Panorama de estudos sobre indexação automática no âmbito da ciência da informação no Brasil (1973-2012)

Renato Fernandes Corrêa
Doutor em Ciências da Computação pela Universidade Federal de Pernambuco (UFPE) – Recife, PE - Brasil.
Professor da Universidade Federal de Pernambuco (UFPE) - Recife, PE – Brasil.
http://lattes.cnpq.br/7536537827447217
E-mail: renato.correa@ufpe.br

Remi Correia Lapa
Mestre em Ciência da Informação pela Universidade Federal de Pernambuco (UFPE) – Recife, PE - Brasil.
Bolsista de Pesquisa da Universidade Federal de Pernambuco (UFPE) - Recife, PE – Brasil.
http://lattes.cnpq.br/4971685740534540
E-mail: rmcrlp@gmail.com

Recebido em: 15/08/2014. Aprovado em: 23/1/2015. Publicado em: 07/08/2015.

Resumo
Apresenta um panorama dos estudos sobre a indexação automática no âmbito da ciência da informação no Brasil, por meio do mapeamento e análise da produção acadêmica e científica nacional no período de 1973 a 2012. Como objetivos específicos, caracteriza o *corpus* de análise quanto aos objetivos e aspectos metodológicos através da análise de conteúdo dos documentos; bem como descreve a institucionalização das pesquisas através de estudo bibliométrico, observando autoria, instituições publicadoras, ano, fonte de informação e instituições acadêmicas. A metodologia tem natureza exploratória e bibliográfica, de caráter quali-quantitativo, pautada nas técnicas de análise bibliométrica e análise de conteúdo. Os resultados bibliométricos apontam a autora Fujita como maior produtora; a revista *Ciência da Informação* como maior publicadora; o século XXI concentrando a maior parte da produção; o periódico como principal meio de divulgação; a instituição Universidade de Brasília como o maior produtor. A análise de conteúdo aponta que 35% dos trabalhos realiza revisão bibliográfica, enquanto 65% investiga a proposição e/ou aplicação de fórmula, método ou sistema de Indexação Automática. Conclui-se que há uma tendência em estudos sobre a indexação automática por extração por meio dos sintagmas nominais e indexação automática por atribuição através de vocabulário controlado.

Palavras-chave: Indexação automática. Método de indexação automática. Sistema de indexação automática. Ciência da informação. Brasil.

Overview of studies about automatic indexing within the Information Science in Brazil (1973-2012)

Abstract
Presents an overview of studies on automatic indexing within the scope of information science in Brazil, through the mapping and analysis of national scientific and academic production from 1973 to 2012. As specific objectives, it characterizes the objectives and methodological aspects of the documents through content analysis; it also describes the institutionalization of research through bibliometric studies, by observing authorship, publishing institutions, year, source of publication and academic institutions. The methodology is of exploratory and bibliographical nature, of qualitative and quantitative character, based on bibliometric and content analysis techniques. The bibliometric results show author Fujita as the largest producer; journal Ciência da Informação as the largest publisher; the XXI century concentrates the most part of the production; the journal as primary means of dissemination; the institution Universidade de Brasília as the largest producer. Content analysis shows that 35% of the papers perform literature review, while 65% investigate a proposition and/or application of a formula, method or automatic indexing system. We conclude that there is a tendency in studies on automatic indexing by extraction through nominal phrases, and on automatic indexing by assignment through controlled vocabulary.

Keywords: *Automatic indexing. Automatic indexing method. Automatic indexing system. Information Science. Brazil.*

Panorama de los estudios sobre indexación automática dentro de la ciencia de la información en Brasil (1973-2012)

Resumen

Presenta un panorama de los estudios sobre indexación automática en ciencia de la información en Brasil, a través de mapeo y análisis de la producción científica y académica nacional entre 1973 y 2012. Como objetivos específicos, caracteriza el corpus de análisis cuanto a los objetivos y aspectos metodológicos de los documentos por medio de análisis de contenido; y describe la institucionalización de investigaciones a través del estudio bibliométrico, señalando autoría, instituciones de publicación, año, fuente de información y las instituciones académicas. La metodología es de naturaleza exploratoria y bibliográfica, de carácter cualitativo y cuantitativo, basado en técnicas de análisis bibliométricos y el análisis de contenido. Resultados bibliométricos muestran al autor Fujita como el mayor productor; la revista Ciencia da Informação como el mayor editor; el siglo XXI concentrando la mayor parte de la producción; la revista como el principal medio de difusión; la institución Universidad de Brasilia como el mayor productor. El análisis de contenido muestra que el 35% de los trabajos hacen revisión de la literatura, mientras que el 65% investiga la proposición y / o aplicación de fórmula, método o sistema automático de indexación. De ello se desprende que hay una tendencia en los estudios sobre la indexación automática por extracción a través de los sintagmas nominales y la indexación automática por atribución a través de un vocabulario controlado.

Palabras clave: *Indexación automática. Método de indexación automática. Sistema de indexación automática. Ciencia de la Información. Brasil.*

INTRODUÇÃO

O estudo da indexação automática decorre da necessidade de melhorar o processo de recuperação da informação diante dos efeitos do fenômeno do aumento da produção científica. A aplicação da indexação automática desenvolveu-se como uma alternativa viável na análise e representação da informação diante do crescimento exponencial do volume de documentos (ROBREDO, 1982) e (NARUKAWA, 2011).

A indexação automática é conceituada por Lancaster (2004), como um processo que ocorre quando o computador é utilizado para substituir, em certa medida, a indexação manual realizada por um indexador.

Um dos pesquisadores pioneiros na temática foi Luhn[1] na década de 1960, que buscando soluções para os problemas relacionados à recuperação da informação, vislumbrou soluções práticas e de baixo custo com a utilização de máquinas (computadores), tornando-se um defensor da indexação automática (PALMQUIST, 1998).

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.

No Brasil, segundo Vieira (1988), 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).

Desde então, vários estudos buscaram alternativas eficazes que possibilitassem a representação automatizada dos assuntos principais dos documentos (SOUZA, 2006; CÂMARA JÚNIOR, 2007; MAIA, 2008; BORGES, 2009; NARUKAWA, 2011). Esses estudos nos conduzem à utilização da indexação automática como mecanismo destinado a facilitar o acesso aos documentos técnico-científicos que fazem parte da memória da instituição e em melhorar a recuperação desses documentos armazenados nos repositórios institucionais.

1 Hans Peter Luhn, especialista da IBM, foi o pioneiro na aplicação da análise estatística de vocabulário para executar uma indexação automática (MORAES, 2002).

Nesse aspecto, a indexação automática ganha um enfoque social, pois auxilia determinado grupo no acesso e uso da informação produzida por uma comunidade específica. "Podemos considerar como "social" qualquer processo de produção/organização/consumo de informação, uma vez que ele acontece entre grupos, segmentos, classes – ou seja, a geração e apropriação de informações só ocorrem no âmbito da sociedade, das relações sociais." (CARDOSO, 1994, p. 107-108).

Segundo Robredo (2005), atualmente 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).

Diante do que foi abordado, percebeu-se uma lacuna na construção de um panorama atual sobre as pesquisas que abordam a indexação automática no âmbito nacional.

Portanto, a **problemática** subjacente à elaboração deste 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. Os **objetivos específicos** se propõem a:
a) levantar um *corpus* com documentos brasileiros da área da ciência da informação a respeito da indexação automática entre os anos de 1973 a 2012;
b) descrever a institucionalização das pesquisas através da identificação dos autores, períodos, instituições publicadoras, fontes de publicação e instituições acadêmicas mais representativas na ciência da informação dentro da temática indexação automática;
c) analisar os objetivos, aspectos metodológicos e sistemas, métodos, fórmulas de indexação automática tratados nos documentos do *corpus* levantado.

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

O presente artigo se encontra estruturado da seguinte forma: na seção 2 são apresentados os fundamentos teóricos necessários para o entendimento dos procedimentos e resultados da pesquisa; a seção 3 descreve a metodologia da pesquisa; a seção 4 apresenta a discussão dos resultados encontrados; e na seção 5 são apresentadas as considerações finais.

INDEXAÇÃO

O termo indexação (*indexing*) pertence à corrente teórica inglesa e é a etapa da representação temática que tem o objetivo de reportar ao conteúdo do documento de modo que possa ser recuperado quando for solicitado em outro momento (FUJITA, 2009).

O termo índice, natural do latim, *index* significa indicador, indício, delator, e o verbo *indicare* significa dar a conhecer, indicar, significar, dar o sinal, conforme a conceituação apresentada por Gaspar (2011). Neste sentido, é função da indexação representar o assunto dos documentos através da elaboração de termos, mostrando "ao pesquisador e ao leitor de uma forma geral, que tópicos, fatos ou outros itens de informação estão tratados nos documentos indexados" (GASPAR, 2011, p. 3).

Para Robredo (1982), o processo de leitura realizado por um indivíduo qualquer apresenta semelhança com o processo de indexação, pois ao ler um texto, o leitor não apresenta interesse pelas letras, mas pela ideia que elas representam quando organizadas em palavras ou em conjuntos de palavras relacionadas. "De fato, a palavra escrita é um signo global, um conceito. O olho – janela do cérebro – reconhece as palavras significativas e suas associações fixando-

se nelas um tempo curto, mas mensurável – o tempo necessário para assegurar a memorização das ideias – pulando, praticamente, as palavras não significativas" (ROBREDO, 1982, p. 237).

INDEXAÇÃO AUTOMÁTICA

Levando em consideração que a indexação é a representação de um documento ou das perguntas feitas pelos usuários, no ato de busca, por meio de linguagem natural ou uma linguagem documentária, a indexação automática seria a execução deste processo por meio de programas ou algoritmos de computador que "varrem" o documento (ou registros de documentos) e realizam a representação do conteúdo sem a intervenção do documentalista (ROBREDO, 1986 apud SILVA; FUJITA, 2004).

Gil Leiva, em sua tese apresentada no ano de 1997, expôs ampla variedade terminológica para a designação de conceitos sobre a automatização da indexação, sendo a expressão *automatic indexing* [indexação automática] a mais utilizada.

Apesar da extensa quantidade de termos encontrada, Gil Leiva (1997) percebeu que esta importante variedade de expressões refere-se apenas a três conceitos diferentes (tradução nossa):

1. Indexação auxiliada por computador – Programas que auxiliam no processo de armazenamento dos termos adquiridos através da indexação intelectual. Estes sistemas procuram facilitar o processo de indexação ao proporcionar através das telas de ajuda, as notas explicativas sobre o uso de um termo, os termos relacionados e permitir a atribuição de termos sem ter que digitá-los, ou mesmo consultar on-line por documentos indexados anteriormente para verificar qualquer aspecto.
2. Indexação semiautomática – Ocorre em sistemas que indexam documentos automaticamente, mas os termos de indexação propostos, se necessário, são validados e editados por um profissional.
3. Indexação automática – Programas que não precisam de validação, ou seja, os termos propostos são armazenados diretamente como descritores do documento indexado.

Ainda assim é encontrada na literatura a aplicação desses conceitos como sinônimos.

Borges (2009, p. 31) menciona que a indexação automática "também chamada de indexação assistida por computador e de indexação semiautomática" é considerada um modelo de extração com características estatísticas e probabilísticas.

Apenas quando a análise e a extração de conceitos são realizadas por programas de computador (o texto precisa estar no formato eletrônico) é que ocorre a indexação automática.

Neste sentindo, a indexação automática pode ser definida como:

> um conjunto de operações, basicamente matemáticas, lingüísticas, de programação, destinadas somente a selecionar certos elementos de um documento sem modificar seu conteúdo. Na Indexação Automática, a análise de conteúdo não é permeada pela interpretação de terceiros, pois os termos significativos são extraídos do texto e ordenados pela sua freqüência de ocorrência. (NASCIMENTO, 2008, p. 24).

A indexação automática de documentos é compreendida por Gomes (1989) e Santos e Ribeiro (2003) como o processo em que o computador efetua a extração de palavras, expressões ou radicais de palavras utilizadas para representar o conteúdo do texto como um todo, sem a intervenção do documentalista.

Para Araújo Júnior (2007), a indexação automática é qualquer procedimento que permita identificar e selecionar os termos que representam o assunto dos documentos sem a intervenção direta do homem.

Diante de tais conceitos, podemos definir a indexação automática como um conjunto de operações realizadas pelo computador, de natureza estatística, linguística, ou de programação, destinado a selecionar termos como elementos descritivos de um documento pelo processamento automático de seu conteúdo.

Quanto ao método de seleção de termos dos documentos, pode-se categorizar a indexação automática em dois tipos: por extração ou por atribuição.

Na indexação automática por extração os termos significativos são extraídos do texto dos documentos e ordenados pela sua frequência de ocorrência (NASCIMENTO, 2008).

A indexação automática por atribuição, segundo Lancaster (2004), consiste numa representação temática por meio de termos selecionados de um vocabulário controlado (tesauro ou lista alfabética), na qual um programa de computador desenvolve para cada termo de indexação um "perfil" de palavras ou expressões, que ao serem encontradas num documento, habilitam a seleção automática do respectivo termo como descritor do documento.

Por se tratar de um estudo com o propósito de elaborar um panorama da pesquisa no âmbito da CI no Brasil, compete neste artigo traçar uma linearidade histórica abordando a(s) característica(s) mais marcante(s) da indexação automática no Brasil no decorrer dos anos, tema que será abordado a seguir na próxima subseção.

INDEXAÇÃO AUTOMÁTICA NO BRASIL

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 (1988), 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), que mantém a publicação da revista *Ciência da Informação* com trabalhos inéditos que tenham relação com a ciência da informação ou que nos resultados apresentem estudos e pesquisas a respeito das atividades do setor de informação em ciência e tecnologia.

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, 1988).

Nos anos 1980 surgem os estudos baseados em referenciais linguísticos, junto 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 linguísticos, 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).

PROCEDIMENTOS METODOLÓGICOS

O presente artigo se desenvolve como pesquisa exploratória e se propõe a realizar uma revisão de literatura com a finalidade de analisar os diversos aspectos referentes ao fato estudado e coletar e analisar a produção científica nacional sobre a indexação automática.

No que diz respeito aos procedimentos técnicos, caracteriza-se como pesquisa bibliográfica, pois se trata do levantamento de toda a bibliografia nacional já publicada (MARCONI; LAKATOS, 2010). Também se caracteriza quanto aos procedimentos técnicos como pesquisa documental, já que parte da análise dos documentos para inferir o panorama das pesquisas sobre indexação automática no Brasil. Para o desenvolvimento do *corpus,* constituído por 69 documentos brasileiros sobre a indexação automática, utilizaram-se livros e capítulos

de livros, dissertações e teses, publicações em periódicos e comunicações em anais de congressos e de seminários, destinados à área de ciência da informação publicados/comunicados nos últimos 40 anos (1973-2012).

Este artigo se formou por meio do mapeamento e discussão da produção acadêmica e científica através de uma abordagem qualitativa sobre a indexação automática no campo da ciência da informação; e, por uma abordagem quantitativa, oriunda de investigação dos resultados das análises bibliométricas e análise de conteúdo, o que fornece ao estudo um caráter teórico.

O método quantitativo utilizado para medir e avaliar o conhecimento científico do *corpus* é a análise bibliométrica. Dentre as vantagens deste método está o fato de amenizar os elementos de julgamento e produzir resultados quantitativos que tendessem a ser a soma de muitos pequenos julgamentos e apreciações realizados por várias pessoas. Como grande parte da produção científica torna-se conhecida através da sua publicação, fica mais fácil a avaliação das atividades de pesquisa por meio desta (CASTRO, 1997).

Para cumprir o objetivo geral, trabalhou-se também com a técnica de análise de conteúdo, por entender que através dela é possível obter, por procedimentos sistemáticos e objetivos de descrição do conteúdo dos documentos, indicadores que permitem a inferência de conhecimentos relativos às condições de produção/recepção (variáveis inferidas) desses documentos (BARDIN, 2004).

Segundo Marconi e Lakatos (2010), a análise de conteúdo realiza uma apreciação da obra e forma um juízo sobre a autoridade do autor e o valor que representa o trabalho e as ideias nele contidas. Para Bardin (2004), a análise de conteúdo do documento é uma operação ou conjunto de operações visando representar o conteúdo de um documento sob uma forma diferente do original, a fim de facilitar, num estado imediato, a sua consulta e referenciação. Para que os dados possam ser analisados e interpretados, é preciso que sejam ordenados e organizados. Para isto, devem ser codificados e tabulados, começando-se o processo pela classificação (RUDIO, 1986).

No primeiro momento, a elaboração do panorama realizar-se-á por meio de um estudo bibliométrico sobre os autores que desenvolveram pesquisas no âmbito nacional sobre a indexação automática; as instituições acadêmicas envolvidas na elaboração da obra; os tipos de fontes de informação utilizadas para publicar os trabalhos; as instituições responsáveis pela publicação do trabalho; e o ano em que o trabalho foi publicado. Em seguida, na análise de conteúdo, procurou-se traçar o panorama sobre as principais ideias pesquisadas ao caracterizar o objetivo dos trabalhos e aspectos metodológicos, como nome do sistema/método/fórmula aplicados e/ou propostos, como ocorreu a avaliação, qual a natureza e a tipologia do *corpus*, como ocorreu a validação e a identificação dos termos, qual o tratamento no texto no momento da entrada de dados, a linguagem de indexação e o método/processo usado na identificação/ponderação/seleção dos termos.

Esses fenômenos da produção científica observados podem apresentar, na concentração dos resultados, o núcleo, a identidade do fenômeno analisado; enquanto os resultados mais afastados da concentração, a dispersão, este pode sinalizar as inovações na área.

RESULTADOS E DISCUSSÕES

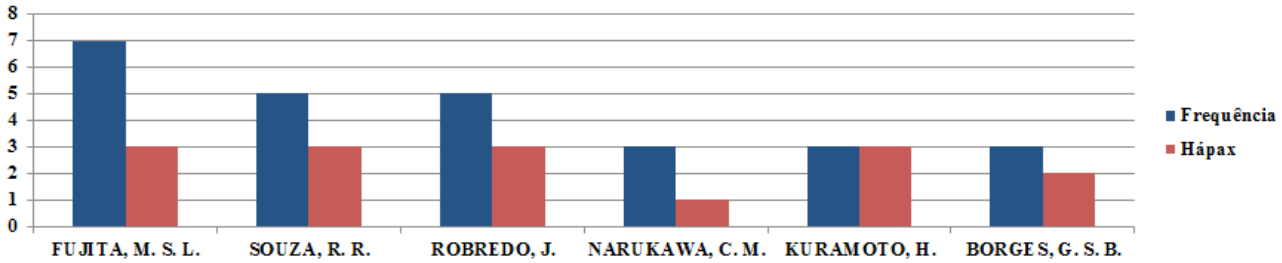
Nesta seção são apresentados os resultados da análise bibliométrica e análise de conteúdo dos documentos que compõem o *corpus* de análise. A primeira subseção procura traçar o panorama da institucionalização das pesquisas por meio da análise bibliométrica, enquanto as demais subseções apresentam a caracterização dos trabalhos através da análise de conteúdo.

INSTITUCIONALIZAÇÃO DAS PESQUISAS

Nesta subseção procura-se traçar o panorama da institucionalização das pesquisas no Brasil sobre a temática indexação automática.

O resultado obtido na análise bibliométrica dos autores obedece à Lei de Lotka, donde em um grupo formado por 78 pesquisadores, apenas 22% (17) se destacam com duas ou mais publicações, e 78% (61) realizaram apenas uma só publicação. E daquele conjunto menor, apenas um seleto grupo, formado por seis pesquisadores, se destacam com três ou mais publicações. Neste restrito grupo, Fujita, M. S. L. se destaca com sete publicações, seguida por Souza, R. R.; e Robredo, J., ambos com cinco publicações. Na sequência, com três publicações, encontramos Narukawa, C. M.; Kuramoto, H.; e Borges, G. S. B., conforme ilustra o gráfico 1. O índice de hápax revela a quantidade de vezes que o autor publicou individualmente.

Gráfico 1 – Relação autor x quantidade de pesquisas publicadas

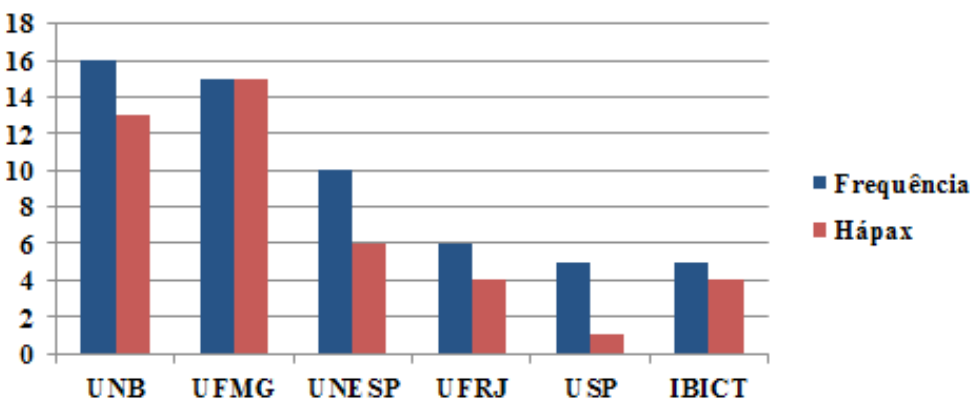


Fonte: desenvolvido pelo autor

Dos autores que realizaram apenas uma publicação, apenas 33% (20) apresentaram o hápax um, ou seja, realizaram uma publicação individualmente, portanto os outros 67% (41) publicaram coletivamente. Esta constatação aponta para o baixo índice de continuidade por parte da maioria dos autores, em que apenas um pequeno grupo de 17 autores (aproximadamente 22%) realizou mais de uma publicação sobre o tema da indexação automática.

Foram levantadas 28 instituições (nacionais e estrangeiras), às quais os autores estavam vinculados no momento da publicação. No gráfico 2 estão representadas apenas as instituições que apresentaram mais de quatro publicações, e ao lado o valor do hápax representando quantos documentos foram produzidos pela instituição individualmente, sem a coparticipação com outras instituições.

Gráfico 2 – Instituições acadêmicas com mais de quatro publicações



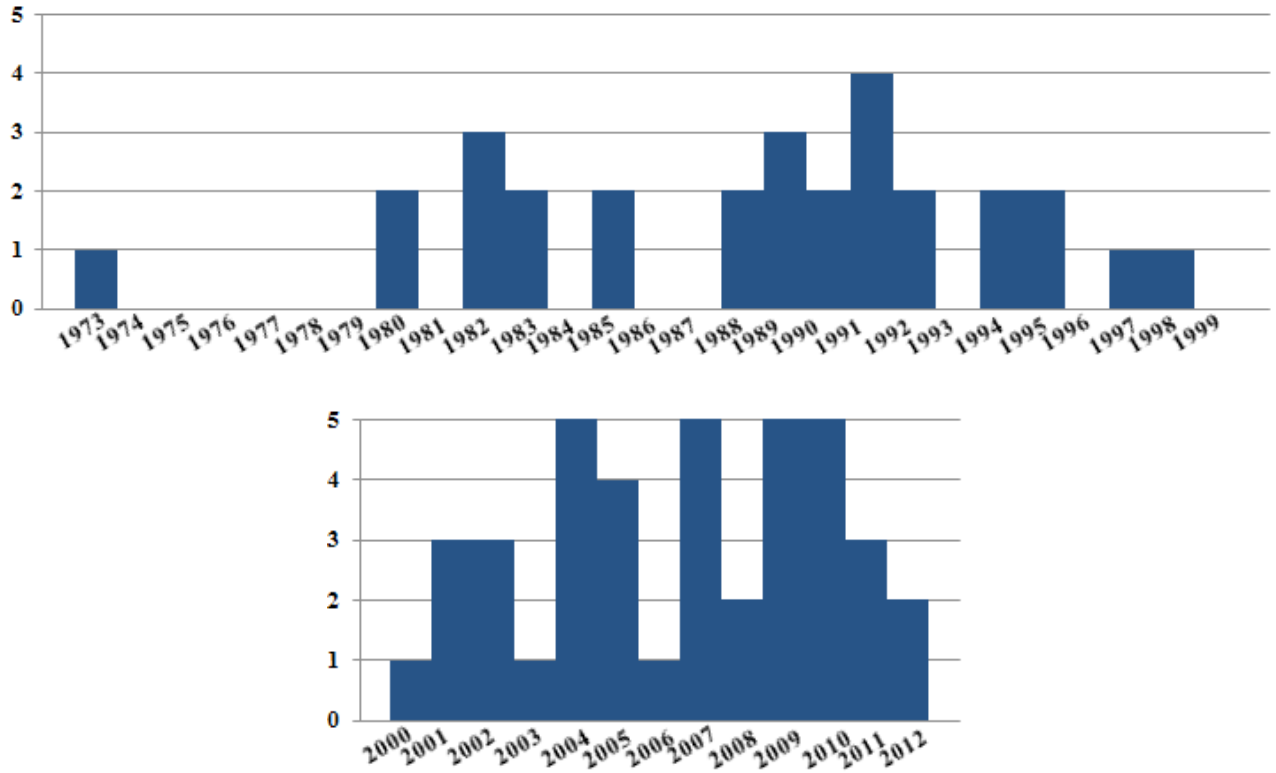
Fonte: desenvolvido pelo autor

Destacam-se a UNB, UFMG e Unesp como as principais instituições produtoras de publicações. Das seis instituições, a única que se destaca por não compor uma universidade é o Ibict, que tem atuado na promoção da popularização da informação científica e tecnológica através do desenvolvimento de produtos e serviços.

Nos últimos dez anos, as três instituições que aparecem em primeiro lugar no gráfico 2 ainda figuram como as maiores produtoras, mas agora a UFMG aparece com 11 trabalhos, a Unesp com sete, e a UnB apenas com três. Porém novas instituições começam a demonstrar sua força, como a USP e a UFRJ, a primeira com quatro e a segunda com três trabalhos. Além disso, percebemos que instituições que outrora não desenvolviam pesquisas sobre a indexação automática na área da ciência da informação começam a evidenciar interesse, como a Universidade Federal de Pernambuco (UFPE) e a UFScar, cada uma com dois trabalhos, e outras realizaram sua primeira publicação, tais como a *Universidade Federal de Uberlândia* (UFU), a Fundação de Amparo à Pesquisa do estado de Minas Gerais (Fapemig), Universidade de Illinois, *Universidade Federal do Rio Grande do Sul (UFGRS), Universidade Federal do Espírito Santo* (*UFES*), e a Faculdade Pitágoras.

Ao dividir as publicações entre aquelas realizadas no século XX e as publicadas no século XXI, constatamos que quantitativamente a diferença é pequena, pois observou-se entre os anos 1970 e 1999 uma frequência de ocorrência constando 31 publicações em contraste com as 40 publicações relacionadas ao período de 2000 a 2012. Entretanto, no século XXI, em apenas 12 anos a quantidade de publicação nacional sobre a indexação automática revelou-se maior (em comparação com os 30 anos do século XX). Também observamos uma oscilação entre a frequência de ocorrência das publicações no decorrer dos anos, ilustrado no gráfico 3, com o registro de alguns picos, formado por cinco publicações, nos anos 2004, 2007, 2009 e 2010.

Gráfico 3 – Frequência de ocorrência dos trabalhos observados através do ano de publicação



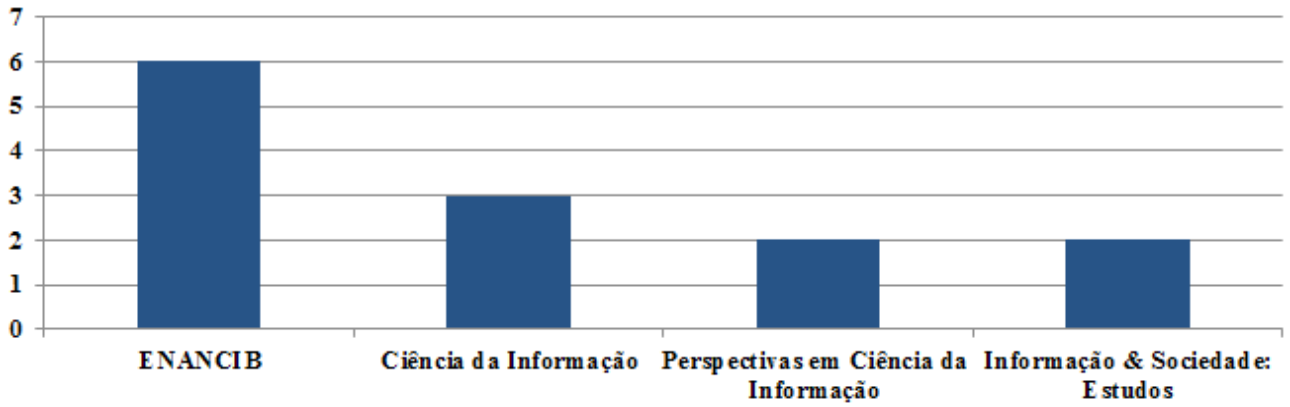
Fonte: desenvolvido pelo autor

Acreditamos que a tendência seja o aumento do número de publicações nos próximos anos, pois de acordo com Estudillo-García (2001) é no século XX, a partir de 1970, que surge um mundo baseado no vertiginoso processo informacional, com o desenvolvimento das tecnologias da informação e da comunicação, devido ao volume e acúmulo de informação, mas é no século XXI que muitas mudanças tecnológicas ocorrem, permitindo que novas pesquisas venham a ser realizadas.

O destaque quanto às fontes de publicação está nas revistas científicas da área de ciência da informação, com uma margem de 62%, correspondendo a 43 documentos dos 69 analisados, em outras palavras, os periódicos constituem uma parcela maior do que a soma de todas as outras fontes de publicação juntas, que chegam a 26 documentos, representando aproximadamente 38% do total. Diante dos dados obtidos, podemos considerar os periódicos como uma fonte de informação indispensável de orientação e pesquisa bibliográfica no campo de estudo sobre a automatização da indexação.

Quanto às instituições publicadoras sobre indexação automática, observando-se os últimos dez anos, percebemos que as principais são o Enancib, com seis trabalhos (um no ano 2005; dois em 2007; um em 2009; e outros dois em 2010); o periódico *Ciência da Informação*, com três publicações (duas em 2004 e uma em 2007); o periódico *Informação & Sociedade*: Estudos, com duas publicações (2008 e 2009); e *Perspectivas em Ciência da Informação*, com duas publicações (2009 e 2010). Ver ilustração no gráfico 4.

Gráfico 4 – Principais instituições publicadoras nos últimos dez anos



Fonte: desenvolvido pelo autor

OBJETIVO DOS TRABALHOS

Através da análise do objetivo dos trabalhos que compõem o *corpus*, foi construído o gráfico 5, no qual observamos que 24 trabalhos (35%) realizaram uma revisão bibliográfica.

Os demais 45 trabalhos, que compõem 65% do *corpus*, propõem e/ou aplicam sistema, método ou fórmula de indexação automática. São 14 os trabalhos que 'somente propõem' algum método, sistema ou fórmula de indexação automática, correspondendo a 20% do total, os outros 31 trabalhos estão divididos entre aqueles que 'somente aplicam' que correspondem a 15 trabalhos (22%), e os que 'aplicam e propõem' algum sistema, método ou fórmula, que correspondem a 16 trabalhos (23%).

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 à 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.

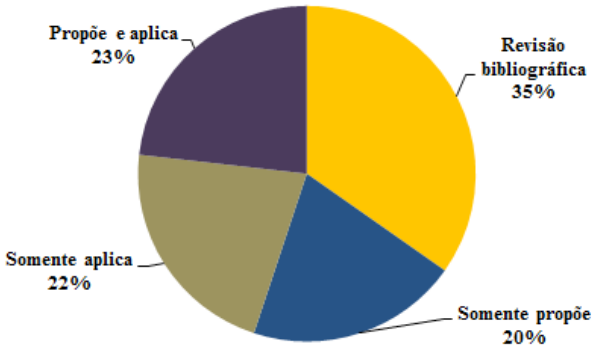
É providencial destacar três trabalhos individuais, pois são pontos de vista poucos explorados que podem estar surgindo para suprimir uma lacuna e representar base para o aparecimento de outras pesquisas: o primeiro de Guedes e Borschiver (2005), que apresenta a aplicação das leis e princípios da bibliometria, com foco nas palavras visando à 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.

Conforme podemos averiguar no gráfico 6, dos 14 trabalhos foram classificados no grupo dos que somente propõem algum método, sistema ou fórmula, nenhum trabalho propôs o uso de fórmula, seis trabalhos (43%), apresentaram a proposta de utilizar algum método, enquanto oito (57%) propuseram o uso de 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 com 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%), conforme pode ser verificado no gráfico 7.

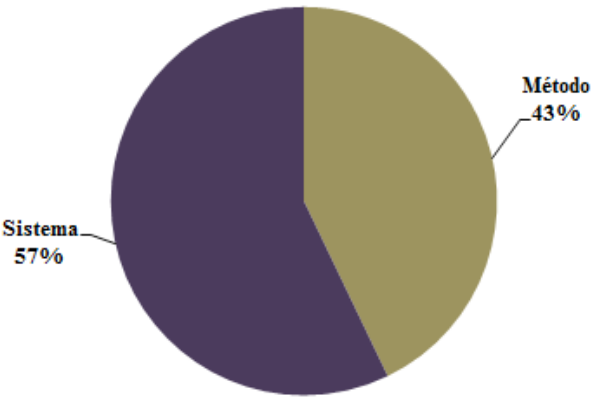
Levando em conta a aplicação, as fórmulas bibliométricas aplicadas foram as leis de Zipf e o ponto T de Goffman; em relação aos sistemas, os mais aplicados foram o Automindex/II e o Sisa, já o método baseado na frequência de ocorrência foi o mais aplicado.

Gráfico 5 – Objetivo dos documentos do corpus



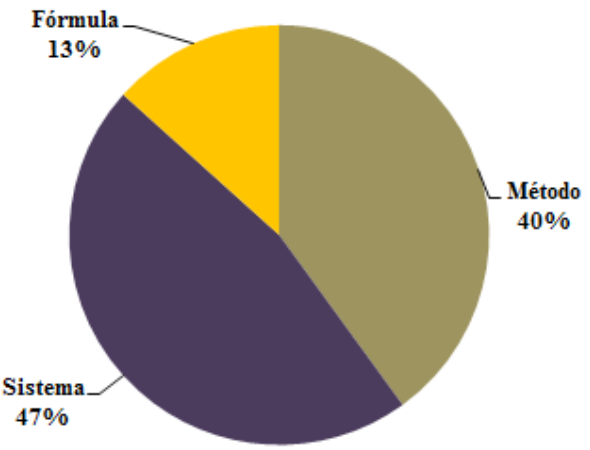
Fonte: desenvolvido pelo autor

Gráfico 6 – Distribuição dos trabalhos que realizaram somente proposição



Fonte: desenvolvido pelo autor

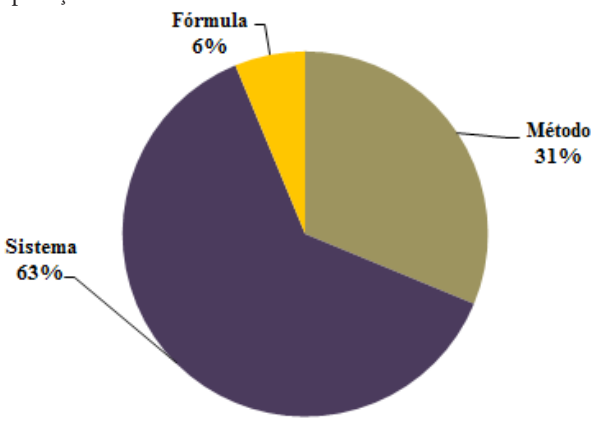
Gráfico 7 – Distribuição dos trabalhos que realizaram somente aplicação



Fonte: desenvolvido pelo autor

O gráfico 8 ilustra a distribuição dos 16 trabalhos que realizam proposições e aplicações de fórmula, sistema ou método de indexação automática, em que um (6%) propõe e aplica fórmula bibliométrica; cinco trabalhos (31%) estão relacionados com os métodos; e 10 (63%) propõem e aplicam sistema de indexação automática.

Gráfico 8 – Distribuição dos trabalhos que realizaram proposição e aplicação



Fonte: desenvolvido pelo autor

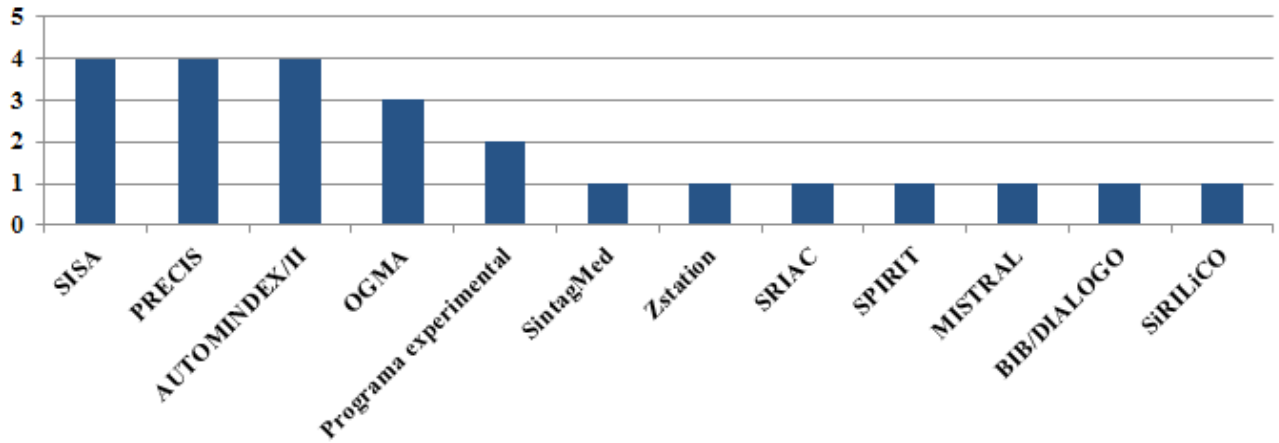
Em razão dos trabalhos que propõem e aplicam, constatou-se um só trabalho referente à adaptação da fórmula de transição de Goffman, no que tange aos sistemas, o Precis e o Ogma foram os mais propostos e aplicados, enquanto aos métodos, todos fizeram menção aos sintagmas nominais.

NOME DO 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 representada no gráfico 9, em que averiguamos que os três sistemas mais pesquisados aparecem empatados com quatro trabalhos cada um (17%), são eles: o sistema Sisa, o Precis e o Automindex/II. Logo em seguida aparece o Ogma, sendo pesquisado por três trabalhos (13%).

Gráfico 9 – Frequência de ocorrência dos sistemas nos documentos do corpus



Fonte: desenvolvido pelo autor

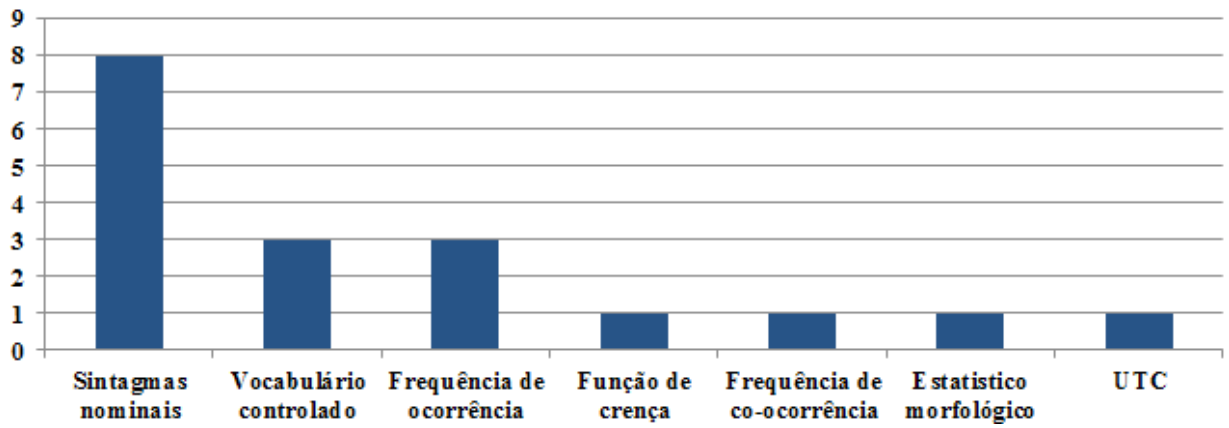
A quinta posição pertence a um sistema nomeado 'Programa experimental', entretanto a verificação de um fato irá modificar este resultado, pois se constatou que este sistema experimental, registrado nos trabalhos de Haller (1983, 1985), foi desenvolvido em um computador Burroughs 6700 do CPD da UnB. Em Vieira (1988), 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. Logo, o sistema outrora classificado como 'programa experimental' e o sistema BIB/Diálogo. E

para finalizar este desfecho, Robredo (1991) explica que o sistema Automindex/II constitui um 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 sendo classificados pelo sistema mais geral, o BIB/Diálogo passa a ser o mais pesquisado, com sete trabalhos.

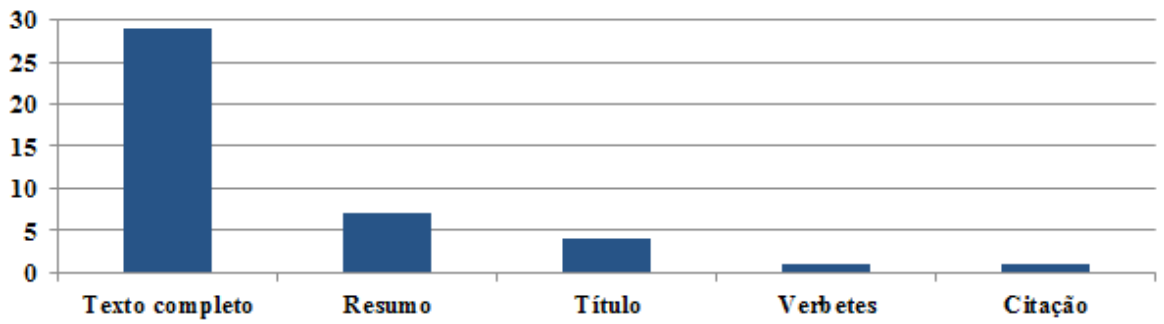
Cada um dos seis sistemas restantes é proposto e/ou aplicado por apenas um trabalho. Assim, os sistemas SintagMed, Zstation, Sriac, Spirit, Mistral e Sirilico, representando 60% dos sistemas observados, não apresentam continuidade nas pesquisas e foram investigados por 25% dos trabalhos, enquanto os 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 certo prosseguimento nas pesquisas sobre esses sistemas.

Gráfico 10 – Frequência de ocorrência dos métodos nos documentos do corpus



Fonte: desenvolvido pelo autor

Gráfico 11 – Natureza do corpus



Fonte: desenvolvido pelo autor

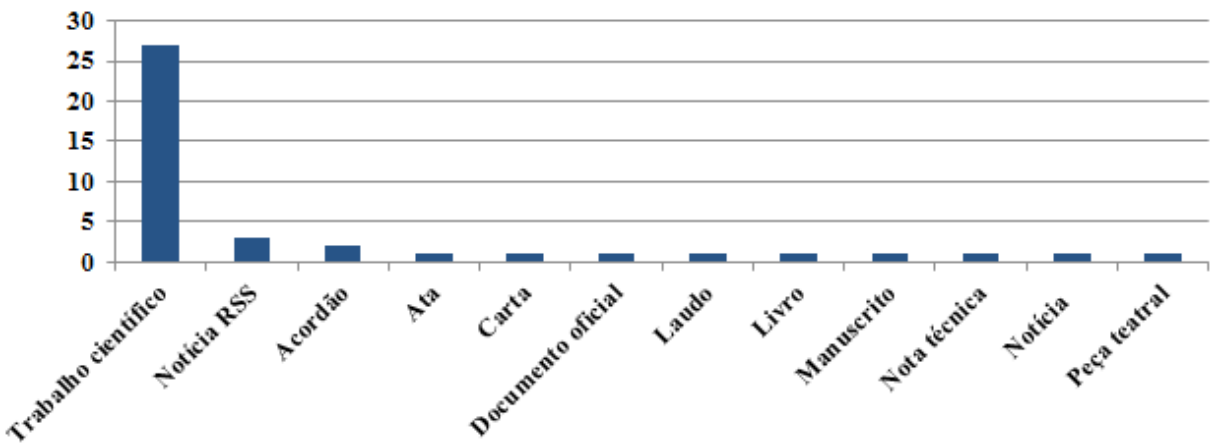
Observando o gráfico 10, constatamos que o método mais pesquisado de indexação automática foram os sintagmas nominais, com oito trabalhos (44%), sendo um trabalho que aplica; três que propuseram e aplicaram; e quatro que propuseram esse método. Empatados na segunda colocação, com três trabalhos (17%) 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: função de crença, frequência de co-ocorrência, estatístico-morfológico e UTC.

NATUREZA E TIPOLOGIA DO *CORPUS*

Quanto à natureza do *corpus* utilizado como objeto da indexação automática, constatou-se que ocorreu uma preferência em 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 do *corpus* empregada. 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 são ilustrados no gráfico 11.

Quanto à tipologia dos documentos do *corpus*, observa-se no gráfico 12 que a distribuição se comporta de acordo com o padrão das distribuições bibliométricas em geral: "poucos com muito e muitos com pouco". Assim, podemos constatar que poucas tipologias ocorreram muitas vezes, enquanto diversas tipologias ocorreram poucas vezes. Verifica-se que a tipologia mais pesquisada com 27 trabalhos (66%) foram os trabalhos científicos, o que pode ser explicado por ser esse tipo de material que mais interessa às instituições que normalmente desenvolvem e/ou financiam tais 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 12 – 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.

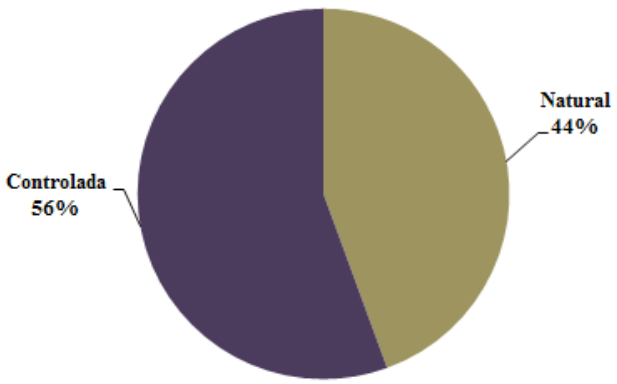
ENTRADA DE DADOS

O gráfico 13 ilustra como ocorreu a 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 13 – Frequência de ocorrência da entrada de dados

Fonte: desenvolvido pelo autor

Gráfico 14 – Natureza da linguagem



Fonte: desenvolvido pelo autor

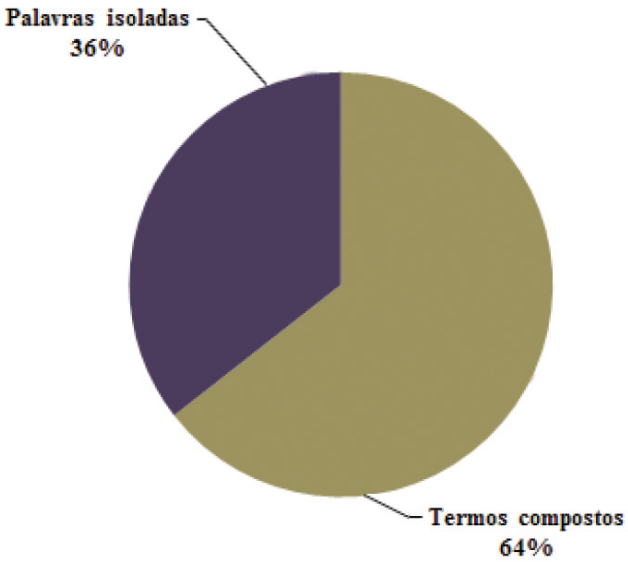
LINGUAGEM DE INDEXAÇÃO

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

Quanto à natureza da linguagem, notamos que ocorreu predominância pela linguagem controlada com 25 trabalhos (56%), em decorrência dos 20 trabalhos (44%) atribuídos à linguagem natural, visualizados no gráfico 14.

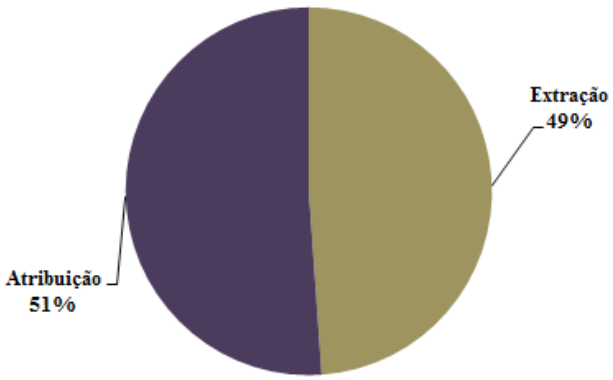
Os dados relacionados aos termos estão ilustrados no gráfico 15, que ressalta a superioridade numérica da extração de termos compostos com 29 trabalhos (64%) em relação à escolha apenas por palavras isoladas, com 16 trabalhos (36%).

Gráfico 15 – Natureza dos termos



Fonte: desenvolvido pelo autor

Gráfico 16 – Identificação dos termos



Fonte: desenvolvido pelo autor

IDENTIFICAÇÃO DOS TERMOS

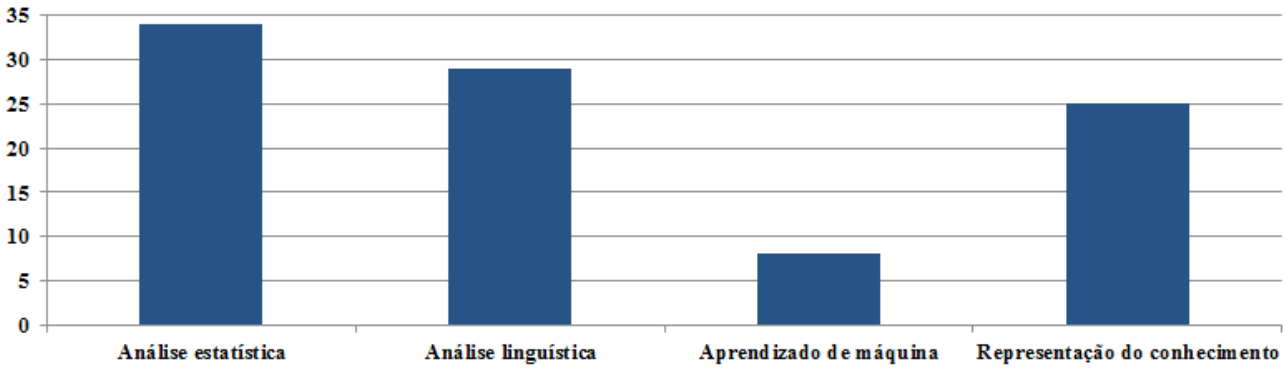
O gráfico 16 caracteriza o processo de identificação dos termos nos trabalhos, em 22 trabalhos (49%), a identificação se deu por meio da extração, enquanto em 23 (53%) por meio da atribuição.

Essa 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.

Quanto à abordagem ou tipo de técnicas utilizadas para identificar os termos utilizados na indexação automática, em um total de 45 documentos que propuseram, aplicaram ou propuseram e aplicaram sistema/método/fórmula, foram observados o 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 17, no qual a abordagem mais pesquisada foi a análise estatística com 34 trabalhos (35%), a segunda abordagem 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 17 – Abordagem na identificação dos termos



Fonte: desenvolvido pelo autor

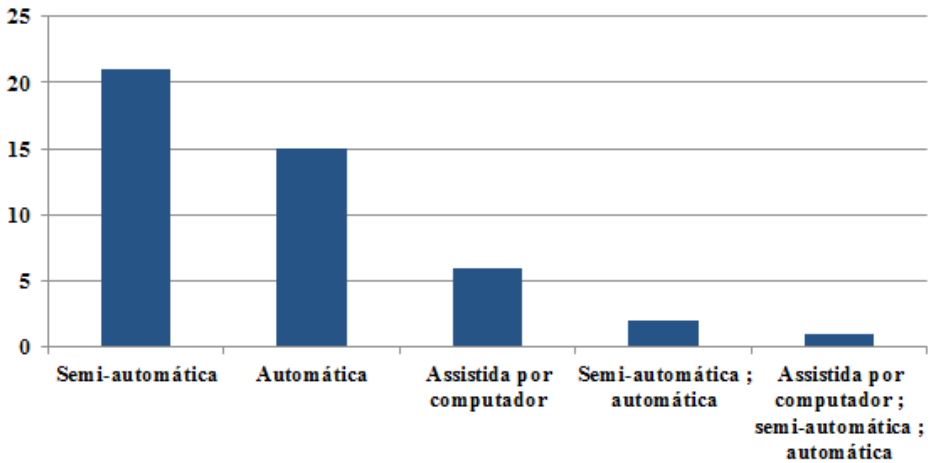
VALIDAÇÃO DOS TERMOS

Analisando o tipo de indexação automática quanto ao método de validação dos termos de indexação nos trabalhos, observamos pelo gráfico 18 que, em primeiro lugar, com 21 trabalhos (47%), estão aqueles que empregaram a indexação semiautomática, 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 semiautomática, quanto automática, e um trabalho (2%) no qual a validação ocorreu através dos três critérios de análise.

A AVALIAÇÃO DA INDEXAÇÃO AUTOMÁTICA

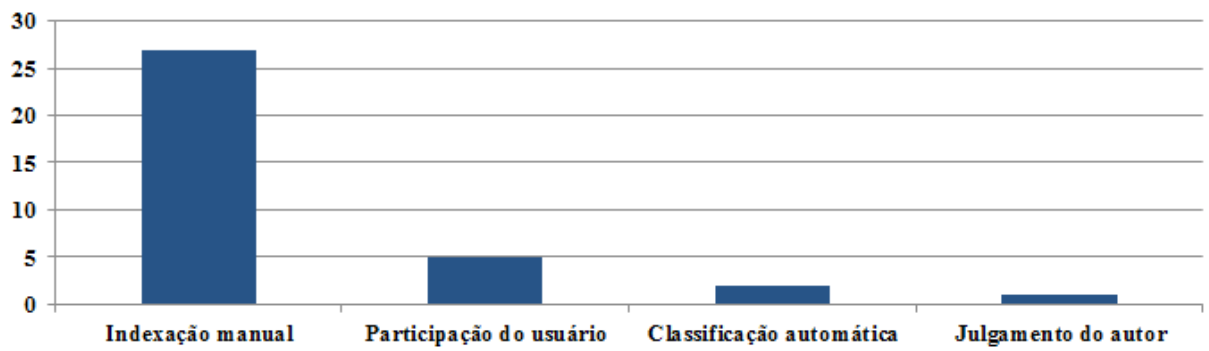
Para a maioria dos autores, o método escolhido para avaliar suas pesquisas sobre o processo automático da indexação foi realizando uma comparação dos resultados obtidos pelo processo automático com os obtidos através do método intelectual de indexação, correspondendo a 27 dos trabalhos (71%), conforme gráfico 19. Não foi possível identificar em dez trabalhos qual o método de avaliação empregado, mas nos 35 documentos analisados ficou evidente que a maioria escolheu como método de avaliação a comparação com a indexação manual.

Gráfico 18 – Validação dos termos



Fonte: desenvolvido pelo autor

Gráfico 19 – 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%). Nesse processo, geralmente se aplicam 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. A avaliação pode ocorrer sobre dois pontos de vista, o do sistema e o do usuário.

Dois trabalhos (11%) utilizam o método de classificação automática, em que aplicam algoritmos de agrupamento ou classificação automática às representações dos documentos obtidas através do processo de indexação automática, e avaliam com base no valor percentual de documentos classificados corretamente.

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

CONSIDERAÇÕES FINAIS

Os resultados indicam que na área de ciência da informação no Brasil, as pesquisas sobre indexação automática estão concentradas em certos autores, instituições acadêmicas, fontes de publicação e instituições publicadoras, além de existir tendência de aumento das publicações sobre esse assunto nos próximos anos. Entretanto, acreditamos que maior nível de institucionalização da pesquisa seja necessário para que ocorram mais contribuições brasileiras para a indexação automática de textos em português.

Fundamentados nos resultados da análise de conteúdo dos documentos do *corpus* desta pesquisa, observamos que a maior parte dos trabalhos propõe e/ou aplica método, sistema ou fórmula de indexação automática. Os demais trabalhos realizam revisão bibliográfica, abordando principalmente a história da indexação automática, seus fundamentos teóricos e evolução dos métodos.

Podemos concluir que o sintagma nominal foi o método mais investigado. Já em relação aos sistemas de indexação automática, quatro se destacam: o BIB/Dialogo (que inclui o Automindex), o Sisa, o Precis e o Ogma.

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.

Já a entrada dos dados apresentou empate técnico entre textos não estruturados em relação ao texto estruturado.

Em relação à linguagem de indexação, foi observada tanto a natureza da linguagem, que demonstrou preferência dos trabalhos pela pesquisa com a linguagem controlada, quanto no que se refere à natureza dos termos, a primazia se encontra no estudo com termos compostos.

Quanto à identificação de termos, foi constatada pequena diferença no percentual de trabalhos que pesquisaram a identificação dos termos por meio da extração (47%) e os que optaram pela atribuição (53%). Há uma explicação provável para isto, 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.

Quanto à abordagem dos métodos de identificação de termos, chegamos à constatação de que o tipo de método mais pesquisado foi a análise estatística, representando que grande parte dos trabalhos recorreu 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. Entretanto, a abordagem linguística e o uso de instrumento de representação do conhecimento também foram muito investigadas nos trabalhos.

Analisando o tipo de validação dos termos, percebemos a preferência pela aplicação da indexação semiautomática. O que pode ser justificado pelo fato, de os 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.

Verificamos que grande parte 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.

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 à realizada pelo especialista humano quando realiza a mesma tarefa.

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

- investigar a análise de citação 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.

REFERÊNCIAS

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.

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

BARDIN, L. *Análise de conteúdo*. Lisboa: Edições 70, 2004.

BORGES, G.S.B. *Indexação automática de documentos textuais*: proposta de critérios essenciais. 2009. 113 f. Dissertação (Mestrado) – Universidade Federal de Minas Gerais, Escola de Ciência da Informação. Minas Gerais, 2009.

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.

CÂMARA JÚNIOR, A.T. *Indexação automática de acórdãos por meio de processamento de linguagem natural.* Brasília, DF, 2007. 142f. Dissertação (Mestrado em Ciência da Informação) – Departamento de Ciência da Informação e Documentação da Universidade de Brasília. Brasília, DF, 2007.

CARDOSO, A.M.P. Retomando possibilidades conceituais: uma contribuição à sistematização do campo da informação social. *Revista da Escola de Ciência da Informação da UFMG*, v.23, n.2, p.107-114, jul./dez.,1994.

CASTRO, C. de M. *Ciência e universidade.* Rio de Janeiro: Zahar, 1997. 56 p.

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, Mariângela Spotti Lopes (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.

GASPAR, L. *Oficina de indexação*: indexação de documentos. Recife, 2011. (Material didático).

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.

GOMES, H. E. O indexador face às novas tecnologias de informação. *Transinformação*, v.1, n.2, p.161-171, maio/ago., 1989.

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.

MAIA, L.C.G. *Uso de sintagmas nominais na classificação automática de documentos eletrônicos*. 2008. 158 f. Tese (Doutorado em Ciência da Informação) - Universidade Federal de Minas Gerais, Belo Horizonte, 2008.

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.

MEIRELES, M.R.G.; CENDÓN, B.V. Aplicação prática dos processos de análise de conteúdo e de análise de citações em artigos relacionados às redes neurais artificiais. *Inf. Inf.* Londrina, v.15, n.2, p.77 – 93, jul./dez., 2010.

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

MORAES, R. Análise de conteúdo. *Revista Educação*. Porto Alegre, v.22, n.37, p.7-32, 1999.

NARUKAWA, C.M. *Estudo de vocabulário controlado na indexação automática*: aplicação no processo de indexação do Sistema de Indización Semiautomatica (SISA). 2011. 222 f. Dissertação (Mestrado) - Faculdade de Filosofia e Ciências, Universidade Estadual Paulista, Marília, 2011.

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. A indexação automática de textos: o presente já entrou no futuro. *Estudos Avançados em Biblioteconomia e Ciência da Informação*, Brasília, v.1, n.1, p.235-274, 1982.

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

RUDIO, F.V. *Introdução ao projeto de pesquisa científica*. Petrópolis: Vozes, 1986.

SANTOS, G. C.; RIBEIRO, C. M. *Acrônimos, siglas e termos técnicos*: arquivística, biblioteconomia, documentação, informática. Campinas: Editora Átomo, 2003.

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

SOUZA, R.R. Uma proposta de metodologia para indexação automática utilizando sintagmas nominais. *Enc. Bibli: R. Eletr. Bibliotecon. Ci. Inf.*, Florianópolis, n. esp., p.42-59, 1. sem. 2006.

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