XI Encontro Nacional de Pesquisa em Ciência da Informação
Inovação e inclusão social: questões contemporâneas da informação
Rio de Janeiro, 25 a 28 de outubro de 2010

GT 2: Organização e Representação do Conhecimento
Modalidade de apresentação: comunicação oral

INDEXAÇÃO AUTOMÁTICA DE DOCUMENTOS TEXTUAIS: PROPOSTA DE CRITÉRIOS ESSENCIAIS

Graciane Bruzinga Borges
Universidade Federal de Minas Gerais

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. Para a análise dos dados, utilizou-se o método *analítico-sintético*, baseado em Dahlberg (1978), e constituído neste trabalho de duas principais etapas: (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. Etapa (2), proposição de 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 e (b) análise da combinação dos critérios utilizados em cada texto e interpretação dos respectivos resultados. Entre os objetivos alcançados, encontram-se: (1) listagem dos critérios encontrados na literatura, (2) caracterização de cada critério, (3) listagem dos critérios mais recorrentes.

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

INTRODUÇÃO

Este artigo apresenta os resultados da pesquisa de mestrado intitulada “Indexação Automática de Documentos Textuais: proposta de critérios essenciais”, de mesma autoria, defendida na Escola de Ciência da Informação da UFMG pelo Programa de Pós-graduação em Ciência da Informação – PPGCI em agosto de 2009. O objetivo do trabalho foi 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. Além disso, desejava-se (1) fornecer parâmetros comparativos para a melhoria da indexação automática de documentos textuais; (2) melhorar o processo da indexação automática através de critérios relevantes para a extração de termos representativos do conteúdo do documento; (3) auxiliar os profissionais da ciência da computação e de áreas afins no desenvolvimento de *softwares* de indexação automática e (4) contribuir com os profissionais e pesquisadores da Biblioteconomia e Ciência da Informação em suas rotinas de trabalho, como indexação de documentos textuais e realização de pesquisas bibliográficas.

Indexar é a 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 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.

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.

O tempo limitado que o indexador possui para realizar a análise de assunto manualmente pode interferir na qualidade da indexação. Para isso, propõem-se as

técnicas de indexação automática. Embora algumas delas não sejam totalmente satisfatórias, elas podem contribuir para 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 é a redução da subjetividade, característica inerente à realização intelectual da atividade.

A literatura da área aponta outros problemas práticos da indexação manual, tais como: (1) diferentes indexadores atribuem diferentes termos a um mesmo documento; (2) 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.

INDEXAÇÃO MANUAL

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. 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. 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.

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. Neste trabalho, considerou-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.

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, podendo comprometer a qualidade da atividade. Segundo Lancaster (1993, p. 20-21), "ao indexador raramente é dado o luxo de poder ler um documento do começo ao fim".

Para a execução desta 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).

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 e (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.

O final do último 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.

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. 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. 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.

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.

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 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)".

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. 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 seqüência de termos de um documento em ordem alfabética e elimina as seqüências 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. 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. 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.

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 de indexação automática de documentos textuais. 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).

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. Os princípios utilizados tentam copiar os que seriam usados por indexadores humanos (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: (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 para a realização do processo de maneira consistente. Semelhante a esse processo, porém com uma preocupação quanto aos aspectos semânticos do texto, pode-se indicar a *indexação por atribuição automática.* Em geral, este processo apresenta dificuldades, pois, para a representação do conteúdo temático, é necessário um controle terminológico, desenvolvendo, para cada termo atribuído, um "perfil" de palavras ou expressões que costumam ocorrer nos documentos (O'CONNOR, 1965 *apud* LANCASTER, 2004).

Outro tipo de indexação automática apontada na literatura 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,* que objetiva a análise das palavras mais relevantes da oração. 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.

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 metodologia apresentada tenta considerar os critérios de indexação de documentos textuais baseada em seu contexto.

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

Metodologia

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.

Identificação dos critérios – Etapa 1

Definição do universo e seleção da amostra de estudo nº 1

O universo de estudo deste trabalho 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.

A amostra de estudo foi composta de 103 (cento e três) pesquisas nacionais e internacionais sobre o assunto publicadas entre a década de 1950 e o ano de 2008. A pesquisa bibliográfica foi realizada de acordo com a seguinte estratégia para seleção dos documentos: (1) Delimitação do objetivo principal da pesquisa e de sua finalidade; (2) Indicação das palavras-chave que utilizadas para delimitação do assunto nos idiomas inglês e português; (3) Determinação dos tipos de documentos que fariam parte da amostra; (4) Seleção das fontes de informação para pesquisa bibliográfica; (5) Delimitação da quantidade da amostra; (6) Indicação do formato e do suporte dos documentos selecionados; (7) Determinação a estratégia de busca; (8) Definição das partes dos documentos a serem consideradas para leitura técnica e (9) Análise da amostra selecionada. A partir da análise desta amostra, foi possível a realização dos procedimentos descritos no estágio a seguir.

Definição do objeto empírico e sistematização dos critérios

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 (QUADRO 1).

QUADRO 1 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.

A partir do guia de observação nº 1 foi possível a execução de duas atividades: (1) Salientar nos textos da amostra nº 1 os aspectos indicados no Quadro 1 e (2) Elaborar, para cada critério, um quadro em que foram expostos os aspectos indicados no guia de observação nº 1. A partir desse procedimento, foram obtidos dois resultados:

Resultado 1: lista de 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 dos 16 critérios com o preenchimento do guia de observação nº 1 para cada um deles.

Análise das combinações dos critérios – Etapa 2

Seleção da amostra de estudo nº 2

O primeiro estágio desta segunda etapa do trabalho 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. 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, que não tem base estatística, sendo definida por critérios pessoais decorrentes da experiência profissional e do conhecimento do setor em exame, sendo usual que corresponda a 10% ou 15% da população alvo (MARCONI e LAKATOS, 1996; LAKATOS, 1991; MATTAR, 1996).

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 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 2.

QUADRO 2 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, abrangeram-se somente os elementos suficientes para apoio no processo de escolha dos melhores critérios para alcance dos objetivos deste trabalho. Assim, cada um dos doze textos selecionados na amostra nº 2 foi sistematizado de acordo com o guia de observação nº 2. As referências dos textos selecionados estão apresentadas no Quadro 3. A numeração precedida de cada texto é referente à sua posição sequencial dentro da amostra nº 1.

QUADRO 3 Amostra de estudo nº 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.
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.
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.
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.
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.
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.

Fonte: desenvolvido pela autora.

A partir dos dados obtidos, foi composta a TAB 1, que apresenta a relação das pesquisas com os critérios por elas utilizados. Os dados para composição desta tabela estavam contidos nos Quadros construídos para os doze textos selecionados na amostra nº 2 e serão assim apresentados para uma melhor visualização da quantidade de textos que utilizou cada critério identificado.

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.

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.

Levando em consideração que a maior parte das pesquisas aponta resultados satisfatórios, o fator mais relevante para a conclusão foi 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 deste trabalho. 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.

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 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.

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 *software*. 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 últimos 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.

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 *software* de indexação automática para o tratamento de documentos textuais e acredita-se que esse conjunto poderá proporcionar uma extração de termos significativos dos documentos indexados, obtendo um resultado semelhante àquele que seria obtido através do trabalho realizado pelo ser humano:

- Formatação de frases-termo (*Word phrase formation*)
- Freqüência absoluta 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)*
- Peso numérico
- Posição do termo no texto (*Term weighting)*
- Vocabulário semântico / vocabulário de cabeçalhos conceituais / Tesauro

CONSIDERAÇÕES FINAIS

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 atualmente 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.

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 *sofwares* 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.

ABSTRACT: 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. 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. 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. For data analysis, we used the analytic-synthetic method, based on Dahlberg (1978), and in this work consists of two main steps: (1), identified the criteria for automatic indexing literature. This step consisted of two stages: (a) selection of texts on automatic indexing, (b) reading of the texts and identifying the criteria for automatic indexing found in them. Step (2), proposing a set of ideal criteria for the automatic indexing process. The stages of this phase were: (a) selecting a sample of texts used in step 1 and (b) analysis of the combination of the criteria used in each text and interpretation of their results. 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.

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

REFERÊNCIAS:

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

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

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.

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).

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

LAKATOS, Eva Maria. **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.

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.

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

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.

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.

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.

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.

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.

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

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.