte das pesquisas aponta resultados satisfatórios, o fator mais relevante para a conclusão será a quantidade de vezes que os critérios foram utilizados em relação ao número total de textos analisados.

Dessa maneira, será apresentado adiante o resultado obtido através da análise realizada ao longo desta pesquisa. Torna-se possível, então, propor um conjunto de critérios considerado ideal para o processo de indexação automática, que, de acordo com os objetivos da pesquisa, busca solucionar o problema proposto inicialmente.

3.4 Discussão e apresentação dos resultados

A partir da análise da TAB 1, é possível a avaliação de alguns aspectos relevantes sobre a utilização dos critérios de indexação automática selecionados na literatura com base na amostra nº 1.

De um total de dezesseis critérios selecionados, 50% destes apresentou uma taxa de utilização acima de 30% em relação ao número total de pesquisas analisadas, que corresponderam a doze pesquisas. Esses critérios são apresentados na TAB. 2.:

Tabela 2 – Relação dos critérios mais utilizados pelas pesquisas indicadas na amostra nº 1

Número do critério	Quantidade de pesquisas que utilizou o critério	Porcentagem	Nome do critério
Critério 7	9	75,00%	Identificação de palavras (Comparação com uso de dicionário)
Critério 1	8	66,67%	Formatação de frases-termo (*Word phrase formation*)

Critério 12	5	41,66%	Posição do termo no texto (*Term weighting)*
Critério 11	5	41,66%	Peso numérico
Critério 8	5	41,66%	Identificação de radicais de palavras *(Word stemming)*
Critério 3	5	41,66%	Freqüência absoluta de ocorrência da palavra no texto
Critério 16	4	33,33%	Vocabulário semântico / vocabulário de cabeçalhos conceituais / Tesauro
Critério 9	4	33,33%	Lista de palavras proibidas / Palavras proibidas *(Stop-list / stop-words)*

Fonte: desenvolvida pela autora.

Julga-se que o critério nº 3, *freqüência absoluta de ocorrência da palavra no texto,* seja relevante para análise de documentos textuais. O critério foi utilizado em cinco das doze pesquisas analisadas, o que corresponde a um total de 41,66%. Embora esse seja um critério que, usualmente, é visto como limitado, por considerar apenas o número de vezes que cada palavra ocorre no texto analisado, ele mostrou um índice considerável de utilização ao longo de cinco das seis décadas analisadas. A *freqüência absoluta de ocorrência da palavra no texto,* apresenta relação direta com três outros critérios:

- *Freqüência de co-ocorrência relativa de termos*, que obteve 25,00% de utilização;
- *Freqüência de co-ocorrência simples de termos*, que obteve 16,66% de utilização;
- *Freqüência relativa de ocorrência da palavra no texto*, que obteve 16,66% de aproveitamento.

De fato, a freqüência de ocorrência relativa e a freqüência de co-ocorrência, simples e relativa, são critérios mais robustos que a freqüência de ocorrência simples, porque consideram, além da quantidade de aparecimento de cada palavra no texto, sua ocorrência na base de dados como um todo e ainda a relação existente entre as palavras que compõem o documento. Assim, o critério de medição da freqüência de ocorrência absoluta de uma palavra em um texto passou a ser utilizado em conjunto com outros critérios que consideram aspectos lingüísticos do texto, como é o caso do critério nº 7, *identificação de palavras (comparação com uso de dicionários)*, que apresentou 75,00% de aproveitamento, e o critério nº 16, *vocabulário semântico / vocabulário de cabeçalhos conceituais / tesauro,* com 33,33% de utilização.

Pode-se acreditar, então, que a parceria da utilização do critério *freqüência absoluta de ocorrência da palavra no texto* com outros critérios que consideram aspectos semânticos pode suprimir o uso de outros critérios puramente estatísticos.

Sobre o critério nº 16, *vocabulário semântico / vocabulário de cabeçalhos conceituais / tesauro,* percebe-se que, embora esse critério vigore entre os mais usados, sua utilização ainda é tímida visto seu grande potencial para o tratamento de aspetos semânticos do texto. Como indicado na seção 2.2 deste texto, o uso de uma taxonomia para o tratamento de documentos textuais poderá ser ótimo recurso visto a carência de *parsers* disponibilizados no mercado, atualmente, que fazem este tipo de análise.

Diferentemente do que era esperado, o critério nº 9, *lista de palavras proibidas / palavras proibidas (stop-list / stop-words),* obteve apenas 33,33% de utilização na amostra analisada. Esperava-se para esse critério, assim como o critério nº 16, um alto índice de utilização, já que foi um dos primeiros desenvolvidos na área. Contudo, considera-se a possibilidade de omissão por parte dos autores dos textos analisados sobre a utilização desse critério em especial, justamente devido ao fato de que sua importância é consensual entre os pesquisadores da área.

Os quatro últimos critérios verificados com índice alto de utilização também podem apresentar um relacionamento. O critério nº 1, *formatação de frases-termo (word phrase formation),* com 66,67% de utilização, e o critério de nº 8, *identificação de radicais de palavras (word stemming),* com 41,66%, são critérios que estão ligados à estrutura de formação da palavra. O primeiro verifica o relacionamento de palavras próximas para a formação de frases ricas em conteúdo representativo do texto. O segundo considera o radical de cada palavra para realização de eliminação, ou consideração, de um grupo de palavras que contenham o radical indicado. Essa verificação é feita com base em uma lista, previamente definida, de radicais de palavras que devem ser descartadas e/ou consideradas posteriormente à verificação do *parser*. Ainda hoje, esses dois critérios são considerados de extrema relevância para análise de documentos textuais, visto que a verificação da estrutura gramatical é a base para a realização de análises semânticas, que se fazem necessárias em um segundo momento.

Finalmente, os dois útimos critérios, *peso numérico* e *posição do termo no texto (term weighting),* que, por coincidência, apresentaram 41,66% de aproveitamento, podem ser associados. Ambos apresentam aspectos de atribuição de grau de importância para determinas palavras do texto. A idéia vigente no primeiro critério é a determinação de valores especiais para grupos de palavras já previamente definidas como relevantes

para aquela área de assunto específica. No segundo critério, a atenção está voltada para a definição de partes do texto potencialmente candidatas a conterem palavras que sejam representativas do documento, como é o caso do título do texto, de seu resumo e de sua conclusão. Atualmente, como indicado para os dois critérios tratados anteriormente a estes, acredita-se que estes dois critérios são considerados relevantes para análise de documentos textuais, visto que prevêem uma redução da análise do texto como um todo para a realização de uma análise baseada em partes específicas do texto e na consideração de palavras com alto grau de relevância relacionado ao assunto tratado.

Os outros 50% de critérios que apresentaram uma taxa de utilização abaixo de 30% em relação ao número total de pesquisas analisadas estão apresentados na TAB. 3.

Tabela 3 – Relação dos critérios menos utilizados pelas pesquisas indicadas na amostra nº 1

Número do critério	Quantidade de pesquisas que utilizou o critério	Porcentagem	Nome do critério
Critério 15	1	8,33%	Tópico frasal
Critério 10	1	8,33%	Palavras destacadas no texto
Critério 2	1	8,33%	Fórmula de transição de Goffman
Critério 14	2	16,66%	Segunda lei de Zipf ou Lei de Zipf-Booth
Critério 13	2	16,66%	Primeira lei de Zipf
Critério 6	2	16,66%	Freqüência relativa de ocorrência da palavra no texto
Critério 5	2	16,66%	Freqüência de co-ocorrência simples de termos
Critério 4	3	25,00%	Freqüência de co-ocorrência relativa de termos

Fonte: desenvolvida pela autora.

Da análise da TAB. 2, fazem-se alguns comentários. Três dos critérios apresentados, o critério nº 2, *fórmula de transição de Goffman*, com 8,33% de aproveitamento, e os critérios nº 13 e 14, *primeira e segunda lei de Zipf ou lei de Zipf-Booth*, respectivamente, ambos com 16,66%, podem ser relacionados entre si devido ao fato de terem como base a análise estatística das palavras do texto. Percebe-se que esses critérios, atualmente, não se fazem mais necessários, visto que, como indicado

anteriormente, a combinação de um critério de análise de freqüência com outros critérios com características de tratamento lingüístico, podem suprir a necessidade da utilização de outros critérios estatísticos em excesso.

Outro critério de pouca representatividade na amostra nº 2, foi o critério nº 10, *palavras destacadas no texto*, com 8,33% de aproveitamento. Essa consideração, para análise do *parser,* embora possa apresentar algum resultado satisfatório, não é consistente o suficiente para ser indicada no resultado final desta pesquisa.

Por último, mas não menos importante, analisamos o critério nº 15, *tópico frasal,* com 8,33% de utilização, ou seja, que foi considerado apenas por uma das doze pesquisas da amostra. Esse é um critério que merece muita atenção, visto ter sido um dos precursores da área. Criado por Hans Peter Luhn, no final da década de 1950, o critério ainda é utilizado atualmente, em conjunto com outros critérios já mencionados, como a identificação de radicais de palavras (*word stemming)* e o uso da freqüência de ocorrência de palavras como medida de relevância para o desenvolvimento de algoritmos, como o Naive Bayes, Log-linear models, o SVMs, entre outros.

Luhn trabalhou intensamente nos laboratórios da IBM no final da década de 1950, bem como outros pioneiros como Baxendale e Edmundson, e pode ser considerado o fundador da disciplina *information retrieval* – recuperação da informação. Em suma, no final da década de 1950 e na década de 1960, discutia-se a realização da sumarização automática de documentos textuais. A partir da década de 1990, e até hoje, percebe-se que o interesse da área está voltado para o tratamento de grandes volumes de informação, principalmente de notícias e reportagens. Ainda assim, tendo apresentado as particularidades do critério nº 15, em especial, este não é considerado relevante para os objetivos desta pesquisa, podendo ser substituído pelo uso de outro critério que irá produzir, além de sua função específica, um resultado semelhante ao uso de tópico frasal para realizar a análise automática de textos.

Finalmente, a partir da análise minuciosa dos critérios observados ao longo do estudo, propõe-se, aqui, um conjunto de 9 (nove) critérios entendidos como ideais para o desenvolvimento de *parser* de indexação automática para o tratamento de documentos textuais. Sugere-se que, para o desenvolvimento desse tipo de *software*, seja considerado o conjunto proposto, pois acredita-se que esse conjunto pode proporcionar uma extração de termos significativos dos documentos indexados, obtendo um resultado semelhante àquele que seria obtido através do trabalho realizado pelo ser humano.

Entende-se que, em um processo acadêmico, as teorias são constantemente revistas e as mais adequadas aos estudos permanecem, são utilizadas e aperfeiçoadas. Assim, acredita-se que os critérios mais utilizados apontam para o melhor conjunto existente.

Seguem, assim, os critérios que foram selecionados:

QUADRO 33

Conjunto de critérios de indexação automática para o desenvolvimento de softwares para análise de conteúdo de documentos textuais

Número do critério	Nome do critério
Critério 1	Formatação de frases-termo (*Word phrase formation*)
Critério 3	Freqüência absoluta de ocorrência da palavra no texto
Critério 7	Identificação de palavras (Comparação com uso de dicionário)
Critério 8	Identificação de radicais de palavras *(Word stemming)*
Critério 9	Lista de palavras proibidas / Palavras proibidas *(Stop-list / stop-words)*
Critério 11	Peso numérico
Critério 12	Posição do termo no texto (*Term weighting)*
Critério 16	Vocabulário semântico / vocabulário de cabeçalhos conceituais / Tesauro

Fonte: desenvolvida pela autora.

Acredita-se que a lista apresentada pode ser uma possível resposta à questão proposta no início do estudo: *quais são os critérios de indexação automática necessários para o desenvolvimento de um software capaz de extrair termos representativos do conteúdo de documentos textuais, aproximando-se ao máximo do trabalho realizado pelo profissional indexador?*

Tendo-se atingido o resultado desejado, passa-se, então, para as considerações finais deste estudo.

4 CONSIDERAÇÕES FINAIS

Neste capítulo, serão apresentadas as reflexões sobre o desenvolvimento e resultados obtidos neste trabalho, assim como as indicações sobre estudos que podem surgir a partir desta pesquisa.

Foi visto que a indexação é o elo entre o que é disponibilizado no sistema e aquilo que é recuperado pelo usuário, de acordo com sua necessidade. Esta atividade tem se tornado cada vez mais intensa, desde que a publicação de documentos textuais sofreu considerável crescimento. Existe, hoje, uma constante produção e busca de informação, o que propicia um cenário no qual se faz necessário organizar as informações, de forma sistemática, para disponibilizá-las ao usuário de maneira satisfatória. Constataram-se deficiências e percalços no processo manual de indexação, o que reafirmou a necessidade de estudos que buscassem encontrar alternativas para esse processo.

Verificou-se que já são muitos os estudos que se voltam para esse tipo de investigação e que há, igualmente, muitos *parsers* com diferentes propostas de indexação automática. Todos objetivam otimizar a atividade de análise de assunto, procurando diminuir o tempo gasto nessa etapa do processo de indexação. Para isso, conta-se com a agilidade apresentada pelo *software*, que consegue fazer uma leitura quase instantânea do texto. Além disso, outros benefícios trazidos pelo uso de programas computacionais no processo são o aumento da coerência obtida nessa etapa e a redução da subjetividade inerente à realização manual do processo.

Constatou-se, a partir da literatura mapeada neste estudo, que a maioria dos documentos produzidos aborda pesquisas que contribuíram para a indexação automática por meio da realização de experimentos práticos, ou seja, com a construção de algoritmos computacionais e com a realização efetiva de testes em documentos textuais.

Em meio aos textos selecionados para realização deste estudo, não se encontrou, em pesquisa alguma, abordagem semelhante à apresentada neste trabalho, tendo este uma proposta diversa das apresentadas anteriormente. Assim, desejou-se o desenvolvimento de um trabalho que pudesse compilar os resultados alcançados nas amostras de estudo, com a intenção de estabelecer um diálogo entre essas pesquisas, proporcionando um meio de se alcançar um resultado a partir de uma síntese de

trabalhos analisados, com base em documentos produzidos ao longo das seis últimas décadas.

A proposta do estudo foi, então, avaliar critérios de indexação automática, com o objetivo de compor uma lista daqueles cuja combinação forneceria o resultado mais satisfatório possível, para a construção de *parsers* destinados ao tratamento de documentos textuais.

Através da investigação de estudos anteriormente produzidos nessa área, identificaram-se os critérios de indexação automática utilizados pelos pesquisadores. A lista desses critérios foi o primeiro dos resultados obtidos. Tal resultado apresenta-se como uma ferramenta bastante útil para pesquisadores e estudantes da BCI, uma vez que se trata da compilação de dados anteriormente dispersos nos diversos trabalhos, tornando-se uma fonte de consulta.

De posse de tal lista de critérios, fez-se uma sistematização detalhada dos mesmos, através da composição de quadros descritivos de cada critério. Esse resultado, o segundo deles, tem ainda grande valia para os estudos realizados na área, pois não se encontra na literatura outra fonte de pesquisa que contenha tal tipo de sistematização. Os critérios foram descritos através da exposição de suas características principais. Essa descrição é importante para os pesquisadores e estudantes da área por contribuir para a melhor compreensão do conteúdo, devido à clareza das informações e de sua disposição através de quadros, permitindo melhor visualização e possibilitando o estudo objetivo dos dados.

O terceiro resultado alcançado foi a indicação dos critérios mais recorrentes nas pesquisas realizadas. Esse resultado teve aplicação direta para o estudo, contribuindo para o alcance da questão proposta.

Esta pesquisa deteve-se na análise dos dados indicados na posição horizontal da TAB 1, ou seja, as linhas, em que está indicado o número de vezes que o critério ocorreu nos textos em análise. Através da observação da tabela, é possível analisar também as colunas, obtendo-se outra informação: o conjunto de critérios utilizado em cada texto. Propõe-se, como estudo futuro, a observação minuciosa desses conjuntos, confrontando-os com os resultados indicados nos quadros, referentes ao grau de satisfação obtido com o uso de tais critérios em cada pesquisa.

Pôde-se perceber também a importância de se considerar aspectos semânticos do texto para que a indexação seja realizada de maneira mais contextualizada e consistente. Acredita-se que a utilização de vocabulário controlado, como uma

taxonomia, embutido nos *parsers* desenvolvidos para o processo de indexação automática pode contribuir e fortalecer o processo, quando associada aos aspectos semânticos. Entretanto, não foi esse o foco deste trabalho. Assim, o aspecto semântico do processo de indexação automática possibilita estudos futuros relevantes para a área.

O que se espera na atualidade, e que deve ser uma das prioridades da área, é que *parsers* de indexação automática sejam ferramentas capazes de, concomitantemente, minimizar a subjetividade do indexador e imitar o raciocínio humano. Eles deverão levar em consideração o contexto semântico, respeitando princípios teóricos consistentes.

Por meio das análises realizadas, da observação dos resultados obtidos pelos pesquisadores em suas respectivas pesquisas e do confronto dos conjuntos de critérios utilizados com os resultados obtidos, forneceram-se parâmetros comparativos para o aumento da qualidade da indexação automática, atingindo-se, assim, o primeiro dos objetivos específicos. O fornecimento desses dados permite a associação do conjunto de critérios utilizados com o grau de satisfação atingido pela pesquisa, oferecendo opções de combinações de critérios possíveis para o desenvolvimento de *softwares*, ainda que limitadas em alguns casos, em que a inserção ou a supressão de determinado critério no conjunto poderia fornecer grau de satisfação ainda mais elevado.

Concomitantemente, a análise e o contraste das combinações abrangem também o segundo dos objetivos específicos: melhorar o processo da indexação automática através de critérios relevantes para a extração de termos representativos do conteúdo do documento. As combinações permitem esse melhoramento no processo.

Por fim, apresentou-se o conjunto final dos critérios, fruto da observação e associação realizadas durante todo o trabalho. No conjunto proposto, encontram-se os critérios mais utilizados nas pesquisas analisadas. Grande parte dessas pesquisas gerou resultado satisfatório, portanto, observando-se a freqüência com que determinado critério foi utilizado, é possível associá-lo a um bom desempenho no processo de indexação.

O resultado final do trabalho responde à questão que constitui o problema proposto por esta pesquisa e está intimamente vinculado ao terceiro dos objetivos específicos, que é auxiliar os profissionais da Ciência da Computação no desenvolvimento de *parsers* de indexação automática.

As pesquisas na área da Lingüística Computacional estão vislumbrando soluções que possibilitarão às máquinas reproduzirem o conhecimento lingüístico e semântico de um ser humano. Para a criação de um *parsers* de indexação automática, é

necessário um conjunto de critérios próprios desse processo para o desenvolvimento de um algoritmo que desempenhe as funções designadas pelo programador. Por sua vez, o programador, profissional responsável pelo desenvolvimento do *software*, necessita de conhecimentos específicos da área a que se destina o programa. Assim, para o desenvolvimento do *parser*, ele precisa saber quais são os critérios a serem utilizados. Para obter tal informação, ele pode contar com o apoio de um profissional da BCI ou pode, ainda, recorrer a outras fontes de informação sobre o assunto.

Acredita-se que os critérios indicados ao final desta pesquisa, se utilizados por criadores de *parsers* de indexação automática no desenvolvimento de seus programas, irão suprir os problemas descritos inerentes à indexação manual, agilizando a seleção de termos representativos dos documentos.

REFERÊNCIAS

ANDER-EGG, Ezequiel. **Introducción a las técnicas de investigación social:** para trabajadores sociales. 7. ed. Buenos Aires: Humanitas, 1978. 335p. parte 2, cap. 6.

ANDERSON, Charles. Indexing with a computer: past and present**. The Indexer**, v. 22, n. 1, p. 23-24, Apr. 2000.

ARAUJO, Vânia M. R. H. Sistemas de informação: nova abordagem teórico-conceitual. **Ci. Inf.**, v. 24, n. 1, 1995.

AUTO JUNIOR, Tavares da C. **Indexação automática de acórdãos por meio de processamento de linguagem natural.** 2007. 142f. Dissertação (Mestrado em Ciência da Informação) – Faculdade de Economia, Administração, Contabilidade e Ciência da Informação e Documentação, Universidade de Brasília. Brasília: Faculdade FAAP, 2007.

BARBOSA, Ricardo R. Perspectivas profissionais e educacionais em biblioteconomia e ciência da informação. **Ci. Inf.**, v. 27, n.1, p. 53-60, jan./abr. 1998.

BAXENDALE, P. B. Machine-made index for technical literature: an experiment. **IBM Journal of Research and Development**, n. 2, p. 354-361, 1958.

BELLUZZO, Regina C. B. et al. Indexação automática de documentos com PRECIS *Software* no serviço de biblioteca e documentação da Faculdade de Odontologia de Bauru (USP). In: SIMPÓSIO LATINO-AMERICANO DE TERMINOLOGIA, 2., Brasília, 1990**. Anais...** Brasília: IBICT, 1990. p. 159-164.

BEST, J. W. **Como investigar em educación.** 2. ed. Madrid: Morata, 1972. cap. 1 e 2.

BORGES, Graciane S. B.; MACULAN, Benildes C. M. S.; LIMA, Gercina Â. B. O. Indexação automática e semântica: estudo da análise do conteúdo de teses e dissertações. **Inf. Soc.: Est.,** João Pessoa, v. 18, n. 2, p. 181-193, maio/ago. 2008.

BORKO, H. Toward a theory of indexing. **Information Processing and Management,** v. 13, p. 355-365, 1977.

BUFREM, Leilah S. A relação inescusável entre lingüística e documentação. **Enc. Bibli: R. Eletr. Biblio. Ci. Inf.**, Florianópolis, n. 19, p. 83-94, 1º sem. 2005.

CAMPOS, Maria Luiza A.; GOMES, Hagar E. Taxonomia e classificação: o princípio de categorização. **DataGramaZero - Rev. Ci. Inf.,** v. 9, n. 4, ago. 2008.

CARVALHO, M. M. M. de. O problema da subjetividade na indexação. **Cad. BAD**, Lisboa, n. 1, p. 9-40, 1995.

CESARINO, Maria Augusta N.; PINTO, Maria Cristina. M. F. Análise de assunto. **Rev. Biblio. Bras.**, Brasília, v. 8, n. 1, p. 32-43, jan./jun. 1980.

CHOWDHURY, G. **Introduction to modern information retrieval**. London: Library Association Publishing, 1999. 452p.

DAHLBERG, Ingetraut. Teoria do conceito. Trad. CAMPOS, Astério Tavares. **Ci. Inf.,** Rio de Janeiro, v. 7, n. 2, p. 101-107, 1978.

DIAS, Eduardo W.; NAVES, Madalena M. L. **Analise de assunto**: teoria e pratica. Brasília: Thesaurus, 2007. 116 p.

ECO, Humberto. **Os limites da interpretação**. Lisboa: Difel, 1992.

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

ELLIS, D. G. **From language to communication.** Hillsdale, NJ: Lawrence Erlbaum, 1992.

FIGUEIREDO, Saulo. **O impacto da taxonomia nas empresas.** [S.l.]: Webinsider, 28 nov. 2006. Disponível em: <http://webinsider.uol.com.br/index.php/2006/11/28/a-importancia-e-oimpacto-da-taxonomia-nas-empresas/>. Acesso em: 24 jul. 2007.

FUGMANN, Robert. **Subject analysis and indexing.** Frankfurt, Main: Indeks Verlag, 1993. v. 1, 250p.

FUJITA, Mariângela S. L. A identificação de conceitos no processo de análise de assunto para indexação. **Rev. Dig. Biblio. Ci. Inf.,** Campinas, v. 1, n. 1, p. 60-90, jul./dez. 2003.

FUJITA, Mariângela S. L. **Leitura documentária do indexador:** aspectos cognitivos e lingüísticos influentes na formação do leitor profissional. 2003. 321f. Tese (Livre-Docência nas disciplinas Análise Documentária e Linguagens Documentárias

Alfabéticas). Faculdade de Filosofia e Ciências, Universidade Estadual Paulista, Marília, 2003b.

FUJITA, Mariângela S. L. **Leitura em análise documentária**. Marília, SP: UNESP/CNPq, 1999. Relatório parcial de pesquisa.

GARVIN, P. L. et al. **Some opinions concerning linguistics and reformation processing**. Washington, D. C.: Center for Applied Linguistics, May 1969. Available from National Technical Information Service. (Reporter PB 190 639).

GIL LEIVA, Isidoro. **La automatización de la indización de documentos**. Madrid: Ediciones Trea, 1999. 221p.

GUEDES, Vânia L. S. Estudo de um critério para indexação automática derivativa de textos científicos e tecnológicos. **Ci. Inf.**, Brasília, v. 23, n. 3, p. 318-326, set./dez. 1994.

GUIRAUD, Pierre. **A semântica**. 2. ed. Rio de Janeiro: DIFEL, 1975.

HALLIDAY, M. A. K. **Spoken and written language**. Trad. Maria Elisa Mascarenhas. Oxford: Oxford University, 1989.

HJELMSLEV, Louis. **Prolegomenos a uma teoria da linguagem**. Trad. J. Teixeira Coelho Netto. São Paulo: Perspectiva, 1975.

HJORLAND, Birger. The concept of 'subject' in Information Science. **Journal of Documentation**, v. 48, n. 2, p. 172-200, Jun. 1992.

LANCASTER, F. W. **Indexação e resumos**: teoria e prática. Brasília: Briquet de Lemos, 2004. 452p.

LANCASTER, F. W. **Indexação e resumos**: teoria e prática. Brasília: Briquet de Lemos, 1993. 347p.

LAKATOS, Eva Maria. **Fundamentos de Metodologia Científica**. 3. ed. rev. e aum. São Paulo: Atlas, 1991.

LIMA, Gercina Â. B. O. Categorização como um processo cognitivo. **Ciências & Cognição**, ano 4, v. 11, p. 156-167, 2007. Disponível em: <www.cienciasecognicao.org>. Acesso em: 9 ago. 2007.

LIMA, Gercina Â. B. O. **Mapa hipertextual (MHTX):** um modelo para organização hipertextual de documentos. 2004. 204f. Tese (Doutorado em Ciência da Informação) - Escola de Ciência da Informação da UFMG, Belo Horizonte. 2004a.

LIMA, Vânia M. A.; BOCATO, Vera R. C. O desempenho terminológico dos descritores em Ciência da Informação do Vocabulário Controlado do SBIi / USP nos processos de indexação manaul, automática e semi-automática. **Persp. Ci. inf.**, Belo Horizonte, v. 14, n. 1, p. 131 -151, jan./abr. 2009.

LOEHRLEIN, Aaron, et. al. A hybrid approach to faceted classification based on analysis of descriptor suffixes. In: Grove, Andrew (Ed.). ANNUAL MEETING OF THE AMERICAN SOCIETY FOR INFORMATION SCIENCE AND TECHNOLOGY, 68., 2005, Charlotte, US. **Proceedings…** Charlotte, US: ASIST, 2005. v. 42, p. 1-25.

LOPES, Ilza L. Uso das linguagens controlada e natural em bases de dados: revisão da literatura. **Ci. Inf.**, Brasília, v. 31, n. 1, p. 41-52, jan./abr. 2002.

LORENA, Ana Carolina; CARVALHO, André C. P. L. F. Uma introdução às support vector machines. **RITA ,** v. XIV, n. 2, 2007. Disponível em: <http://www.seer.ufrgs.br/index.php/rita/article/viewFile/rita_v14_n2_p43-67/3543>. Acesso em: 21 jul. 2009.

MAMFRIM, Flávia P. B. Representação de conteúdo via indexação automática em textos integrais em língua portuguesa. **Ci. Inf.**, Brasília, v. 20, n. 2, p. 191-203, jul./dez. 1991.

MARCONI, M. D. A.; LAKATOS, E. M. **Técnicas de pesquisa:** planejamento e execução de pesquisas, amostragens e técnicas de pesquisas, elaboração, análise e interpretação de dados. 3.ed. São Paulo: Atlas, 1996.

MARON, M. E. On Indexing, retrieval and the meaning of about. **Journal of the American Society for Information Science**, n. 28, n. 1, p. 38-43, 1977.

MATTAR, F. N. **Pesquisa de Marketing**. São Paulo: Altas, 1996.

MOENS, Marie-Francine. **Automatic indexing and abstracting of document texts**. [S.l.]: Springers, 2000. 284p. (The Information Retrieval Series).

MOREIRO GONZÁLEZ, José Antonio. **El contenido de los documentos textuales**: su análisis y representación mediante el lenguaje natural. Gijón: Ediciones Trea, 2004.

NAVARRO, Sandrelei. Interface entre lingüística e indexação: uma revisão de literatura. **Rev. Bras. Biblio. Doc.**, São Paulo, v.21, n. 1/2, p. 46-62, jan./jun. 1988.

NAVES, Madalena M. L. **Curso de indexação**: princípios e técnicas de indexação, com vistas à recuperação da informação. Belo Horizonte: UFMG, Biblioteca Universitária, 2004. Material didático. 23p.

NAVES, Madalena M. L. Estudo de fatores interferentes no processo de análise de assunto. **Persp. Ci. inf.**, Belo Horizonte, v. 6, n. 2, p. 189 -203, jul./dez. 2001.

NEVES, Dulce A. B.; DIAS, Eduardo W.; PINHEIRO, Ângela Maria V. Uso de estratégias metacognitivas na leitura do indexador. **Ci. Inf.**, Brasília, v. 35, n. 3, p. 141-152, set./dez. 2006. Disponível em: <http://www.scielo.br/pdf/ci/v35n3/v35n3a14.pdf>. Acesso em: 21 jun. 2009.

O'CONNOR, J. Automatic subject recognition in scientific papers: an empirical study. **Journal of the Association for Computing Machinery,** n. 12, p. 490-515, 1965.

OLIVEIRA, Elias et al. Um modelo algébrico para representação, indexação e classificação automática de documentos digitais. **Rev. Bras. Biblio. Doc.**, Nova Série, São Paulo, v. 3, n. 1, p. 73-98, jan./jun. 2007.

OTHERO, Gabriel Á.; MENUZZI, Sérgio M. **Lingüística computacional:** teoria e prática. São Paulo: Parábola, 2005. 126p.

PACAK, M.; PRATT, A. W. The function of semantics in automated language processing. In: INTERNATIONAL ACM SIGIR CONFERENCE ON INFORMATION STORAGE AND RETRIEVAL, Session: Natural language processing and query systems, 1971, College Park, Maryland. **Proceedings…** New York, NY: University of Maryland, Apr. 1971. p. 5-18.

PALEGRINA, J. A. **Dicionário de informática**. [S.l.]: DicWeb, 2008. Disponível em: <http://www.dicweb.com/cc.htm>. Acesso em: 26 abr. 2009.

PÉREZ, Cláudia C. C.; GASPERIN, Caroline; VIEIRA, Renata. **Extração semi-automática de conhecimento a partir de textos.** São Leopoldo: PIPCA – UNISINOS/RS, 2003. 10p. Disponível em: <http://www.inf.unisinos.br/~renata/laboratorio/publicacoes/enia2003-submitted.pdf>. Acesso em: 18 ago. 2008.

PICKLER, Maria Elisa V. Web semântica: ontologias como ferramentas de representação do conhecimento. **Persp. Ci. Inf.**, Belo Horizonte, v. 12, n. 1, p. 65-83,

jan./abr. 2007. Disponível em: <http://www.scielo.br/scielo.php?script=sci_arttext&pid=S1413-99362007000100006&lng=en&nrm=iso&tlng=en>. Acesso em: 24 jul. 2007.

PRESLEY, M. What should comprehension instruction be the instruction of?. In: KAMIL, M. et al. (Ed.). **Handbook of reading research.** New Jersey: Lawrence Erlbaum, 2000.

RECTOR, Monica; YUNES, Eliana. **Manual de semântica.** Rio de Janeiro: Ao Livro Técnico, 1980. 171p.

ROBREDO, Jaime. A indexação automática de textos: o presente já entrou no futuro. In: Machado, U. D. (Org.). **Estudos Avançados em Ciência da Informação,** Brasília, DF: Associação dos Bibliotecários do Distrito Federal, 1982. v. 1, p. 235-274.

ROBREDO, Jaime. **Documentação de hoje e de amanhã**: uma abordagem revisitada e contemporânea da Ciência da Informação e de suas aplicações biblioteconômicas, documentárias, arquivísticas e museológicas. 4. ed. Brasília: Reproart, 2005. 409p.

ROBREDO, Jaime. Indexação e recuperação da informação na era das publicações virtuais. **Comunicação e Informação**, Goiânia, v. 2, n. 1, p. 83-97, jan./jun. 1999.

SALTON, Gerard. **Automatic information organization and retrieval***.* New York, NY: McGraw-Hill, 1968. 480p.

SALTON, Gerard. Automatic text analysis. **Science***,* v. 168, n. 3929, p. 335-343, 17 Apr. 1970.

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

SALTON, Gerard. **The SMART retrieval system-experiments in automatic document processing.** Upper Saddle River, NJ: Prentice-Hall, 1971.

SILVA, Antônio C. As teorias do signo e as significações lingüísticas. **Partes, a sua revista virtual**, São Paulo, ano 3, n. 39, nov. 2003. Disponível em: <http://www.partes.com.br/ed39/teoriasignosreflexaoed39.htm>. Acesso em: 10 jul. 2007.

SILVA, Maria R; FUJITA, Mariângela S. L. A prática da indexação: análise da evolução de tendências teóricas e metodológicas. **Transinformação**, Campinas, v. 16, n. 2, p. 133-161, maio/ago. 2004.

SIMÕES, Darcilia. **Estudos semânticos n. 1:** o que é semântica?. Rio de Janeiro: Instituto de Letras, UERJ, 2002. Disponível em: <www.darciliasimoes.pro.br/aulas/docs/estudos_semanticos01.doc>. Acesso em: 13 ago. 2008.

SOUZA, José Maria P..; BENÍCIO, Maria Helena D. Análise mutivariada: um exemplo usando modelo Log-linear. **Rev. Saúde públ.,** São Paulo, v. 19, p. 236-245. Disponível em: <http://www.scielo.br/pdf/rsp/v19n3/07.pdf>. Acesso em: 21 jul. 2009.

SOUZA, Renato R. Sistemas de recuperação de informações e mecanismos de busca na web: panorama atual e tendências. **Perspect. Ci. Inf.**, Belo Horizonte, v. 11 n. 2, p. 161-173, maio/ago. 2006.

SVENONIUS, Elaine. **The intellectual foundation of information organization.** Massachusetts: MIT Press, 2000. 264 p.

SVENONIUS, Elaine. Subject indexing: principles and practices in the 90's. In: IFLA SATELLITE MEETING HELD, New Series, 1993, LISBON, Portugal. **Proceedings...** Ubcim Publications, 1995, v. 15, 302 p**.**

SWANSON, D. R. Searching natural language text by computer. **Science**, v. 132, n. 3434, p. 1099-1104, 21 Oct. 1960.

TAXONOMIA. In: GLOSSÁRIO NETIC. [S.l.]: Portal NETIC - Núcleo de Estudos em Tecnologias para Informação e Conhecimento, [200-]. Disponível em: <http://www.netic.com.br/ glossario.html#T>. Acesso em: 24 jul. 2007.

NATIONAL LIBRARY OF MEDICINE (U.S.). The principles of MEDLARS. Washington, D.C.: Superintendent of Documents, 1970. 77p.

UNISIST. Princípios de indexação. Tradução de Maria Cristina M. F. Pinto. **Rev. Esc. Biblio.**, Belo Horizonte, v. 1, n. 10, p. 83-94, mar. 1981. Título original: Indexing principles.

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

WIVES, Leandro K. **Indexação de documentos textuais.** 1997. 19f. Trabalho Monográfico - Disciplina de Sistemas de Banco de Dados (Programa de Pós-Graduação em Ciência da Computação) - Instituto de Informática, Universidade Federal do Rio Grande do Sul, Porto Alegre, 1997. Orientadora: Prof. Lia Golendziner.

WIVES, Leandro K. **Tecnologias de descoberta de conhecimento em textos aplicadas à inteligência competitiva.** 2002. 116 f. Trabalho para Exame de Qualificação - (Programa de Pós-Graduação em Ciência da Computação) - Instituto de Informática, Universidade Federal do Rio Grande do Sul, Porto Alegre, 2002. Orientador: José Palazzo Moreira de Oliveira.

REFERÊNCIAS CONSULTADAS

MACULAN, B. C. M. S. **Manual de normalização:** padronização de documentos acadêmicos do NITEG/UFMG e do PPGCI/UFMG. Belo Horizonte: UFMG, 2008. 50p. Apostila. Disponível em: <http://www.eci.ufmg.br/normalizacao>. Acesso em: 10 maio 2009.

SILVA, Edna Lúcia; MENEZES, Estera M. **Metodologia da pesquisa e elaboração de dissertação.** 3. ed. Florianópolis: UFSC, 2001. Disponível em: <http://projetos.inf.ufsc.br/arquivos/Metodologia%20da%20Pesquisa%203a%20edicao.pdf>. Acesso em: 18 ago. 2008.

ANEXOS

Anexo A – Amostra de estudo n. 1

TOTAL DE DOCUMENTOS DA AMOSTRA n. 1: 103.

DÉCADA DE 1950 / TOTAL DE DOCUMENTOS: 2
•[1958] BAXENDALE, P. B. Machine-made index for technical literature: an experiment. **IBM Journal of Research and Development**, n. 2, p. 354-361, 1958. •[1959] MARON, M. E.; KUHNS, J. L.; RAY, L. C. Probabilistic indexing**:** a statistical approach to the library problem. In: NATIONAL MEETING OF THE ASSOCIATION FOR COMPUTING MACHINERY, 14., ACM, 1959, Cambridge, Massachusetts. **Proceedings…** New York, NY: ACM, 1959. p.1-2.
DÉCADA DE 1960 / TOTAL DE DOCUMENTOS: 5
•[1960] SWANSON, Don R. Searching natural language text by computer. **Science**, v. 132, n. 3434, p. 1099-1104, 1960. •[1965] O'CONNOR, J. Automatic subject recognition in scientific papers: an empirical study. **J. ACM***,* n. 12, p. 490-515, 1965. •[1968] SALTON, G. **Automatic information organization and retrieval**. New York, NY: McGraw-Hill, 1968. 480p. •[1969] EDMUNDSON, H. P. New methods in automatic extracting. **J. ACM***,* v. 16, n. 2, p. 264-285, Apr. 1969.
DÉCADA DE 1970 / TOTAL DE DOCUMENTOS: 8
•[1970] SALTON, G. Automatic text analysis. **Science***,* v. 168, n. 3929, p. 335-343, 17 Apr. 1970. •[1971] SALTON, G. **The SMART retrieval system-experiments in automatic document processing.** Upper Saddle River, NJ: Prentice-Hall, 1971. •[1973] SALTON**,** G. Recent studies in automatic text analysis and document retrieval**.**

Journal of the Association for Computing Machinery, v. 20, n. 2, p. 258-27, Apr. 1973.

•[1973] SCHANK, Roger C.; RIEGER III, Charles J. Inference and the computer understanding of natural language. **Artificial intelligence**, v. 5, p. 373-412, May 1973.

•[1977] BORKO, H. Toward a theory of indexing. **Information Processing and Management,** v. 13, p. 355-365, 1977.

•[1977] MARON, M. E. On Indexing, retrieval and the meaning of about. **Journal of the American Society for Information Science**, n. 28, n. 1, p. 38-43, 1977.

•[1977] MARTAU, Mariza C. et al. **Aplicação da microfilmagem e do índice kwic no tratamento de alguns materiais especiais**. In: CONGRESSO BRASILEIRO DE BIBLIOTECONOMIA E DOCUMENTAÇÃO, 9., 1977, Porto Alegre. **Anais...** Porto Alegre: CBBD, 1977. p. 206-217.

DÉCADA DE 1980 / TOTAL DE DOCUMENTOS: 10

•[1980] ROBREDO, Jaime; FERREIRA, José A. De P. Conceituação de um programa para indexação automática de textos. **R. Bibliotecon.**, Brasília, v. 8, n. 2, p. 254-265, jul./dez. 1980.

•[1982] CROFT, W. B; RUGGLES, L. The implementation of a document retrieval system. In: ANNUAL ACM CONFERENCE ON RESEARCH AND DEVELOPMENT IN INFORMATION RETRIEVAL**,** 1982, West Berlin, Germany**. Proceedings**... New York, NY: Springer-Verlag New York, 1982. p. 28 - 37 .

•[1982] 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, 1982. p. 3-18**.**

•[1983] SALTON, Gerard. **Introduction to moder information retrieval**. McGraw-Hill. 1983.

•[1983] STANFEL, Larry E. Applications of clustering to information system design. **Information Processing and Management**, v. 19, n. 1, p. 37-50, 1983.

•[1985] HALLER, Johan. **Indexação automática de textos**. **R. Bibliotecon**., Brasília, v. 13, n. 1, p. 27-32, jan./jun. 1985.

•[1988] BISWAS, Subal C.; SMITH, Fred. Term and cross reference structure in computerized deep structure indexing system. **Int. Classif**., n. 3, v. 15, p.139-144, 1988.

•[1988] SALTON, Gerard. Syntactic approaches to automatic book indexing. In: ANNUAL MEETING ON ASSOCIATION FOR COMPUTATIONAL LINGUISTICS, 26., 1988, Buffalo, NY. **Proceedings…** Morristown, NJ, USA: ACL, 1988. p. 204-210.

•[1988] VIEIRA, Simone B. Análise comparativa entre indexação automática e manual da literatura brasileira de ciência da informação. **R. Bibliotecon.**, Brasília, v. 16, n. 1, p. 83-94, 1988.

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

•[1989] SALTON, Gerard; SMITH, Maria. On the application of syntactic methodologies in automatic text analysis. In: BELKIN, N. J.; RIJSBERGEN, C.,J. Van (Eds.). ANNUAL INTERNATIONAL ACMSIGIR CONFERENCE ON RESEARCH AND DEVELOPMENT IN INFORMATION RETRIEVAL, 12., 1989, Cambridge, MA. **Proceedings…** New York, NY, v. 23, n. SI, Jun. 25-28, 1989. p. 137-150.

DÉCADA DE 1990 / TOTAL DE DOCUMENTOS: 32

•[1990] BELLUZZO, Regina Célia Baptista et al. Indexação automática de documentos com PRECIS *Software* no serviço de biblioteca e documentação da Faculdade de Odontologia de Bauru (USP). In: SIMPÓSIO LATINO-AMERICANO DE TERMINOLOGIA, 2., Brasília, 1990**. Anais...** Brasília: IBICT, 1990. p. 159-164.

•[1990] CRAVEN, Timothy C. Linked phrase indexing. **Information Processing and Management,** v. 14, p. 469-476, 1990.

•[1990] JONES, Kevin P. Natural-language processing and automatic indexing: a reply. **The Indexer**, v. 17, n. 2, p. 114-115, Oct. 1990.

•[1990] KORYCINSKI, C.; NEWELL, Alan F. Natural-language processing and automatic indexing. **The Indexer**, v. 17, n. 1, p. 21-29, Apr. 1990.

•[1990] MULVANY, Nancy C. *Software* tools for indexing: what we need. **The Indexer**, v. 17, n. 2, p. 108-113, Oct. 1990.

•[1990] SALTON, Gerard; BUCKLEY, Chris; SMITH, Maria. On the application of syntactic methodologies in automatic text analysis**. Information Processing and management,** v. 26, n. 1, p. 73-92, 1990.

•[1991] MAMFRIM, Flávia P. B. Representação de conteúdo via indexação automática em textos integrais em língua portuguesa. **Ci. Inf.** Brasília, v. 20, n. 2, p. 191-203, jul./dez. 1991.

•[1991] ROBREDO, Jaime. Indexação automática de textos: uma abordagem

otimizada e simples. **Ci. Inf.**, Brasília, v. 20, n. 2, p. 130-136, jul./dez. 1991.

•[1991] RUGE, Gerda; SCHWARTZ, Christoph. Term associations and computational linguistics. **Int. Classif.**, v. 18, n. 1, p. 19-25, 1991.

•[1991] SACCONI, L. A. **Nossa gramática**: teoria. São Paulo, Brasil: Atual. 1991.

•[1992] BRITTO, Marcílio. Sistemas de informação em linguagem natural: em busca de uma indexação automática. **Ci. Inf.**, Brasília, v. 21, n. 3, p. 223-232, set/dez 1992.

•[1992] MORRIS, Andrew H.; KASPER, George M.; ADAMS, Dennis A. The effects and limitations of automated text condensing on reading comprehension performance. **Information Systems Research**, v. 3, n. 1, p. 17-35, Mar. 1992.

•[1993] ALBRECHTSEN, H. Subject analysis and indexing: from automated indexing to domain analysis. **The Indexer**, v. 18, n. 4, p. 219-243, Oct.1993.

•[1993] LANCASTER, F. W. **Indexação e resumos**: teoria e prática. Brasília: Briquet de Lemos, 1993. 347p.

•[1994] GUEDES, Vânia L. S. Estudo de um critério para indexação automática derivativa de textos científicos e tecnológicos. **Ci. Inf**., Brasília, v. 23, n. 3, p. 318-326, set./dez. 1994.

•[1994] SALTON, Gerard et. al. Automatic analysis, theme generation, and summarization of machine-readable texts. **Science**, v. 264, p. 1421-1426, Jun. 1994.

•[1995] COHEN, Jonathan D. Highlights: language- and domain-independent automatic indexing terms for abstracting. **Journal Of The American Society For Information Science**, v. 46, n. 3, p. 162-174, 1995.

•[1995] GODIN, Robert; KERHERVÉ, Brigitte; TURNER, James. Classification and automatic indexing in a persistent object environment. In: THE 6th ASIS SIG/CR CLASSIFICATION RESEARCH WORKSHOP, 6., 1995, Chicago, Illinois. **Proceedings…** Chicago, Illinois: ASIS SIG/CR, 1995. p. 37-43.

•[1995] JORGENSEN, Corinne. Classifying images: criteria for grouping as revealed in a sorting task. In: THE 6th ASIS SIG/CR CLASSIFICATION RESEARCH WORKSHOP, 6., 1995, Medford, NJ. **Proceedings…** Medford, NJ: Information Today, 1995. p. 45-64.

•[1995] MCKEOWN, Kathleen; ROBIN, Jacques; KUKICH, Karen. Generating concise natural language summaries. **Information Processing and Management: an International Journal**, New York, v. 31, n. 5, p. 703-733, Sep. 1995.

•[1995] MOLINA, Maria Pinto. Documentary abstracting: toward a methodological

model. **Journal of the American Society for Information Science**, v. 46, n. 3, p. 225-234, 1995.

•[1995] VILES, Charles L; FRENCH, James C. Dissemination of collection wide information in a distributed information retrieval system. In: ANNUAL INTERNATIONAL ACM SIGIR CONFERENCE ON RESEARCH AND DEVELOPMENT IN INFORMATION RETRIEVAL**,** 18, 1995, Seattle, Washington. USA**. Proceedings...** New York, NY, USA: ACM, 1995. p. 12 - 20 .

•[1996] KURAMOTO, Helio. Uma abordagem alternativa para o tratamento e a recuperação de informação textual: os sintagmas nominais. **Ci. Inf.**, Brasília, v. 25, n. 2, p. 182-192, maio/ago. 1996.

•[1997] MOENS, Marie-Francine; UYTTENDAELE, Caroline; DUMORTIER, Jos. Abstracting of legal cases: the SALOMON experience. In: INTERNATIONAL CONFERENCE ON ARTIFICIAL INTELLIGENCE AND LAW, 6., 1997, Melbourne, Australia. New York, NY: 1997. p. 114-122.

•[1997] ROBREDO, Jaime. Indexação automática e infometria: um casamento que está dando certo. In: 18º CONGRESSO BRASILEIRO DE BIBLIOTECONOMIA E DOCUMENTAÇÃO, 18., 1997, São Luiz. **Anais...** São Luiz: CBBD, 1997. p. 27-31.

•[1997] WIVES, Leandro K. **Indexação de documentos textuais.** 1997. 19f. Trabalho Monográfico - Disciplina de Sistemas de Banco de Dados (Programa de Pós-Graduação em Ciência da Computação) - Instituto de Informática, Universidade Federal do Rio Grande do Sul, Porto Alegre, 1997. Orientadora: Prof. Lia Golendziner.

•[1998] CHEN, Hsinchun; ZHANG, Yin; HOUSTON, Andrea L. Semantic indexing and searching using a hopfield net. **Journal Of Information Science,** v. 24, n. 1, p. 3-18, 1998.

•[1998] HJORLAND, Birger. Information Retrieval, text composition and semantics. **Knowledge Organisation** v. 25, n. 1 e 2, p. 16-31, 1998.

•[1998] MAUER, Peg. Embedded indexing. [**Key Words**, Jan./Feb. 1998 and Sep./Oct. 1998].12p. Disponível em: <http://www.asindexing.org/site/Mauer_EmbeddedIndexing.pdf>. Acesso em: 22 ago. 2008.

•[1998] MOENS, Marie-Francine; DUMORTIER, Jos. Automatic abstracting of magazine articles: the creation of 'highlight' abstracts. In: ANNUAL ACM CONFERENCE ON RESEARCH AND DEVELOPMENT IN INFORMATION RETRIEVAL, 21., ACM SIGIR, 1998, Melbourne, Australia. **Proceedings…** New York, NY: ACM, 1998. p. 359-360.

•[1998] ROBREDO, Jaime; CUNHA, Murilo B. Aplicação de técnicas infométricas para identificar a abrangência do léxico básico que caracteriza os processos de indexação e

recuperação da informação. **Ci. Inf.**, Brasília, v. 27, n. 1, p. 11-27, jan./abr. 1998.

•[1999] GIL LEIVA, Isidoro. **La automatización de la indización de documentos**. Madrid: Ediciones Trea, 1999. 221p.

•[1999] MULVANY, Nancy C. *Software* tools for indexing: revisited. **The Indexer**, v. 21, n. 4, p. 160-163, Oct. 1999.

•[1999] ROBREDO, Jaime. Indexação e recuperação da informação na era das publicações virtuais. **Comunicação e Informação**, Goiânia, v. 2, n. 1, p. 83-97, jan./jun. 1999.

•[1999] SEBASTIANI, Fabrizio. A tutorial on automated text categorization. In: AMANDI, Analia; ZUNINO, Alejandro (Eds.). ARGENTINIAN SYMPOSIUM ON ARTIFICIAL INTELLIGENCE,1.,1999, Buenos Aires, AR. **Proceedings…** Buenos Aires: ASAI-99, 1999. p. 7-35.

DÉCADA DE 2000 / TOTAL DE DOCUMENTOS: 44

•[2000] ANDERSON, Charles. Indexing with a computer: past and present**. The Indexer**, v. 22, n. 1, p. 23-24, April 2000.

•[2000] MAUER, Peg. Embedded indexing: pros and cons for the indexer. **The Indexer**, v. 22, n. 1, p. 27-28, Apr. 2000.

•[2000] MOENS, Marie-Francine. **Automatic indexing and abstracting of document texts**. [S.l.]: Springers, 2000. 284p. (The Information Retrieval Series).

•[2000] ROSS, Jan. The impact of technology on indexing. **The Indexer**, v. 22, n. 1, p. 25-26, Apr. 2000.

•[2000] SALTON, Gerard. Automatic text indexing using complex identifiers. In: CONFERENCE ON DOCUMENT PROCESSING SYSTEMS, DocProcess, 2000, Santa Fe, New Mexico, United States. **Proceedings…** New York, NY: ACM, 2000. p. 135-144.

•[2001] CAROPRESO, Maria Fernanda; MATWIN, Stan; SEBASTIANI, Fabrizio. A learner-independent evaluation of the usefulness of statistical phrases for automated text categorization. In: **Text atabases and document management:** theory and practice. Hershey, PA, USA: IGI Publishing, 2001. p. 78-102. Disponível em: <http://nmis.isti.cnr.it/sebastiani/Publications/TD01a.pdf>. Acesso em: 20 set. 2008.

•[2001] MOENS, Marie-Francine, BUSSER, Rik de. Generic topic segmentation of document texts. In: ANNUAL ACM CONFERENCE ON RESEARCH AND DEVELOPMENT IN INFORMATION RETRIEVAL, 24., Sep. 9-12, 2001, New Orleans,

Louisiana, USA. **Proceedings...** New York, NY: ACM SIGIR, 2001. p. 418-419.

•[2001] WIVES, Leandro K. et al. **Aplicando métodos de descoberta de conhecimento em textos em documentos sobre a mortalidade pública**. Porto Alegre: PGCC, Instituto de Informática da UFRGS, 2001. 11p. (Material de Aula).

•[2002] DE BUSSER, Rik; ANGHELUTA, Roxana; MOENS, Marie-Francine. Semantic case role detection for information extraction. In: INTERNATIONAL CONFERENCE ON COMPUTATIONAL LINGUISTICS, 19., Taipei, Taiwan, 2002. **Proceedings…** Morristown, NJ, USA: Association for Computational Linguistics (ACL), 2002. p. 1-5.

•[2002] GIORGETTI, Daniela; PRODANOF, Irina; SEBASTIANI, Fabrizio. **Mapping an automated survey coding task into a probabilistic text categorization framework.** Berlin, Heidelberg: Springer, 2002. v. 2389, p. 369-386.

•[2002] JOACHIMS, Thorsten; SEBASTIANI, Fabrizio. Guest editors' introduction to the special issue on automated text categorization. **Journal of Intelligent Information Systems**, v. 18, n. 1-3, p. 103-105, Mar./May 2002.

•[2002] PEIXOTO, Maria D. F. et al. **Categorização de textos**. Armazenamento e Pesquisa de Informação, Universidade da Beira Interior, Departamento de Informática, Covilhã, Portugal, [2002]. 19p. Disponível em: <http://www.di.ubi.pt/~api/text_categorization.pdf>. Acesso em: 19 ago. 2008.

•[2002] SEBASTIANI, Fabrizio**.** Machine learning in automated text categorization. **ACM Computing Surveys**, v. 34, n. 1, p. 1–47, Mar. 2002.

•[2002] WIVES, Leandro Krug. **Tecnologias de descoberta de conhecimento em textos aplicadas à inteligência competitiva.** 2002. 116 f. Trabalho para Exame de Qualificação - (Programa de Pós-Graduação em Ciência da Computação) - Instituto de Informática, Universidade Federal do Rio Grande do Sul, Porto Alegre, 2002. Orientador: José Palazzo Moreira de Oliveira.

•[2003] DEBOLE, Franca; SEBASTIANI, Fabrizio. Supervised term weighting for automated text categorization. In: SYMPOSIUM ON APPLIED COMPUTING, Information access and retrieval, Melbourne, Florida, 2003. **Proceedings…** New York, NY: ACM, 2003. p. 784-788.

•[2003] GALHO, Thaís Silva; MORAES, Silvia Maria Wanderley. Categorização automática de documentos de texto utilizando lógica difusa. In: XV SALÃO E XIII FEIRA DE INICIAÇÃO CIENTÍFICA DA UFRGS, 15. e 13., 2003, Porto Alegre. **Anais...** Porto Alegre: UFRGS, 2003. p. 93-94.

•[2003] GIORGETTI, Daniela; SEBASTIANI, Fabrizio. Automating survey coding by multiclass text categorization techniques. **Journal of the American Society for Information Science and Technology**, New York, NY, v. 54, n. 14, p. 1269-1277, Dec. 2003.

•[2003] MOENS, Marie-Francine; ANGHELUTA, Roxana. Concept extraction from legal cases: the use of a statistic of coincidence. In: INTERNATIONAL CONFERENCE ON

ARTIFICIAL INTELLIGENCE AND LAW, Short paper session, 9., 2003, Scotland, United Kingdom. **Proceedings…** New York, NY: ACM, 2003. p. 142-146.

•[2003] NARDIELLO, Pio; SEBASTIANI, Fabrizio; SPERDUTI, Alessandro. Discretizing continuous attributes in adaboost for text categorization. In: EUROPEAN CONFERENCE ON INFORMATION RETRIEVAL, ECIR-03, 25., 2003. **Proceedings…** Berlin, Heidelberg: Springer, 2003. v. 2633 (Lecture Notes in Computer Science).

•[2003] PÉREZ, Cláudia C. C.; GASPERIN, Caroline; VIEIRA, Renata. **Extração semi-automática de conhecimento a partir de textos.** São Leopoldo: PIPCA – UNISINOS/RS, 2003. 10p. Disponível em: <http://www.inf.unisinos.br/~renata/laboratorio/publicacoes/enia2003-submitted.pdf>. Acesso em: 18 ago. 2008.

•[2004] HONORATO, Daniel de F. et al. Utilização da indexação automática para auxílio à construção de uma base de dados para a extração de conhecimento aplicada à doenças pépticas. In: I WORKSHOP DE COMPUTAÇÃO, 1., 2004, Palhoça. **Anais…** Palhoça: WORKCOMP-SUL, 2004. p. 1-9.

•[2004] LANCASTER, F. W. **Indexação e resumos**: teoria e prática. Brasília: Briquet de Lemos, 2004. 452p.

•[2004] PÉREZ, Cláudia C. C., VIEIRA, Renata. Aquisição de conhecimento a partir de textos para construção de mapas conceituais. In: II WORKSHOP DE TESES E DISSERTAÇÕES EM INTELIGÊNCIA ARTIFICIAL, 2., 2004, São Leopoldo. **Anais...** São Leopoldo, RS: PIPCA – UNISINOS/RS, 2004. 10p.

•[2004] SILVA, Cassiana F.; VIEIRA, Renata; OSÓRIO, Fernando S. Uso de informações linguísticas em categorização de textos utilizando redes neurais artificiais. In: VIII SIMPÓSIO BRASILEIRO DE REDE NEURAIS, 8., 2004, São Luís. **Anais...** São Luís: SBRN , 2004. v. 1, p. 1-6.

•[2004] WIVES, Leandro K. **Utilizando *conceitos* como descritores de textos para o processo de identificação de conglomerados (*clustering*) de documentos.** 2004. 136f. Tese (Doutorado em Ciência da Computação) – Instituto de Informática, Universidade Federal do Rio Grande Do Sul. Porto Alegre: 2004.

•[2005] ESULI, Andrea; SEBASTIANI, Fabrizio. Determining the semantic orientation of terms through gloss classification. In: CONFERENCE ON INFORMATION AND KNOWLEDGE MANAGEMENT, Paper session IR-8 (information retrieval): sentiment and genre classification, 14., Bremen, Germany, 2005. **Proceedings…** New York, NY: ACM, 2005. p. 617-624.

•[2005] LAPUENTE María J. L. Indización Automática. In: **Hipertexto:** el nuevo concepto de documento en la cultura de la imagen. 2005. 184f. Tesis (Tesis Doctoral) – Universidad Complutense de Madrid. Madrid: UCM, 2005. Disponível em: <http://www.hipertexto.info/documentos/indiz_automat.htm>. Acesso em: 20 ago.

2008.

•[2005] MOENS, Marie-Francine. Combining structured and unstructured information in a retrieval model for accessing legislation. In: INTERNATIONAL CONFERENCE ON ARTIFICIAL INTELLIGENCE AND LAW, Legal knowledge bases, legislation, 10., June 2005, Bologna, Italy. **Proceedings…** New York: ACM, 2005. p. 141-145.

•[2005] SEBASTIANI, Fabrizio. Text categorization. In: RIVERO, Laura C.; DOORN, Jorge H.; FERRAGGINE, Viviana E. (eds.). **Encyclopedia of database technologies and applications.** Hershey, US: Idea Group Publishing, 2005. p. 683-687.

•[2005] SOUZA, Renato R. **Uma proposta de metodologia para indexação automática utilizando sintagmas nominais**. 2005. 215f. Tese (Doutorado em Ciência da Informação) – Escola de Ciência da Informação da Universidade de Federal de Minas Gerais. Belo Horizonte, 2005.

•[2006] AVANCINI, Henri et al. Automatic expansion of domain-specific lexicons by term categorization. **ACM Transactions on Speech and Language Processing**, v. 3, n. 1, p. 1-30, May 2006.

•[2006] ESULI, Andrea; FAGNI, Tiziano; SEBASTIANI, Fabrizio. **MP-Boost:** a multiple-pivot boosting algorithm and its application to text categorization**.** Berlin / Heidelberg: Springer, 2006. v. 4209, p. 1-12. (Lecture Notes in Computer Science).

•[2006] LOEHRLEIN, Aaron et al. A hybrid approach to faceted classification based on analysis of descriptor suffixes. In: PROCEEDINGS of the American Society for Information Science and Technology, v. 42, n. 1, Oct. 2006. Disponível em: <http://elvis.slis.indiana.edu/upload/CSKD/204.pdf>. Acesso em: 23 ago. 2008.

•[2006] SEBASTIANI, F. Classification of text, automatic. In: BROWN, Keith, (Ed.). 2nd ed. **Encyclopedia of Language and Linguistics**, Oxford: Elsevier, 2006. v. 2, p. 457-462.

•[2006] SOUZA, Renato R. Sistemas de recuperação de informações e mecanismos de busca na web: panorama atual e tendências. **Perspect. Ciênc. Inf.**, Belo Horizonte, v. 11 n. 2, p. 161-173, Maio/Ago. 2006.

•[2006] SOUZA, Renato R. Uma proposta de metodologia para indexação automática utilizando sintagmas nominais. **Enc. Bibli: R. Eletr. Bibliotecon. Ci. Inf.**, Florianópolis, n. esp., 1º sem. 2006.

•[2007] ARANHA, Christian N. **Uma abordagem de pré-processamento automático para mineração de textos em português:** sob o enfoque da inteligência computacional. 2007. 144 f. Tese (Doutorado em Engenharia Elétrica) – Pontifícia Universidade Católica do Rio de Janeiro, Rio de Janeiro, 2007.

•[2007] AUTO JUNIOR, Tavares da C. **Indexação automática de acórdãos por meio de processamento de linguagem natural.** 2007. 142f. Dissertação (Mestrado em

Ciência da Informação) – Faculdade de Economia, Administração, Contabilidade e Ciência da Informação e Documentação, Universidade de Brasília. Brasília: Faculdade FAAP, 2007.

•[2007] ESULI, Andrea; SEBASTIANI, Fabrizio. Random-walk models of term semantics: an application to opinion-related properties. In: LANGUAGE AND TECHNOLOGY CONFERENCE, 3., Poznan, PL, 5-7 Oct. 2007. **Proceedings...** Poznan, PL: LTC 2007. 5p.

•[2007] FREDDY, Angel; VIERA, Godoy; VIRGIL, Johnny. Uma revisão dos algoritmos de radicalização em língua portuguesa. **Information Research**, v. 12, n. 3, p. 26-26, Apr. 2007.

•[2007] MOENS, Marie-Francine et al. Automatic detection of arguments in legal texts. In: INTERNATIONAL CONFERENCE ON ARTIFICIAL INTELLIGENCE AND LAW, NLP and text mining, 11., 2007, Stanford, California. **Proceedings…** New York, NY: ACM, 2007. p. 225-230.

•[2007] OLIVEIRA, Elias et al. Um modelo algébrico para representação, indexação e classificação automática de documentos digitais. **Rev. Bras. Biblio. Doc.**, Nova Série, São Paulo, v. 3, n. 1, p. 73-98, jan./jun. 2007.

•[2007] SHLOMO, Argamon et al. Automatically determining attitude type and force for sentiment analysis. In: PROCEEDINGS OF THE 3RD LANGUAGE AND TECHNOLOGY CONFERENCE (LTC'2007*)*, 3., Poznan, PL, **Proceedings...** Poznan, PL: LTC, Oct. 2007. p. 369-373. Disponível em: <http://nmis.isti.cnr.it/sebastiani/Publications/Publications.html#2005>. Acesso em: 22 set. 2008.

•[2008] FIGUEIREDO, Fabio S. **Construção de evidências para classificação automática de textos.** 2008. 73f. Dissertação (Mestrado em Ciência da Computação) – Instituto de Ciência Exatas da Universidade de Federal de Minas Gerais. Belo Horizonte: Departamento de Ciência da Computação, 2008.

Anexo B – Amostra de estudo n. 2

TOTAL DE DOCUMENTOS DA AMOSTRA N. 2: 12

DÉCADA DE 1950 / TOTAL DE DOCUMENTOS: 2
TEXTO N. 1: [1958] BAXENDALE, P. B. Machine-made index for technical literature: an experiment. **IBM Journal of Research and Development**, n. 2, p. 354-361, 1958. **TEXTO N. 2:** [1959] MARON, M. E.; KUHNS, J. L.; RAY, L. C. Probabilistic indexing**:** a statistical approach to the library problem. In: NATIONAL MEETING OF THE ASSOCIATION FOR COMPUTING MACHINERY, 14., ACM, 1959, Cambridge, Massachusetts. **Proceedings…** New York, NY: ACM, 1959. p.1-2.
DÉCADA DE 1960 / TOTAL DE DOCUMENTOS: 2
TEXTO N. 3: [1960] SWANSON, Don R. Searching natural language text by computer. **Science**, v. 132, n. 3434, p. 1099-1104, 1960. **TEXTO N. 6:** [1969] EDMUNDSON, H. P. New methods in automatic extracting. **J. ACM***,* v. 16, n. 2, p. 264-285, Apr. 1969.
DÉCADA DE 1970 / TOTAL DE DOCUMENTOS: 2
TEXTO N. 7: [1970] SALTON, Gerard. Automatic text analysis. **Science***,* v. 168, n. 3929, p. 335-343, 17 Apr. 1970. **TEXTO N. 9:** [1973] SALTON, Gerard. Recent studies in automatic text analysis and document retrieval. **Journal of the Association for Computing Machinery**, v. 20, n. 2, p. 258-278, April 1973.
DÉCADA DE 1980 / TOTAL DE DOCUMENTOS: 2
TEXTO N. 16: [1982] 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, 1982. p. 3-18**.** **TEXTO N. 24:** [1989] SALTON, Gerard; SMITH, Maria. On the application of syntactic methodologies in automatic text analysis. In: BELKIN, N. J.; RIJSBERGEN, C.,J. Van

(Eds.). ANNUAL INTERNATIONAL ACMSIGIR CONFERENCE ON RESEARCH AND DEVELOPMENT IN INFORMATION RETRIEVAL, 12., 1989, Cambridge, MA. **Proceedings…** New York, NY, v. 23, n. SI, Jun. 25-28, 1989. p. 137-150.

DÉCADA DE 1990 / TOTAL DE DOCUMENTOS: 2

TEXTO N. 54: [1998] MOENS, Marie-Francine; DUMORTIER, Jos. Automatic abstracting of magazine articles: the creation of 'highlight' abstracts. In: ANNUAL ACM CONFERENCE ON RESEARCH AND DEVELOPMENT IN INFORMATION RETRIEVAL, 21., ACM SIGIR, 1998, Melbourne, Australia. **Proceedings…** New York, NY: ACM, 1998. p. 359-360.

TEXTO N. 55: [1998] ROBREDO, Jaime; CUNHA, Murilo Bastos da. Aplicação de técnicas infométricas para identificar a abrangência do léxico básico que caracteriza os processos de indexação e recuperação da informação. **Ci. Inf.**, Brasília, v. 27, n. 1, p. 11-27, jan./abr. 1998.

DÉCADA DE 2000 / TOTAL DE DOCUMENTOS: 2

TEXTO N. 80: [2004] HONORATO, Daniel F. et al. Utilização da indexação automática para auxílio à construção de uma base de dados para a extração de conhecimento aplicada à doenças pépticas. In: I WORKSHOP DE COMPUTAÇÃO, 1., 2004, Palhoça. **Anais…** Palhoça: WORKCOMP-SUL, 2004. p. 1-9.

TEXTO N. 101: [2007] OLIVEIRA, Elias et al. Um modelo algébrico para representação, indexação e classificação automática de documentos digitais. **Rev. Bras. Biblio. Doc.**, Nova Série, São Paulo, v. 3, n. 1, p. 73-98, jan./jun. 2007.

Livros Grátis

(http://www.livrosgratis.com.br)

Milhares de Livros para Download:

Baixar livros de Literatura
Baixar livros de Literatura de Cordel
Baixar livros de Literatura Infantil
Baixar livros de Matemática
Baixar livros de Medicina
Baixar livros de Medicina Veterinária
Baixar livros de Meio Ambiente
Baixar livros de Meteorologia
Baixar Monografias e TCC
Baixar livros Multidisciplinar
Baixar livros de Música
Baixar livros de Psicologia
Baixar livros de Química
Baixar livros de Saúde Coletiva
Baixar livros de Serviço Social
Baixar livros de Sociologia
Baixar livros de Teologia
Baixar livros de Trabalho
Baixar livros de Turismo