

A INDEXAÇÃO AUTOMÁTICA NO BRASIL NO ÂMBITO DA CIÊNCIA DA INFORMAÇÃO (1973-2012): indicadores bibliométricos de autoria

Remi Lapa (UFPE)

rmcrlp@gmail.com

Renato Fernandes Correa (UFPE)

fc_renato@yahoo.com.br

EIXO TEMÁTICO: Produção e produtividade científica

MODALIDADE: Apresentação oral

1 INTRODUÇÃO

O objetivo geral deste trabalho é traçar um panorama da produção científica brasileira sobre a indexação automática no período de 1973 e 2012 por meio da bibliometria, a fim de levantar os autores e instituições acadêmicas que mais produziram sobre o tema, bem como analisar a distribuição temporal das publicações.

A justificativa para a realização de tal pesquisa está na necessidade de conhecer e socializar como está estruturada a produção científica nacional sobre a indexação automática, a fim de contribuir para o avanço das pesquisas nesta temática.

Além da introdução este trabalho está dividido em outras quatro seções: na segunda seção aborda-se a fundamentação teórica e um breve panorama da indexação automática no Brasil; na terceira seção estrutura-se os procedimentos metodológicos; na quarta encontram-se os resultados e discussões; enquanto na quinta apresentam-se as considerações finais.

2 INDEXAÇÃO AUTOMÁTICA NO BRASIL

A indexação automática pode ser definida como um conjunto de operações, basicamente matemáticas, linguísticas, de programação, destinadas somente a selecionar certos elementos descritivos de um documento sem modificar seu conteúdo.

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

Um outro tipo de indexação automática apresentada por Lancaster (2004) é a indexação automática por atribuição, que 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 termos a ser indexado um “perfil” de palavras ou expressões.

No Brasil, segundo Santos (2009), o uso da indexação automática iniciou-se no final dos anos 60, com a utilização do programa KWIC 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).

No final da década de 70, as pesquisas sobre o tema em território nacional ocorrem por meio de estudos individuais, realizados em cursos de pós-graduação, concentrando-se na análise de frequência, análise semântica e sintática do termo (VIEIRA, 1988, p. 54).

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 *Système Syntaxique et Probabiliste d'Indexation et de Recherche d'Informaticos Textuelles* (SPIRIT) para documentos em língua portuguesa (SANTOS, 2009).

Para Santos (2009), 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 (KURAMOTO, 1995; BORGES; MACULAN; LIMA, 2008).

3 PROCEDIMENTOS METODOLÓGICOS

Esta pesquisa tem como procedimento metodológico o emprego de análise bibliométrica em relação ao *corpus* de análise, constituído de 69 documentos, observando-os através dos campos: ano, autoria, instituição acadêmica, fonte de publicação, e instituição publicadora, como forma de produzir resultados que descrevam os trabalhos.

A metodologia consiste em um estudo bibliográfico de caráter quantitativo sobre a produção literária no Brasil a respeito da indexação automática de textos escritos no idioma português em âmbito brasileiro, tais como: livros, artigos de periódicos científicos, anais publicados em congressos e seminários, e literaturas cinzentas, obtidos através de processo

realizada por meio de pesquisa nas bases de dados virtuais BRAPCI¹, Google Acadêmico² e PERI³, e na biblioteca da UFPE, onde os documentos foram localizados através das expressões de busca ‘indexação automática’, ‘automatização da indexação’ e ‘indexação semi-automática’.

Para a realização da análise bibliométrica o *corpus* analisado foi classificado dentro de seis categorias: *autor*, que consta do nome do(s) autor(es) responsáveis pelo desenvolvimento intelectual da obra; *instituição acadêmica*, que apresenta o nome da(s) Instituição(ões)/Organização(ões) envolvidas direta ou indiretamente na elaboração da obra; *título*, que contém o título do trabalho; *ano*, que exibe o ano da publicação do trabalho; *fonte de publicação*, que registra o tipo da fonte onde o trabalho foi publicado; e *instituição publicadora*, que mostra o Periódico/Evento/Editora onde o documento foi publicado. Estas categorias serão representadas, respectivamente, pelas legendas: **AUT**, **INST**, **TIT**, **ANO**, **FONT**, **IPUB**. Sendo que a categoria TIT (título) consta somente como forma de orientação, não possuindo valor analítico.

O procedimento seguinte será a formatação de uma base de dados através do *software Microsoft Word*, em seguida utilizar-se-á o *software DataView*⁴ com a finalidade de contar e organizar de forma automática os elementos do arquivo, além de realizar a análise por meio da escolha de pares de campos específicos. Por último, as tabelas contendo as contagens serão exportadas para o *software Microsoft Excel* para a elaboração de gráficos e tabelas.

4 RESULTADOS E DISCUSSÕES

O resultado obtido na análise bibliométrica dos autores obedece a 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 única publicação. E daquele conjunto menor, apenas um seletor grupo formado por 35%, representado 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

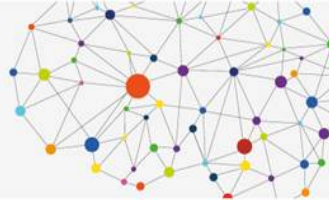
¹ <http://www.brapci.ufpr.br>

² <http://scholar.google.com.br/schhp?hl=pt-BR&tab=ws>

³ <http://bases.eci.ufmg.br/peri.htm>

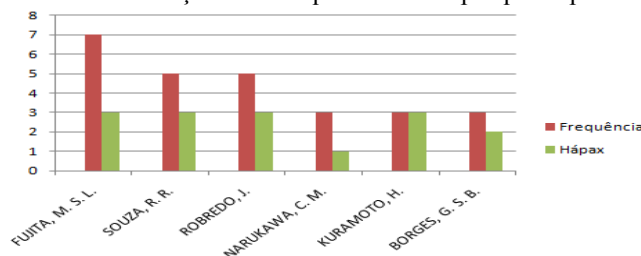
⁴ http://www.slb.com/services/characterization/software/data_utilities/dataview.aspx

Recife, PE, 14 a 16 de maio de 2014



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

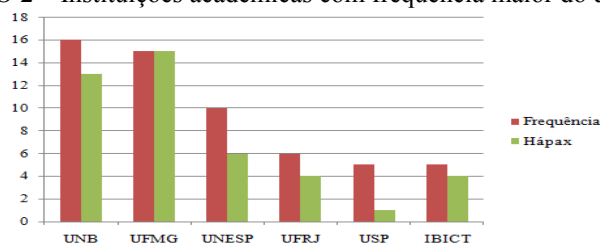


Fonte: desenvolvido pelo autor

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, onde 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). No GRÁFICO 2 estão representadas apenas as instituições que apresentaram uma frequência de ocorrência maior do que quatro e ao lado o valor do hápax representando quantos documentos foram produzidos pela instituição individualmente, sem a co-participação com outras instituições. Destacam-se a UNB, UFMG e UNESP como as principais produtoras de publicações.

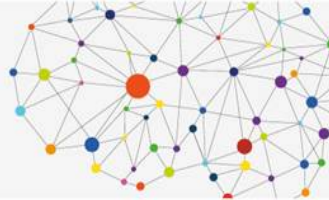
GRÁFICO 2 – Instituições acadêmicas com frequência maior do que quatro



Fonte: desenvolvido pelo autor

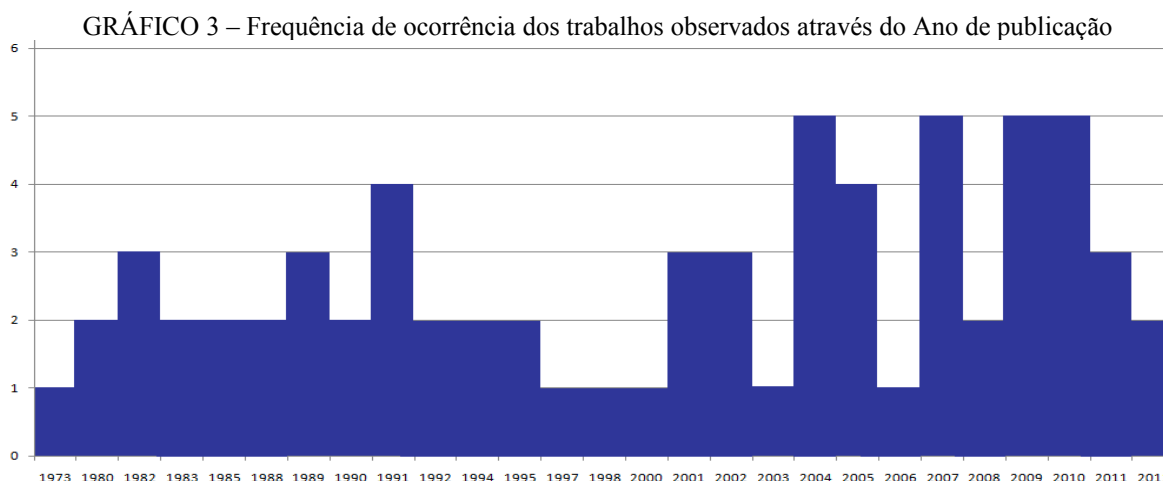
Nos últimos dez anos, as três instituições que aparecem em primeiro lugar no GRÁFICO 2, ainda figuram como as maiores produtoras, mas agora a UFMG aparece com 11 trabalhos, a UNESP com 7, e a UnB apenas com 3. Porém novas instituições começam a demonstrar sua força, como a USP e a UFRJ, a primeira com 4 e a segunda com 3 trabalhos.

Além disso, percebemos que instituições que outrora não desenvolviam pesquisas sobre a seleção automática de descritores na área da Ciência da Informação começam a evidenciar interesse, como a Universidade Federal de Pernambuco (UFPE) e a UFScar, cada



uma com 2 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.

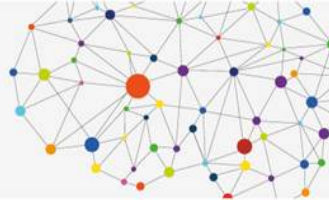
Ao dividir as publicações entre aquelas realizadas no século XX e as publicadas no século XXI constatamos que quantitativamente a diferença é pequena, pois observou-se entre os anos 1970 e 1999 uma frequência de ocorrência constando 31 publicações em contraste com as 40 publicações relacionadas ao período 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.



Fonte: desenvolvido pelo autor

Acreditamos que a tendência é aumentar o 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.

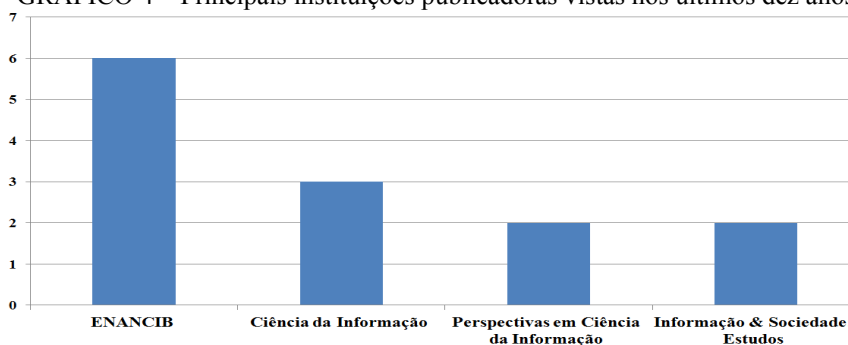
O destaque quanto às fontes de publicação está nos periódicos produzidos pelas revistas científicas da área de ciência da informação, com uma margem de 62%,



correspondendo a 43 documentos dos 69 analisados, em outras palavras, os periódicos correspondem a um valor maior do que a soma de todas as outras fontes de publicação juntos, 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 as 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 vistas nos últimos dez anos



Fonte: desenvolvido pelo autor

5 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 uma tendência de aumento das publicações sobre este assunto nos próximos anos. Como trabalhos futuros sugerimos que sejam realizados análises de conteúdo no *corpus* de análise procurando-se identificar os objetivos dos trabalhos e seus aspectos metodológicos.

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.
- LANCASTER, F. W. **Indexação e resumos**: teoria e prática. 2. ed. ver. atual. Brasília: Briquet de Lemos, 2004.
- 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.
- SANTOS, V. N. **Indexação automática de documentos textuais**: iniciativas dos grupos de pesquisa de universidades públicas brasileiras. 2009. 72f. Trabalho de Conclusão de Curso (Bacharelado em Biblioteconomia) – Departamento de Biblioteconomia e Documentação, Universidade de São Paulo, 2009.
- VIEIRA, S. B. Indexação automática e manual: revisão de literatura. **Ci. Inf.**, Brasília, v. 17, n. 1, p. 43-57, jan./jun. 1988.
- 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.
- 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.
- ESTUDILLO-GARCÍA, Joel. Elementos que conforman la sociedad de la información. **Investigación Bibliotecológica**, v. 15, n. 31, p. 161-194, jul./dic. 2001.