Desenvolvimento de Softwares de Indexação Automática: breve Avaliação dos Principais Critérios

Graciane Silva Bruzinga Borges
Escola de Ciência da Informação/Universidade Federal de Minas Gerais, Email: gracianebruzinga@gmail.com
Gercina Ângela de Lima
Escola de Ciência da Informação/Universidade Federal de Minas Gerais, Email: limagercina@gmail.com

Resumo

Este estudo apresenta um resultado de pesquisa sobre critérios utilizados na construção de softwares para indexação automática. O objetivo principal foi realizar um mapeamento panorâmico, a partir da análise de literatura da área, desde a década de 1950 até o ano de 2008, para verificar quais critérios foram apontados pelos autores como relevantes para o desenvolvimento dos *softwares*. Como suporte teórico e metodológico analisou-se: a semântica e a sintaxe; a Linguística computacional e o Tratamento de documentos textuais para fins de recuperação da informação. Para tal, utilizou-se do procedimento metodológico de Análise de Conteúdo, identificando os critérios de indexação automática desenvolvidos e utilizados no período através de relato de experiência dos próprios pesquisadores autores. Priorizaram-se aqueles que têm como preocupação central o tratamento das questões semânticas do documento textual. Como resultados finais, obteve-se o levantamento dos principais critérios e a proposição de possíveis combinações entre eles, visando auxiliar aos profissionais na primeira etapa do processo de indexação, que trata da extração de termos relevantes para representação de assuntos. Tornou-se possível, desta forma, a utilização dos critérios que estavam dispersos na literatura através de relatos de experiências e que nem sempre são divulgadas nas áreas de interseção com a Ciência da Informação - CI, tais como a Linguística e a Ciência da Computação. Entre os objetivos alcançados, encontram-se: (1) listagem dos critérios encontrados na literatura; (2) caracterização de cada critério e (3) listagem dos critérios mais recorrentes. Obteve-se um conjunto de critérios ideais para o desenvolvimento de *softwares* de extração automática.

Palavras-chave: Indexação Automática. Critérios de Indexação Automática. *Software* de Indexação automática. Representação da informação.

Introdução

O trabalho apresenta uma análise da literatura sobre indexação automática desde a década de 1950 a 2008, com o intuito de mapear os principais critérios para a construção de *softwares* desta natureza, observando como se deu a evolução da área. O estudo é proveniente da pesquisa de mestrado da primeira autora, defendida na Escola da Ciência da Informação da Universidade Federal de Minas Gerais (BORGES, 2009).

O processo de indexação corresponde à atividade de representar um documento através de uma descrição abreviada de seu conteúdo, com o intuito de sinalizar sua essência. Essa representação é feita a partir da análise de assunto do texto-fonte, que preferencialmente, deveria ser feita por especialistas da área, que tivessem um olhar atento para as metodologias e procedimentos provenientes da Ciência da Informação e da Biblioteconomia. Na prática, o resultado do processo de indexação deverá apresentar termos ou expressões significativas que irão possibilitar o acesso ao documento original, ou

seja, irão possibilitar a sua recuperação em uma base de dados ou em Sistema de Recuperação da Informação - SRI.

No âmbito das tecnologias para representação da informação, a indexação automática veio como alternativa para resolver os problemas da indexação manual, também denominada como indexação intelectual, desde a década 1950 com os estudos de H. P. Luhn. Embora a indexação automática possa não apresentar resultados totalmente satisfatórios, suas soluções podem contribuir para significativas melhoras no processo de indexação manual. Soluções estas que almejam realizar automaticamente a extração inicial de termos (palavras ou expressões) do documento indexado, deixando para o profissional o trabalho de selecionar aqueles mais adequados para representar seu conteúdo. Além disso, a técnica permite a redução da subjetividade, característica inerente à realização intelectual da atividade.

Indexação Automática, um olhar histórico

Também chamada de *indexação assistida por computador* e de *indexação semi-automática*, a indexação automática é considerada uma técnica 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 citado por 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 (MOREIRO GONZÁLEZ, 2004, p.3 citado por BUFREM, 2005).

De acordo com os trabalhos de Luhn (1957), e de Baxendale (1958), registra-se o início das pesquisas sobre indexação automática baseada em frequência de ocorrência de palavras no texto. Baxendale (1958 citado por 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.

Ainda no decorre da década de 1950, desenvolveram-se métodos relativamente simples para a construção de índices a partir de textos, especialmente utilizando as palavras que ocorrem nos títulos dos documentos. O *Keyword in Context* – KWIC (Palavra-chave no Contexto) 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. O programa reconhece as palavras que não são palavras-chaves, baseando-se em uma lista de palavras proibidas, e impede que elas sejam adotadas na entrada. O *Keyword out of Context* – KWOC (Palavra-chave fora do Contexto) é um método 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.

Além do KWIC e do KWOC, podemos citar o *Selective Listing in Combination* – SLIC (Listagem Seletiva em Combinação), criado por J. R. Sharp, em 1966, que organiza a sequência de termos de um documento em ordem alfabética e elimina as sequências redundantes, e o método *Preserved Context Indexing System* – PRECIS, criado pelo Dr. Derek Austin, em 1968, e que produz o índice impresso baseado na ordem alfabética e na alteração sistemática de termos para que ocupem a posição de entrada (LANCASTER, 2004). Outro importante sistema desenvolvido foi o *Nested Phrase Indexing System* – NEPHIS (Sistema de Indexação de Frase Encaixada), criado por T. C. Craven, em 1977, e corresponde a um índice articulado de assunto.

De acordo com Garvin (1969 citado por SALTON, 1973) e Salton (1973), já na década de 1960, percebia-se a intrínseca relação entre processamento da informação e aspectos linguísticos. Os esforços deviam ser voltados para estudos das propriedades estruturais e semânticas das línguas naturais. Contudo, percebe-se que grande parte das metodologias linguísticas da época geralmente produzia resultados decepcionantes.

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. A partir da década de 1970, percebe-se uma intensificação das pesquisas na área, destacando-se dois dos importantes experimentos do período: (1) desempenho do SRI MEDlars, que operava no National Library of Medicine, em Washington, e (2) SRI experimental SMART, criado por Gerard Salton enquanto trabalhava na universidade de Cornell (SALTON, 1973).

Quanto aos tipos de indexação automática conhecidos destaca-se a *indexação por extração automática*. Nesse processo, palavras ou expressões que aparecem no texto são extraídas para representar seu conteúdo como um todo. Os princípios utilizados tentam copiar os que seriam usados por indexadores humanos (LANCASTER, 2004).

Os sistemas baseados em indexação por extração automática realizam, basicamente, as seguintes tarefas: (1) contar palavras num texto; (2) cotejá-las com uma lista de palavras proibidas; (3) eliminar palavras não significativas (artigos, preposições, conjunções, etc.) e (4) ordenar as palavras de acordo com sua freqüência. Percebe-se que esse tipo de indexação apresenta limitações. Semelhante a esse processo, porém com uma preocupação quanto aos aspectos semânticos do texto, pode-se indicar a *indexação por atribuição automática* (O'CONNOR, 1965 citado por LANCASTER, 2004).

Para Jaime 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.

Outro tipo de indexação automática destacada é a *identificação automática de palavras full text,* 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 (idem),* que objetiva a análise das palavras mais relevantes da oração. Há, ainda, a *indexação automática semântica (idem),* que se baseia no princípio

de que o documento já possui estruturas de formatação para a indicação da semântica dos termos. Arrisca-se a dizer que para a obtenção de um tratamento automático adequado é necessário o desenvolvimento de algoritmos que considerem a semântica e a sintaxe do conteúdo desses documentos.

Metodologia de Levantamento e Descrição dos Critérios de Indexação Automática

Utilizou-se o método de estudo caracterizado por Análise de Conteúdo que foi implementado a partir da divisão do trabalho em duas etapas principais, conforme segue.

A Etapa I da pesquisa, denominada *Identificação dos Critérios de Indexação Automática*, foi subdividida em dois estágios:

A) Definição do universo de pesquisa e da amostra de estudo nº 1: o universo de estudo do trabalho foi caracterizado por artigos técnico-científicos sobre indexação automática que apresentavam resultados de pesquisa ou de experimentos. Os documentos deveriam conter, necessariamente, metodologia científica e apontamento de resultados conclusivos quanto à pertinência dos critérios de indexação automática utilizados. Deste universo, fez-se o recorte de uma amostra composta por 103 (cento e três) pesquisas nacionais e internacionais sobre o assunto publicadas entre a década de 1950 e o ano de 2008. A partir desta análise, foi possível a realização dos procedimentos descritos no estágio a seguir.

B) Definição do objeto empírico e sistematização dos critérios: os textos da amostra foram disponibilizados na versão impressa permitindo a manipulação física dos documentos 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 um *Guia de observação nº1* como instrumento de pesquisa, que direcionou o estudo para identificação dos seguintes aspectos em cada um dos textos da primeira amostra: (1) Nome do critério conforme definido pelo autor; (2) Objetivo do critério; (3) Descrição do critério; (4) Fontes de identificação; (5) Análise do critério mediante elaboração de parágrafo síntese indicando aplicações, vantagens e desvantagens observadas pelo autor.

Já a Etapa II da pesquisa, denominada *Análise das combinações dos critérios*, foi também subdividida em dois estágios, foram eles:

A) Seleção da amostra de estudo nº 2: constituiu-se de um recorte de 12 (doze) textos a partir da amostra de estudo nº 1, obtendo-se, assim, os seguintes documentos (QUADRO 1). Optou-se pela definição de uma amostragem do tipo não-probabilística – subjetiva, que não tem base estatística, sendo definida por critérios decorrentes da experiência profissional e do conhecimento da área em exame, sendo usual que corresponda a 10% ou 15% da população alvo (MARCONI; LAKATOS, 1996; LAKATOS, 1991; MATTAR, 1996).

QUADRO 1

Amostra de estudo nº 2

Década de 1950	(BAXENDALE, 1958) & (MARON; KUHNS; RAY, 1959).
Década de 1960	(SWANSON, 1960) & (EDMUNDSON, 1969).

Década de 1970	(SALTON, 1970) & (SALTON, 1973).
Década de 1980	(ROBREDO, 1982b) & (SALTON; SMITH, 1989).
Década de 1990	(MOENS; DUMORTIER, 1998) & (ROBREDO; CUNHA, 1998).
Década de 2000	(HONORATO et al., 2004) & (OLIVEIRA, 2007).

Fonte: desenvolvido pelas autoras.

B) Interpretação dos critérios: verificação da utilização prática dos critérios identificados na primeira etapa observando suas respectivas ocorrências e combinações na amostra de estudo nº 2. Para tal, fez-se uso do Guia de observação nº2, registrando através deste os seguintes aspectos presentes na amostra em questão, foram eles: (1) Título da pesquisa; (2) Objetivos da pesquisa; (3) Nome do pesquisador autor responsável; (4) Período de realização do trabalho; (5) Localidade de realização/aplicação da pesquisa; (6) Listagem dos critérios utilizados; (7) *Software* utilizado e/ou desenvolvidos na pesquisa; (8) Comparação com indexação manual; (9) Registros de métodos comparativos entre os critérios mencionados; (10) Tipo de documento analisado; (11) Área do conhecimento em que se contextualiza o trabalho; (12) Identificação dos resultados como satisfatórios ou insatisfatórios de acordo com análise dos próprios autores pesquisadores e (13) Numeração do texto conforme amostra de estudo nº 1.

Desta forma tendo procedido, obteve-se, satisfatoriamente, os resultados almejados que serão apresentados na sequência.

Critérios de Indexação Automática: apresentação e análise dos resultados

Em decorrência da aplicação minuciosa do método já descrito anteriormente, passa-se à apresentação dos resultados alcançados.

Resultados da Etapa I da pesquisa

A) Listagem de dezesseis critérios identificados na literatura:

- ✓ *CRITÉRIO 1* - Formatação de frases-termo (Word phrase formation);
- ✓ *CRITÉRIO 2* - Fórmula de transição de Goffman;
- ✓ *CRITÉRIO 3* - Freqüência absoluta de ocorrência da palavra no texto;
- ✓ *CRITÉRIO 4* - Freqüência de co-ocorrência relativa de termos;
- ✓ *CRITÉRIO 5* - Freqüência de co-ocorrência simples de termos;
- ✓ *CRITÉRIO 6* - Freqüência relativa 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 (Stop-list/stop-words);
- ✓ *CRITÉRIO 10* - Palavras destacadas no texto;
- ✓ *CRITÉRIO 11* - Peso numérico;
- ✓ CRITÉRIO 12 - Posição do termo no texto (Term weighting);
- ✓ *CRITÉRIO 13* - Primeira lei de Zipf;
- ✓ *CRITÉRIO 14* - Segunda lei de Zipf ou Lei de Zipf-Booth;
- ✓ *CRITÉRIO 15* - Tópico frasal;

- ✓ *CRITÉRIO 16* - Vocabulário semântico / Cabeçalhos conceituais / Tesauro.

B) Sistematização dos 16 critérios identificados na literatura. 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 essenciais para apoio no processo de escolha dos melhores critérios para o desenvolvimento de *softwares* de indexação automática. Assim, de acordo com detalhamento apresentado no Guia de observação nº 1, seguem os dados obtidos (QUADROS 2 a 17).

QUADRO 2

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

Objetivo do critério:	Formatação de frases-termo a partir da união de palavras adjacentes.
Descrição:	O critério pretende formar novos termos, buscando solucionar o problema dos termos abrangentes, pois as ideias estão agrupadas em contextos, e palavras compostas, geralmente, categorizam melhor os assuntos, tornando-os mais específicos.
Fontes de identificação:	(CROFT; RUGGLES, 1982; SALTON, 1983; WIVES, 1997, p. 8).

Fonte: desenvolvido pelas autoras.

A utilização de palavras mais específicas permite que o sistema recupere documentos de forma mais precisa, devido ao fato destas palavras aparecerem em menor quantidade no documento como um todo. Documentos de contextos específicos utilizam termos igualmente específicos. Em uma consulta que pretenda recuperar documentos que tratem de *programas computacionais,* por exemplo, além da consulta utilizando-se da composição "programa computacional", recomenda-se a utilização da frase-termo "programa de computador". (CROFT; RUGGLES, 1982), (SALTON, 1983), (WIVES, 1997, p. 8).

Deve-se tomar o cuidado para não confundir o conceito de *frase-termo* com a utilização de 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*.

QUADRO 3

Formula de transição de Goffman

Objetivo do critério:	Identificar as palavras representativas do conteúdo do documento em um ponto específico do texto.
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. O pesquisador sugeriu um ponto do texto onde haveria a transição das palavras de alta frequência para as palavras de baixa frequência. A fórmula é bibliométrica: $T = \frac{-1 + \sqrt{1 + 8\ I1}}{2}$ Onde: **I1-** número de palavras que ocorrem uma única vez; **8** - constante atribuída à língua inglesa; **2** - constante matemática da fórmula de Baskara, para resolução de equação de 2º grau.
Fontes de identificação:	(LANCASTER, 1993, p. 287-288).

Fonte: desenvolvido pelas autoras.

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 frequência aproximada. Assim, a partir desta frequência são contidas as palavras entre o ponto T e a palavra de maior frequência. Este mesmo número de palavras é projetado para baixo do Ponto T, definindo uma região (LANCASTER, 1993). Embora baseado exclusivamente em uma análise estatística, o critério expande a análise puramente baseada na frequência das palavras dispersas por todo o texto para uma análise onde se identifica uma região potencial para verificação de termos representativos do documento.

QUADRO 4

Frequência absoluta de ocorrência de termos

Objetivo do critério:	Ordenar as palavras de acordo com sua frequência de ocorrência no texto.
Descrição:	Palavras no topo da lista são candidatas mais fortes para representarem o conteúdo. São considerados: o número absoluto de palavras, a extensão do texto e a frequência acima de determinado limiar.
Fontes de identificação:	(LANCASTER, 1993).

Fonte: desenvolvido pelas autoras.

O critério considera apenas o próprio documento indexado, havendo dificuldade para se definir o ponto de corte da lista gerada. Mesmo depois de se utilizar listas de palavras sem significado aparente (*stop-list)*, algumas palavras podem ocorrer

frequentemente no texto e, ainda assim, não serem bons descritores do mesmo, devido ao fato também ocorrerem com alta frequência na base de dados como um todo.

QUADRO 5

Frequência de co-ocorrência relativa de termos

Objetivo do critério:	Identificar termos relacionados nos documentos indexados.
Descrição:	Considera-se o total de vezes que os termos ocorrem no texto e na base como um todo, a fim de recuperar textos que tratem de assuntos semelhantes. Se os termos A e B co-ocorram 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. Supondo que A ocorre 50 vezes, e B ocorra 25 vezes, e ambos co-ocorram 20 vezes, o fator de associação será forte, pois é 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 equação: $$R = \frac{a\ e\ b}{a\ ou\ b}$$ Quando R excede um limiar preestabelecido, os dois termos são aceitos como relacionados.
Fontes de identificação:	(LANCASTER, 1993, p.294).

Fonte: desenvolvido pelas autoras.

Não se calcula o grau de associação entre dois termos com base na frequência simples, mas na frequência de co-ocorrência relativa à frequência de ocorrência de cada termo no documento. Há dificuldade para definição do ponto de corte da lista e para análise dos termos representativos, o critério considera não apenas o documento, mas a base de dados como um todo.

QUADRO 6

Frequência de co-ocorrência simples de termos

Objetivo do critério:	Identificar termos relacionados nos documentos indexados.
Descrição:	A fim de recuperar textos que tratem de assuntos semelhantes, considera-se que quanto mais frequentemente

	dois termos ocorrem juntos, maior a probabilidade deles tratem de assunto similar. 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.
Fontes de identificação:	(LANCASTER, 1993, p.294).

Fonte: desenvolvido pelas autoras.

O critério considera apenas o documento para análise da ocorrência dos termos, e não a base de dados na qual o documento está armazenado, sendo capaz de identificar associações diretas (X e Y tendem a ocorrer juntos) e associações indiretas entre termos. Supondo 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 co-ocorram nos documentos, é possível supor que há uma relação entre D e T (provavelmente são sinônimos) (LANCASTER, 1993).

QUADRO 7

Frequência relativa de ocorrência de termos

Objetivo do critério:	Selecionar palavras ou expressões que ocorram num documento com mais frequência do que sua taxa de ocorrência na base de dados com um todo.
Descrição:	A frequência com que uma palavra ocorre na base de dados como um todo é ainda mais importante que a frequência 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.
Fontes de identificação:	(LANCASTER, 1993, p. 287-288).

Fonte: desenvolvido pelas autoras.

O critério ordena as palavras de acordo com sua frequência de ocorrência no documento indexado e também na base de dados como um todo. Há possibilidade de haver documentos em que o assunto principal seja também um assunto que ocorre

sistematicamente na base de dados, 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, o mesmo ocorre ao longo de todo o texto (introdução, desenvolvimento, metodologia e conclusão). Uma lista de termos extraídos usando-se o critério de *frequência relativa* será diferente de uma lista de termos onde se usou a *frequência absoluta*, porém não de forma radical. Provavelmente, desaparecerão os termos que ocorrem com muita frequência num documento e também na base de dados.

QUADRO 8

Identificação de palavras (Comparação com uso de dicionário)

Objetivo do critério:	Identificar as palavras nos documentos a partir da análise de sequências de caracteres no texto.
Descrição:	Salton (1973) aconselha fazer um *dictionary lookup*, ou seja, comparar as sequências de caracteres retiradas do texto com um dicionário a fim de validar se estas palavras realmente existem.
Fontes de identificação:	(WIVES, 1997, p. 6-7).

Fonte: desenvolvido pelas autoras.

Processo de validação bastante útil, especialmente quando o documento apresenta muitos caracteres inválidos ou palavras com erros gramaticais. As sequências 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 sequências 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. Um simples analisador léxico que identifique sequências de caracteres e monte palavras pode ser utilizado (WIVES, 1997). Contudo, há possibilidade de o dicionário não contemplar um termo relevante e este não ser analisado e/ou corrigido pelo critério.

QUADRO 9

Identificação de radicais de palavras *(Word stemming)*

Objetivo do critério:	Reduzir variações de uma mesma palavra a uma representação única, em tese: isolar o semantema das palavras dos seus morfemas, assim como na linguística.
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 *radica*l. Todas as palavras que possuem o mesmo radical e, portanto, significados similares são reconhecidas pelo mesmo identificador, facilitando a consulta.
Fontes de identificação:	(FREDDY; VIERA; VIRGIL, 2007; SACCONI, 1991; WIVES, 1997, p. 8).

Fonte: desenvolvido pelas autoras.

Uma maneira de identificar os radicais das palavras é remover seus sufixos e prefixos, assim como eliminar seus plurais. A desvantagem deste método é que ele pode acabar utilizando palavras muito abrangentes, não recuperando documentos específicos. 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 radicais 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.*

QUADRO 10

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

Objetivo do critério:	Impedir que palavras que aparecem intensamente em todos os documentos da base de dados sejam indexadas.
Descrição:	Consiste na listagem de 'todas' as palavras que não devem ser indexadas por não serem representativas aos conteúdos dos documentos. A esta estrutura foi atribuído o nome de s*top-list*, e as palavras presentes nesta lista são conhecidas como *stop-words*. É possível também a eliminação de preposições, que são termos utilizados para se fazer o encadeamento de ideias inerentes à linguagem, e não ao conteúdo dos documentos.
Fontes de identificação:	(WIVES, 1997, p. 7).

Fonte: desenvolvido pelas autoras.

O tempo gasto para elaboração de uma *stop-list* consistente é relativamente grande, além desta dificuldade, 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. Contudo, considera-se ainda que como o uso de uma *stop-list* torna-se possível a eliminação de

palavras proibidas, como artigos, preposições, conjunções, etc., sendo que essa eliminação reduz consideravelmente o tempo de processamento do restante do texto.

QUADRO 11

Palavras destacadas no texto

Objetivo do critério:	Identificar palavras ou expressões destacadas no texto como fortes candidatas a serem representativas do conteúdo
Descrição:	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.
Fontes de identificação:	(LANCASTER, 1993).

Fonte: desenvolvido pelas autoras.

O destaque de palavras no texto é feito pelo próprio autor com a intenção de enfatizar determinado aspecto do seu conteúdo, o que aumenta a probabilidade de se encontrar fortes candidatas para a representação do documento. Há, contudo, a possibilidade de extração de termos que foram destacados com um enfoque negativo.

QUADRO 12

Peso numérico *(Term weighting)*

Objetivo do critério:	Atribuir *pesos* ou *graus* de relação entre uma palavra e os documentos em que ela ocorre.
Descrição:	Consiste em identificar a frequência de determinada palavra em um documento (*term requency*) e o número de documentos em que esta palavra ocorre (*inverse document frequency*). 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 também por peso. A partir daí, é possível atribuir um valor de relação entre esta palavra e o documento através da fórmula: $$Peso\ td = \frac{Freq\ td}{DocFreq\ t}$$ Onde: ***Peso td*** = grau de relação entre o termo *t* e o documento *d*; ***Freq td*** = número de vezes que o termo *t* aparece no documento *d*; ***DocFreq t*** = número de documentos que o termo *t* aparece.
Fontes de identificação:	(LANCASTER, 1993; SALTON, 1983; VILES; FRENCH, 1995; WIVES, 1997).

Fonte: desenvolvido pelas autoras.

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. Para cada termo do documento é necessário calcular a sua relação utilizando-se a fórmula mencionada e 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. Cada documento possui um vetor com pares de elementos na forma {(palavra1, peso1), (palavra2, peso2), ... , (palavra n, peso n)}. Caso uma palavra não exista em um documento, seu valor de frequência é zero (0). Ao final, os pesos são somados, e os documentos, listados por ordem decrescente de pesos. 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 (LANCASTER, 1993; SALTON, 1983; VILES; FRENCH, 1995; WIVES, 1997).

QUADRO 13

Posição do termo no texto

Objetivo do critério:	Analisar partes específicas do texto, diminuindo o tempo gasto com o processamento.
Descrição:	Consiste na análise de apenas partes do documento consideradas relevantes. 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.
Fontes de	(LANCASTER, 1993).

identificação:	

Fonte: desenvolvido pelas autoras.

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.

QUADRO 14
Primeira lei de Zipf

Objetivo do critério:	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 frequência. De acordo com a lei, se as palavras de um texto suficientemente longo forem colocadas em ordem decrescente de frequência, poder-se-á verificar que a ordem de série das palavras (R) multiplicada por sua frequência (F) produz uma constante (K): R x F = K
Fontes de identificação:	(MAMFRIM, 1991).

Fonte: desenvolvido pelas autoras.

O critério não considera aspectos semânticos para análise do documento, embora o critério apresente limitações e, principalmente, por ser de natureza exclusivamente estatística, é a base para outros critérios que pretendem analisar o texto de maneira contextualizada.

QUADRO 15
Segunda lei de Zipf ou Lei de Zipf-Booth

Objetivo do critério:	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 frequência: $\frac{In}{I1} = \frac{2}{n \times (n+1)}$ 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.
Fontes de identificação:	(MAMFRIM, 1991).

Fonte: desenvolvido pelas autoras.

Assim como a *Primeira lei de Zipf, o* critério não considera aspectos semânticos para análise do documento e, embora seja de natureza exclusivamente estatística, é a base para outros critérios que pretendem analisar o texto de maneira contextualizada.

QUADRO 16

Tópico frasal (Palavras sugestivas)

Objetivo do critério:	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. 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.
Fontes de identificação:	(BAXENDALE, 1958; LANCASTER, 1993).

Fonte: desenvolvido pelas autoras.

Embora o tempo de análise do texto por meio dos *softwares* de indexação automática sofra considerável redução, há a possibilidade de que partes diferentes daquelas definidas como tópicos frasais sejam representativas.

QUADRO 17

Vocabulário semântico / Vocabulário de cabeçalhos conceituais / Tesauro

Objetivo do critério:	Cotejar os termos que ocorrem nos títulos dos documentos com um vocabulário semântico.
Descrição:	Formado por termos de uma área específica, os quais são ligados a um vocabulário de cabeçalhos conceituais, é possível cotejar termos que aparecem nos títulos de artigos com um vocabulário

	semântico, os quais devem ser ligados a um vocabulário de cabeçalhos conceituais.
Fontes de identificação:	(LANCASTER, 1993; VLEDUTS-STOKOLOV, 1987, citado por LANCASTER, 2004, p.291).

Fonte: desenvolvido pelas autoras.

Desse modo, os cabeçalhos conceituais podem ser atribuídos pelo computador com base em palavras/expressões que ocorrem nos títulos dos documentos. *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. O critério é capaz de atribuir 61% dos cabeçalhos que um ser humano, possivelmente, atribuiria.

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.

Resultados da Etapa II da pesquisa

Quantificação dos critérios de indexação automática *mais* utilizados nas pesquisas correspondentes à amostra nº1. Do total de dezesseis critérios selecionados, 50% destes apresentou uma taxa de utilização acima de 30% em relação ao número total de pesquisas analisadas. Esses critérios são apresentados na Tabela 1.

Tabela 1 – Relação dos critérios mais na amostra nº 1

Nome do critério	Pesquisas em que se utilizou o critério	%
Identificação de palavras (Comparação com uso de dicionário)	9	75,00%
Formatação de frases-termo (*Word phrase formation*)	8	66,67%
Posição do termo no texto (*Term weighting)*	5	41,66%
Peso numérico	5	41,66%
Identificação de radicais de palavras *(Word stemming)*	5	41,66%
Frequência absoluta de ocorrência da palavra no texto	5	41,66%
Vocabulário semântico/vocabulário de cabeçalhos conceituais/Tesauro	4	33,33%
Lista de palavras proibidas/Palavras proibidas *(Stop-list / stop-words)*	4	33,33%

Fonte: desenvolvido pelas autoras.

Julga-se que o critério nº 3, *freqüência absoluta de ocorrência da palavra no texto,* seja relevante para análise de documentos textuais. Embora esse seja um critério que, usualmente, é visto como limitado, por considerar apenas o número de vezes que cada palavra ocorre no texto, ele mostrou um índice considerável de utilização ao longo do

período pesquisado. Este critério apresenta relação direta com três outros critérios: *Freqüência de co-ocorrência relativa de termos*; *Freqüência de co-ocorrência simples de termos* e *Freqüência relativa de ocorrência da palavra no texto*.

De fato, a frequência de ocorrência relativa e a frequência de co-ocorrência, simples e relativa, são critérios mais robustos que a frequência 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 frequência de ocorrência absoluta de uma palavra em um texto passou a ser utilizado em conjunto com outros critérios que consideram aspectos linguísticos, como é o caso do critério nº 7, *identificação de palavras (comparação com uso de dicionários)* e o critério nº 16, *vocabulário semântico / vocabulário de cabeçalhos conceituais / tesauro*.

Sobre o critério nº 16, *vocabulário semântico / vocabulário de cabeçalhos conceituais / tesauro,* percebe-se que esse critério vigora entre os mais usados, apresentando grande potencial para o tratamento semânticos do texto.

Diferentemente do que era esperado, o critério nº 9, *lista de palavras proibidas / palavras proibidas (stop-list / stop-words),* que foi um dos primeiros desenvolvidos na área, obteve baixo índice de utilização na amostra. Contudo, considera-se a possibilidade de omissão por parte dos autores quanto à utilização do mesmo, uma vez que sua importância seja, até certo ponto, consensual no campo.

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)* e o critério de nº 8, *identificação de radicais de palavras (word stemming)* 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. 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 *software*. Atualmente, esses dois critérios são considerados de grande relevância, visto que a verificação da estrutura gramatical é a base para a realização de análises semânticas.

Finalmente, os dois últimos critérios, *peso numérico* e *posição do termo no texto (term weighting)* podem ser associados. Ambos apresentam aspectos de atribuição de grau de importância para determinas palavras do texto. A ideia do 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, como é o caso do título do texto, de seu resumo e de sua conclusão.

Quantificação dos critérios de indexação automática *menos* utilizados na amostra nº1. Os 50% restante de critérios que apresentaram uma taxa de utilização inferior a 30% em relação ao número total de pesquisas analisadas estão apresentados na Tabela 2.

Tabela 2 – Relação dos critérios menos utilizados na amostra nº 1

Nome do critério	Quantidade de pesquisas que utilizou o	%

	critério	
Tópico frasal	1	8,33%
Palavras destacadas no texto	1	8,33%
Fórmula de transição de Goffman	1	8,33%
Segunda lei de Zipf ou Lei de Zipf-Booth	2	16,66%
Primeira lei de Zipf	2	16,66%
Freqüência relativa de ocorrência da palavra no texto	2	16,66%
Freqüência de co-ocorrência simples de termos	2	16,66%
Freqüência de co-ocorrência relativa de termos	3	25,00%

Fonte: desenvolvido pelas autoras.

Da análise da Tabela 2, fazem-se alguns comentários. Três dos critérios apresentados, o critério nº 2, *fórmula de transição de Goffman*, e os critérios nº 13 e 14, *primeira e segunda lei de Zipf ou lei de Zipf-Booth*, podem ser relacionados entre si devido ao fato de terem como base a análise estatística das palavras do texto. Percebe-se que esses critérios 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 linguístico, 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*. Essa consideração, para análise do *software,* embora possa apresentar algum resultado satisfatório, não é consistente o suficiente.

Por último, analisamos o critério nº 15, *tópico frasal.* Embora tendo sido considerado apenas por uma das doze pesquisas da amostra, se trata de dos critérios precursores da área e deve ser considerado em trabalhos específicos.

Desta forma, a partir da análise minuciosa dos critérios observados ao longo do estudo, propõe-se um conjunto de 9 (nove) critérios entendidos como ideais para o desenvolvimento de *software* de indexação automática para o tratamento de documentos textuais. Acredita-se que esse conjunto poderá proporcionar uma extração de termos significativos dos documentos indexados, obtendo um resultado semelhante, ou próximo, àquele que seria obtido através do trabalho realizado pelo ser humano:

1. Formatação de frases-termo (*Word phrase formation*);
2. Freqüência absoluta de ocorrência da palavra no texto;
3. Identificação de palavras (Comparação com uso de dicionário);
4. Identificação de radicais de palavras *(Word stemming);*
5. Lista de palavras proibidas / Palavras proibidas *(Stop-list / stop-words);*
6. Peso numérico;
7. Posição do termo no texto (*Term weighting);*
8. Vocabulário semântico / vocabulário de cabeçalhos conceituais / Tesauro.

Considerações Finais

Acredita-se 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. Atualmente, há grande produção bibliográfica e na busca por informação o usuário exige respostas rápidas e relevantes. Constatou-se deficiências e percalços no processo manual de indexação, o que reafirmou a necessidade de estudos que busquem encontrar alternativas para esse atividade. Destaca-se a importância de se considerar aspectos semânticos do texto para que a indexação seja realizada de maneira contextualizada e consistente. Acredita-se que a utilização de vocabulários controlados, embutidos nos *softwares*, pode melhorar o processo.

Recentemente, os autores Lapa e Correa (2014) apresentaram relevante panorama dos estudos sobre a indexação automática no Brasil, no período de 1973 a 2012. Os autores afirmaram que há uma tendência em estudos sobre a indexação automática por meio dos sintagmas nominais e que, com o uso de novas tecnologias, procura-se desenvolver uma identificação automática dos termos por meio da atribuição.

Desta forma, considera-se que este trabalho trouxe resultados positivos para as áreas de Ciência da Informação e Ciência da Computação, pois pode auxiliar aos profissionais envolvidos no desenvolvimento de *softwares* para a indexação automática para o tratamento de documentos textuais.

Development of Automatic Indexing Software: brief evaluation of main criteria

Abstract

This study presents a result of a research about the criteria used in an automatic indexing to build up software. The main objective was to go through a wide map starting from the literature review analysis of the area since 1950 until 2008 to verify which elements were pointed out by the authors as relevant to the softwares development. As a theoretical and methodological support there were analyzed: the semantic and syntax elements; the Computational Linguistic and the processing of textual documents directed to information retrieval. In order to make it possible the Content Analysis methodological procedure was established identifying the automatic indexing criteria developed and used during the period through the experience report of the own authors and researchers. The speeches that have as the central concern the treatment of the textual document semantic matters were prioritized. As final results, the main criteria were pointed out and the proposition of possible combinations among them, in order to give support to the professionals on the first step of the indexing process, that refers to the extraction of relevant terms of the subject representation. Consequently, it was possible the use of the criteria that were spread in the literature, through the experience report that are not always launched in the intersection areas of the Information Science – IC, such as Linguistics and Computer Science. Among the goals that have been reached, there are: (1) the criteria's list found in the literature; (2) the criteria characterization; (3) a list of the more repetitive criteria that were found. As a result, there was formed a group of ideal criteria to the development of the automatic extraction software.

Keywords: *Automatic Indexing. Criteria of Automatic Indexing. Automatic Indexing Software. Representation of information.*

REFERÊNCIAS

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

BORGES, G. S. Bruzinga. **Indexação automática de documentos textuais**: critérios essenciais. 2009. 111 f. Dissertação (Mestrado em Ciência da Informação)- Escola de Ciência da Informação, Universidade Federal de Minas Gerais, Belo Horizonte, 2009.

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 .

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

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.

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.

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.

LAKATOS, Eva Maria. MARCONI, M. de A. **Fundamentos de metodologia científica.** 3. ed. rev. e aum. São Paulo: Atlas, 1991.

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.

LAPA, Remi; CORREA, Renato. Indexação automática no âmbito da Ciência da Informação no Brasil. **Informação & Tecnologia (ITEC***),* Marília/João Pessoa, n. 1, v. 2, p. 59-76, jul./dez., 2014.

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.; 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.

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

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.

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

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.

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; 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.

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

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

SALTON, Gerard. Introduction to moder information retrieval. McGraw-Hill. 1983.

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; 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.

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

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 .

WIVES, Leandro K. **Indexação de documentos textuais.** 1997. 19 f. 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.