Indexação Automática no Âmbito da Ciência da Informação no Brasil

Remi Lapa
Universidade Federal de Pernambuco (UFPE), Brasil. E-mail: rmcrlp@gmail.com

Renato Correa
Universidade Federal de Pernambuco (UFPE), Brasil. E-mail: fc_renato@yahoo.com.br

Resumo

Apresenta um panorama dos estudos sobre a Indexação Automática por meio do mapeamento e análise da produção acadêmica e científica nacional da área de ciência da informação no período de 1973 a 2012. Como objetivos específicos, procura coletar um *corpus* de análise e caracterizar as publicações quanto aos objetivos e aspectos metodológicos. A metodologia consiste em um estudo bibliográfico aprofundado de caráter qualitativo e quantitativo, bem como análise de conteúdo da produção literária no Brasil a respeito da indexação automática de textos escritos no idioma português, dentre livros, artigos de periódicos científicos, anais publicados e literaturas cinzentas. Os principais resultados encontrados foram: 35% dos trabalhos realizaram revisão bibliográfica, enquanto 65% investigaram a indexação automática por meio de fórmula, método ou sistema, dos quais 23% realizaram proposição e aplicação, 20% a proposição, e 22% realizaram aplicação; os sistemas como o objeto de estudo mais pesquisado, e a comparação com a indexação manual como o método de avaliação mais usado; o texto completo como a natureza do *corpus* mais pesquisado; o trabalho científico como a tipologia do *corpus* mais estudada; a indexação semi-automática como procedimento mais aplicado na validação dos termos; o processo de atribuição como o meio mais adotado para identificar os termos; o texto não estruturado como a entrada de dados preferida nos sistemas; a linguagem natural como a natureza da linguagem, os termos compostos como a natureza dos termos mais pesquisados; a análise estatística como o método de seleção dos termos mais utilizado. Concluímos 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.

Palavras-chave: Indexação Automática. Recuperação da Informação. Sistemas de Recuperação da Informação. Ciência da Informação. Brasil.

1 Introdução

Neste artigo, abordamos a Indexação Automática como um processo circunscrito ao campo da Ciência da Informação (CI), que investiga a geração, coleta, organização, interpretação, armazenamento, recuperação, disseminação, transformação e uso da informação, com ênfase particular, na aplicação de tecnologias modernas nestas atividades (CAPURRO; HJØRLAND, 2007).

De acordo com Moraes (2002), os problemas relacionados com a recuperação da informação tornaram-se o foco de interesse para Hans Peter Luhn, especialista da *International Business Machines* (IBM) e pioneiro na aplicação da análise estatística de vocabulário para executar uma indexação automática. Luhn procurou soluções práticas e de baixo custo, o que o levou a utilização de máquinas para resolvê-los, tornando-se um defensor da Indexação Automática (PALMQUIST, 1998, tradução nossa).

Com a intenção de reverter os efeitos negativos causados pela produção de grandes volumes de informações, e almejando obter a informação confiável, de fácil acesso, com um tempo de resposta

reduzido e com um custo acessível, o tratamento físico e de conteúdo dos documentos assumem um papel fundamental, pois, analisam, traduzem e representam a forma e o assunto dos documentos com a finalidade de auxiliar na recuperação de informações (ALVES; CAFÉ, 2010).

Nesse sentido, segundo Fujita (2009, p. 22) "estudos vem sendo desenvolvidos acerca da teoria da indexação, sua natureza, procedimentos, estruturas e características de seu produto final, o índice", visando melhor tratamento temático e recuperação da informação.

Segundo Robredo (2005), existe uma preocupação em oferecer um acesso mais rápido à literatura técnico-científica utilizando o computador no processamento de dados e informações. Sua aplicação advém da necessidade em indexar grandes volumes de informações, em um tempo curto para manter as bases de dados atualizadas, o que torna inviável pensar na indexação manual (humana ou intelectual) como única forma de analisar e codificar o conteúdo dos documentos (ROBREDO, 2005).

A **problemática** subjacente à este artigo está em descrever e analisar a produção científica sobre a indexação automática no Brasil entre os anos 1973 e 2012.

Destarte, esta pesquisa tem por **objetivo geral** apresentar o panorama da pesquisa no âmbito da CI no Brasil referentes aos estudos sobre a Indexação Automática no período 1973 – 2012. Para tanto este artigo possui como **objetivos específicos**: levantar um *corpus* contendo a produção brasileira da área da Ciência da Informação a respeito da indexação automática entre os anos de 1973 a 2012; e analisar os objetivos e aspectos metodológicos das publicações que compõem o *corpus* selecionado.

A **justificativa** para a realização de tal pesquisa está no valor da informação obtida através da análise do conjunto de documentos selecionados, permitindo, deste modo, distinguir as tendências e realizar projeções sobre futuras pesquisas.

2 Indexação Automática no Brasil

A indexação automática pode ser definida como um conjunto de operações, basicamente matemáticas, linguísticas, de programação, destinadas a selecionar termos como elementos descritivos de um documento pelo processamento de seu conteúdo.

Na indexação automática por extração o processamento do conteúdo não é permeado pela interpretação de terceiros, pois os termos significativos são extraídos do texto e ordenados pela sua frequência de ocorrência (NASCIMENTO, 2008).

Outro tipo, é a indexação automática por atribuição, que segundo Lancaster (2004), consiste numa representação temática por meio de termos selecionados de um vocabulário controlado (tesauro ou lista alfabética), onde um programa de computador desenvolve para cada termo a ser indexado um "perfil" de palavras ou expressões.

As primeiras propostas de indexação automática ocorreram nos anos 60, segundo estudo desenvolvido por Cesarino e Pinto (1980), e eram totalmente baseadas em métodos estatísticos de ocorrência de palavras.

No Brasil, segundo Vieira (1988b), a aplicação da indexação automática tem seu início no final dos anos 60, com a utilização do programa KWIC (*Keyword In Context*) para elaborar os índices das bibliografias especializadas publicados pelo Instituto Brasileiro de Bibliografia e Documentação (IBBD), atual Instituto Brasileiro de Informação em Ciência e Tecnologia (IBICT).

Na década de 70, as pesquisas de indexação automática em território nacional ocorrem através de estudos individuais, realizados em cursos de pós-graduação, concentrando-se na análise de frequência (VIEIRA, 1988b).

Nos anos de 1980 surgem os estudos baseados em referenciais linguísticos, conjuntamente com uma abordagem estatística, como por exemplo, o estudo de Andreewski e Ruas (1983) que trata da adaptação do sistema francês *Systéme Syntaxique et Probabiliste d´Indexation et de Recherche d´Informaticos Textuelles* (SPIRIT) para documentos em língua portuguesa (GIL LEIVA, 1997).

O uso de referenciais lingüísticos, mais exatamente de critérios sintático-semânticos, tal como a proposta de uso de sintagmas nominais como unidades de análise, estão presentes nos trabalhos de alguns autores brasileiros a partir da década de 90 (KURAMOTO, 1995; SOUZA, 2006; BORGES; MACULAN; LIMA, 2008).

3 Procedimentos Metodológicos

Este estudo formou-se por meio do mapeamento e da discussão da produção acadêmica e científica sobre a Indexação Automática no campo da Ciência da Informação no Brasil através de uma abordagem qualitativa e quantitativa.

O estudo desenvolveu-se como pesquisa exploratória, pois tem como finalidade "proporcionar maior familiaridade com o problema, com vistas a torná-lo mais explícito ou a construir hipóteses" (GIL, 2010, p. 27).

No que diz respeito aos procedimentos técnicos, se caracteriza como pesquisa bibliográfica, pois se trata do levantamento e análise de toda bibliografia nacional já publicada no idioma português, seja no formato de livros, artigos de periódicos científicos, anais publicados em congressos e seminários, e literaturas cinzentas (MARCONI; LAKATOS, 2010). O corpus foi levantado por meio de buscas nas bases de dados virtuais: Base de Dados Referencial de Artigos de Periódicos em Ciência da Informação (BRAPCI[1]), Google Acadêmico[2] e Base PERI[3], e na biblioteca da UFPE, onde os documentos foram localizados através das expressões de busca "indexação automática", "automatização da indexação" e "indexação semi-automática".

Visando obter um panorama das pesquisas sobre o tema desenvolveu-se uma revisão de literatura com a finalidade de analisar os diversos aspectos referentes ao tema estudado e a análise de conteúdo dos documentos do *corpus* levantado. O *corpus* constitui-se de 69 documentos, que foram categorizados quanto ao objetivo, objeto de investigação, nome do sistema/método/fórmula, avaliação da indexação automática, natureza e tipologia do *corpus* processado, validação e identificação dos termos, forma que ocorreu a entrada dos dados, linguagem de indexação e a abordagem utilizada na identificação/ponderação/seleção de termos.

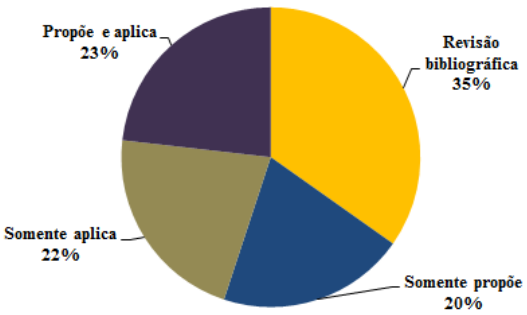
4 Resultados e Discussões

Esta seção estará subdividida em dez subseções apresentando as categorias examinadas na análise de conteúdo.

4.1 Análise do objetivo dos documentos do corpus

De acordo com o GRÁFICO 1 observamos que 24 trabalhos (35%) realizaram uma revisão bibliográfica.

GRÁFICO 1 – Objetivo dos documentos do *corpus*



Fonte: desenvolvido pelo autor.

São 14 os trabalhos que 'propõem' algum método, sistema ou fórmula de indexação automática, correspondendo a

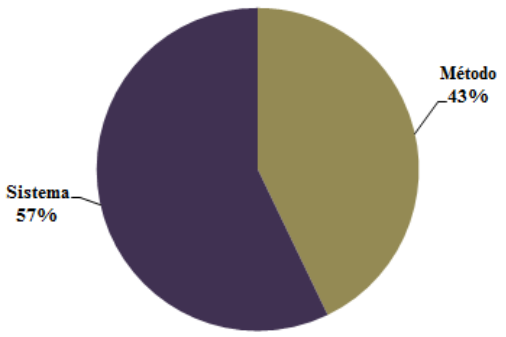
[1] http://www.brapci.ufpr.br
[2] http://scholar.google.com.br/schhp?hl=pt-BR&tab=ws
[3] http://bases.eci.ufmg.br/peri.htm

20% do total, os outros 31 trabalhos estão divididos entre aqueles que 'aplicam', 15 trabalhos (22%), e os que 'aplicam e propõem', 16 trabalhos (23%), algum sistema, método ou fórmula.

Procurando-se identificar como o tema da indexação automática é abordado nos trabalhos classificados na categoria Revisão Bibliográfica constatamos que a maioria dos trabalhos apresenta os fundamentos teóricos da indexação automática, sua evolução histórica e desenvolvimento teórico metodológico; um segundo grupo se concentra em abordar o embasamento filosófico e conceitual subjacente a Web Semântica e suas contribuições na automação da indexação na internet, por meio de motores de busca; o terceiro grupo é formado por trabalhos que discutem as vantagens e desvantagens do uso da indexação automática em comparação com a manual.

É providencial destacar três trabalhos individuais, pois são pontos de vista poucos explorados que podem estar surgindo para suprimir uma lacuna e representar base para o desenvolvimento de outras pesquisas: o primeiro de Guedes e Borschiver (2005), que apresenta a aplicação das leis e princípios da bibliometria, com foco nas palavras, na indexação automática; outro de Barreto (2007), que aborda a aplicação da indexação automática em vídeos; e Kochani, Boccato e Rubi (2011), que mencionam o desenvolvimento de uma política de indexação em sistemas automatizados.

GRÁFICO 2 – Distribuição dos trabalhos que realizaram proposição



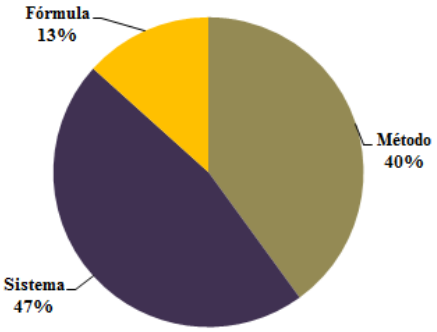
Fonte: desenvolvido pelo autor.

Conforme podemos averiguar no GRÁFICO 2, 14 trabalhos foram classificados no grupo dos que somente propõem algum método, sistema ou fórmula. Destes, nenhum trabalho propôs o uso de fórmula, seis trabalhos (43%), apresentaram a proposta de utilizar algum método, enquanto oito (57%) propuseram algum sistema de indexação automática.

Analisando trabalhos classificados na categoria proposição, verificamos que em relação ao sistema, o mais proposto foi o AUTOMINDEX/II; enquanto a extração dos Sintagmas Nominais foi o método mais proposto.

Dos 15 trabalhos (22%) que foram classificados como tendo o objetivo de somente aplicar fórmula, sistema ou método de indexação automática, dois (13%) aplicam fórmula; seis aplicam métodos (40%) e sete aplicam sistema (47%), ressaltados no GRÁFICO 3.

GRÁFICO 3 – Distribuição dos trabalhos que realizaram aplicação



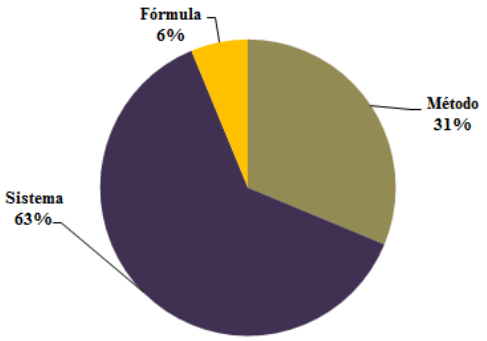
Fonte: desenvolvido pelo autor.

Levando em conta a aplicação, as fórmulas bibliométricas aplicadas foram as leis de Zipf e o ponto T de Goffman, em relação aos sistemas os mais aplicados foram o AUTOMINDEX/II e o *Sistema de Indización Semi-automática* (SISA), já o método baseado na frequência de ocorrência foi o mais aplicado.

O GRÁFICO 4 ilustra a distribuição dos 16 trabalhos que realizam proposições e aplicações de fórmula, sistema ou método de indexação automática, onde um (6%) propõe e aplica uma fórmula matemática; cinco (31%) estão relacionados com os

métodos; e 10 (63%) propõem e aplicam sistemas de indexação automática.

GRÁFICO 4 – Distribuição dos trabalhos que realizaram proposição e aplicação



Fonte: desenvolvido pelo autor.

Em razão dos trabalhos que propõem e aplicam constatou-se um único trabalho referente à adaptação da fórmula de transição de Goffman, no que tange aos sistemas o *PREserved Context Indexing System* (PRECIS) e o OGMA foram os mais propostos e aplicados enquanto aos métodos, todos fizeram menção aos Sintagmas Nominais.

4.2 Nome Sistema/Método/Fórmula

As fórmulas localizadas nas pesquisas são as Leis de Zipf e o Ponto T de Goffman, ambas consistem em fórmulas bibliométricas relacionadas com a frequência de ocorrência de palavras em textos, e aparecem como assunto em apenas três trabalhos.

Foram constatados 12 sistemas de indexação automática durante a análise dos trabalhos, e a relação entre o nome do sistema e sua frequência de ocorrência nos trabalhos está representado no GRÁFICO 5, de onde averiguamos que os três sistemas mais pesquisados para representar automaticamente os descritores aparecem empatados com quatro trabalhos (17%) cada um, são eles: o sistema SISA, o PRECIS e o AUTOMINDEX/II. Logo em seguida aparece o OGMA, sendo pesquisado por três trabalhos (13%).

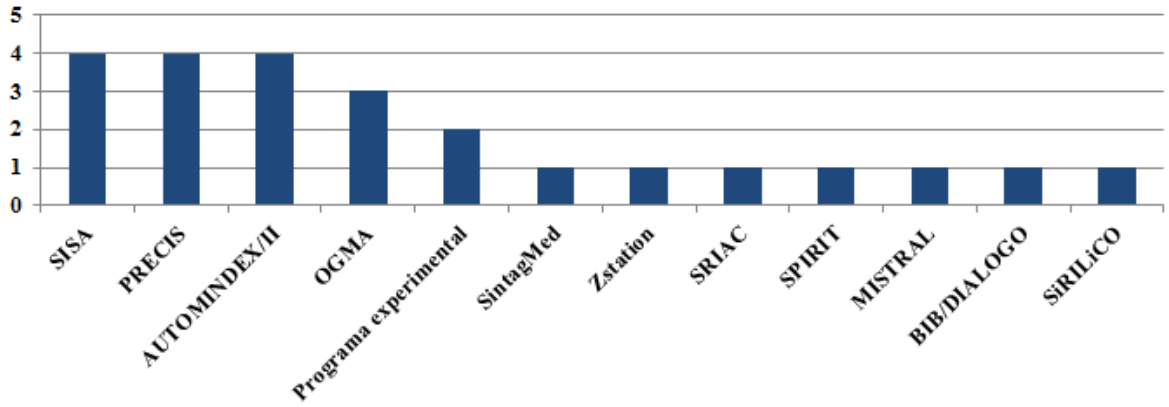
A quinta posição pertence a um sistema nomeado de 'Programa experimental', entretanto a verificação de um fato irá modificar este resultado, pois se constatou que este sistema experimental, registrado nos trabalhos de Haller (1983, 1985), foi desenvolvido em um computador Burroughs 6700 do Centro de Processamento de Dados da Universidade de Brasília (UnB). Em Vieira (1988b), comenta-se que foi utilizado o sistema BIB/DIÁLOGO, implementado no Departamento de Biblioteconomia da UnB, para computadores Burroughs B6700, e terminais Burroughs, modelo TVA 800/10. Desta forma, o sistema outrora classificado como 'programa experimental' trata-se do sistema BIB/DIÁLOGO. E para finalizar este desfecho, Robredo (1991) explica que o sistema AUTOMINDEX/II, constitui-se num subsistema do sistema BIB/DIÁLOGO, o qual já no início dos anos 80 é utilizado em estudos desenvolvidos por Robredo, então professor titular do Departamento de Biblioteconomia da Faculdade de Ciências Sociais Aplicadas da UnB.

Portanto, considerando que os sistemas 'programa experimental', BIB/DIÁLOGO e AUTOMINDEX/II fazem parte do mesmo sistema, e classificados pelo sistema mais geral, o BIB/DIÁLOGO, este passa a ser o mais pesquisado com sete trabalhos.

Cada um dos seis sistemas restantes é proposto, aplicado, ou proposto e aplicado por apenas um trabalho. Desta forma os sistemas SintagMed, Zstation, Sistema de Recuperação de Informação Assistida por Computadores (SRIAC), SPIRIT, MISTRAL e Sistema de Recuperação de Informação baseado em teorias da Linguística Computacional e Ontologia (SiRILiCo), representando 60% dos sistemas observados, não apresentam continuidade nas pesquisas e foram investigados por 25% dos trabalhos, enquanto os sistemas SISA, PRECIS, BIB/DIALOGO (programa experimental e AUTOMINDEX/II) e OGMA bancando os outros 40% foram pesquisados por 75% dos trabalhos, o que demonstra um certo

prosseguimento nas pesquisas sobre estes sistemas.

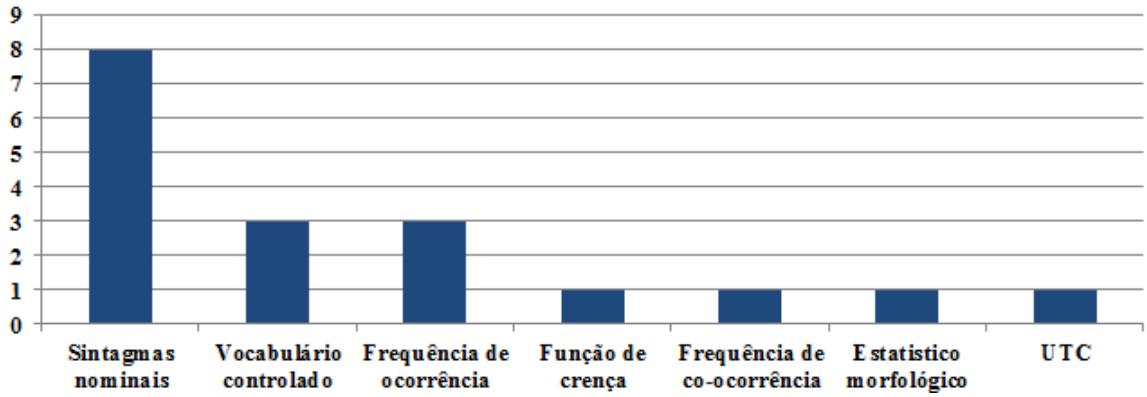
GRÁFICO 5 – Frequência de ocorrência dos sistemas nos documentos do *corpus*



Fonte: desenvolvido pelo autor

Avaliando os resultados referentes ao GRÁFICO 6 obtemos que o método mais pesquisado de indexação automática foram os sintagmas nominais, com 8 trabalhos (44%), dividido entre um trabalho que aplica; três que propuseram e aplicaram; e quatro que propuseram este método. Empatados na segunda colocação com três trabalhos (17%) cada um, estão os métodos que utilizam o vocabulário controlado e a frequência de ocorrência. Este com três trabalhos que aplicam, e aquele com um trabalho que aplica e dois trabalhos que propõem e aplicam o método. Os demais métodos apareceram uma vez, são eles: função de crença, frequência de co-ocorrência, estatístico-morfológico e Unidades Terminológicas Complexas (UTC).

GRÁFICO 6 – Frequência de ocorrência dos métodos nos documentos do *corpus*

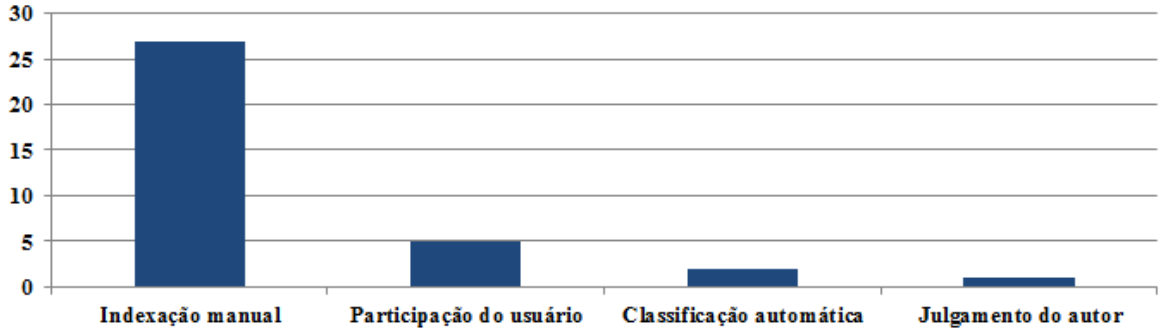


Fonte: desenvolvido pelo autor

4.3 A avaliação da indexação automática

Para a grande maioria dos autores o método escolhido para avaliar suas pesquisas sobre o processo automático da indexação foi realizando uma comparação dos resultados obtidos pelo processo automático com os obtidos através do método intelectual de indexação, correspondendo a 27 dos trabalhos (71%), conforme GRÁFICO 7. Não foi possível identificar em dez trabalhos qual o método de avaliação empregado, mas nos 35 documentos analisados ficou evidente que a maioria optou como método de avaliação a comparação com a indexação manual.

GRÁFICO 7 – Frequência de ocorrência dos métodos de avaliação



Fonte: desenvolvido pelo autor.

Em segundo lugar, encontra-se o método de avaliar através da participação do usuário na comparação de índices ou Sistemas de Recuperação da Informação, com seis trabalhos (14%). Neste processo geralmente se aplica aos usuários, questionários e entrevistas estruturadas ou se avalia através de uma busca experimental comparada e simulada, com o objetivo de identificar as dificuldades e/ou facilidades através da reação dos usuários na utilização do índice. Esta avaliação pode ocorrer sobre dois pontos de vista, o do sistema e o do usuário.

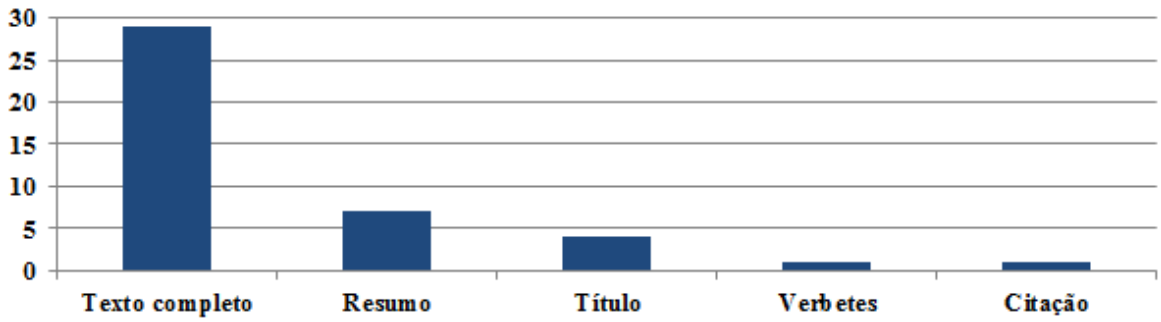
Dois trabalhos (11%) utilizam o método de classificação automática, que aplicam algoritmos de agrupamento e classificação para apresentar um valor percentual indicando quantos documentos foram classificados corretamente.

Um trabalho (1%) compara os resultados dos termos obtidos dos documentos através das modificações das fórmulas bibliométricas, onde o autor é quem fica encarregado de analisar se os termos obtidos são satisfatórios.

4.4 Natureza do corpus

Quanto à natureza do *corpus*, constatou-se que ocorreu uma preferência em se realizar pesquisas quanto à indexação automática do texto completo dos documentos, foram 29 trabalhos (69%). Não foi possível identificar em três trabalhos qual a natureza empregada.

GRÁFICO 8 – Natureza do corpus



Fonte: desenvolvido pelo autor.

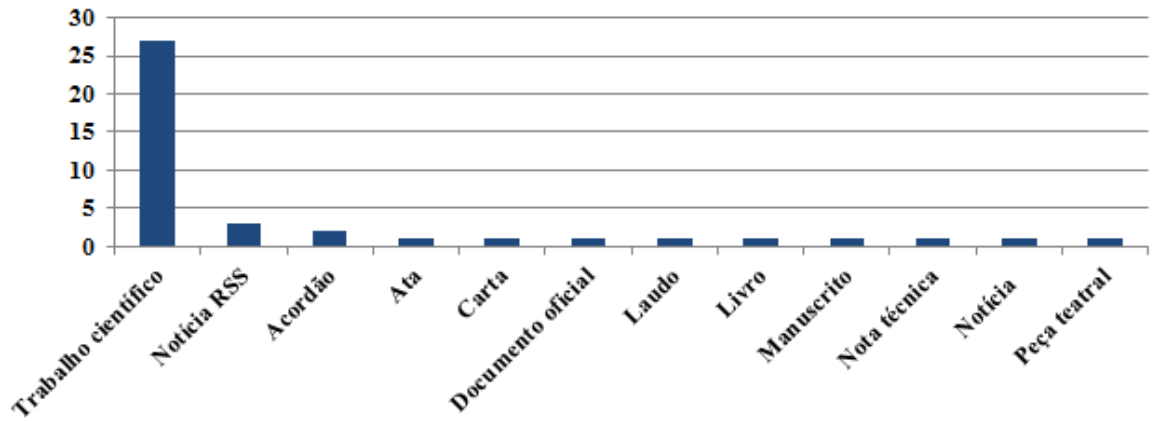
Os demais 31% estão divididos entre sete trabalhos que optaram utilizar os resumos como seu *corpus* de pesquisa, quatro trabalhos que utilizaram como *corpus* de análise os títulos. Tanto os verbetes quanto as citações constam como *corpus* de investigação em apenas um trabalho, cada. Os resultados estão ressaltados visualmente no GRÁFICO 8.

4.5 Tipologia do corpus

Observa-se que a distribuição referente à tipologia do *corpus* se comporta de acordo com o padrão das distribuições bibliométricas em geral: "poucos com muito e muitos com pouco". Assim, no GRÁFICO 9 podemos constatar que poucas tipologias ocorreram muitas vezes, enquanto diversas tipologias ocorreram poucas vezes.

Os dados ilustrados no GRÁFICO 9 ajudam a compreender qual foi o comportamento da tipologia do *corpus* pesquisado nos trabalhos analisados. Verifica-se que a tipologia mais pesquisada com 27 trabalhos (66%) foram os trabalhos científicos, o que pode ser justificado por ser este tipo de material o que mais interessa às instituições, que normalmente desenvolvem e/ou financiam estas pesquisas, isto é, as Universidades Públicas, e por este motivo a importância da natureza do *corpus* incide sobre os trabalhos científicos que normalmente são produzidos na própria instituição. Não foi possível identificar em quatro trabalhos qual a tipologia do *corpus*.

GRÁFICO 9 – Tipologia do corpus



Fonte: desenvolvido pelo autor.

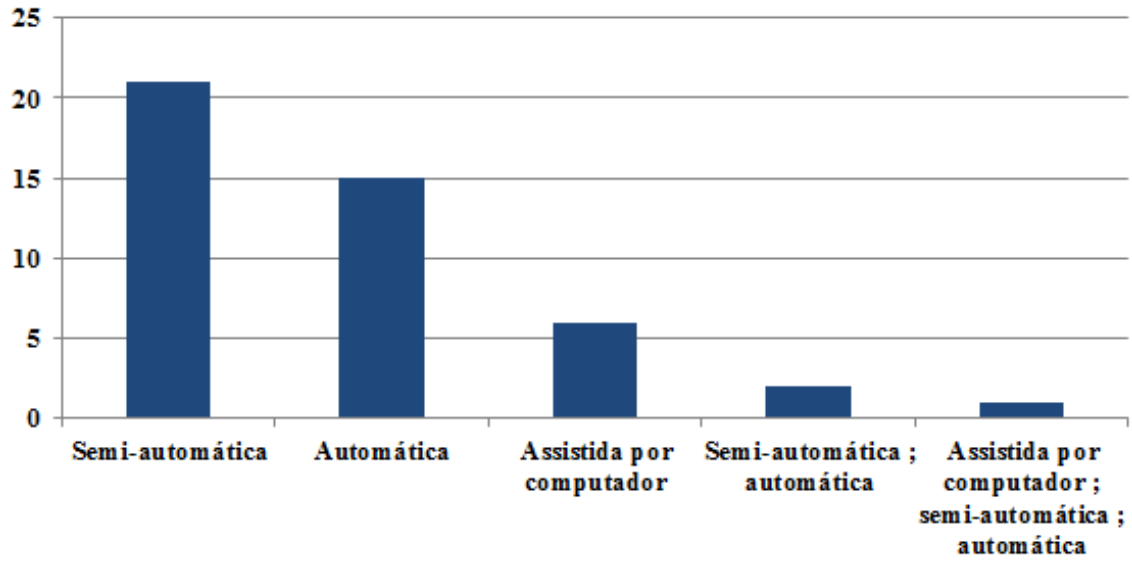
As demais tipologias, que juntas somam 14 trabalhos (34%), quase metade em relação à tipologia mais estudada, acabam refletindo realidades específicas caracterizando necessidades isoladas.

4.6 Validação dos termos

Analisando o tipo de indexação que os trabalhos aplicaram ou propuseram para validar os termos de suas pesquisas, observamos através do GRÁFICO 10 que em primeiro lugar, com 21 trabalhos (47%), estão aqueles que empregaram a indexação semi-automática, seguida pelos trabalhos que empregaram a indexação automática, com 15 trabalhos (33%), e um pouco atrás, se encontram seis trabalhos (13%) assistidos pelo computador.

Dois trabalhos (4%) declaram que a validação dos termos aconteceu tanto através de uma indexação semi-automática, quanto automática. E um trabalho (2%) que a validação ocorreu através dos três critérios de análise.

GRÁFICO 10 – Validação dos termos

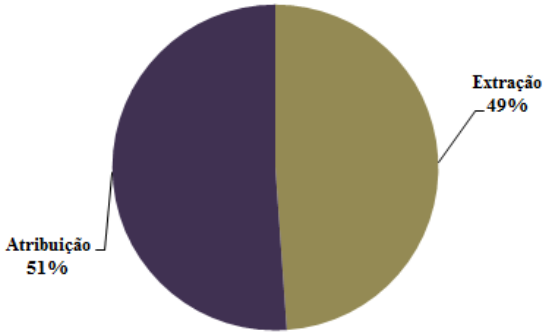


Fonte: desenvolvido pelo autor

4.7 Identificação dos termos

O GRÁFICO 11 demonstra que o processo de identificação foi realizado em 22 trabalhos (49%) por meio da extração, enquanto em 23 (51%) por meio da atribuição.

GRÁFICO 11 – Identificação dos termos



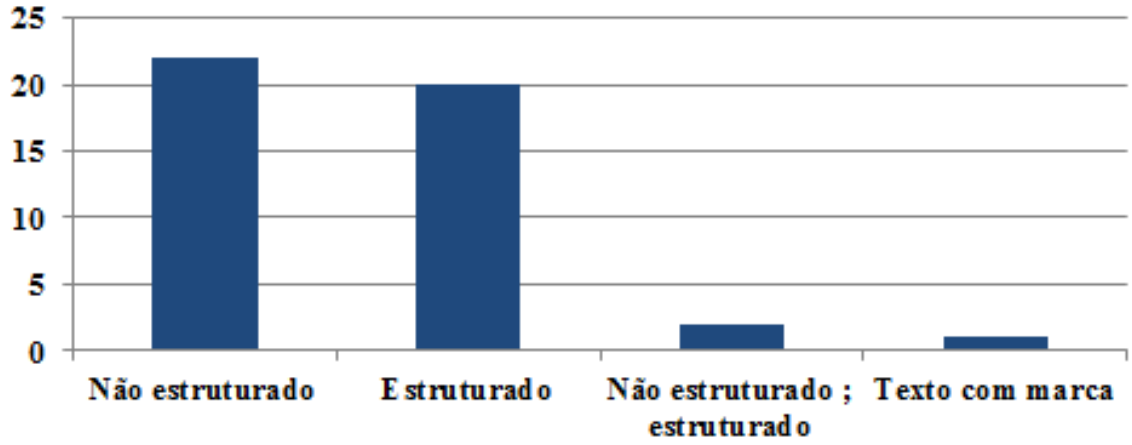
Fonte: desenvolvido pelo autor.

Esta pequena diferença pode ser explicada, pois apesar da dificuldade do computador em realizar o processo de obter um termo através da atribuição, a chegada de novas tecnologias e de pesquisas sobre aplicação de tesauros e vocabulários controlados motivou pesquisas sobre a atribuição.

4.8 Forma que ocorreu a entrada dos dados

O GRÁFICO 12 ilustra como ocorreu à entrada dos dados nos sistemas, métodos e fórmulas investigados nos trabalhos que compõem o *corpus* deste artigo. Destes, 22 (49%) realizaram a entrada dos dados através de um texto não estruturado, 20 (44%) correspondem aos trabalhos que estruturaram o texto de alguma forma antes da inserção dos dados para análise, dois (aproximadamente 4%) foram os que alegaram trabalhar com a entrada de dados de forma não estruturada e em outro momento com o texto estruturado, e um trabalho (2%) cita utilizar a marcação nos textos.

GRÁFICO 12– Frequência de ocorrência da forma como ocorreu a entrada de dados



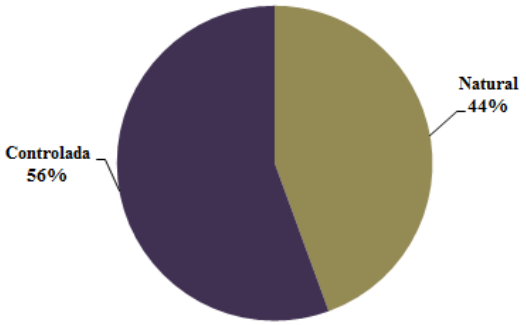
Fonte: desenvolvido pelo autor.

4.9 Linguagem de indexação

Nesta subseção procurou-se descrever os trabalhos quanto a natureza da linguagem de indexação (natural ou controlada) e dos termos (palavras isoladas ou termos compostos).

Quanto à natureza da linguagem notamos que ocorreu uma predominância pela linguagem controlada com 25 trabalhos (56%) em decorrência dos 20 trabalhos (44%) atribuídos à linguagem natural, visualizados no GRÁFICO 13.

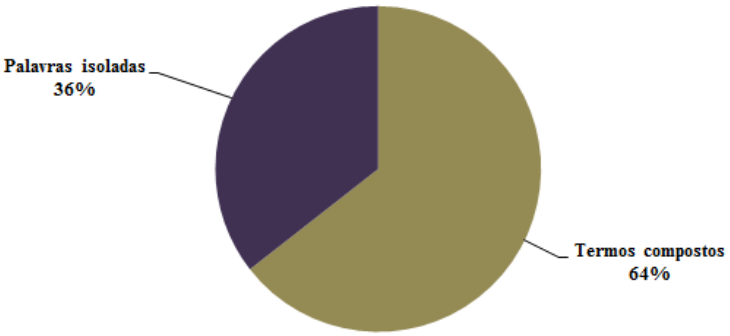
GRÁFICO 13 – Natureza da linguagem



Fonte: desenvolvido pelo autor.

Os dados relacionados aos termos estão ilustrados no GRÁFICO 14, que ressalta a superioridade numérica da extração de termos compostos com 29 trabalhos (64%) em relação à escolha apenas por palavras isoladas com 16 trabalhos (36%)

GRÁFICO 14 – Natureza dos termos



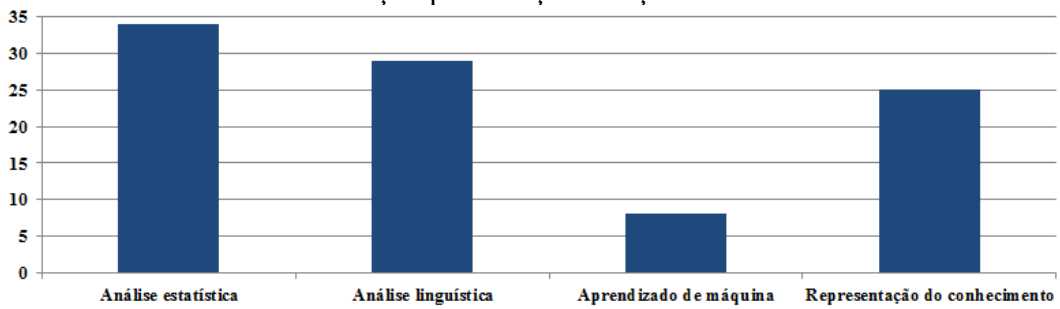
Fonte: desenvolvido pelo autor.

4.10 Abordagem utilizada na identificação / ponderação / seleção de termos

Quanto à abordagem ou tipo de técnicas utilizadas para identificar, ponderar ou selecionar os termos utilizados na indexação automática (em um total de 45 documentos que propuseram, aplicaram ou propuseram e aplicaram sistema/método/fórmula) foram observados quanto ao uso da análise estatística, análise linguística, aprendizado de máquina e representação do conhecimento, sendo classificados nas categorias e mensurados quanto à frequência em que apareceram nos trabalhos.

O resultado pode ser observado no GRÁFICO 15, onde o método mais pesquisado foi a análise estatística com 34 trabalhos (35%), o segundo método foi a análise linguística, aparecendo em 29 dos trabalhos (30%), a representação do conhecimento está em 25 trabalhos (26%), enquanto o aprendizado de máquina condiz com oito trabalhos (9%).

GRÁFICO 15 – Frequência de ocorrência do método/processo de identificação/ponderação/seleção de termos



Fonte: desenvolvido pelo autor.

5 Considerações Finais

Fundamentados nos resultados da análise de conteúdo dos documentos do *corpus* desta pesquisa, observamos que quanto aos objetivos 65% dos trabalhos investigaram algum sistema, método ou fórmula de indexação automática. Além disso, percebemos que 35% dos trabalhos foram classificados como revisão bibliográfica, e que grande parte destes abordou a história da indexação automática, apontando seus fundamentos teóricos e evolução dos métodos.

Podemos concluir que os sintagmas nominais foram os métodos mais investigados. Em relação aos sistemas de indexação automática, quatro se destacam: o BIB/DIALOGO (incluindo o AUTOMINDEX), o SISA, o PRECIS e o OGMA.

Verificamos que a grande concentração dos trabalhos utilizou como método de avaliação a comparação com a indexação manual. Dessa forma, eles procuram avaliar se a implantação do sistema automático trará benefícios, obtendo resultados equivalentes em menos tempo.

Quanto à natureza e a tipologia do *corpus*, identificamos que a preocupação da maioria dos autores está concentrada em indexar o texto completo de trabalhos científicos.

Analisando o tipo de validação dos termos, percebemos a preferência pela aplicação da indexação semi-automática. O que pode ser justificado pelo fato dos processos totalmente automáticos ainda serem falhos e apresentarem limitações tecnológicas. Entretanto, a diferença em relação ao processo automático, na segunda posição, não é muito grande, podendo ser interpretada como um esforço no desenvolvimento de uma indexação automática de qualidade.

Com relação à pequena diferença existente entre os trabalhos que pesquisaram a identificação dos termos por meio da extração (49%) e os que optaram pela atribuição (51%). Uma possível justificativa para tal fato, é que a chegada de novas tecnologias e de pesquisas sobre aplicação de tesauros e vocabulários controlados motivou pesquisas sobre a atribuição, apesar da dificuldade em fazer com que o computador execute o processo de obter um termo através da atribuição.

Já a entrada dos dados apresentou um empate técnico entre textos não estruturados em relação ao texto estruturado. Em relação à linguagem de indexação de indexação, foi observada tanto a natureza da linguagem, que demonstrou uma preferência dos trabalhos pela pesquisa com a linguagem controlada, quanto a natureza dos termos, no qual a primazia encontra-se no estudo com termos compostos.

Quanto à categoria dos tipos de métodos de identificação, ponderação e seleção dos termos, chegamos à constatação de que o tipo de método mais pesquisado foi a análise estatística representando que uma grande parte dos trabalhos recorreram a, um ou mais dos seguintes processos: radicalização, eliminação de *stopwords*, análise de posição de ocorrência (localização), análise de frequência de ocorrência, análise de co-ocorrência, peso numérico, dicionário de raízes e/ou matriz binária.

O movimento ininterrupto da ciência continuará incentivando os pesquisadores a continuarem produzindo novas pesquisas. Consequentemente, eles identificarão e explicitarão outros caminhos (ou mesmo aqueles já trilhados, mas sobre uma ótica diferente), para que se chegue a um modelo automático de indexação de termos com qualidade igual ou superior a realizada pelo especialista humano quando realizam a mesma tarefa.

Em vista disso, este trabalho aponta trabalhos futuros na área da Ciência da Informação, sobre a indexação automática, que dariam continuidade ao trabalho desenvolvido nesta pesquisa. Como sugestão para trabalhos futuros, apontamos:

- Investigar a análise de citação no corpus levantado, por permitir

identificar características e mapear a comunicação científica;

- Realizar uma análise da produção internacional sobre a indexação automática;
- Mapear e discutir a produção acadêmica sobre a indexação automática em diversos campos do conhecimento (ciência da informação, ciência da computação, linguística), diferentes fontes de informação, épocas e lugares, elaborando seu Estado da Arte.

Automatic Indexing in Information Science Area in Brazil

Abstract

This work presents an overview of studies about Automatic Indexing through the mapping and analysis of Brazilian scientific and academic production in information science area over the period 1973-2012. Its specific objectives are to collect a corpus, analyze and characterize the publications by observing its goals and methodological aspects. The methodology consists of a detailed bibliographical study and contents' analysis on literary production in Brazil about the automatic indexing of texts written in Portuguese language. The corpus for the realization of content analysis consists of documents in the Portuguese language, such as books, journal articles, proceedings and gray literature. The most significant results show that: 35% of the publications performs a literature review, while 65% researches systems, methods, and formulas of automatic indexation, where 23% proposes and applies, 20% proposes, and 22% applies automatic indexing instruments; the systems are the object of study more researched; the comparison with manual indexing is the most used method of evaluation; the full text is the most researched corpus' nature; scientific publication is the corpus' typology most studied; semi-automatic indexing is the most applied term validation procedure; the attributing process was the most adopted to identify terms; the unstructured text is data input preferred of the systems; natural language is the nature of indexing language, the compound term is the nature of terms most searched; statistical analysis is the method most used to term selection. We conclude that there is a tendency in studies on automatic indexing by means of noun phrases and the application of new technologies in automatic indexing by term assignment.

Keywords: Automatic Indexing. Information Retrieval. Information Retrieval Systems. Information Science. Brazil.

Referências

ALVES, J. C.; CAFÉ, L. M. A. Análise focada em metadados sob a luz do padrão MTD-BR. **Em Questão**, Porto Alegre, v. 16, n. 2, p. 179-202, jul./dez. 2010.

ANDREEWSKI, A.; RUAS, V. Indexação automática baseada em métodos linguísticos e estatísticos e sua aplicabilidade a língua portuguesa. **Ciência da Informação**, Brasília, v. 12, n. 1, p. 61-73, 1983.

BARRETO, J. S. Desafios e avanços na recuperação automática da informação audiovisual. **Ciência da Informação**, Brasília, v. 36, n. 3, p. 17-28, set./dez. 2007.

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

CAPURRO, R.; HJØRLAND, B. O conceito de informação. **Perspectivas em Ciência da Informação**, v. 12, n. 1, p. 148-207, jan./abr. 2007.

CESARINO, M. A. N.; PINTO, M. C. M. F. Análise de assunto. **Revista de Biblioteconomia de Brasília**, Brasília, v. 8, n. 1, p. 254-263, p. 32-43, jan./jun. 1980.

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

GIL, A. C. **Como elaborar projetos de pesquisa**. 5. ed. São Paulo: Atlas, 2010.

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

GUEDES, V.; BORSCHIVER, S. Bibliometria: uma ferramenta estatística para a gestão da informação e do conhecimento, em sistemas de informação, de comunicação e de avaliação científica e tecnológica. In: ENCONTRO NACIONAL DE CIÊNCIA DA INFORMAÇÃO, 6., 2005, Salvador. **Anais...** Salvador: ICI/UFBA, 2005.

KOCHANI, A. P.; BOCCATO, V. R. C.; RUBI, M. P. Política de indexação para sistemas automatizados de coordenadorias de comunicação em ambientes universitários. In: CONGRESSO BRASILEIRO DE BIBLIOTECONOMIA, DOCUMENTAÇÃO E CIÊNCIA DA INFORMAÇÃO, 24., 2011, Maceió. **Anais...** São Paulo: FEBAB, 2011.

KURAMOTO, H. Uma abordagem alternativa para o tratamento e a recuperação de informação textual: os sintagmas nominais. **Ciência da Informação**, Brasília, v. 25, n. 2, p. 1-18, 1995.

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

MARCONI, M. A.; LAKATOS, E. M. **Metodologia do trabalho científico:** procedimentos básicos, pesquisa bibliográfica, projeto e relatório, publicações e trabalhos científicos. 7. ed., 5. reimpr. São Paulo: Atlas, 2010.

MORAES, A. F. de. Os pioneiros da ciência da informação nos EUA. **Informação & Sociedade: estudos**, João Pessoa, v. 12, n. 2, 2002.

NASCIMENTO, G. F. C. L. **Folksonomia como estratégia de indexação dos bibliotecários no Del.icio.us.** 2008. 104f. Dissertação (Mestrado em Ciência da Informação) – Programa de Pós-Graduação em Ciência da Informação, Universidade Federal da Paraíba, João Pessoa, 2008.

PALMQUIST, R. A. **Class lecture notes:** Luhn and automatic indexing – references to the early years of automatic indexing and information retrieval. Organizing and providing acess to information – LIS 391D.2 – Spring, 1998.

ROBREDO, J. **Documentação de hoje e de amanhã**. 4. ed. rev. ampl. Brasília, DF: Ed. Do Autor, 2005.

ROBREDO, J. Indexação automática de textos: uma abordagem otimizada e simples. **Ciência da Informação**, Brasília, v. 20, n. 2, p. 130-136, jul./dez. 1991.

SOUZA, R. R. Uma proposta de metodologia para indexação automática utilizando sintagmas nominais. **Encontros Bibli: Revista Eletrônica de Biblioteconomia e Ciência da Informação**, Florianópolis, n. esp., p.42-59, 1. Sem. 2006.

VIEIRA, S. B. Indexação automática e manual: revisão de literatura. **Ciência da Informação**, Brasília, v. 17, n. 1, p. 43-57, jan./jun. 1988.